

Statistics for Biology and Health

Series Editors:

Mitchell Gail

Klaus Krickeberg

Jonathan M. Samet

Anastasios Tsiatis

Wing Wong

For further volumes:

<http://www.springer.com/series/2848>

Eric Vittinghoff • David V. Glidden
Stephen C. Shiboski • Charles E. McCulloch

Regression Methods in Biostatistics

Linear, Logistic, Survival, and Repeated
Measures Models

Second edition

 Springer

Eric Vittinghoff
Department of Epidemiology
and Biostatistics
University of California, San Francisco
Parnassas Ave. 500
94143 San Francisco California
MU-420 West
USA

David V. Glidden
Department of Epidemiology
and Biostatistics
University of California, San Francisco
Parnassas Ave. 500
94143 San Francisco California
MU-420 West
USA

Stephen C. Shiboski
Department of Epidemiology
and Biostatistics
University of California, San Francisco
Parnassas Ave. 500
94143 San Francisco California
MU-420 West
USA

Prof. Charles E. McCulloch
Department of Epidemiology
and Biostatistics
University of California, San Francisco
Berry 185
94107 San Francisco California
Suite 5700
USA

ISSN 1431-8776
ISBN 978-1-4614-1352-3 e-ISBN 978-1-4614-1353-0
DOI 10.1007/978-1-4614-1353-0
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011945441

© Springer Science+Business Media, LLC 2004, 2012

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*For Rupert & Jean; Kay & Minerva;
Caroline, Erik & Hugo; and J.R.*

Preface

In the second edition of *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, we have substantially revised and expanded the core chapters of the first edition, and added two new chapters. The first of these, Chap. 9, on strengthening causal inference, introduces potential outcomes, average causal effects, and two primary methods for estimating these effects, what we call *potential outcomes estimation* and inverse probability weighting. It also covers propensity scores in detail, then more briefly discusses time-dependent exposures, controlled and natural direct effects, instrumental variables, and principal stratification. The second, Chap. 11, on missing data, explains why this is a problem, classifies missingness by mechanism, and discusses the shortcomings of some simple approaches. Its focus is on three primary approaches for dealing with missing data: maximum likelihood estimation, multiple imputation, and inverse weighting, and lays out in detail when each of these approaches is most appropriate.

Among the core chapters of the first edition, Chap. 5, on logistic regression, has substantial new sections on models for ordinal and multinomial outcomes, as well as exact logistic regression. Chapter 6, on survival analysis, has an in-depth new section on competing risks, as well as new coverage of interval censoring and left truncation. Chapter 7, on repeated measures analysis, introduces recently developed methods for distinguishing between- and within-cluster effects, and for estimating the effects of fixed and time-dependent covariates (TDCs) on change. Chapter 8, on generalized linear models, adds coverage of negative binomial as well as zero-inflated and zero-truncated models for counts. Chapters 4–8 all now cover restricted cubic splines, take a new approach to mediation, and provide methods for sample size, power, and detectable effect calculation. Chapter 10, on predictor selection, has expanded coverage of developing and assessing models for prediction, as well as a new section on *directed acyclic graphs*. Our summary in Chap. 13 includes a new discussion of multiple comparisons and updated coverage of software packages. All Stata examples have been updated. As before, Stata, SAS, and Excel datasets and Stata do-files for most examples are provided on the website for the book, <http://www.biostat.ucsf.edu/vgsm>. We also posted implementations of analyses for time-dependent exposures too complicated for inclusion in the text.

At UCSF, we have used the first edition for a two-quarter course on regression methods for clinical researchers and epidemiologists, the first quarter covering linear and logistic models and predictor selection, and the second covering survival and repeated measures analysis. The new chapter on strengthening causal inference is the basis of new quarter-long course, and the new missing data chapter will play an important role in a more advanced quarter-long course next year. The new breadth of coverage of the second edition should make it more widely useful in year-long biostatistics courses for students like ours, MPH students, and for masters-level courses in biostatistics.

Finally, we gratefully acknowledge the very important contributions made by Professors Joseph Hogan of Brown University, Michael Hudgens of the University of North Carolina, Barbara McKnight of the University of Washington, and Maya Peterson of the University of California, Berkeley, who generously provided detailed, insightful reviews of the two new chapters. Any remaining errors and shortcomings are of course entirely ours.

San Francisco, CA, USA

Eric Vittinghoff
David V. Glidden
Stephen C. Shiboski
Charles E. McCulloch

Preface to the First Edition

The primary biostatistical tools in modern medical research are single-outcome, multiple-predictor methods: multiple linear regression for continuous outcomes, logistic regression for binary outcomes, and the Cox proportional hazards model for time-to-event outcomes. More recently, generalized linear models (GLMs) and regression methods for repeated outcomes have come into widespread use in the medical research literature. Applying these methods and interpreting the results require some introduction. However, introductory statistics courses have no time to spend on such topics and hence they are often relegated to a third or fourth course in a sequence. Books tend to have either very brief coverage or to be treatments of a single topic and more theoretical than the typical researcher wants or needs.

Our goal in writing this book was to provide an accessible introduction to multipredictor methods, emphasizing their proper use and interpretation. We feel strongly that this can only be accomplished by illustrating the techniques using a variety of real data sets. We have incorporated as little theory as feasible. Further, we have tried to keep the book relatively short and to the point. Our hope in doing so is that the important issues and similarities between the methods, rather than their differences, will come through. We hope this book will be attractive to medical researchers needing familiarity with these methods and to students studying statistics who would like to see them applied to real data. The methods we describe are, of course, the same as those used in a variety of fields, so non-medical readers will find this book useful if they can extrapolate from the predominantly medical examples.

A prerequisite for the book is a good first course in statistics or biostatistics or an understanding of the basic tools: paired and independent samples t -tests, simple linear regression and one-way analysis of variance (ANOVA), contingency tables and χ^2 (chi-square) analyses, Kaplan–Meier curves, and the logrank test.

We also think it is important for researchers to know how to interpret the output of a modern statistical package. Accordingly, we illustrate a number of the analyses with output from the Stata statistics package. There are a number of other packages that can perform these analyses, but we have chosen this one because of its accessibility and widespread use in biostatistics and epidemiology.

We begin the book with a chapter introducing our viewpoint and style of presentation and the big picture as to the use of multipredictor methods. Chapter 2 presents descriptive numerical and graphical techniques for multipredictor settings and emphasizes choice of technique based on the nature of the variables. Chapter 3 briefly reviews the statistical methods we consider prerequisites for the book.

We then make the transition in Chap. 4 to multipredictor regression methods, beginning with the linear regression model. This chapter also covers confounding, mediation, interaction, and model checking in the most detail. In Chap. 5, we turn to binary outcomes and the logistic model, noting the similarities to the linear model. Ties to simpler, contingency table methods are also noted. Chapter 6 covers survival outcomes, giving clear indications as to why such techniques are necessary, but again emphasizing similarities in model building and interpretation with the previous chapters. Chapter 7 looks at the accommodation of correlated data in both linear and logistic models. Chapter 8 extends Chap. 5, giving an overview of GLMs.

In the second edition, new sections of Chaps. 4–8 deal with pooled and exact logistic regression (Chap. 5), competing risks (Chap. 6), and time-varying predictors and separating between and within cluster information (Chap. 7). Chapters 4–8, also now conclude with short sections on calculating sample size, power, and minimum detectable effects.

The next three chapters, two of them new in the second edition, cover broader issues. Chapter 9 looks more closely at making causal inferences, using the models discussed in Chaps. 4–8, as well as alternatives including propensity scores and instrumental variables. Chapter 10 deals with predictor selection, with expanded treatment of methods for prediction problems. Chapter 11 considers missing data and methods for dealing with it, including maximum likelihood models, multiple imputation, and complete case analysis, the problematic default.

Finally, Chap. 12 is a brief introduction to the analysis of complex surveys. The text closes with a summary, Chap. 13, attempting to put each of the previous chapters in context. Too often it is hard to see the forest for the trees of each of the individual methods. Our goal in this final chapter is to provide guidance as to how to choose among the methods presented in the book and also to realize when they will not suffice and other techniques need to be considered.

San Francisco, CA, USA

Eric Vittinghoff
David V. Glidden
Stephen C. Shiboski
Charles E. McCulloch

Contents

1	Introduction	1
1.1	Example: Treatment of Back Pain	1
1.2	The Family of Multipredictor Regression Methods	2
1.3	Motivation for Multipredictor Regression	3
1.3.1	Prediction	3
1.3.2	Isolating the Effect of a Single Predictor	3
1.3.3	Understanding Multiple Predictors	4
1.4	Guide to the Book	4
2	Exploratory and Descriptive Methods	7
2.1	Data Checking	7
2.2	Types of Data	8
2.3	One-Variable Descriptions	9
2.3.1	Numerical Variables	9
2.3.2	Categorical Variables	16
2.4	Two-Variable Descriptions	17
2.4.1	Outcome Versus Predictor Variables	17
2.4.2	Continuous Outcome Variable	18
2.4.3	Categorical Outcome Variable	21
2.5	Multivariable Descriptions	22
2.6	Summary	25
2.7	Problems	25
3	Basic Statistical Methods	27
3.1	<i>t</i> -Test and Analysis of Variance	27
3.1.1	<i>t</i> -Test	28
3.1.2	One- and Two-Sided Hypothesis Tests	28
3.1.3	Paired <i>t</i> -Test	29
3.1.4	One-Way Analysis of Variance	30
3.1.5	Pairwise Comparisons in ANOVA	30
3.1.6	Multi-way ANOVA and ANCOVA	31
3.1.7	Robustness to Violations of Normality Assumption	31

3.1.8	Nonparametric Alternatives	32
3.1.9	Equal Variance Assumption	32
3.2	Correlation Coefficient	33
3.2.1	Spearman Rank Correlation Coefficient	34
3.2.2	Kendall's τ	34
3.3	Simple Linear Regression Model	35
3.3.1	Systematic Part of the Model	35
3.3.2	Random Part of the Model	36
3.3.3	Assumptions About the Predictor	37
3.3.4	Ordinary Least Squares Estimation	38
3.3.5	Fitted Values and Residuals	39
3.3.6	Sums of Squares	39
3.3.7	Standard Errors of the Regression Coefficients	40
3.3.8	Hypothesis Tests and Confidence Intervals	40
3.3.9	Slope, Correlation Coefficient, and R^2	42
3.4	Contingency Table Methods for Binary Outcomes	42
3.4.1	Measures of Risk and Association for Binary Outcomes	43
3.4.2	Tests of Association in Contingency Tables	46
3.4.3	Predictors with Multiple Categories	48
3.4.4	Analyses Involving Multiple Categorical Predictors	50
3.4.5	Collapsibility of Standard Measures of Association ...	52
3.5	Basic Methods for Survival Analysis	54
3.5.1	Right Censoring	54
3.5.2	Kaplan–Meier Estimator of the Survival Function	55
3.5.3	Interpretation of Kaplan–Meier Curves	57
3.5.4	Median Survival	58
3.5.5	Cumulative Event Function	59
3.5.6	Comparing Groups Using the Logrank Test	60
3.6	Bootstrap Confidence Intervals	62
3.7	Interpretation of Negative Findings	64
3.8	Further Notes and References	65
3.9	Problems	65
3.10	Learning Objectives	66
4	Linear Regression	69
4.1	Example: Exercise and Glucose	70
4.2	Multiple Linear Regression Model	72
4.2.1	Systematic Part of the Model	72
4.2.2	Random Part of the Model	73
4.2.3	Generalization of R^2 and r	75
4.2.4	Standardized Regression Coefficients	75
4.3	Categorical Predictors	76
4.3.1	Binary Predictors	76

4.3.2	Multilevel Categorical Predictors	77
4.3.3	The F -Test	81
4.3.4	Multiple Pairwise Comparisons Between Categories ..	81
4.3.5	Testing for Trend Across Categories	84
4.4	Confounding	89
4.4.1	Range of Confounding Patterns	90
4.4.2	Confounding Is Difficult to Rule Out	91
4.4.3	Adjusted Versus Unadjusted $\hat{\beta}$ s	92
4.4.4	Example: BMI and LDL	93
4.5	Mediation	94
4.5.1	Indirect Effects via the Mediator	95
4.5.2	Overall and Direct Effects	95
4.5.3	Percent Explained	96
4.5.4	Example: BMI, Exercise, and Glucose	96
4.5.5	Pitfalls in Evaluating Mediation	97
4.6	Interaction	99
4.6.1	Example: Hormone Therapy and Statin Use	100
4.6.2	Example: BMI and Statin Use	102
4.6.3	Interaction and Scale	105
4.6.4	Example: Hormone Therapy and Baseline LDL	106
4.6.5	Details	107
4.7	Checking Model Assumptions and Fit	108
4.7.1	Linearity	109
4.7.2	Normality	116
4.7.3	Constant Variance	119
4.7.4	Outlying, High Leverage, and Influential Points	124
4.7.5	Interpretation of Results for Log Transformed Variables	128
4.7.6	When to Use Transformations	129
4.8	Sample Size, Power, and Detectable Effects	130
4.8.1	Calculations Using Standard Errors Based on Published Data	133
4.9	Summary	135
4.10	Further Notes and References	135
4.10.1	Generalized Additive Models	136
4.11	Problems	136
4.12	Learning Objectives	138
5	Logistic Regression	139
5.1	Single Predictor Models	140
5.1.1	Interpretation of Regression Coefficients	144
5.1.2	Categorical Predictors	146
5.2	Multipredictor Models	150
5.2.1	Likelihood Ratio Tests	154
5.2.2	Confounding	156

5.2.3	Mediation.....	158
5.2.4	Interaction.....	160
5.2.5	Prediction.....	165
5.2.6	Prediction Accuracy	166
5.3	Case-Control Studies	168
5.3.1	Matched Case-Control Studies	171
5.4	Checking Model Assumptions and Fit	173
5.4.1	Linearity	173
5.4.2	Outlying and Influential Points.....	175
5.4.3	Model Adequacy	177
5.4.4	Technical Issues in Logistic Model Fitting	179
5.5	Alternative Strategies for Binary Outcomes	180
5.5.1	Infectious Disease Transmission Models	181
5.5.2	Pooled Logistic Regression	183
5.5.3	Regression Models Based on Risk Differences and Relative Risks.....	186
5.5.4	Exact Logistic Regression	188
5.5.5	Nonparametric Binary Regression	189
5.5.6	More Than Two Outcome Levels	190
5.6	Likelihood.....	192
5.7	Sample Size, Power, and Detectable Effects.....	194
5.8	Summary	199
5.9	Further Notes and References	200
5.10	Problems	200
5.11	Learning Objectives.....	202
6	Survival Analysis	203
6.1	Survival Data	203
6.1.1	Why Linear and Logistic Regression Would not Work.....	203
6.1.2	Hazard Function	204
6.1.3	Hazard Ratio	205
6.1.4	Proportional Hazards Assumption	207
6.2	Cox Proportional Hazards Model	207
6.2.1	Proportional Hazards Models	207
6.2.2	Parametric Versus Semi-parametric Models.....	208
6.2.3	Hazard Ratios, Risk, and Survival Times	211
6.2.4	Hypothesis Tests and Confidence Intervals.....	212
6.2.5	Binary Predictors	213
6.2.6	Multilevel Categorical Predictors	213
6.2.7	Continuous Predictors	217
6.2.8	Confounding	218
6.2.9	Mediation.....	219
6.2.10	Interaction	220
6.2.11	Model Building	222

6.2.12	Adjusted Survival Curves for Comparing Groups.....	222
6.2.13	Predicted Survival for Specific Covariate Patterns	224
6.3	Extensions to the Cox Model.....	225
6.3.1	Time-Dependent Covariates	225
6.3.2	Stratified Cox Model	228
6.4	Checking Model Assumptions and Fit.....	231
6.4.1	Log-Linearity of the Hazard Function	231
6.4.2	Proportional Hazards	232
6.5	Competing Risks Data.....	239
6.5.1	What Are Competing Risks Data?	239
6.5.2	Notation for Competing Risks Data.....	240
6.5.3	Summaries for Competing Risk Data	241
6.6	Some Details	247
6.6.1	Bootstrap Confidence Intervals	247
6.6.2	Prediction.....	248
6.6.3	Adjusting for Nonconfounding Covariates	248
6.6.4	Independent Censoring	249
6.6.5	Interval Censoring	249
6.6.6	Left-Truncation	250
6.7	Sample Size, Power, and Detectable Effects.....	252
6.8	Summary	256
6.9	Further Notes and References	256
6.10	Problems	257
6.11	Learning Objectives.....	259
7	Repeated Measures and Longitudinal Data Analysis.....	261
7.1	A Simple Repeated Measures Example: Fecal Fat	262
7.1.1	Model Equations for the Fecal Fat Example	264
7.1.2	Correlations Within Subjects	264
7.1.3	Estimates of the Effects of Pill Type	266
7.2	Hierarchical Data	267
7.2.1	Example: Treatment of Back Pain	267
7.2.2	Example: Physician Profiling	267
7.2.3	Analysis Strategies for Hierarchical Data	268
7.3	Longitudinal Data.....	270
7.3.1	Analysis Strategies for Longitudinal Data	271
7.3.2	Analyzing Change Scores	273
7.4	Generalized Estimating Equations	276
7.4.1	Example: Birthweight and Birth Order Revisited	277
7.4.2	Correlation Structures	279
7.4.3	Working Correlation and Robust Standard Errors.....	281
7.4.4	Tests and Confidence Intervals	282
7.4.5	Use of <code>xtgee</code> for Clustered Logistic Regression	284
7.5	Random Effects Models	284
7.6	Re-Analysis of the Georgia Babies Data Set	286

7.7	Analysis of the SOF BMD Data.....	288
7.7.1	Time Varying Predictors.....	289
7.7.2	Separating Between- and Within-Cluster Information ..	291
7.7.3	Prediction.....	293
7.7.4	A Logistic Analysis.....	294
7.8	Marginal Versus Conditional Models	295
7.9	Example: Cardiac Injury Following Brain Hemorrhage	296
7.9.1	Bootstrap Analysis.....	298
7.10	Power and Sample Size for Repeated Measures Designs	301
7.10.1	Between-Cluster Predictor	301
7.10.2	Within-Cluster Predictor.....	303
7.11	Summary	304
7.12	Further Notes and References	305
7.12.1	Missing Data	305
7.12.2	Computing	306
7.13	Problems	306
7.14	Learning Objectives.....	308
8	Generalized Linear Models.....	309
8.1	Example: Treatment for Depression	309
8.1.1	Statistical Issues	310
8.1.2	Model for the Mean Response	311
8.1.3	Choice of Distribution	312
8.1.4	Interpreting the Parameters.....	312
8.1.5	Further Notes.....	313
8.2	Example: Costs of Phototherapy	314
8.2.1	Model for the Mean Response	315
8.2.2	Choice of Distribution	315
8.2.3	Interpreting the Parameters.....	316
8.3	Generalized Linear Models.....	316
8.3.1	Example: Risky Drug Use Behavior	317
8.3.2	Modeling Data with Many Zeros	318
8.3.3	Example: A Randomized Trial to Reduce Risk of Fracture	321
8.3.4	Relationship of Mean to Variance.....	323
8.3.5	Non-Linear Models	324
8.4	Sample Size for the Poisson Model	325
8.5	Summary	328
8.6	Further Notes and References	328
8.7	Problems	329
8.8	Learning Objectives.....	330
9	Strengthening Causal Inference.....	331
9.1	Potential Outcomes and Causal Effects.....	332
9.1.1	Average Causal Effects	332
9.1.2	Marginal Structural Model	333

9.1.3	Fundamental Problem of Causal Inference	333
9.1.4	Randomization Assumption	334
9.1.5	Conditional Independence	334
9.1.6	Marginal and Conditional Means	335
9.1.7	Potential Outcomes Estimation	336
9.1.8	Inverse Probability Weighting	337
9.2	Regression as a Basis for Causal Inference	337
9.2.1	No Unmeasured Confounders	338
9.2.2	Correct Model Specification	338
9.2.3	Overlap and the Positivity Assumption	338
9.2.4	Lack of Overlap and Model Misspecification	339
9.2.5	Adequate Sample Size and Number of Events	341
9.2.6	Example: Phototherapy for Neonatal Jaundice	341
9.3	Marginal Effects and Potential Outcomes Estimation	344
9.3.1	Marginal and Conditional Effects	344
9.3.2	Contrasting Conditional and Marginal Effects	346
9.3.3	When Marginal and Conditional Odds-Ratios Differ	346
9.3.4	Potential Outcomes Estimation	347
9.3.5	Marginal Effects in Longitudinal Data	350
9.4	Propensity Scores	352
9.4.1	Estimation of Propensity Scores	352
9.4.2	Effect Estimation Using Propensity Scores	355
9.4.3	Inverse Probability Weights	356
9.4.4	Checking for Propensity Score/Exposure Interaction ..	358
9.4.5	Addressing Positivity Violations Using Restriction	359
9.4.6	Average Treatment Effect in the Treated (ATT)	360
9.4.7	Recommendations for Using Propensity Scores	362
9.5	Time-Dependent Treatments	364
9.5.1	Models Using Time-dependent IP Weights	365
9.5.2	Implementation	367
9.5.3	Drawbacks and Difficulties	368
9.5.4	Focusing on New Users	369
9.5.5	Nested New-User Cohorts	370
9.6	Mediation	370
9.7	Instrumental Variables	373
9.7.1	Vulnerabilities	375
9.7.2	Structural Equations and Instrumental Variables	377
9.7.3	Checking IV Assumptions	377
9.7.4	Example: Effect of Hormone Therapy on Change in LDL	378
9.7.5	Extension to Binary Exposures and Outcomes	379
9.7.6	Example: Phototherapy for Neonatal Jaundice	380
9.7.7	Interpretation of IV Estimates	382
9.8	Trials with Incomplete Adherence to Treatment	382
9.8.1	Intention-to-Treat	382

9.8.2	As-Treated Comparisons by Treatment Received	384
9.8.3	Instrumental Variables	385
9.8.4	Principal Stratification	385
9.9	Summary	387
9.10	Further Notes and References	387
9.11	Problems	391
9.12	Learning Objectives.....	394
10	Predictor Selection	395
10.1	Prediction.....	396
10.1.1	Bias–Variance Trade-off and Overfitting	397
10.1.2	Measures of Prediction Error.....	397
10.1.3	Optimism-Corrected Estimates of Prediction Error	398
10.1.4	Minimizing Prediction Error Without Overfitting.....	401
10.1.5	Point Scores	404
10.1.6	Example: Risk Stratification of Patients with Heart Disease	405
10.2	Evaluating a Predictor of Primary Interest.....	407
10.2.1	Including Predictors for Face Validity	408
10.2.2	Selecting Predictors on Statistical Grounds	408
10.2.3	Interactions With the Predictor of Primary Interest ...	409
10.2.4	Example: Incontinence as a Risk Factor for Falling ...	409
10.2.5	Directed Acyclic Graphs	410
10.2.6	Randomized Experiments	416
10.3	Identifying Multiple Important Predictors	418
10.3.1	Ruling Out Confounding Is Still Central	418
10.3.2	Cautious Interpretation Is Also Key	419
10.3.3	Example: Risk Factors for Coronary Heart Disease ...	420
10.3.4	Allen–Cady Modified Backward Selection	420
10.4	Some Details	421
10.4.1	Collinearity	421
10.4.2	Number of Predictors	422
10.4.3	Alternatives to Backward Selection	424
10.4.4	Model Selection and Checking	425
10.4.5	Model Selection Complicates Inference	425
10.5	Summary	427
10.6	Further Notes and References	427
10.7	Problems	428
10.8	Learning Objectives.....	429
11	Missing Data	431
11.1	Why Missing Data Can Be a Problem	432
11.1.1	Missing Predictor in Linear Regression	432
11.1.2	Missing Outcome in Longitudinal Data	434

11.2	Classifications of Missing Data	437
11.2.1	Mechanisms for Missing Data	438
11.3	Simple Approaches to Handling Missing Data	442
11.3.1	Include a Missing Data Category	442
11.3.2	Last Observation or Baseline Carried Forward	442
11.4	Methods for Handling Missing Data	444
11.5	Missing Data in the Predictors and Multiple Imputation	444
11.5.1	Remarks About Using Multiple Imputation	446
11.5.2	Approaches to Multiple Imputation	447
11.5.3	Multiple Imputation for HERS	449
11.6	Deciding Which Missing Data Mechanism May Be Applicable	451
11.7	Missing Outcomes, Missing Completely at Random	452
11.8	Missing Outcomes, Covariate-Dependent Missing Completely at Random	452
11.9	Missing Outcomes for Longitudinal Studies, Missing at Random	453
11.9.1	ML and MAR	455
11.9.2	Multiple Imputation	456
11.9.3	Inverse Probability Weighting	456
11.10	Technical Details About Maximum Likelihood and Data Which are Missing at Random	458
11.10.1	An Example of the EM Algorithm	458
11.10.2	The EM Algorithm Imputes the Missing Data	460
11.10.3	ML Versus MI with Missing Outcomes	461
11.11	Methods for Data that are Missing Not at Random	461
11.11.1	Pattern Mixture Models	461
11.11.2	Multiple Imputation Under MNAR	463
11.11.3	Joint Modeling of Outcomes and the Dropout Process	463
11.12	Summary	463
11.13	Further Notes and References	464
11.14	Problems	465
11.15	Learning Objectives	467
12	Complex Surveys	469
12.1	Overview of Complex Survey Designs	470
12.2	Inverse Probability Weighting	471
12.2.1	Accounting for Inverse Probability Weights in the Analysis	473
12.2.2	Inverse Probability Weights and Missing Data	473
12.3	Clustering and Stratification	474
12.3.1	Design Effects	474
12.4	Example: Diabetes in NHANES	475

12.5	Some Details	477
12.5.1	Ignoring Secondary Levels of Clustering	477
12.5.2	Other Methods of Variance Estimation	477
12.5.3	Model Checking	478
12.5.4	Postestimation Capabilities in Stata	478
12.5.5	Other Statistical Packages for Complex Surveys	479
12.6	Summary	479
12.7	Further Notes and References	479
12.8	Problems	480
12.9	Learning Objectives	480
13	Summary	481
13.1	Introduction	481
13.2	Selecting Appropriate Statistical Methods	482
13.3	Planning and Executing a Data Analysis	483
13.3.1	Analysis Plans	483
13.3.2	Choice of Software	484
13.3.3	Data Preparation	484
13.3.4	Record Keeping and Reproducibility of Results	484
13.3.5	Data Security	485
13.3.6	Consulting a Statistician	485
13.3.7	Use of Internet Resources	486
13.4	Further Notes and References	486
13.4.1	Multiple Hypothesis Tests	486
13.4.2	Statistical Learning	487
	References	489
	Index	501

Chapter 1

Introduction

The book describes a family of statistical techniques that we call *multipredictor* regression modeling. This family is useful in situations where there are multiple measured factors (also called predictors, covariates, or independent variables) to be related to a single outcome (also called the response or dependent variable). The applications of these techniques are diverse, including those where we are interested in prediction, isolating the effect of a single predictor, or understanding multiple predictors. We begin with an example.

1.1 Example: Treatment of Back Pain

Korff et al. (1994) studied the success of various approaches to treatment for back pain. Some physicians treat back pain more aggressively, with prescription pain medication and extended bed rest, while others recommend an earlier resumption of activity and manage pain with over-the-counter medications. The investigators classified the aggressiveness of a sample of 44 physicians in treating back pain as low, medium, or high, and then followed 1,071 of their back pain patients for two years. In the analysis, the classification of treatment aggressiveness was related to patient outcomes, including cost, activity limitation, pain intensity, and time to resumption of full activity.

The primary focus of the study was on a single categorical predictor, the aggressiveness of treatment. Thus for a continuous outcome like cost, we might think of an analysis of variance (ANOVA), while for a categorical outcome we might consider a contingency table analysis and a χ^2 -test. However, these simple analyses would be incorrect at the very least because they would fail to recognize that multiple patients were *clustered* within physician practice and that there were *repeated outcome measures* on patients.

Looking beyond the clustering and repeated measures (which are covered in Chap. 7), what if physicians with more aggressive approaches to back pain also

tended to have older patients? If older patients recover more slowly (regardless of treatment), then even if differences in treatment aggressiveness have no effect, the age imbalance would nonetheless make for poorer outcomes in the patients of physicians in the high-aggressiveness category. Hence, it would be misleading to judge the effect of treatment aggressiveness without correcting for the imbalances between the physician groups in patient age and, potentially, other prognostic factors—that is, to judge without *controlling for confounding*. This can be accomplished using a model which relates study outcomes to age and other prognostic factors as well as the aggressiveness of treatment. In a sense, multipredictor regression analysis allows us to examine the effect of treatment aggressiveness while *holding the other factors constant*.

1.2 The Family of Multipredictor Regression Methods

Multipredictor regression modeling is a family of methods for relating multiple predictors to an outcome, with each member of the family suitable for a different type of outcome. The cost outcome, for example, is a numerical measure and for our purposes can be taken as *continuous*. This outcome could be analyzed using the linear regression model, though we also show in Chap. 8 why a *generalized linear model* (GLM) might be a better choice.

Perhaps the simplest outcome in the back pain study is the yes/no indicator of moderate-to-severe activity limitation; a subject's activities are limited by back pain or not. Such a categorical variable is termed *binary* because it can only take on two values. This type of outcome is analyzed using the logistic regression model, presented in Chap. 5.

In contrast, pain intensity was measured on a scale of ten equally spaced values. The variable is numerical and could be treated as continuous, although there were many tied values. Alternatively, it could be analyzed as a categorical variable, with the different values treated as ordered categories, using the proportional-odds or continuation-ratio models, both extensions of the logistic model and briefly covered in Chap. 5.

Another potential outcome might be time to resumption of full activity. This variable is also continuous, but what if a patient had not yet resumed full activity at the end of the follow-up period of two years? Then the time to resumption of full activity would only be known to exceed two years. When outcomes are known only to be greater than a given value (like two years), the variable is said to be *right-censored*—a common feature of time-to-event data. This type of outcome can be analyzed using the Cox proportional hazards model, the primary topic of Chap. 6.

Furthermore, in the back pain example, study outcomes were measured on groups, or clusters, of patients with the same physician, and on multiple occasions for each patient. To analyze such *hierarchical* or *longitudinal* outcomes, we need to use extensions of the basic family of regression modeling techniques suitable for

repeated measures data, described in Chap. 7. Related extensions are also required to analyze data from complex surveys, briefly covered in Chap. 12.

The various regression modeling approaches, while differing in important statistical details, also share important similarities. Numeric, binary, and categorical predictors are accommodated by all members of the family, and are handled in a similar way: on some scale, the systematic part of the outcome is modeled as a linear function of the predictor values and corresponding *regression coefficients*. The different techniques all yield estimates of these coefficients that summarize the results of the analysis and have important statistical properties in common. This leads to unified methods for selecting predictors and modeling their effects, as well as for making inferences to the population represented in the sample. Finally, all the models can be applied to the same broad classes of practical questions involving multiple predictors.

1.3 Motivation for Multipredictor Regression

Multipredictor regression can be a powerful tool for addressing three important practical questions. These questions, which provide the framework for our discussion of predictor selection in Chap. 10, include *prediction*, *isolating the effect of a single predictor*, and *understanding multiple predictors*.

1.3.1 *Prediction*

How can we identify which patients with back pain will have moderate-to-severe limitation of activity? Multipredictor regression is a powerful and general tool for using multiple measured predictors to make useful predictions for future observations. In this example, the outcome is binary and thus a multipredictor logistic regression model could be used to estimate the predicted probability of limitation for any possible combination of the observed predictors. These estimates could then be used to classify patients as likely to experience limitation or not. Similarly, if our interest was future costs, a continuous variable, we could use a linear regression model to predict the costs associated with new observations characterized by various values of the predictors. In developing models for this purpose, we need to avoid *over-fitting*, and to *validate* their predictiveness in actual practice.

1.3.2 *Isolating the Effect of a Single Predictor*

In settings where multiple, related predictors contribute to study outcomes, it will be important to consider multiple predictors even when a single predictor is of interest. In the von Korff study, the primary predictor of interest was how

aggressively a physician treated back pain. But incorporation of other predictors was necessary to minimize *confounding*, so that we could plausibly consider a causal interpretation of the estimated effects of the aggressiveness of treatment. Estimating causal effects from observational data is difficult, and sometimes requires special methods, including *potential outcomes estimation* and *propensity scores*. These approaches depend on the assumption that there are no unmeasured confounders. Causal estimation using *instrumental variables* depends on different but equally stringent assumptions. We consider these specialized methods in Chap. 9.

1.3.3 Understanding Multiple Predictors

Multipredictor regression can also be used when our aim is to identify multiple independent predictors of a study outcome—independent in the sense that they appear to have an effect over and above other measured variables. Especially in this context, we may need to consider other complexities of how predictors jointly influence the outcome. For example, the effect of injuries on activity limitation may in part operate through their effect on pain; in this view, pain *mediates* the effect of injury and should not be adjusted for, at least initially. Alternatively, suppose that among patients with mild or moderate pain, younger age predicts more rapid recovery, but among those with severe pain, age makes little difference. The effects of both age and pain severity will both potentially be misrepresented if this *interaction* is not taken into account. Fortunately, all the multipredictor regression methods discussed in this book easily handle interactions, as well as mediation and confounding, using essentially identical techniques. Though certainly not foolproof, multipredictor models are well suited to examining the complexities of how multiple predictors are associated with an outcome of interest.

1.4 Guide to the Book

This text attempts to provide practical guidance for regression analysis. We interweave real data examples from the biomedical literature in the hope of capturing the reader's interest and making the statistics as easy to grasp as possible. Theoretical details are kept to a minimum, since it is usually not necessary to understand the theory to use these methods appropriately. We avoid formulas and keep mathematical notation to a minimum, instead emphasizing selection of appropriate methods and careful interpretation of the results.

This book grew out a two-quarter sequence in multipredictor methods for physicians beginning a career in clinical research, with a focus on techniques appropriate to their research projects. For these students, mathematical explication is an ineffective way to teach these methods. Hence our reliance on real-world examples and heuristic explanations.

Our students take the course in the second quarter of their research training. A beginning course in biostatistics is assumed and some understanding of epidemiologic concepts is clearly helpful. However, Chap. 3 presents a review of topics from a first biostatistics course, and we explain epidemiologic concepts in some detail throughout the book.

Although theoretical details are minimized, we do discuss techniques of practical utility that some would consider advanced. We treat extensions of basic multi-predictor methods for repeated measures and hierarchical data, for data arising from complex surveys, and for the broader class of *generalized linear models*, of which logistic regression is the most familiar example. In addition, we consider alternative approaches to estimating the causal effects of an exposure or treatment from observational data, including *propensity scores* and *instrumental variables*. We address model checking as well as model selection in considerable detail, including specialized methods for avoiding over-fitting in selecting prediction models. And we consider how missing data arise, and the conditions under which maximum likelihood methods for repeated measures as well as multiple imputation of the missing values can successfully deal with it.

The orientation of this book is to *parametric* methods, in which the systematic part of the model is a simple function of the predictors, and substantial assumptions are made about the distribution of the outcome. In our view, parametric methods are usually flexible and robust enough, and we show how model adequacy can be checked. The Cox proportional hazards model covered in Chap. 6 is a *semi-parametric* method which makes few assumptions about an important component of the systematic part of the model, but retains most of the efficiency and many of the advantages of fully parametric models. *Generalized additive models*, briefly reviewed in Chap. 5, go an additional step in this direction. However, fully *nonparametric* regression methods in our view entail losses in efficiency and ease of interpretation which make them less useful to researchers. We do recommend a popular bivariate nonparametric regression method, LOWESS, but only for exploratory data analysis.

Our approach is also to encourage exploratory data analysis as well as thoughtful interpretation of results. We discourage focusing solely on P -values, which have an important place in statistics but also important limitations. In particular, P -values measure the strength of the evidence for an effect, but not its size. Furthermore, they can be misleading when data-driven model selection has been carried out. In our view, data analysis profits from considering the estimated effects, using confidence intervals (CIs) to quantify their precision. In prediction problems, P -values are a poor guide to *prediction error*, the proper focus of interest, and over-reliance of them can lead to over-fitting.

We recommend that readers begin with Chap. 2, on exploratory methods. Since Chap. 3 is largely a review, students may want to focus only on unfamiliar material. Chapter 4, on multipredictor regression methods for continuous outcomes, introduces most of the important themes of the book, which are then revisited in later chapters, and so is essential reading. Similarly, Chap. 9 covers causal inference, Chap. 10 addresses predictor selection, and Chap. 11 deals with missing data, all

topics common to the entire family of regression techniques. Chapters 5 and 6 cover regression methods specialized for binary and time-to-event outcomes, while Chaps. 7, 8, and 12 cover extensions of these methods for repeated measures, counts, and other special types of outcomes, and complex surveys. Readers may want to study these chapters as the need arises. Finally, Chap. 13 reprises the themes considered in the earlier chapters and is recommended for all readers.

For interested readers, Stata code and selected datasets used in examples and problems, plus errata, are posted on the website for this book:

<http://www.biostat.ucsf.edu/vgsm>

Chapter 2

Exploratory and Descriptive Methods

Before beginning any sort of statistical analysis, it is imperative to take a preliminary look at the data with three main goals in mind: first, to check for errors and anomalies; second, to understand the distribution of each of the variables on its own; and third, to begin to understand the nature and strength of relationships among variables. Errors should, of course, be corrected, since even a small percentage of erroneous data values can drastically influence the results. Understanding the distribution of the variables, especially the outcomes, is crucial to choosing the appropriate multipredictor regression method. Finally, understanding the nature and strength of relationships is the first step in building a more formal statistical model from which to draw conclusions.

2.1 Data Checking

Procedures for data checking should be implemented before data entry begins, to head off future headaches. Many data entry programs have the capability to screen for egregious errors, including values that are out the expected range or of the wrong “type.” If this is not possible, then we recommend regular checking for data problems as the database is constructed.

Here are two examples we have encountered recently. First, some values of a variable defined as a proportion were inadvertently entered as percentages (i.e., 100 times larger than they should have been). Although they made up less than 3% of the values, the analysis was completely invalidated. Fortunately, this simple error was easily corrected once discovered. A second example involved patients with a heart anomaly. Those whose diagnostic score was poor enough (i.e., exceeded a numerical threshold) were to be classified according to the type of anomaly. Data checks revealed missing classifications for patients whose diagnostic score exceeded the

threshold, as well as classifications for patients whose score did not, complicating planned analyses. Had the data been screened as they were collected, this problem with study procedures could have been avoided.

2.2 Types of Data

The proper description of data depends on the nature of the measurement. The key distinction for statistical analysis is between numerical and categorical variables. The number of diagnostic tests ordered is a numerical variable, while the gender of a person is categorical. Systolic blood pressure (SBP) is numerical, whereas the type of surgery is categorical.

A secondary but sometimes important distinction within numerical variables is whether the variable can take on a whole continuum or just a discrete set of values. So SBP would be continuous, while number of diagnostic tests ordered would be discrete. Cost of a hospitalization would be continuous, whereas number of mice able to successfully navigate a maze would be discrete. More generally,

Definition: A numerical variable taking on a continuum of values is called *continuous* and one that only takes on a discrete set of values is called *discrete*.

A secondary distinction sometimes made with regard to categorical variables is whether the categories are ordered or unordered. So, for example, categories of annual household income (<\$20,000, \$20,000–\$40,000, \$40,000–\$100,000, >\$100,000) would be ordered, while marital status (single, married, divorced, widowed) would be unordered. More exactly,

Definition: A categorical variable is *ordinal* if the categories can be logically ordered from smallest to largest in a sense meaningful for the question at hand (we need to rule out silly orders like alphabetical); otherwise it is unordered or *nominal*.

Some overlap between types is possible. For example, we may break a numerical variable (such as exact annual income in dollars and cents) into ranges or categories. Conversely, we may treat a categorical variable as a numerical score, for example, by assigning values one to five to the ordinal responses Poor, Fair, Good, Very Good, and Excellent.

Most of the analysis methods we will describe for numerical scores (e.g., linear regression or t-tests) have interpretations based on average scores. So assigning scores to a categorical variable is effective if average scores are readily interpretable. This may well be the case for scoring the categories Poor through Excellent as 1 through 5: an average value of 3.5 is between Good and Very Good. It might be a less effective strategy ordinal categorical variables such as the modified Rankin Scale, a scale used to assess disability following a stroke. For that scale, 0 represents no symptoms, 1 and 2 slight disability, 3 and 4 moderate disability, 5 severe disability, and 6 is dead. Consider two sets of three patients, the first set with scores of 0, 6, and 6 and the second with scores of 4, 4, and 4. Both have averages of 4, but the

first set would generally be considered as having worse outcomes since two of the patients died. In such a case, summarizing with the average, and hence treating the variable as numeric, may not be appropriate.

In the following sections, we present each of the descriptive and exploratory methods according to the types of variables involved.

2.3 One-Variable Descriptions

We begin by describing techniques useful for examining a single variable at a time. These are useful for uncovering mistakes or extreme values in the data and for assessing distributional shape.

2.3.1 Numerical Variables

We can describe the distribution of numerical variables using either numerical or graphical techniques.

2.3.1.1 Example: Systolic Blood Pressure

The western collaborative group study (WCGS) was a large epidemiological study designed to investigate the association between the “type A” behavior pattern and coronary heart disease (CHD) ([Rosenman et al. 1964](#)). We will revisit this study later in the book, focusing on the primary outcome, but for now we want to explore the distribution of SBP.

2.3.1.2 Numerical Description

As a first step, we obtain basic descriptive statistics for SBP. Table [2.1](#) gives detailed summary statistics for the SBP variable, `sbp`. Several features of the output are worth consideration. The largest and smallest values should be scanned for outlying or incorrect values, and the mean (or median) and standard deviation should be assessed as general measures of the location and spread of the data. Secondary features are the skewness and kurtosis, though these are usually more easily assessed by the graphical means described in the next section. Another assessment of skewness is a large difference between the mean and median. In *right-skewed* data, the mean is quite a bit larger than the median, while in *left-skewed* data, the mean is much smaller than the median. Of note, in this dataset, the largest observation is more than six standard deviations above the mean!

Table 2.1 Numerical description of systolic blood pressure

. summarize sbp, detail

systolic BP				

Percentiles		Smallest		
1%	104	98		
5%	110	100		
10%	112	100	Obs	3154
25%	120	100	Sum of Wgt.	3154
50%	126	Largest	Mean	128.6328
			Std. Dev.	15.11773
75%	136	210		
90%	148	210	Variance	228.5458
95%	156	212	Skewness	1.204397
99%	176	230	Kurtosis	5.792465

2.3.1.3 Graphical Description

Graphs are often the quickest and most effective way to get a sense of the data. For numerical data, three basic graphs are most useful: the histogram, boxplot, and normal quantile–quantile (or Q–Q) plot. Each is useful for different purposes. The histogram easily conveys information about the location, spread, and shape of the frequency distribution of the data. The boxplot is a schematic identifying key features of the distribution. Finally, the normal Q–Q plot facilitates comparison of the shape of the distribution of the data to a normal (or bell-shaped) distribution.

The histogram displays the frequency of data points falling into various ranges as a bar chart. Figure 2.1 shows a histogram of the SBP data from WCGS. Generated using an earlier version of Stata, the default histogram uses five intervals and labels axes with the minimum and maximum values only. In this figure, we can see that most of the measurements are in the range of about 100 to 150, with a few extreme values around 200. The percentage of observations in the first interval is about 47.4%.

However, this is not a particularly well-constructed histogram. With over 3,000 data points, we can use more intervals to increase the definition of the histogram and avoid grouping the data so coarsely. Using only five intervals, the first two including almost all the data, makes for a loss of information, since we only know the value of the data in those large “bins” to the limits of the interval (in the case of the first bin, between 98 and 125), and learn nothing about how the data are distributed within those intervals. Also, our preference is to provide more interpretable axis labeling. Figure 2.2 shows a modified histogram generated using the current version of Stata that provides much better definition as to the shape of the frequency distribution of SBP.

The key with a histogram is to use a sufficient number of intervals to define the shape of the distribution clearly and not lose much information, without using so many as to leave gaps, give the histogram a ragged shape, and defeat the goal of summarization. With 3,000 data points, we can afford quite a few bins. A *rough*

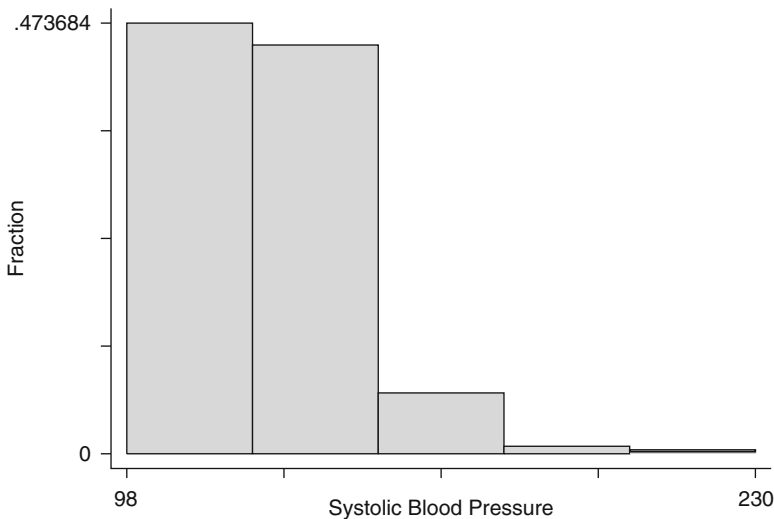


Fig. 2.1 Histogram of the systolic blood pressure data

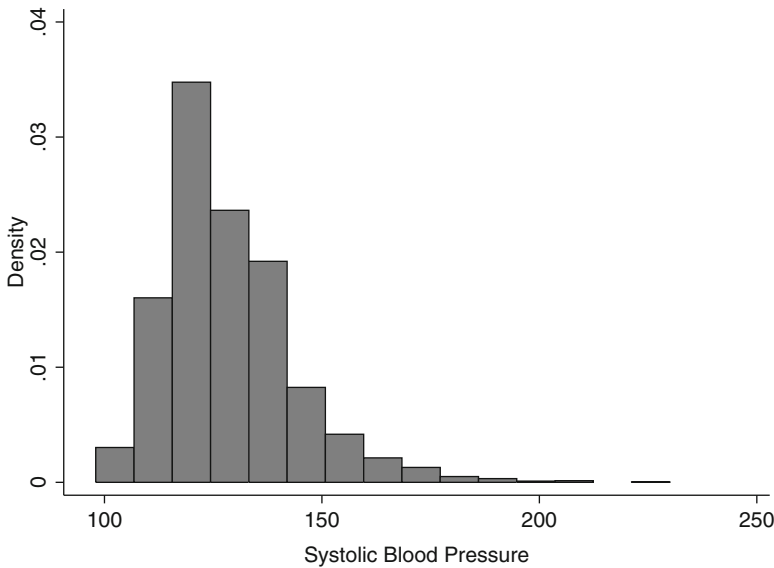


Fig. 2.2 Histogram of the systolic blood pressure data using 15 intervals

rule of thumb is to choose the number of bins to be about $1 + 3.3 \log_{10}(n)$, ([Sturges 1926](#)) where n is the sample size (so this would suggest 12 or 13 bins for the WCGS data). More than 20 or so are rarely needed. Figure 2.2 uses 15 bins and provides a clear definition of the shape as well as a fair bit of detail.

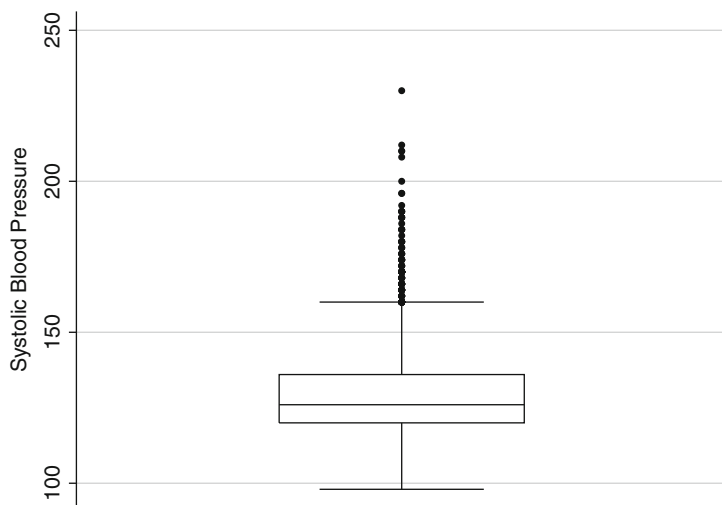


Fig. 2.3 Boxplot of the systolic blood pressure data

A boxplot represents a compromise between a histogram and a numerical summary. The boxplot in Fig. 2.3 graphically displays information from the summary in Table 2.1, specifically the minimum, maximum, and 25th, 50th (median), and 75th percentiles. This retains many of the advantages of a graphical display while still providing fairly precise numerical summaries. The “box” displays the 25th and 75th percentiles (the lower and upper edges of the box) and the median (the line across the middle of the box). Extending from the box are the “whiskers” (this colorful terminology is due to the legendary statistician John Tukey, who liked to coin new terms). The bottom whisker extends to the minimum data value, 98, but the maximum is above the upper whisker. This is because Stata uses an algorithm to try to determine if observations are “outliers,” that is, values a large distance away from the main portion of the data. Data points considered outliers (they can be in either the upper or lower range of the data) are plotted with symbols and the whisker only extends to the most extreme observation not considered an outlier.

Boxplots convey a wealth of information about the distribution of the variable:

- Location, as measured by the median
- Spread, as measured by the height of the box (this is called the interquartile range or IQR)
- Range of the observations
- Presence of outliers
- Some information about shape

This last point bears further explanation. If the median is located toward the bottom of the box, then the data are *right-skewed* toward larger values. That is, the

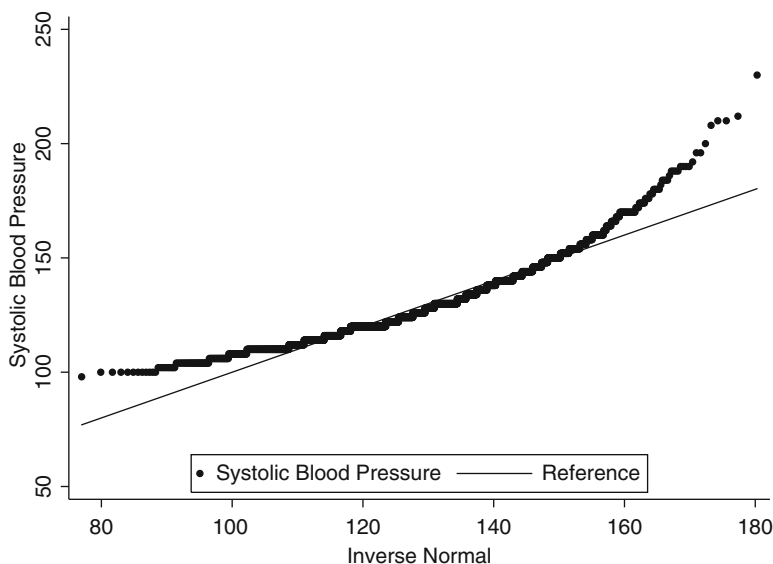


Fig. 2.4 Normal Q–Q plot of the systolic blood pressure data

distance between the median and the 75th percentile is greater than that between the median and the 25th percentile. Likewise, right-skewness will be indicated if the upper whisker is longer than the lower whisker or if there are more outliers in the upper range. Both the boxplot and the histogram show evidence for right-skewness in the SBP data. If the direction of the inequality is reversed (more outliers on the lower end, longer lower whisker, median toward the top of the box), then the distribution is *left-skewed*.

Our final graphical technique, the normal Q–Q plot, is useful for comparing the frequency distribution of the data to a normal distribution. Since it is easy to distinguish lines that are straight from ones that are not, a normal Q–Q plot is constructed so that the data points fall along an approximately straight line when the data are from a normal distribution, and deviate *systematically* from a straight line when the data are from other distributions. Figure 2.4 shows the Q–Q plot for the SBP data. The line of the data points shows a distinct curvature, indicating the data are from a nonnormal distribution.

The shape and direction of the curvature can be used to diagnose the deviation from normality. Upward curvature, as in Fig. 2.4, is indicative of right-skewness, while downward curvature is indicative of left-skewness. The other two common patterns are S-shaped. An S-shape as in Fig. 2.5 indicates a *heavy-tailed* distribution, while an S-shape like that in Fig. 2.6 is indicative of a *light-tailed* distribution.

Heavy- and light-tailed are always in reference to a hypothetical normal distribution with the same spread. A heavy-tailed distribution has more observations in

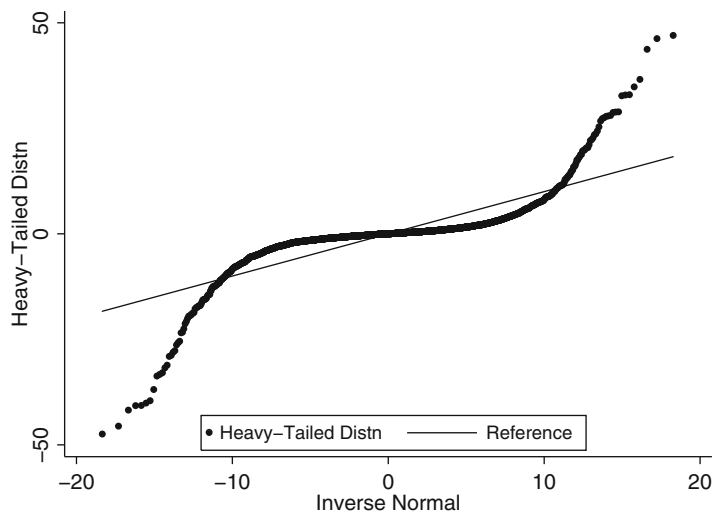


Fig. 2.5 Normal Q-Q plot of data from a heavy-tailed distribution

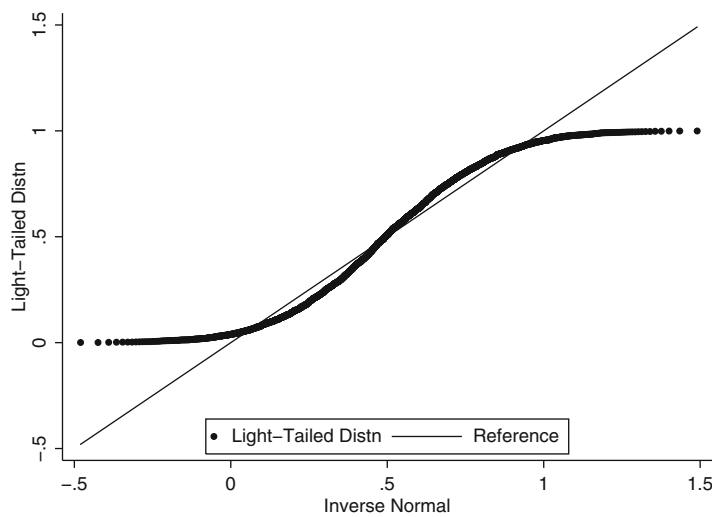


Fig. 2.6 Normal Q-Q plot of data from a light-tailed distribution

the middle of the distribution and way out in the tails, and fewer a modest way from the middle (simply having more in the tails would just mean a larger spread). Light-tailed means the reverse: fewer in the middle and far out tails and more in the mid-range. Heavy-tailed distributions are generally more worrisome than light-tailed since they are more likely to include outliers.

Table 2.2 Effect of a $\log_{10}$ transformation

Value	Difference	$\log_{10}$ value	Difference
0.01	0.09	-2	1
0.1	0.9	-1	1
1	9	0	1
10	90	1	1
100	900	2	1
1,000	–	3	–

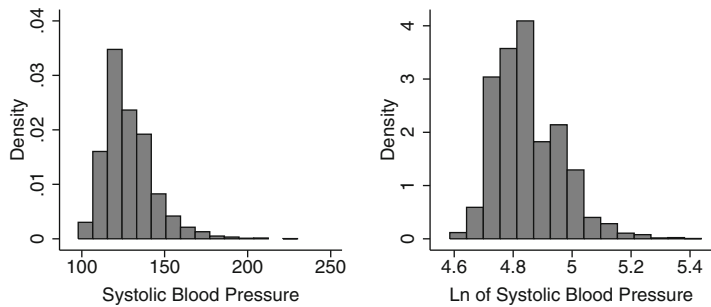


Fig. 2.7 Histograms of systolic blood pressure and its natural logarithm

2.3.1.4 Transformations of Data

A number of the techniques we describe in this book require the assumption of approximate normality or, at least, work better when the data are not highly skewed or heavy-tailed, and do not include extreme outliers. A common method for dealing with these problems is to transform such variables. For example, instead of the measured values of SBP, we might instead use the logarithm of SBP. We first consider why this works and then some of the advantages and disadvantages of transformations.

Transformations affect the distribution of values of a variable because they emphasize differences in a certain range of the data, while de-emphasizing differences in others. Consider a table of transformed values, as displayed in Table 2.2. On the original scale the difference between 0.01 and 0.1 is 0.09, but on the $\log_{10}$ scale, the difference is 1. In contrast, the difference between 100 and 1,000 on the original scale is 900, but this difference is also 1 on the $\log_{10}$ scale. So a log transformation de-emphasizes differences at the upper end of the scale and emphasizes those at the lower end. This holds for the natural log as well as $\log_{10}$ transformation. The effect can readily be seen in Fig. 2.7, which displays histograms of SBP on the original scale and after natural log transformation.

The log-transformed data is distinctly less right-skewed, even though some skewness is still evident. Essentially, we are viewing the data on a different scale of measurement.

There are a couple of other reasons to consider transforming variables, as we will see in later sections and chapters: transformations can simplify the relationships

Table 2.3 Frequencies of behavior patterns

tabulate behpat			
behavioral			
pattern (4			
level)		Freq.	Percent
	-----		Cum.
A1		264	8.37
A2		1325	42.01
B3		1216	38.55
B4		349	11.07
	-----		100.00
Total		3154	100.00

between variables (e.g., by making a curvilinear relationship linear), can remove interactions, and can equalize variances across subgroups that previously had unequal variances.

A primary objection to the use of transformations is that they make the data less interpretable. After all, who thinks about medical costs in log dollars? In situations where there is good reason to stay with the original scale of measurement (e.g., dollars), we may prefer alternatives to transformation including GLMs and weighted analyses. Or we may appeal to the robustness of normality-based techniques: many perform extremely well even when used with data exhibiting fairly serious violations of the assumptions.

In other situations, with a bit of work, it is straightforward to express the results on the original scale when the analysis has been conducted on a transformed scale. For example, Sect. 4.7.5 gives the details for log transformations in linear regression.

A compromise when the goal is, for example, to test for differences between two arms in a clinical trial is to plan ahead to present basic descriptive statistics in the original scale, but perform tests on a transformed scale more appropriate for statistical analysis. After all, a difference on the transformed scale is still a difference between the two arms.

Finally, we remind the reader that different scales of measurement just take a bit of getting used to: consider pH.

2.3.2 Categorical Variables

Categorical variables require a different approach, since they are less amenable to graphical analyses and because common statistical summaries, such as mean and standard deviation, are inapplicable. Instead we use tabular descriptions. Table 2.3 gives the frequencies, percents, and cumulative percents for each of the behavior pattern categories for the WCGS data. Note that cumulative percentages are really only useful with ordinal categorical data (why?).

When tables are generated by the computer, there is usually little latitude in the details. However, when tables are constructed by hand, thought should be given to their layout; [Ehrenberg \(1981\)](#) is recommended reading. Three easy-to-follow

Table 2.4 Characteristics of top medical schools

School	Rank	NIH research (\$10 millions)	Tuition (\$ thousands)	Average MCAT
Harvard	1	68	30	11.1
Johns Hopkins	2	31	29	11.2
Duke	3	16	31	11.6
Penn	4(Tie)	33	32	11.7
Washington U.	4(Tie)	25	33	12.0
Columbia	6	24	33	11.7
UCSF	7	24	20	11.4
Yale	8	22	30	11.1
Stanford	9(Tie)	19	30	11.1
Michigan	9(Tie)	20	29	11.0

Source: US News and World Report (<http://www.usnews.com>, 12/6/01)

suggestions from that article are to arrange the categories in a meaningful way (e.g., not alphabetically), report numbers to two effective digits, and to leave a gap every three or four rows to make it easier to read across the table. Table 2.4 illustrates these concepts. With the table arranged in order of the rankings, it is easy to see values that do not follow the pattern predicted by rank, for example, out-of-state tuition.

2.4 Two-Variable Descriptions

Most of the rest of this book is about the relationships among variables. An example from the WCGS is whether behavior pattern is related to SBP. In investigating the relationships between variables, it is often useful to distinguish the role that the variables play in an analysis.

2.4.1 Outcome Versus Predictor Variables

A key distinction is whether a variable is being predicted by the remaining variables, or whether it is being used to make the prediction. The variable singled out to be predicted from the remaining variables we will call the *outcome variable*; alternate and interchangeable names are *response variable* or *dependent variable*. The variables used to make the prediction will be called *predictor variables*. Alternate and equivalent terms are *covariates* and *independent variables*. We slightly prefer the outcome/predictor combination, since the term *response* conveys a cause-and-effect interpretation, which may be inappropriate, and *dependent/independent* is confusing with regard to the notion of statistical independence. (“Independent variables do not have to be independent” is a true statement!).

Table 2.5 Correlation coefficient for systolic blood pressure and weight

```
. correlate sbp weight (obs=3154)
```

	sbp	weight
sbp	1.0000	
weight	0.2532	1.0000

In the WCGS example, we might hypothesize that change in behavior pattern (which is potentially modifiable) might cause change in SBP. This would lead us to consider SBP as the outcome and behavior pattern as the predictor.

2.4.2 Continuous Outcome Variable

As before, it is useful to consider the nature of the outcome and predictor variables in order to choose the appropriate descriptive technique. We begin with continuous outcome variables, first with a continuous predictor and then with a categorical predictor.

2.4.2.1 Continuous Predictor

When both the predictor and outcome variables are continuous, the typical numerical description is a correlation coefficient and its graphical counterpart is a scatterplot. Again considering the WCGS study, we will investigate the relationship between SBP and weight.

Table 2.5 shows the Stata command and output for the correlation coefficient, while Fig. 2.8 shows a scatterplot. Both the graph and the numerical summary confirm the same thing: there is a weak association between the two variables, as measured by the correlation of 0.25. The graph conveys important additional information. In particular, there are quite a few outliers, including an especially anomalous data point with high blood pressure and the lowest weight in the dataset.

The Pearson correlation coefficient r , more fully described in Sect. 3.2, is a scale-free measure of association that does not depend on the units in which either SBP or weight is measured. The correlation coefficient varies between -1 and 1 , and correlations of absolute value 0.7 or larger are considered strong associations in many contexts. In fields where data are typically noisy, including our SBP example, much smaller correlations may be considered meaningful.

It is important to keep in mind that the Pearson correlation coefficient only measures the strength of the *linear* relationship between two variables. To determine whether the correlation coefficient is a reasonable numerical summary of the association, a graphical tool that helps to assess linearity in the scatterplot is a *scatterplot smoother*. Figure 2.9 shows a scatterplot smooth superimposed on the

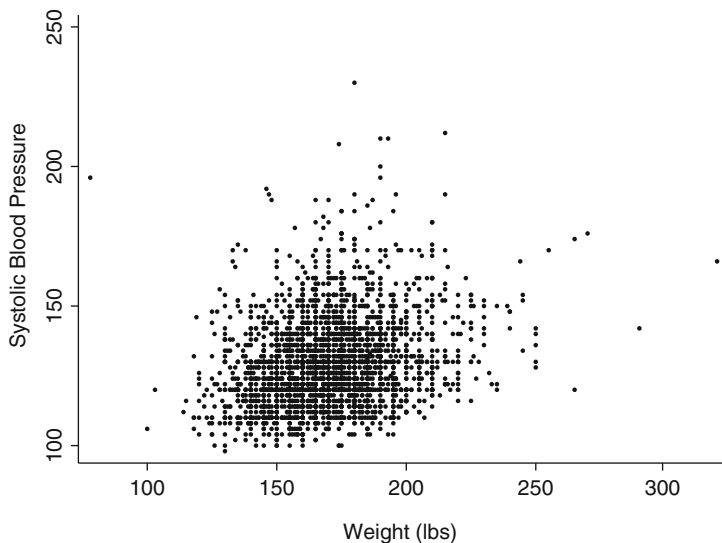


Fig. 2.8 Scatterplot of systolic blood pressure versus weight

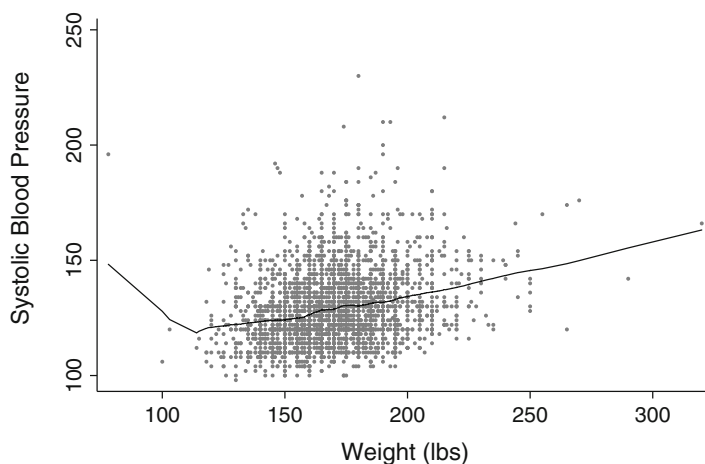


Fig. 2.9 LOWESS smooth of systolic blood pressure versus weight

graph of SBP versus weight. The figure was generated by the Stata command `lowess sbp weight, bw(0.25)` (with a few embellishments to make it look nicer). This uses the LOWESS technique to draw a smooth (but not necessarily straight) line representing the average value of the variable on the y -axis as a function of the variable on the x -axis. LOWESS is short for LOcally WEighted Scatterplot Smoother. The `bw(0.25)` option specifies that for estimation of the height of the curve at each point, 25% of the data nearest that point should be used. This is all just a fancy way of drawing a flexible curve through a cloud of points.

Table 2.6 Summary data for systolic blood pressure by behavior pattern

. bysort behpat: summarize sbp

-> behpat = A1

Variable	Obs	Mean	Std. Dev.	Min	Max
sbp	264	129.2462	15.29221	100	200

-> behpat = A2

Variable	Obs	Mean	Std. Dev.	Min	Max
sbp	1325	129.8891	15.77085	100	212

-> behpat = B3

Variable	Obs	Mean	Std. Dev.	Min	Max
sbp	1216	127.5551	14.78795	98	230

-> behpat = B4

Variable	Obs	Mean	Std. Dev.	Min	Max
sbp	349	127.1547	13.10125	102	178

Figure 2.9 shows that the relationship between SBP and weight is very close to linear. The small upward bend at the far left of the graph is mostly due to the outlying observation at the lowest weight and is a warning as to the instability of LOWESS (or any scatterplot smoother) at the edges of the data.

Choice of bandwidth is somewhat subjective. Small bandwidths like 0.05 often give very bumpy curves, which are hard to interpret. At the other extreme, bandwidths too close to one force the curve to be practically a straight line, obviating the advantage of using a scatterplot smoother. See Problem 2.6.

2.4.2.2 Categorical Predictor

With a continuous outcome and a categorical predictor, the usual strategy is to apply the same numerical or graphical methods used for one-variable descriptions of a continuous outcome, but to do so separately within each category of the predictor. As an example, we describe the distribution of SBP in WCGS, within levels of behavior pattern. Table 2.6 shows the most direct way of doing this in Stata. Alternatively, the `table` command can be used to make a more compact display, with command options controlling which statistics are listed. The results are shown in Table 2.7.

Side-by-side boxplots, as shown in Fig. 2.10, are an excellent graphical tool for examining the distribution of SBP in each of the behavior pattern categories and

Table 2.7 Descriptive statistics for systolic blood pressure by behavior pattern

```
. table behpat, contents(mean sbp sd sbp min sbp max sbp)
```

Behavioral Pattern	mean(sbp)	sd(sbp)	min(sbp)	max(sbp)
A1	129.2462	15.29221	100	200
A2	129.8891	15.77085	100	212
B3	127.5551	14.78795	98	230
B4	127.1547	13.10125	102	178

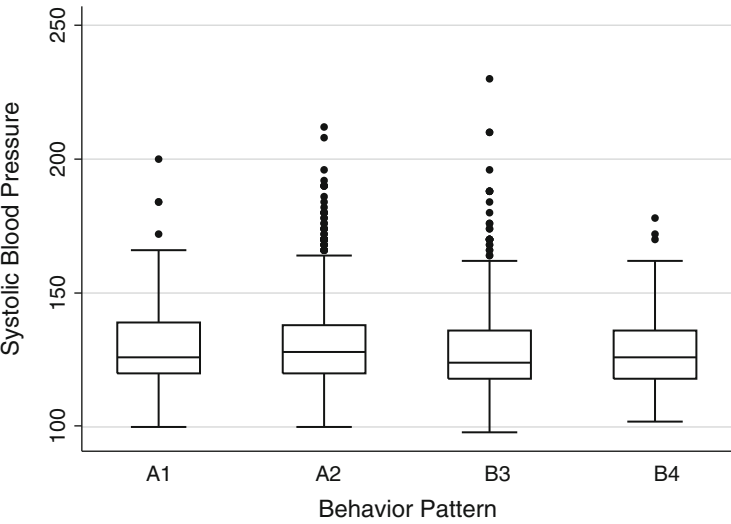


Fig. 2.10 Boxplots of systolic blood pressure by behavior pattern

making comparisons among them. The four boxplots are quite similar. They each have about the same median, interquartile range, and a slight right-skewness. At least on the basis of this figure, there appears to be little relationship between SBP and behavior pattern.

2.4.3 Categorical Outcome Variable

With a categorical outcome variable, the typical method is to tabulate the outcome within levels of the predictor variable. To do so first requires breaking any continuous predictors into categories. Suppose, for example, we wished to treat behavior pattern as the outcome variable and weight as the predictor. We might first divide weight into four categories: ≤ 140 pounds, $> 140\text{--}170$, $> 170\text{--}200$, and > 200 . As with histograms, we need enough categories to avoid loss of information, without

Table 2.8 Behavior pattern by weight category

. tabulate behpat wghtcat, column

behavioral pattern (4 level)	wghtcat				Total
	< 140	140-170	170-200	> 200	
A1	20 8.62	125 8.13	98 8.37	21 9.86	264 8.37
A2	100 43.10	612 39.79	514 43.89	99 46.48	1325 42.01
B3	90 38.79	610 39.66	443 37.83	73 34.27	1216 38.55
B4	22 9.48	191 12.42	116 9.91	20 9.39	349 11.07
Total	232 100.00	1538 100.00	1171 100.00	213 100.00	3154 100.00

defining categories that include too few observations. Familiar clinical categories are often useful (e.g., glucose <110, 110–125, >125). In Table 2.8, we have requested percentages for each column to facilitate the comparison of the percentages in each behavior pattern between the weight categories. Row percentages or percentages out of the total of 3,154 could also have been requested.

In choosing cutoff points for categorical variables, it is entirely fair to look at the distribution of that variable to try to obtain, for example, roughly equal sample sizes in each of the categories. Splitting the data into 3, 4, 5, or 10 groups of equal size is a common approach. However, fishing for cutpoints that prove a point is an easy way to arrive at misleading conclusions.

A different strategy with a categorical outcome and a continuous predictor is to “turn the problem around” and treat the continuous variable as the outcome, using the methods of the previous section. If the only goal is to determine whether the two variables are associated, this may suffice. But when the categorical variable is clearly the outcome, this may lead to awkward models and hard-to-interpret conclusions.

2.5 Multivariable Descriptions

Description of more than two or three variables simultaneously quickly becomes difficult. One approach is to look at pairwise associations, e.g., for categorical variables, looking at a series of two-way tables, taking each pair of variables in turn. If a number of the variables are continuous, a correlation matrix (giving all the pairwise correlations) or a scatterplot matrix (giving all the pairwise plots) can be generated. Table 2.9 and Fig. 2.11 show these for the variables SBP, age, weight, and height. The correlation matrix shows that SBP is very weakly correlated with age and weight and essentially uncorrelated with height.

Table 2.9 Correlation matrix for systolic blood pressure, age, weight, and height

```
. correlate sbp age weight height (obs=3154)
```

	sbp	age	weight	height
sbp	1.0000			
age	0.1657	1.0000		
weight	0.2532	-0.0344	1.0000	
height	0.0184	-0.0954	0.5329	1.0000

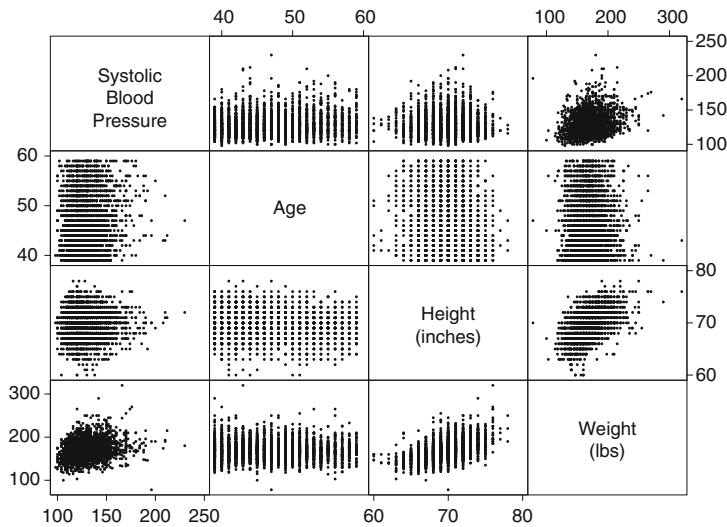


Fig. 2.11 Scatterplot matrix of systolic blood pressure, age, weight, and height

The scatterplot matrix supports the correlation calculation. If one of the variables is clearly the outcome variable, it is useful to list this variable first in the command. That way the first row of the matrix shows the outcome variable on the *y*-axis, plotted against each of the predictor variables on the *x*-axis. The matrix of scatterplots for these four variables additionally displays the modest positive correlation between weight and height, indicating the people come in all sizes and shapes!

Multi-way tables that go beyond pairwise relationships can be generated with multiple categorical variables. For example, Table 2.10 shows whether or not the subject had a coronary event (*chd69*), by behavior pattern within weight category. Options in the Stata command are used to obtain the row and column totals. With some study, it is possible to extract information from this three-way table, but it is more difficult than with a one- or two-way table. An advantage of a three-way table is the ability to assess *interaction*, the topic of Sect. 4.6. That is, is the relationship between CHD and behavior pattern the same for each weight category?

Table 2.10 CHD events and behavior pattern by weight category

. table chd69 behpat wghtcat, row col

CHD event	wghtcat and behavioral pattern (4 level)									
	< 140					140-170				
	A1	A2	B3	B4	Total	A1	A2	B3	B4	Total
no	18	93	84	22	217	115	559	582	184	1,440
yes	2	7	6		15	10	53	28	7	98
Total	20	100	90	22	232	125	612	610	191	1,538

CHD event	wghtcat and behavioral pattern (4 level)									
	170-200					> 200				
	A1	A2	B3	B4	Total	A1	A2	B3	B4	Total
no	81	438	422	108	1,049	20	87	67	17	191
yes	17	76	21	8	122	1	12	6	3	22
Total	98	514	443	116	1,171	21	99	73	20	213

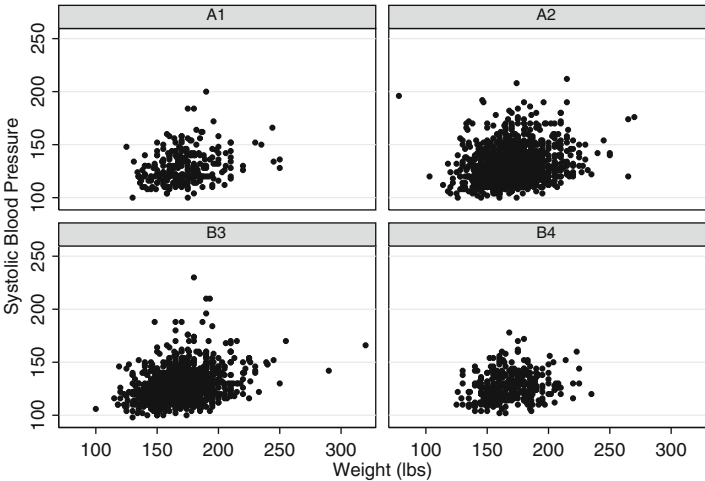


Fig. 2.12 Scatterplot of SBP versus weight by behavior pattern

Analogous graphical displays are also possible. For example, we could look at the relationship between SBP and weight separately by behavior pattern, as displayed in Fig. 2.12. This indicates that the relationship seems to be the same for each behavior pattern, indicating a lack of interaction.

2.6 Summary

Exploratory summaries and graphs are a crucial first step in any data analysis. They provide an opportunity to uncover unusual or anomalous data points which may affect the analysis. Summaries and graphs uncover properties of the data (for instance, skewness) which are useful for informing which model families may fit the data best. Finally, exploring the strength of relationships between variables through graphs provides compelling summaries of the relationships as well as guidance for building regression models.

2.7 Problems

Problem 2.1. Classify each of the following variables as numerical or categorical. Then further classify the numerical variables as continuous or discrete, and the categorical variables as ordinal or nominal.

- (1) Gender
- (2) Race
- (3) Age (in years)
- (4) Age in categories (0–20, 21–35, 36–45, 45–60, 60–85, 85+)
- (5) Zipcode
- (6) Toxicity (mild, moderate, life-threatening, dead)
- (7) Number of hospitalizations in the past year
- (8) Change in HIV-RNA
- (9) Weeks on treatment
- (10) Treatment (placebo versus estrogen)

Problem 2.2. Generate pseudo-random data from a normal distribution using a computer program or statistics package. In Stata, this can be done using the `generate` command and the function `invnorm(uniform())`. Now generate a normal Q–Q plot for these data. Do this for several samples of size 10, 50, and 200. How well do the Q–Q plots approximate straight lines? This is valuable practice for judging how well an actual dataset can be expected to approximate a straight line.

Problem 2.3. Generate pseudo-random samples of size 50 from a normal distribution (see Problem 2.2 for how to do this in Stata). Construct histograms of the data using 5, 7, and 15 bins. What do you notice? Do the shapes look like a normal distribution?

Problem 2.4. Warfarin is a drug used to prevent blood clots, for example in patients with irregular heartbeat and after heart surgery. However, too much warfarin can cause unusual bleeding or bruising, so calibration of the dose is important. A study contrasting calibration times (in hours) in two ethnic groups had the following results. For the sample of 19 Caucasians, the times were 2, 4, 6, 7, 8, 9, 10, 10,

12, 14, 16, 19, 21, 24, 26, 30, 35, 44, and 70; for the 18 Asian–Americans, the times were 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 12, 19, and 32.

- (1) Display the data numerically to compare the two ethnic groups.
- (2) Display the data graphically to compare the two ethnic groups.
- (3) Describe the distribution of the data within ethnic group.
- (4) Log transform the data and repeat the graphical display. How do the displays with and without log transformation compare?
- (5) Can you think of other variables you might want to adjust to help understand the ethnic differences better?

Problem 2.5. The timing of various stages in the contraction of the heart, determined by electro-cardiogram (EKG), can be used to diagnose heart problems. A commonly measured time interval in the contraction of the ventricles is the so-called QRS wave. A study was conducted to see if longer QRS times were related to the ability to induce rapid heart rhythms (called inducible ventricular tachycardia or IVT), which have been associated with adverse outcomes. In a study of 53 subjects, the 18 with IVT had QRS times (in milliseconds) of 70, 75, 86, 90, 96, 102, 110, 114, 116, 117, 120, 130, 136, 142, 145, 152, 170, and 182. The 35 patients without IVT had QRS times of 40, 50, 65, 70, 76, 78, 80, 82, 85, 88, 88, 89, 90, 94, 95, 96, 98, 98, 100, 102, 105, 107, 109, 110, 114, 115, 120, 125, 130, 135, 138, 150, 165, 170, and 180.

- (1) Display the data numerically to help understand whether QRS time is related to IVT.
- (2) Display the data graphically to help understand whether QRS time is related to IVT.
- (3) QRS time is commonly considered as abnormal if the value is greater than 120 ms. Generate a numerical display to help understand if abnormal QRS is related to IVT.
- (4) What are the advantages and disadvantages of treating QRS as binary (above 120 ms) instead of continuous?

Problem 2.6. Using the WCGS dataset, generate a LOWESS (or equivalent) scatterplot smooth of SBP versus weight, comparable to Fig. 2.9. Next try the plot with bandwidths of 0.05, 0.15, and 0.50. How do they compare? Which is most useful for judging the linearity or lack of linearity of the relationship? The WCGS data are available at <http://www.biostat.ucsf.edu/vgsm>.

Chapter 3

Basic Statistical Methods

Statistical analyses involving multiple predictors are generalizations of simpler techniques developed for investigating associations between outcomes and single predictors. Although many of these should be familiar from basic statistics courses, we review some of the key ideas and methods here as background for the methods covered in the rest of the book and to introduce some basic notation.

Sections 3.1–3.3 review basic methods for continuous outcomes, including the t -test and one-way ANOVA, the correlation coefficient and the linear regression model for a single predictor. Section 3.4 focuses on contingency table methods for investigating associations between binary outcomes and categorical predictors, including a discussion of basic measures of association. Section 3.5 introduces descriptive methods for survival time outcomes, including Kaplan–Meier survival curves and the logrank test. In Sect. 3.6, we introduce the use of the bootstrap as a method to obtain CIs for estimates in situations where traditional methods are inappropriate. Finally, Sect. 3.7 discusses the importance of properly interpreting negative findings from statistical analyses, focusing on the use of point estimates and CIs rather than P -values.

3.1 t -Test and Analysis of Variance

The t -test and one-way ANOVA are basic tools for assessing the statistical significance of differences between the average values of a continuous outcome across two or more samples. Both the t -test and one-way ANOVA can be seen as methods for assessing the association of a categorical predictor—binary in the case of the t -test, with more than two levels in the case of one-way ANOVA—with a continuous outcome. Both are based in statistical theory for normally distributed outcomes, but work well for many other types of data; and both turn out to be special cases of linear regression models.

Table 3.1 *t*-Test of difference in average glucose by exercise

. t-test glucose if diabetes == 0, by(exercise)

Two-sample t-test with equal variances

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
no	1191	97.36104	.2868131	9.898169	96.79833	97.92376
yes	841	95.66825	.3258672	9.450148	95.02864	96.30786
combined	2032	96.66043	.2162628	9.74863	96.23631	97.08455
diff		1.692789	.4375862		.8346243	2.550954

Degrees of freedom: 2030

Ho: mean(no) - mean(yes) = diff = 0

Ha: diff < 0

Ha: diff != 0

Ha: diff > 0

t = 3.8685

t = 3.8685

t = 3.8685

P < t = 0.9999

P > |t| = 0.0001

P > t = 0.0001

3.1.1 *t*-Test

The basic *t*-test is used in comparing two independent samples. The *t*-statistic on which the test is based is the difference between the two sample averages, divided by the standard error of that difference. The *t*-test is designed to work in small samples, whereas *Z*-tests are not. Table 3.1 shows the result of a *t*-test comparing average fasting glucose levels among women without diabetes, according to exercise. This is the first of many examples in Chaps. 3 and 4 using data from the heart and estrogen/progestin study (HERS), a clinical trial of hormone therapy (HT) for prevention of recurrent heart attacks and death among 2,763 post-menopausal women with existing coronary heart disease (CHD) (Hulley et al. 1998). Average glucose is 97.4 mg/dL among the 1,191 women who do not exercise as compared to 95.7 mg/dL among the 841 women who do. The difference of 1.7 mg/dL is statistically significant ($P = 0.0001$) in the two-sided test shown in the column headed `Ha: diff != 0` (`!=` is Stata notation for “not equal to.”) The *P*-value gives the probability—under the null hypothesis that mean glucose levels are the same in the two populations being compared—of observing a *t*-statistic more extreme, or larger in absolute value, than the observed value.

3.1.2 One- and Two-Sided Hypothesis Tests

In clinical research, unlike some other areas of science, two-sided hypothesis tests are almost always used. In the two-sided *t*-test, we are testing the null hypothesis (H_0) of equal population means against the alternative hypothesis (H_a) that the one

mean is either smaller or larger than the other. The two-sided test is appropriate, for example, when a new treatment might turn out to be beneficial *or* to have adverse effects.

In contrast, only one of these alternatives is considered in a one-sided test. As a result, the smaller of the one-sided P -values is half the magnitude of the two-sided P -value. The resulting advantage of the one-sided test is that at a given significance level, less evidence in favor of the alternative hypothesis is required to reject the null. For example, using a one-sided test in a sample of 100 observations, we would declare statistical significance at the 5% level if the t -statistic exceeds 1.66; using a two-sided test it would need to exceed 1.98 (in absolute value). A direct benefit is that a somewhat smaller sample size is required when a study is designed to be analyzed using a one-sided test.

Use of a one-sided test is sometimes motivated by prior information that makes only one of the alternatives of interest. An example might be in testing an existing treatment known to be safe for evidence of benefit on a new endpoint. One-sided tests are also used in *noninferiority* trials comparing a new to a standard treatment; in this setting the alternative hypothesis is that the new treatment performs almost as well or better than the standard treatment, as against the null hypothesis of clearly performing worse.

However, in part because they make it possible to reject the null hypothesis on weaker evidence, one-sided tests are not commonly used in clinical research. Even in noninferiority trials where one-sided tests are clearly appropriate, a standard text on the conduct of clinical trials (Friedman et al.1998) recommends that the tests be carried out at a significance level of 2.5%. Thus to claim noninferiority, the same strength of evidence would be required as in a two-sided test. Furthermore, Fleiss (1988) argues that the other alternative *ought* generally to be of interest, and that in treatment trials adverse effects can rarely be ruled out with sufficient certainty to justify a one-sided test. We endorse this conservative view, and recommend using two-sided tests unless a one-sided test is strongly motivated by specific reasons.

The Stata `t-test` command gives P -values for both one-sided test as well as the two-sided test. In Table 3.1, the one-sided P -value on the right ($H_a: \text{diff} > 0$) gives the probability (again, under the null hypothesis) of observing a t -statistic larger than the observed value, while the one on the left ($H_a: \text{diff} < 0$) gives the probability of observing one that is smaller. In this example, there is strong evidence ($P = 0.0001$) that the mean glucose level is higher in the population of women who do not exercise, as compared to those who do, and essentially no evidence ($P = 1.0$) that it is smaller.

3.1.3 Paired *t*-Test

The paired t -test is for use in settings where individuals or observations are linked across the two samples. Examples include measurements taken at two time points on the same individuals, or on other naturally linked pairs, as in a clinical trial where

one eye is treated and the other serves as a control. In this case, the two samples are not independent and failure to take account of the pairwise relationships wastes information and is potentially erroneous.

The paired t -test procedure first computes the pairwise differences for each individual or linked pair. In the first example, this is the change in the outcome from the first time point to the second, and in the second, the difference between the outcomes for the treated and control eyes. Then a t -test is used to assess whether the population mean of these paired differences differs from zero. An increase in power results because between-individual variability is eliminated in the first step. The paired t -test is also implemented using the `t-test` command in Stata. The more complicated case where we want to examine the influence of some other factor on within-individual changes is covered in Sect. 7.3.

3.1.4 One-Way Analysis of Variance

Suppose that we need to compare sample averages across the arms of a clinical trial with multiple treatments, or more generally across more than two independent samples. For this purpose, one-way ANOVA and the F -test take the place of the t -test. The F -test, presented in more detail in Sect. 4.3, assesses the null hypothesis that the mean value of the outcome is the same across all the populations sampled from, against the alternative that the means differ in at least two of the populations. For example, the one-way ANOVA shown in Table 3.2, the F -test for Between groups ($P = 0.0371$), suggests that mean SBP differs by ethnicity in the population represented in the HERS cohort.

3.1.5 Pairwise Comparisons in ANOVA

The statistically significant F -test in the one-way ANOVA indicates the overall importance of ethnicity for predicting SBP. In addition, Stata implements the Bonferroni, Scheffé, and Sidak procedures for assessing the statistical significance of all possible pairwise differences between groups, without inflation of the overall or family-wise type-I error rate (FER), which can arise from testing multiple null hypotheses. These and other methods for controlling the FER are discussed in Sects. 4.3.4 and 13.4.1. All three methods implemented in the `oneway` command show that the difference in average SBP between the African American and white groups is statistically significant after correction for multiple comparisons, but that the other pairwise differences are not; we show the Scheffé result.

Table 3.2 One-way ANOVA assessing differences in SBP by ethnicity

. oneway sbp ethnicity, tabulate scheffe

ethnicity	Summary of systolic blood pressure		
	Mean	Std. Dev.	Freq.
White	134.78376	18.831686	2451
Afr Amer	138.23394	19.992518	218
Other	135.18085	21.259767	94
Total	135.06949	19.027807	2763

Source	Analysis of Variance			F	Prob > F
	SS	df	MS		
Between groups	2384.26992	2	1192.13496	3.30	0.0371
Within groups	997618.388	2760	361.455938		
Total	1000002.66	2762	362.057443		

Comparison of systolic blood pressure by ethnicity
(Scheffe)

Row Mean-		
Col Mean	White	Afr-Amer
Afr-Amer	3.45018	
	0.037	
Other	.397089	-3.05309
	0.980	0.429

3.1.6 *Multi-way ANOVA and ANCOVA*

Multi-way ANOVA is an extension of the one-way procedure to deal simultaneously with more than one categorical predictor, while analysis of covariance (ANCOVA) is commonly defined as an extension of ANOVA that includes continuous as well as categorical predictors. The *t*- and *F*-tests retain their central importance in these procedures. However, one-way ANOVA and the *t*-test implicitly estimate the different population means by the sample averages; in contrast, the population means in multi-way ANOVA and ANCOVA are usually *modeled*. Thus these procedures are most easily understood as multipredictor linear regression models, which are covered in Chap. 4.

3.1.7 *Robustness to Violations of Normality Assumption*

The *t*- and *F*-tests are fairly robust to violations of the normality assumption, especially in larger samples. By robust we mean that the type-I error rate, or probability of rejecting the null hypothesis when it holds, is not seriously affected. They are primarily sensitive to outliers, which tend to decrease efficiency and make it harder to detect real differences between groups. Thus the effect is conservative,

in the sense of making it more likely that we will accept the null hypothesis when some real difference exists.

Large samples reduce sensitivity of the t -test to the assumption that the outcome is normally distributed because the distribution of the difference between the sample averages, which directly underlies the test, converges to a normal distribution even when the outcome itself has some other distribution. If violations of the normality assumption are mild to moderate, samples of 50–100 may be large enough, in particular with equal group sizes, but considerably larger samples might be needed with severe violations. Analogous large-sample behavior holds for the regression coefficients estimated in multipredictor linear models as well as the other regression models that are the primary topic of this book.

3.1.8 *Nonparametric Alternatives*

One commonly recommended solution for violations of the normality assumption is to use nonparametric Wilcoxon rank-sum or Kruskal–Wallis tests rather than the t -test or one-way ANOVA. Two other nonparametric methods are discussed below in Sect. 3.2 on the correlation coefficient.

While they avoid specific parametric distributional (i.e., normality) assumptions, these methods are not assumption-free. For example, the Wilcoxon and Kruskal–Wallis tests are based on the assumption that the outcome distributions being compared differ in *location* (mean and/or median) but not in *scale* (variance) or *shape*, as might be captured by a histogram, and can give misleading results if these assumptions are violated. Furthermore, these two tests do not provide an interpretable measure of the strength of the association. More generally, nonparametric methods sometimes result in loss of efficiency, and do not easily accommodate multiple predictors, unlike the regression methods which are the focus of this book.

Nonparametric tests are most useful for unadjusted between-group comparisons where the P -value is of primary interest, in particular for variables with skewed distributions that cannot be normalized by transformation, or outliers that must be retained for substantive reasons.

3.1.9 *Equal Variance Assumption*

When sample sizes are unequal, the t -test is less robust to violations of the assumption of equal variance across samples than to violations of normality. Violations of this assumption can seriously affect the type-I error rate, not always in a conservative direction, and large samples do not make the test any more robust. In contrast, the overall F -test in ANOVA loses efficiency, but the type-I error rate is generally not increased. However, subsequent pairwise comparisons using t -tests remain vulnerable.

Table 3.3 *t*-Test allowing for unequal variances

. t-test glucose if diabetes == 0, by(exercise) unequal

Two-sample t-test with unequal variances

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
no	1191	97.36104	.2868131	9.898169	96.79833	97.92376
yes	841	95.66825	.3258672	9.450148	95.02864	96.30786
combined	2032	96.66043	.2162628	9.74863	96.23631	97.08455
diff		1.692789	.4341096		.8413954	2.544183

Satterthwaite's degrees of freedom: 1858.33

Ho: mean(no) - mean(yes) = diff = 0

Ha: diff < 0

Ha: diff != 0

Ha: diff > 0

t = 3.8995

t = 3.8995

t = 3.8995

P < t = 1.0000

P > |t| = 0.0001

P > t = 0.0000

In the two-sample case, this problem is easily addressed using a version of the *t*-test for unequal variances. This is based on a modified estimate of the standard error of the difference in sample averages. In the analysis shown in Table 3.1, the standard deviation of glucose is 9.9 mg/dL among women who do not exercise, as compared to 9.5 mg/dL among the women who do. In this case, the re-analysis allowing for unequal variances, shown in Table 3.3, gives qualitatively the same result ($P = 0.0001$). We recommend systematic use of this version of the *t*-test, since the increase in robustness comes at very little cost in efficiency. Analogous extensions of ANOVA in which the variance is allowed to vary by group are also possible, though not implemented in the Stata `one-way` or `anova` commands.

3.2 Correlation Coefficient

The Pearson correlation coefficient r is a scale-free measure of linear association between two variables x and y , and is defined as follows:

$$r(x, y) = \frac{\text{Cov}(x, y)}{\text{SD}(x)\text{SD}(y)}$$
$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n - 1)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)}}. \tag{3.1}$$

In (3.1), $\text{Cov}(x, y)$ is the sample covariance of x and y , $\bar{x}$ and $\bar{y}$ are their sample means, $\text{SD}(x)$ and $\text{SD}(y)$ their standard deviations, and n is the sample size. The covariance reflects the degree to which observations on the two variables differ from

their respective means in the same degree and direction. Dividing $\text{Cov}(x, y)$ by the standard deviations of x and y in (3.1) gives the correlation $r(x, y)$, which is scale-free in the sense that it always takes on values between -1 and 1 and does not vary with the units of measurement used for either variable (Problem 3.2).

The correlation coefficient is a measure of *linear* association, in a sense that will become clearer in Sect. 3.3 on the simple linear model. Values of r near zero denote the absence of linear association, while values near 1 mean that x and y increase almost in lockstep, their paired values in a scatterplot falling close to a straight line with positive slope. Correlations between -1 and zero mean that y tends to *decrease* as x increases. Note that powerful *nonlinear* associations between x and y —for example, if y is proportional to x^2 —are often consistent with correlations near zero; in the example, this can happen if $\bar{x} \approx 0$.

3.2.1 Spearman Rank Correlation Coefficient

Like the t -test (and the coefficients of the linear regression model described below), the correlation coefficient is sensitive to outliers. In this case, a robust alternative is the Spearman correlation coefficient, which is equivalent to the Pearson coefficient applied to the *ranks* of x and y . This measure of correlation also takes on values between -1 and 1 . By rank, we mean position in the ordered sequence of the values of a variable; if x takes on values $1.2, 0.5, 18.3$, and 2.7 , then the ranks of these values are $2, 1, 4$, and 3 , respectively. Thus the rank of the outlier 18.3 is only 1 unit larger than the rank of the next largest value 2.7 , the same distance that separates the ranks of any two sequential values of x , thus depriving the outlier of undue influence in estimating the correlation between x and y . Ties are handled by computing the average rank of the tied values. Ranks are used in a range of nonparametric methods, in no small part because of their robustness when the data include outliers. Their disadvantage is that any information contained in the measured values of the outcome beyond the ranks is lost.

3.2.2 Kendall's τ

Another rank-based alternative to Pearson's correlation coefficient is Kendall's τ , defined as the difference in the number of concordant and discordant pairs of data points, as a proportion of the number of evaluable pairs. In the absence of ties, the pair of data points (x_i, y_i) and (x_j, y_j) for observations i and j is concordant if $x_i > x_j$ and $y_i > y_j$, or if $x_i < x_j$ and $y_i < y_j$, and discordant otherwise. It is easy to see that we need only know the ranks of the x and y values, not their actual values, to evaluate the conditions for concordance. If the numbers of concordant and discordant pairs are about equal, then $\tau \approx 0$; essentially this means that the fact

that $x_i > x_j$ gives little information about whether $y_i > y_j$. But as the proportion of concordant pairs grows, τ approaches 1, reflecting the fact that the ordering of the x pairs is highly associated with the ordering of the y pairs. Conversely, if most pairs are discordant, then τ approaches -1 ; again, the orderings of the x and y pairs are highly associated. Kendall's τ is sometimes used as a measure of correlation for time-to-event outcomes.

3.3 Simple Linear Regression Model

Here we present the simple linear regression model with a continuous outcome and a single continuous predictor variable.

3.3.1 Systematic Part of the Model

The main purpose of this model is to determine how the average value of the continuous outcome y varies with the value of a single predictor x . The average values of the outcome are assumed to lie on a “regression line” or “line of means.” Figure 3.1 shows values of baseline SBP by age in the HERS trial of hormone therapy. To make the idea of a line of means more concrete, the square symbols in the plot show the average SBP within each decile of age. Naturally, there is some noise in these local means, although much less than in the raw data. Moreover, the continuous regression line, assumed to be linear, captures the increasing trend rather

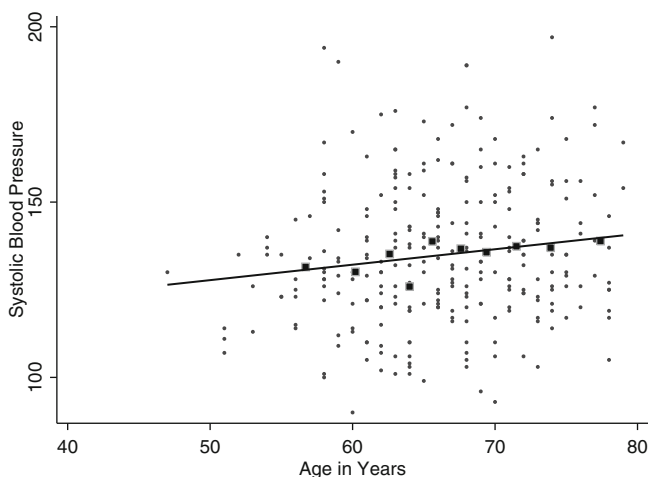


Fig. 3.1 Linear regression model for SBP and age

well. Its slope represents the systematic dependence of the outcome on the predictor, and is thus usually the focus of interest.

The formula for the regression line is simple and has interpretable parameters:

$$\begin{aligned} E[y|x] &= \text{average value of SBP for a given age} \\ &= \beta_0 + \beta_1 \text{age} \\ &= 105.7 + 0.44 \text{age}. \end{aligned} \tag{3.2}$$

In (3.2), $E[y|x]$ is shorthand for the *Expected* or average value of the outcome y at a given value of the predictor x . β_1 gives the slope of the regression line, and is interpretable as the change in average SBP for a one-year increase in age. The estimate of β_1 from the sample shown in the plot suggests that among women with heart disease, average SBP increases 0.44 mmHg for each one-year increase in age. This estimate is the best fitting value in a sense explained below in Sect. 3.3.4.

It is also easy to see that the estimate of the intercept parameter $\beta_0 = 105.7$ gives the average value of the outcome when age is zero. While not meaningless in this case, these data obviously provide no direct information about SBP at age zero. This illustrates the more general point that while regression models are often approximately true within the range of the observed data, extrapolation is usually risky. “Centering” the predictor by subtracting off a value within the range of the data can resolve this problem. One reasonable choice in this example would be the sample average age of 67; then the centered age variable would have value zero for women at age 67, and the new intercept, 135.2 mmHg, estimates average SBP among women this age. The slope estimate is unaffected by centering the age variable.

3.3.2 *Random Part of the Model*

It is also clear from Fig. 3.1 that at any given age, SBP varies considerably. Possible sources of this variability include measurement error, diurnal patterns, and a potentially broad range of unmeasured determinants of SBP, including the immediate circumstances when the measurement was made. These factors are combined in an error term ε , so that for observation i

$$\begin{aligned} \text{SBP}_i &= \text{mean SBP for subjects of age}_i + \text{error}_i \\ &= \beta_0 + \beta_1 \text{age}_i + \varepsilon_i. \end{aligned} \tag{3.3}$$

The statistical assumptions of the linear regression model concern the distribution of ε . Specifically, we assume that $\varepsilon_i \sim \text{i.i.d } \mathcal{N}(0, \sigma_\varepsilon^2)$, meaning that ε is independently and identically distributed and has a

- Normal distribution
- Mean zero at every value of age
- Constant variance σ_ε^2 at every value of age
- Values that are statistically independent

In Sect. 4.7, we will see that the first assumption may sometimes be relaxed. The second assumption is important to checking whether the relationship between a numerical predictor and the outcome is linear, as assumed in (3.2), (3.3), and Fig. 3.1; violations can be examined and repaired using methods also introduced in Sect. 4.7. The third assumption, of constant variance, is sometimes called *homoscedasticity*; data which violate this assumption are called *heteroscedastic*, and can be dealt with using methods also discussed in Sect. 4.7 as well as Chap. 8. Chapters 7 and 12 introduce methods for data where the fourth assumption, of independence, does not hold. Some examples include samples with repeated measures on individuals, cluster samples where patients are selected from within a sample of physician practices, and complex survey samples such as the national health and nutrition examination survey (NHANES).

3.3.3 Assumptions About the Predictor

In contrast to the outcome, no distributional assumptions are made about the predictor in the linear regression model. In the case of the linear model with a single continuous predictor, we do not assume that the predictor has a normal distribution, although we will see in Sect. 4.7 that outlying values of the predictor can cause trouble in some circumstances. In addition, binary, categorical, and discrete numeric variables including counts are easily accommodated as predictors in these models.

Although we do not need to make assumptions about the distribution of the predictor, these models do perform better when it is relatively variable. For example, it would be more difficult to estimate the age trend in average SBP if the sample were limited to women aged 65–70. For binary and categorical predictors, the analogous limitation is that the subgroups defined by the predictor should not be too small. The impact of the variability of the predictor, or lack of it, is reflected in the standard error of the regression coefficient, as shown below in Sect. 3.3.7.

Finally, when we want to assess the relationship of the outcome with the true values of the predictor, we effectively assume that the predictors are measured without error. This is often not very realistic, and the effects of violations are the subject of ongoing statistical research. Random measurement errors unrelated to the outcome result in attenuation of estimated slope coefficients toward zero, sometimes called *regression dilution bias* (Frost and Thompson 2000). Despite some loss of efficiency, reasonable estimation is often possible in the presence of mild-to-moderate error in the measurement of the predictors. Moreover, for prediction of new outcomes, values of the predictor measured with error may suffice.

Table 3.4 OLS regression of SBP on age

```
. reg SBP age
```

Source	SS	df	MS	Number of obs =	276
Model	2179.70702	1	2179.70702	F(1, 274) =	5.58
Residual	106991.347	274	390.47937	Prob > F =	0.0188
				R-squared =	0.0200
				Adj R-squared =	0.0164
Total	109171.054	275	396.985652	Root MSE =	19.761

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.4405286	.186455	2.36	0.019	.0734621 .8075952
_cons	105.713	12.40238	8.52	0.000	81.2969 130.129

3.3.4 Ordinary Least Squares Estimation

The model (3.3) refers to the population of women with heart disease from which the sample shown in Fig. 3.1 was drawn. The regression line in the figure is an estimate of the population regression line that was found using *ordinary least squares* (OLS). Of all the lines that could be drawn though the scatterplot of the data to represent the trend in SBP with increasing age, the OLS estimate minimizes the sum of the squared vertical differences between the data points and the line.

Since the regression line is uniquely determined by β_0 and β_1 , the intercept and slope parameters, fitting the regression model amounts to finding estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ which meet the OLS criterion. In addition to being easy to compute, these OLS estimates have desirable statistical properties. If model assumptions hold, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates of the population parameters.

Definition: An estimate is *unbiased* if, over many repeated samples drawn from the population, the average value of the estimates based on the different samples would equal the population value of the parameter being estimated.

OLS estimates are also minimally variable and well behaved in large samples when the distributional assumptions concerning ε are not precisely met. However, a drawback of the OLS estimation criterion is sensitivity to outliers, which arises from squaring the vertical differences (Problem 3.1). Section 4.7 will show how to diagnose and deal with influential points.

Table 3.4 shows Stata results for an OLS regression of SBP on age. The estimate of β_1 , the slope coefficient (Coef.) for age, is 0.44 mmHg per year, and the intercept estimate $\hat{\beta}_0$ is 105.7 mmHg (_cons).

3.3.5 Fitted Values and Residuals

The OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ in turn determine the *fitted value* $\hat{y}$ corresponding to every data point:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \quad (3.4)$$

It should be plain that the fitted value $\hat{y}_i$ lies on the estimated regression line at the point where $x = x_i$. For a woman at the average age of 67, the fitted value is

$$105.713 + 0.4405286 \times 67 = 135.2 \text{ mmHg}. \quad (3.5)$$

The *residuals* are defined as the difference between observed and fitted values of the outcome:

$$r_i = y_i - \hat{y}_i. \quad (3.6)$$

The residuals are the sample analog of ε , the error term introduced earlier in Sect. 3.3, and as such are particularly important in fitting the model, in estimating the variability of the parameter estimates, and in checking model assumptions and fit (Sect. 4.7).

3.3.6 Sums of Squares

Various *sums of squares* are central to understanding OLS estimation and to reading the Stata regression model output in Table 3.4. First is the *total sum of squares* (TSS):

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad (3.7)$$

where $\bar{y}$ is the sample average of the outcome y . TSS captures the total variability of the outcome about its mean. In Table 3.4, TSS = 109,171 and appears in the row and column labeled `Total` and `SS` (for Sum of Squares), respectively.

In an OLS model, TSS is split into two components. The first is the *model sum of squares* (MSS), or the part of the variability of the outcome about its mean that can be accounted for by the model:

$$\text{MSS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \quad (3.8)$$

The second component of outcome variability, the part that cannot be accounted for by the model, is the *residual sum of squares* (RSS):

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (3.9)$$

By definition, RSS is minimized by the fitted regression line. In Table 3.4, MSS and RSS appear in the rows labeled `Model` and `Residual` of the `SS` column. The identity $TSS = MSS + RSS$ is a central property of OLS, but more difficult to prove than it may seem.

3.3.7 Standard Errors of the Regression Coefficients

MSS and RSS also play an important role in estimating the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ and in testing the null hypothesis of central interest, $H_0: \beta_1 = 0$. These standard errors depend on the variance of ε —that is, the variance of the outcome about the regression line—which is estimated in our single predictor model by

$$\widehat{\text{Var}}(\varepsilon) = \hat{\sigma}_{y|x}^2 = \text{RSS}/(n - 2). \quad (3.10)$$

In Table 3.4, $\hat{\sigma}_{y|x}^2$ equals 390.5, and appears in the column and row labeled `MS` (for Mean Square) and `Residual`, respectively.

The variance of $\hat{\beta}_1$ is estimated by

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}_{y|x}^2}{(n - 1)\hat{\sigma}_x^2}, \quad (3.11)$$

where $\hat{\sigma}_x^2$ is the sample variance of the predictor x . The square root of the variance of an estimate is referred to as its *standard error*, or $\text{SE}(\hat{\beta})$. In Table 3.4, the standard error of the estimated slope coefficient for `age`, found in the column labeled `Std. Err.`, is approximately 0.186.

From the numerator and denominator of (3.11), it is clear that the variance of the slope estimate *increases* with the residual outcome variance not explained by the model, but *decreases* with larger sample size and with the variance of the predictor (as we pointed out earlier in Sect. 3.3.3). In our example of SBP and age, estimation of the trend in age is helped by the relatively large age range in the sample. It should also be intuitively clear that the precision of the slope estimate is increased in samples where the data are tightly clustered about the regression line—in other words, if the residual variance of the outcome is small. Figure 3.1 shows that this is not the case with our example; SBP varies widely about the regression line at every value of age.

3.3.8 Hypothesis Tests and Confidence Intervals

When the outcome is normally distributed, the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have a normal distribution, and the ratio of the slope estimate to its standard error has a t -distribution with $n - 2$ degrees of freedom. This leads directly to a test of

the null hypothesis of no slope: that is, $H_0: \beta_1 = 0$, or in substantive terms, no systematic relationship between predictor and outcome. In Table 3.4, the t -statistic and corresponding P -value for age are shown in the columns labeled t and $P > |t|$. In the example, we are able to reject the null hypothesis that SBP does not change with age at the usual 5% level of significance ($P = 0.019$).

The t -distribution also leads to 95% CIs for the population parameter β_1 , shown in Table 3.4 in the columns labeled [95% Conf. Interval]. The confidence interval does not include 0, in accord with the result of the t -test of the null hypothesis. Under the assumptions of the model, a CI computed this way would, on average, include the population value of the parameter in 95 of 100 random samples. In a more intuitive interpretation, we could exclude with 95% confidence age trends in SBP smaller than 0.07 mmHg/year or larger than 0.81 mmHg/year.

3.3.8.1 Relationship Between Hypothesis Tests and Confidence Intervals

Hypothesis tests and CIs provide overlapping information about the parameter or association being assessed. Common ground is that when a two-sided test is statistically significant at $P < 0.05$, then the corresponding 95% CI will exclude the null parameter value. However, the P -value, especially if it is small, does give a more direct sense of the strength of the evidence against the null hypothesis. Likewise, only the confidence interval provides information about the range of parameter values that are consistent with the data. In Sect. 3.7 below, we argue that CIs are particularly important in the interpretation of negative findings—that is, cases where the null hypothesis is not rejected. Both the P -value and the CI are important for understanding statistical results in depth, and getting beyond the simple question of whether or not an association is statistically significant. This overlapping relationship between hypothesis tests and CIs holds in many settings in addition to linear regression.

3.3.8.2 Hypothesis Tests and Confidence Intervals in Large Samples

The hypothesis tests and CIs in this section follow from basic statistical theory for data with normally distributed outcomes. However, linear regression models are commonly used with outcomes that are at best approximately normal, even after transformation. Fortunately, in large samples the t -tests and CIs for $\hat{\beta}_0$ and $\hat{\beta}_1$ are valid even when the underlying outcome is not normal. How large a sample is required depends on how far and in what way the outcome departs from normality. If the outcome is uniformly distributed, meaning that every value in its range is equally likely, then the t -tests and CIs may be valid with as few as 30–50 observations. However, with long-tailed outcomes, samples of at least 100 and sometimes much larger may be required for hypothesis tests and CIs to be valid.

3.3.9 Slope, Correlation Coefficient, and R^2

The slope coefficient β_1 in a simple linear model is systematically related to the Pearson correlation coefficient r , reviewed in Sect. 3.2:

$$r = \beta_1 \sigma_x / \sigma_y, \quad (3.12)$$

where σ_x and σ_y are the standard deviations of the predictor and outcome, respectively. Thus we can get r from β_1 by factoring out the scales on which x and y are measured (Problem 3.3), scales which are reflected in the standard deviations. Furthermore, the t -test of $H_0: \beta_1 = 0$ is equivalent to a test of $H_0: r = 0$.

However, the correlation coefficient is not simply interchangeable with the slope coefficient in a simple linear model. In particular, the slope coefficient distinguishes the roles of the predictor x and outcome y , with differing assumptions applying to each, and would change if those roles were reversed, but $r(x, y) = r(y, x)$. Note that reversing the roles of predictor and outcome becomes even more problematic with multipredictor models. In addition, the slope coefficient β_1 depends on the units in which both predictor and outcome are measured, so that if either or both were measured in different units, β_1 would change. For example, our estimate of the age trend in SBP would be 4.4 mmHg per decade if age were measured in ten-year units. While both versions are interpretable, this dependence on the scale of both predictor and outcome can make it difficult to assess the strength of the association. In addition, the dependence on scale would make it hard to judge whether age is a stronger predictor of SBP than other variables. From this point of view, the scale-free correlation coefficient r is easier to interpret.

The correlation coefficient r and thus the slope coefficient β_1 are also systematically related to the *coefficient of determination* R^2

$$R^2 = r^2 = \frac{\text{MSS}}{\text{TSS}}. \quad (3.13)$$

R^2 is interpretable as the proportion of the total variability of the outcome (TSS) that is accounted for by the model (MSS). As such, it is useful for comparing models (Sect. 10.2). In Table 3.4, the value of R-squared is only 0.0200, which you can easily verify is equal to $\text{MSS}/\text{TSS} = 2,179/109,171$. This shows that age only explains a very small proportion of the variability of SBP, even though it is a statistically significant predictor in a sample of moderate size.

3.4 Contingency Table Methods for Binary Outcomes

In Chap. 2, we reviewed exploratory techniques for categorical outcome variables. We expand that review here to include contingency table methods for assessing associations between binary outcomes and categorical predictors.

Table 3.5 Two-by-two contingency table for CHD and arcus

. cs chd69 arcus, or

	arcus senilis		
	Exposed	Unexposed	Total
Cases	102	153	255
Noncases	839	2058	2897
Total	941	2211	3152
Risk	.1083953	.0691995	.080901
	Point estimate		[95% Conf. Interval]
Risk difference	.0391959	.0166915	.0617003
Risk ratio	1.566419	1.233865	1.988603
Attr. frac. ex.	.3616011	.1895387	.4971343
Attr. frac. pop	.1446404		
Odds ratio	1.63528	1.257732	2.126197 (Cornfield)

chi2(1) = 13.64 Pr>chi2 = 0.0002

3.4.1 Measures of Risk and Association for Binary Outcomes

In the WCGS (Rosenman et al. 1964) of CHD introduced in Chap. 2, an association of interest to the original investigators was the relationship between CHD risk and the presence/absence of corneal arcus senilis among participants upon entry into the study. Because each participant could be unambiguously classified as having developed CHD or not during the ten-year course of the study, the indicator variable that takes on the value one or zero according to whether or not participants developed the disease is a legitimate binary outcome for the analysis. Corneal arcus is a whitish annular deposit around the iris that occurs in a small percentage of older adults, and is thought to be related to serum cholesterol level. Table 3.5 presents the results of a basic two-by-two table analysis for this example. The results were obtained using the `cs` command in Stata, which provides a number of useful quantities in addition to a simple crosstabulation of the binary CHD outcome `chd69` with the binary indicator of the presence of arcus.

The Risk estimates (0.108 and 0.069) summarize outcome risk for individuals with and without arcus and are simply the observed proportions of individuals with CHD in these groups at the baseline visit of the study. The output also includes several standard epidemiological measures of association between outcome risk and the predictor variable, along with corresponding 95% CIs. These are numerical comparisons of the risk estimates between the two groups defined by the predictor.

The Risk difference or *excess risk* is defined as the difference between the estimated risk in the groups defined by the predictor. For the table, we can verify that the risk difference is

$$0.1084 - 0.0692 = 0.039$$

The `Risk ratio` or *relative risk* is the ratio of these risks—for the example in the table,

$$0.1084/0.0692 = 1.57.$$

The `Odds ratio` is the ratio between the corresponding odds in the two groups. The odds of an outcome occurring are computed as the probability of occurrence divided by the complementary probability that the event does not occur. Since the denominators of these two probabilities are identical, the odds can be also be calculated as the ratio of the number of outcomes to nonoutcomes. Frequently used in games of chance, “even odds” obtains when these two probabilities are equal.

In Table 3.5, the odds of CHD occurrence in the two arcus groups are $0.1084/(1 - 0.1084) = 102/839$ and $0.0692/(1 - 0.0692) = 153/2058$, respectively. The ratio of these two numbers yields the estimated odds ratio (1.635) comparing the odds of CHD occurrence among participants with arcus to the odds of those without this condition. Although the odds ratio is somewhat less intuitive as a risk measure than the risk difference and relative risk, we will see that it has properties that make it useful in a wide range of study designs, and (in Chap. 5) that it is fundamental in the definition and interpretation of the *logistic regression* model.

Finally, note that Table 3.5 provides two auxiliary summary measures of *attributable risk* (i.e., `Attr. frac. ex.` and `Attr. frac. pop`), which estimate the fraction of outcomes which can be attributed to the predictor in the subgroup with the predictor (sometimes referred to as “exposed” individuals) and in the overall population, respectively. Although these measures can easily be estimated from the data in the table, their validity and interpretability depends on a number of factors, including study design and the causal connections between measured and unmeasured predictors and the outcome. See Rothman and Greenland (1998) for further discussion of these measures.

In the last example, we saw that the observed outcome proportions for groups defined by different values of a predictor are the fundamental components of the three summary measures of association: the excess risk, relative risk, and odds ratio. To discuss these further, it will be useful to have symbolic definitions. Following the notation introduced in Sect. 3.3 for a continuous outcome measure, we will denote the binary outcome variable CHD by y , and let the values 1 and 0 represent individuals with and without the outcome, respectively. We will symbolize the outcome probability for an individual associated with a particular value x of a single predictor as

$$P(x) = \Pr(y = 1|x)$$

and estimate this using the proportion of individuals with the outcome $y = 1$ among all those in the sample with the value x of the predictor. For example, $P(0)$ and $P(1)$ symbolize the outcome probability or risk associated with two levels of the binary predictor `arcus` in Table 3.5 (where we follow the usual convention that individuals possessing the characteristic have the values $x = 1$, and individuals without the characteristic have $x = 0$). The following equation defines all three summary risk measures introduced above using this notation:

$$\begin{aligned}
ER &= P(1) - P(0) \\
RR &= P(1)/P(0) \\
OR &= \frac{P(1)/[1 - P(1)]}{P(0)/[1 - P(0)]},
\end{aligned} \tag{3.14}$$

where ER , RR , and OR denote the excess risk, relative risk, and odds ratio, respectively.

Like the correlation coefficient, these measures provide a convenient single number summary of the direction and magnitude of the association. The major distinction between them is that the ER is a measure of the difference in risk between the two groups (with no difference indicated by a value of zero), while both the RR and OR compare the risks in relative terms (with no difference indicated by a value of one). Note that because the component risks range between zero and one, the ER can take on values between -1 and 1 . By contrast, the RR and OR range between 0 and ∞ .

Relative measures are appealing because they are dimensionless, and convey a clear impression of how outcome risk is increased/decreased by exposure. The RR in particular is favored by epidemiologists because of its interpretability as a ratio of risks. However, relative measures are less desirable when the goal is to convey the “importance” of a particular risk in absolute terms: In the example, the estimated RR for the risk of CHD is approximately 1.6 times higher for men with arcus. The ER tells us that this corresponds to a 4% difference in absolute risk. Note that if the risk of the outcome were ten times lower in both groups, we would have the same estimated RR , but the corresponding ER would also be ten times smaller (or 0.4%).

A further feature of the RR worth remembering is that its maximum value is constrained by the level of risk in the comparison group. For example, if $\Pr(0) = 0.5$, $RR \leq 2$ must hold. The OR has the advantages of a relative measure, and in addition is not constrained by the level of the risk in the reference group. However, being based on the odds of the outcome rather than the probability, the OR lacks the intuitive interpretation of RR . The only exception is when the outcome risk is quite small. For such rare outcomes, the OR closely approximates the RR and can be interpreted similarly. (This property can be seen from the above definition by noting that if outcome risk is close to zero, then $[1 - \Pr(0)]$ and $[1 - \Pr(1)]$ will both be approximately one.) Unfortunately, the odds ratio is often inappropriately reported as a relative risk even when this condition is not met ([Holcomb et al. 2001](#)). Because the value of the OR is always more extreme than the value of the RR (except when both equal one), this can be misleading. For these reasons, we recommend that the measure of association reported in research findings be that actually used in the analysis.

A final important property of all three measures of association introduced above is that their interpretation depends on the underlying study design. In the WCGS example, the outcome risks represent the *incidence proportion* of CHD over the entire duration of the study (approximately ten years). The measures of association in the table should be interpreted accordingly. By contrast, the sexually transmitted infection example mentioned at the beginning of this chapter

referred to a cross-sectional sample. Outcome risk in this setting is measured by the *prevalence* of the outcome among the groups defined by the predictor. In this case, the terms “prevalence odds,” “prevalence ratio,” and “excess prevalence” provide unambiguous alternative labels for *OR*, *RR*, and *ER*, respectively.

The relative merits of the *ER*, *RR*, and *OR* are discussed at length in most epidemiology textbooks (e.g., [Rothman and Greenland 1998](#)). For our purposes, they are equally valid and the choice is dependent on the nature and goals of the research investigation. In fact, for prospective and cross-sectional study designs, we will see that we can freely convert between measures. (Case-control designs are a special case which will be covered in Sect. 5.3.) However, from the standpoint of regression modeling, we will see in Chap. 5 that the *OR* has clear advantages.

3.4.2 Tests of Association in Contingency Tables

Addressing the research question posed in the example presented in Table 3.5 involves more than simply summarizing the degree of the observed association between CHD and arcus. We would also like to account for uncertainty in our estimates before concluding that the association reflects more than just a chance finding in this particular sample of individuals. The 95% CIs associated with the measures of association in the table help in this regard. For example, the fact that the confidence interval for the odds ratio excludes the value 1.0 allows us to conclude that the true value for this measure is greater than one, and indicates a statistically significant positive association between the presence of arcus and CHD occurrence. This corresponds to testing the null hypothesis that the true odds ratio is equal to one, with the alternative hypothesis being that this odds ratio is different than one. The fact that the value of one is excluded from the CI corresponds to rejection of this hypothesis at the 5% significance level. Of course, establishing the possible causal connection between these two variables is a more complex issue.

The χ^2 (*chi-squared*) test of association is an alternative way to make inferences about an observed association. Note that the result of this test (presented in Table 3.5) agrees with the conclusions drawn for the 95% CIs for the various measures of association. The statistic addresses the null hypothesis of no association, and is computed using the squared differences between the observed proportions of individuals in each cell of the two-way table and the corresponding proportions that would be expected if the null hypothesis were true. Large values of the statistic indicate departure from this hypothesis, and the associated *P*-value is computed using the χ^2 distribution with degrees of freedom specified. The χ^2 statistic for a two-by-two table is less appealing as a measure of association than the alternative measures discussed above. However, in cases where predictors have more than two levels (as discussed below) and a single summary measure of association cannot be calculated, the χ^2 statistic is useful as a global indicator of whether or not an association may be present.

Table 3.6 Female partner’s HIV status by AIDS diagnosis of male partner

. cs hivp aids, or exact

	AIDS diag. in male [1=yes/0=no]			
	Exposed	Unexposed	Total	
Cases	3	4	7	
Noncases	2	22	24	
Total	5	26	31	
Risk	.6	.1538462	.2258065	
	Point estimate		[95% Conf. Interval]	
Risk difference	.4461538		-.0050928	.8974005
Risk ratio	3.9		1.233644	12.32933
Attr. frac. ex.	.7435897		.1893933	.9188926
Attr. frac. pop	.3186813			
Odds ratio	8.25		1.200901	57.1864 (Cornfield)
+-----				
1-sided Fisher's exact P = 0.0619				
2-sided Fisher's exact P = 0.0619				

The validity of the χ^2 test is dependent on available sample size; like many commonly used statistical tests, the validity of the reference χ^2 distribution for the test statistic is approximate, with the approximation improving with increasing number of observations. Consequently, for small sample sizes, approximate P -values and associated inferences may be unreliable. An alternative in these cases is to base inferences on *exact* methods. Table 3.6 presents an example from a cross-sectional study of sexual transmission of human immunodeficiency virus (HIV) in monogamous female partners of males infected from contaminated blood products (O’Brien et al. 1994). The outcome of this study was HIV status of the female partner at recruitment. Males were known to have been infected first (via medical records) and exposure of females was limited to contact with male partners. The available sample size ($n = 31$) was limited by the availability of couples meeting the strict eligibility criteria.

Table 3.6 addresses the hypothesis that more rapid disease progression in the males (as indicated by an AIDS diagnosis occurring at or before the time of recruitment of the couple) is associated with sexual transmission of HIV to the female (represented by the binary indicator `hivp`). In addition to observed counts, the table includes proportions of the outcome by AIDS diagnosis in the male partners, and the measures of association described above. The table also presents the results of Fisher’s exact test. Similar to the χ^2 test, the Fisher test addresses the hypothesis of independence of outcome and predictor. However, the P -value is computed exactly, conditioning on the observed marginal totals. The P -value for the χ^2 test applied to the data in Table 3.6 (not shown) is 0.029. Similarly, the lower 95% confidence limits for the RR and OR exclude the value one, also indicating

Table 3.7 CHD events by age in WCGS cohort

```
. tabulate chd69 agec, col chi2
```

CHD event	35-40	41-45	46-50	51-55	56-60	Total
no	512 94.29	1,036 94.96	680 90.67	463 87.69	206 85.12	2,897 91.85
yes	31 5.71	55 5.04	70 9.33	65 12.31	36 14.88	257 8.15
Total	543 100.00	1,091 100.00	750 100.00	528 100.00	242 100.00	3,154 100.00

Pearson chi2(4) = 46.6534 Pr = 0.000

a statistically significant association. By contrast, the (two-sided) P -value for the Fisher’s exact test for Table 3.6 is 0.062, indicating failure to reject the hypothesis of independence at the 5% level.

A commonly cited rule-of-thumb is that the Fisher’s exact test should be used whenever any of the expected cell counts are less than 5. Note that Fisher’s exact test applies to tables formed by variables with more than two categories. Although it can almost always be used in place of the χ^2 test, the associated computations can be lengthy for large sample sizes, especially for tables with dimensions larger than 2×2 . Given the increased speed of modern desktop computers and the availability of more computationally efficient algorithms, we recommend using the exact P -value whenever it can easily be computed (i.e., in a matter of minutes) or is provided, and especially in cases where either actual or expected minimum cell counts are less than 5.

3.4.3 Predictors with Multiple Categories

In the WCGS study discussed above, one potentially important predictor of CHD risk is age at entry into the study. Despite the fact that this can be considered as a continuous variable for the purpose of analyses, we might begin investigating the relationship by grouping age into multiple categories and summarizing CHD risk in the resulting groups. Table 3.7 shows the results obtained by dividing subjects into five-year age intervals using a constructed five-level categorical variable AGE_C. With the exception of the first two columns, the estimated percentages of individuals with CHD in the second row of the table clearly increase with increasing age. In addition, the accompanying χ^2 test indicates that age and CHD risk are associated.

As mentioned above, the conclusion of association based on the χ^2 test does not reveal anything about the nature of the relationship between these variables. More insight could be gained by computing measures of association between age and CHD risk. However, unlike the two-by-two table case, the fact that age is represented

Table 3.8 Odds ratios for CHD events by age group

. tabodds chd69 agec, or

agec	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
35-40	1.000000	.	.	.	.
41-45	0.876822	0.32	0.5692	0.557454	1.379156
46-50	1.700190	5.74	0.0166	1.095789	2.637958
51-55	2.318679	14.28	0.0002	1.479779	3.633160
56-60	2.886314	18.00	0.0000	1.728069	4.820876

Test of homogeneity (equal odds): chi2(4) = 46.64
Pr>chi2 = 0.0000

Score test for trend of odds: chi2(1) = 40.76
Pr>chi2 = 0.0000

with five levels means that a single measure will not suffice here. In fact, odds ratios can be computed to compare any two age groups. For example, the *ER*, *RR*, and *OR* comparing CHD risk in 56 to 60-year-olds with that in 35 to 40-year-olds are calculated by applying the formulas in (3.14) as follows:

$$ER = (36/242) - (31/543) = 0.092$$
$$RR = \frac{36/242}{31/543} = 2.606$$
$$OR = \frac{36/242}{\frac{206/242}{31/543}} = 2.886.$$

(3.15)

The results in Table 3.8 further reinforce our observation that CHD risk is increasing with increasing age. The odds ratios in the table are all computed using the youngest age group as the reference category. The pattern of increase in estimated odds ratios mirrors that seen in Table 3.7. Note that each odds ratio in the table is accompanied by a 95% confidence interval and associated hypothesis test. In addition, two global tests providing additional information are provided: The Test of homogeneity addresses the null hypothesis that odds ratios do not differ across age categories. In this case, the *P*-value indicates rejection, confirming the observed difference in the odds ratios mentioned above. Since age can be viewed as a continuous variable, and the categorical version considered here is ordinal, more specific alternatives to nonhomogeneity of odds are of greater scientific interest. The Score test for trend in Table 3.8 addresses the alternative hypothesis that there is a linear trend in the odds of CHD with increasing age categories. The statistically significant results indicate support for this hypothesis, and represent a stronger conclusion than nonhomogeneity. Note that this test is not applicable to nominal categorical variables.

Despite the useful information gained from the analysis in Tables 3.7 and 3.8, we may be concerned that our conclusions depend on the arbitrary choice of

grouping age into five categories. Increasing the number of age categories may provide more information on how risk varies with age, but will also reduce the number of individuals in each category and lead to more variable estimates of risk in each group. This dilemma is one of the primary motivations for introducing a regression model for the dependence of outcome risk on a continuous predictor variable. Another motivation (which will be explored briefly below and more fully in Chap. 5) arises when we consider the joint effects on risk of multiple (categorical and/or continuous) predictor variables.

3.4.4 Analyses Involving Multiple Categorical Predictors

A common feature of observational clinical and epidemiological studies is that investigators do not experimentally control the distributions of characteristics of interest among participants in the sample. Unlike randomized trials in which random allocation serves to balance the distributions of characteristics across treatment arms, observational data are usually characterized by differing distributions across subgroups defined by predictors of primary interest. For example, observational studies of the relationship between dietary factors and cancer typically adjust for age since it is frequently related to both diet and cancer risk. A fundamental part of drawing inferences regarding the relationship between the outcome and key predictors in observational studies is to consider the potential influence of these other characteristics. This topic will be covered in detail for regression models in Chaps. 4–6, 9, and 10. Here we give a brief introduction for binary outcomes and categorical predictors.

Consider the cross-tabulation of a binary indicator 20-year mortality and self-reported smoking presented in Table 3.9. These data represent women participating in a health survey in Whickham, England, in 1972–1974 (Vanderpump et al. 1996). Deaths were ascertained via follow-up of participants over a 20-year period. The results indicate a statistically significant negative association between smoking and mortality (where Cases denote deceased women).

Before concluding that this somewhat unintuitive inverse relationship between smoking and mortality may reflect a real association in the population being studied, we need to consider the possibility that it may be due to the influence of other characteristics of women in the sample. The standard approach for controlling for the influence of additional categorical predictors in contingency tables is via a *stratified* analysis, where a relationship of interest is examined in subgroups defined by a additional variable (or variables).

Table 3.10 presents the same analysis stratified by a three-level categorical variable `agegrp` representing three categories of participant age (as ascertained in the original survey). The age-specific odds ratios and associated 95% CIs indicate a positive (but not statistically significant) association between smoking and vital status in two of the three age groups. The crude odds ratio reproduces the result obtained in Table 3.9, while the age-adjusted (M-H combined,

Table 3.9 Twenty-year vital status by smoking behavior

. cs vstatus smoker [freq = nn], or

	smoker		
	Exposed	Unexposed	Total
Cases	139	230	369
Noncases	443	502	945
Total	582	732	1314
Risk	.2388316	.3142077	.2808219
	Point estimate		[95% Conf. Interval]
Risk difference	-.075376	-.1236536	-.0270985
Risk ratio	.7601076	.6347365	.9102415
Prev. frac. ex.	.2398924	.0897585	.3652635
Prev. frac. pop	.1062537		
Odds ratio	.6848366	.5354784	.8758683 (Cornfield)

chi2(1) = 9.12 Pr>chi2 = 0.0025

Table 3.10 Twenty-year vital status by smoking behavior, stratified by age

. cs vstatus smoker [freq = nn], or by(agegrp)

agegrp	OR	[95% Conf. Interval]		M-H Weight	
18-44	1.776666	.8727834	3.615113	5.568471	(Cornfield)
45-64	1.320359	.8728567	1.997089	19.55856	(Cornfield)
64+	1.018182	.4240727	2.43359	4.772727	(Cornfield)
Crude	.6848366	.5354784	.8758683		
M-H combined	1.357106	.9710409	1.896662		

Test of homogeneity (M-H) chi2(2) = 0.945 Pr>chi2 = 0.6234

Test that combined OR = 1:
Mantel--Haenszel chi2(1) = 3.24
Pr>chi2 = 0.0719

or *Mantel–Haenszel*) estimate is computed via a weighted average of the age-specific estimates, where the stratum-specific weights are given in the right table margin (M-H Weight). Because this estimate is based on separate estimates made in each age stratum, the weighted average adjusts for the influence of age.

Comparison of the crude estimate with the adjusted estimate reveals that adjusting for age reverses the direction (and alters the significance) of the unadjusted result. Considering that none of the stratum-specific estimates indicate reduced risk associated with smoking, the crude estimate is surprising. This seemingly paradoxical result is often referred to as *Simpson’s paradox*. To aid in further interpretation, Table 3.10 also includes results from two hypothesis tests of properties of the stratum-specific and combined odds ratios. The *test of homogeneity* addresses the null hypothesis that the three age-specific odds ratios are identical. Rejection of this hypothesis would provide evidence that the stratum-specific odds ratios differ, and may indicate a differential effect of smoking on mortality across different age

groups. This phenomenon is also known as *interaction* or *effect modification*. In this case, the results indicate that the data do not support rejecting the null hypothesis in favor of the alternative hypothesis of differing age-specific odds ratios. We conclude that there is no strong evidence of interaction and that the age-specific odds ratios are similar. However, note that if we base the analysis in Table 3.10 on the relative risk rather than the odds ratio, the P -value for the test of homogeneity equals 0.045, indicating the presence of interaction. This illustrates that the presence or absence of statistical interaction may reflect our choice to work with a particular measure of association rather than some underlying causal phenomenon.

The second test result presented in Table 3.10 addresses the null hypothesis that the true age-adjusted (“combined”) odds ratio for the association between vital status and smoking is different than one. This hypothesis is meaningful if we have already failed to reject the hypothesis of homogeneity. In this case, we have already concluded that we do not have strong evidence that the age-specific odds ratios differ, and the results of the test for an age-adjusted association indicate failure to reject the null hypothesis at the 5% significance level. We conclude that the observed unadjusted negative association between vital status and smoking is at least partially explained by age adjustment. In fact, adjusting for age results in a positive association between smoking and vital status, that is more in accordance with our expectations that smokers may experience more health problems.

The results of the Whickham example are an instance of a more general phenomenon in observational studies known as *confounding*. In the example, the seemingly paradoxical finding of a positive association (albeit not statistically significant) after adjustment for age can be explained by differences between age groups in the proportion of women who were smokers (women in the intermediate age group were more likely to smoke than women in the other groups), and the fact that mortality was much higher in the older women. Of course, other measured or unmeasured factors may also influence the relationship between smoking and vital status. A complete analysis would consider these. Also, it would be a good idea to consider alternate measures of age and smoking if available (e.g., treating them as continuous variables in a regression model). The phenomena of confounding and interaction will be discussed extensively in the regression context in the remaining chapters of the book.

3.4.5 Collapsibility of Standard Measures of Association

Following the discussion in the previous section, it is tempting to conclude that in situations where interaction can be ruled out, the presence of confounding can be assessed via observed differences between the crude and adjusted measures of association obtained from the Mantel–Haenszel approach for stratified contingency tables. Conversely, agreement between the stratum-specific estimates and the crude (unadjusted) estimate would seem to imply a lack of confounding.

There are two primary issues to consider when assessing absence/presence of confounding based on comparing unadjusted and adjusted association measures: the first is that because confounding is fundamentally tied to the causal interpretation given the associations involved, its presence can never be confirmed solely on statistical grounds. In the Whickam example from Table 3.10, interpreting age as a confounder of the smoking–mortality association as measured by odds ratios seems plausible. However, in many situations, the direction of the causal link between a risk factor and a suspected confounder is less clear. In these settings, observed differences between crude and adjusted association measures may reflect causal relationships other than confounding. Section 4.5 provides examples of *mediation* of the causal effects of an exposure variable on an outcome by an intermediate variable, and points out that this cannot be distinguished from confounding solely by observing differences between crude and adjusted measures of association.

The second issue is that different measures of association may exhibit different properties with respect to adjustment and pooling across strata, and these properties complicate simple interpretation of observed differences between pooled and adjusted measures. Intuitively, we might expect that in the absence of confounding and interaction, the association between a binary outcome and a single binary predictor at levels defined by a third categorical predictor would be homogeneous, and that the observed association in the strata would equal the crude association from the pooled table ignoring the third variable. A measure of association with this property is called *strictly collapsible*. Both the risk difference and the relative risk are collapsible in this sense. However, the odds ratio is not strictly collapsible. In some situations, the crude odds ratio may differ from the corresponding stratum specific and adjusted measures even when confounding is demonstrably absent.

Noncollapsibility of the odds ratio is illustrated in Table 3.11, in which the odds ratios measuring the association between a binary outcome variable Y and a binary predictor X are equal in strata defined by a third binary variable Z , and also equal to the adjusted measure. Yet, the crude odds ratio ignoring Z is different from the stratum specific measures, even though there is no marginal association between X and Z (i.e., confounding cannot be present). Note that both the crude and adjusted odds ratios are valid measures in this example. The crude measure is interpreted as the *marginal* odds ratio for the association between Y and X , while the adjusted measure is interpreted as the *conditional* odds ratio for a fixed value of Z .

We will see in Chap. 5 that noncollapsibility is also manifested in logistic regression models for binary outcomes, where regression coefficients have a log odds ratio interpretation, and in proportional hazards regression models for survival outcomes (Chap. 6), with coefficients interpretable as log hazard ratios. Note that in the case of rare outcomes, the close correspondence between odds ratios and relative risks noted above minimizes this distinction, and these cases analyses based on either measure will agree closely. Chapter 9 is entirely devoted to the topic of making valid causal inferences using data from observational studies, and provides a framework for understanding confounding that further clarifies the issues raised here.

Table 3.11 Example illustrating inequality of the odds ratio for the association between a binary outcome Y and a binary predictor X when stratified by a binary variable Z versus pooled across values of Z

```
. tabulate Y X if Z==0
```

Y	X		Total
	0	1	
0	20	10	30
1	25	25	50
Total	45	35	80

```
. tabulate Y X if Z==1
```

Y	X		Total
	0	1	
0	25	25	50
1	10	20	30
Total	35	45	80

```
. cs Y X, or by(Z)
```

Z	OR	[95% Conf. Interval]		M-H Weight
0	2	.7897239	5.05171	3.125 (Cornfield)
1	2	.7897239	5.05171	3.125 (Cornfield)
Crude	1.653061	.8873163	3.079631	
M-H combined	2	1.028901	3.887644	
Test of homogeneity (M-H) chi2(1) = 0.000 Pr>chi2 = 1.0000				

```
Test that combined OR = 1:
```

```
Mantel-Haenszel    chi2(1) =    4.18
Pr>chi2 =    0.0409
```

3.5 Basic Methods for Survival Analysis

In the previous section, we considered binary outcomes—that is, whether or not an event has occurred. Survival data represent an extension in which we take into account the time until the event occurs—or until the end of follow-up, if the event has not yet occurred at that point. These more complex outcomes are studied using techniques collectively known as *survival analysis*. The term reflects the origin of these methods in demographic studies of life expectancy.

3.5.1 Right Censoring

To illustrate the special characteristics of survival data, we consider a study of 6-mercaptopurine (6-MP) as maintenance therapy for children in remission from

Table 3.12 Weeks in remission among leukemia patients

Placebo:	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17 22, 23
6-MP:	6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*

acute lymphoblastic leukemia (ALL) (Freireich et al. 1963). Forty-two patients achieved remission from induction therapy and were then randomized in equal numbers to 6-MP or placebo. The survival time studied was from randomization until relapse. At the time of the analysis, all 21 patients in the placebo group had relapsed, whereas only 9 of 21 patients in the 6-MP group had.

One crucial characteristic of these survival times is that for the 12 patients in the 6-MP group who remained in remission at the time of the analysis, the exact time to relapse was unobserved; it was only known to exceed the follow-up time. For example, one patient had only been under observation for six weeks, so we only know that the relapse time is longer than that. Such a survival time is said to be *right-censored*—“right” because on a graph the relapse time would lie somewhere to the right of the censoring time of six weeks.

Definition: A survival time is said to be *right-censored* at time t if it is only known to be greater than t .

Table 3.12 displays follow-up times in the leukemia study. Asterisks mark the right-censored remission times.

Because of the censoring, we could not validly estimate the effects of 6-MP on time to relapse simply by comparing average follow-up times in the two groups (say, with a t -test). This simple approach would not work because the right-censored follow-up times in the 6-MP group are shorter, possibly much shorter, than the actual unobserved times to relapse for these patients. Furthermore, five of the right-censored values in the 6-MP group exceed the largest follow-up time in the placebo group; to ignore this would be throwing away valuable evidence for the effectiveness of the treatment. Survival analysis makes it possible to analyze right-censored data like these without bias or losing information contained in the length of the follow-up times.

3.5.2 *Kaplan–Meier Estimator of the Survival Function*

Suppose we would like to describe the probability of remaining in remission during each of the first ten weeks of the leukemia study. This probability is called the *survival function*.

Definition: The *survival function* at time t , denoted $S(t)$, is the probability of being event-free at t ; equivalently, the probability that the survival time is greater than t .

Table 3.13 Follow-up table for placebo patients in the leukemia study

Week of follow-up	No. followed	No. relapsed	No. censored	Conditional prob. of remission	Survival function
1	21	2	0	$19/21 = 0.91$	0.91
2	19	2	0	$17/19 = 0.90$	$0.90 \times 0.91 = 0.81$
3	17	1	0	$16/17 = 0.94$	$0.94 \times 0.81 = 0.76$
4	16	2	0	$14/16 = 0.88$	$0.88 \times 0.76 = 0.67$
5	14	2	0	$12/14 = 0.86$	$0.86 \times 0.67 = 0.57$
6	12	0	0	$12/12 = 1.00$	$1.00 \times 0.57 = 0.57$
7	12	0	0	$12/12 = 1.00$	$1.00 \times 0.57 = 0.57$
8	12	4	0	$8/12 = 0.67$	$0.67 \times 0.57 = 0.38$
9	8	0	0	$8/8 = 1.00$	$1.00 \times 0.38 = 0.38$
10	8	0	0	$8/8 = 1.00$	$1.00 \times 0.38 = 0.38$

We will first show how the survival function can be estimated for the 21 placebo patients. Because there is no right-censoring in the placebo group, we could simply estimate the survival function by the sample proportion in remission for each week. However, we will use a more complicated method because it accommodates right-censored data. This method depends on writing the survival function in any given week as a chain of conditional probabilities.

In Table 3.13 the placebo data are summarized by consecutive one-week intervals. The number of subjects who remain both in remission and in follow-up at the start of the week is given in the second column. The third and fourth columns list the numbers who relapse and who are censored during the week, respectively. Since none are censored, the number in follow-up is reduced only during weeks when a patient relapses. From the table, we see that in the first week, 19 of 21 patients remained in remission, so a natural estimate of the probability of being in remission in the first week is $19/21 = 0.91$. In the second week, 2 of the 19 placebo patients still in remission in the first week relapsed, and the remaining 17 remained in remission. Thus the probability of not relapsing in the second week, conditional on not having relapsed in the first, is estimated by $17/19 = 0.90$. It follows that the overall probability of remaining in remission in the second week is estimated by $19/21 \times 17/19 = 17/21 = 0.81$. Likewise, the probability of remaining in remission in the third week is estimated by $19/21 \times 17/19 \times 16/17 = 16/21 = 0.76$. In this case where there is no censoring, our chain of conditional probabilities reduces to the overall sample proportion in remission at the end of every week. You can easily verify that after ten weeks, the survival function estimate given by the chain of conditional probabilities is equal to the sample proportion still in remission.

Now we show how the survival function estimate based on the chain of conditional probabilities accommodates the censoring in the 6-MP group, as shown in Table 3.14. The problem we have to address is that two 6-MP subjects are censored prior to week 10. Since it is unknown whether they would have relapsed before the end of that week, we can no longer estimate the survival function at week 10 by the sample proportion still in remission at that point.

Table 3.14 Follow-up table for 6-MP patients in the leukemia study

Week of follow-up	No. followed	No. relapsed	No. censored	Condition. prob. of remission	Survival function
1	21	0	0	$21/21 = 1.00$	1.00
2	21	0	0	$21/21 = 1.00$	$1.00 \times 1.00 = 1.00$
3	21	0	0	$21/21 = 1.00$	$1.00 \times 1.00 = 1.00$
4	21	0	0	$21/21 = 1.00$	$1.00 \times 1.00 = 1.00$
5	21	0	0	$21/21 = 1.00$	$1.00 \times 1.00 = 1.00$
6	21	3	1	$18/21 = 0.86$	$0.86 \times 1.00 = 0.86$
7	17	1	0	$16/17 = 0.94$	$0.94 \times 0.86 = 0.81$
8	16	0	0	$16/16 = 1.00$	$1.00 \times 0.81 = 0.81$
9	16	0	0	$16/16 = 1.00$	$1.00 \times 0.81 = 0.81$
10	16	0	1	$16/16 = 1.00$	$1.00 \times 0.81 = 0.81$

The rows of Table 3.14 for weeks 6 and 7 show how the method works with right-censored data. In week 6, three patients are observed to relapse, and one is censored (by assumption at the end of the week). Thus the probability of remaining in remission in week 6, conditional on having remained in remission in week 5, is $18/21 = 0.86$. Then we estimate the probability of remaining in remission in week 7, conditional on having remained in remission in week 6, as $16/17$: in short, the patient censored during week 6 has disappeared from the denominator, and does not contribute to the calculations for any subsequent week. Using this method for dealing with the censored observations, the conditional probabilities can still be estimated. As a result, we obtain a valid estimate of the probability of remaining in remission at the end of week 10, even though it is unknown whether the two censored patients remained in remission at that time. This approach allows us to extrapolate the survival experience of censored observation by those followed longer. This method requires modification in the case of *competing risks data* (Sect. 6.5) where *cumulative incidence functions* define the probability of failure in the presence of other causes of failure.

In essence, we have estimated the survival functions in the placebo and 6-MP groups using the well-known Kaplan–Meier estimator to deal with right censoring. In this example, the follow-up times have been grouped into weeks, but the method also applies to cases where they are observed more exactly. In Sect. 6.6.4, we examine the important assumption of *independent censoring* which underlies these procedures.

3.5.3 Interpretation of Kaplan–Meier Curves

Plots of the Kaplan–Meier estimates of $S(t)$ for the 6-MP and placebo groups in the leukemia study are shown in Fig. 3.2. Note that the curves drop at observed relapse times and are flat in the intervening periods. As a result, we can infer periods of

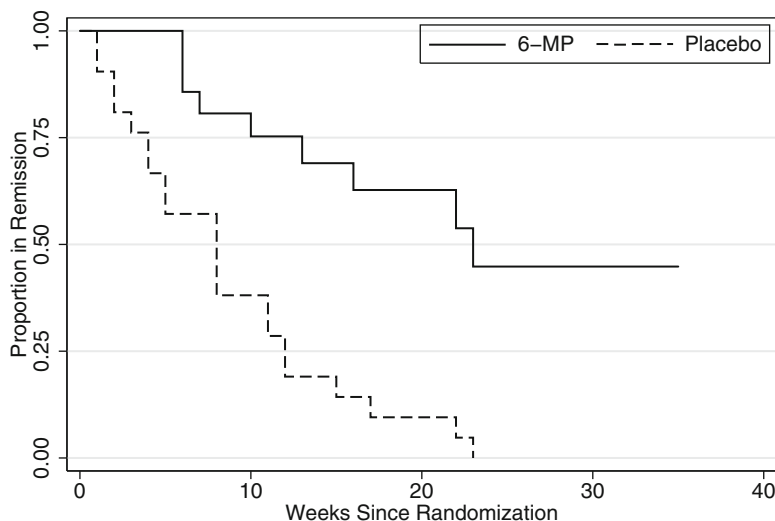


Fig. 3.2 Survival curves by treatment for leukemia patients

high risk, when the survival curve descends rapidly, as well as periods of lower risk, when it remains relatively flat. In particular, placebo patients appear to be at high risk of relapse in the first five weeks.

In addition, the estimated survival function for the 6-MP group is above the placebo curve over the entire follow-up period, giving evidence for higher probability of remaining in remission, or equivalently longer times in remission and lower risk of relapse in patients treated with 6-MP. In Sect. 3.5.6 below, we show how to test the null hypothesis that the survival functions are the same in the two groups.

3.5.4 Median Survival

The Kaplan–Meier results may also be used to obtain estimates of the median survival time, defined as the time at which half the relevant population has experienced the outcome event. In the absence of censoring, with every survival time observed exactly, the median survival time could be simply estimated by the sample median of survival times: that is, the earliest time at which half the study participants have experienced the event. From Table 3.13, we can see that median time to relapse is eight weeks in the placebo group—the first week in which at least half the sample (12/21) have relapsed.

In the presence of censoring, however, we need to use the Kaplan–Meier estimate $\hat{S}(t)$ to estimate the median. In this case, the median survival time is estimated by the earliest time at which the Kaplan–Meier curve dips below 0.50. In the leukemia

example, Fig. 3.2 shows that estimated median time to relapse is 23 weeks for 6-MP group, as compared to eight weeks for placebo—more evidence for the effectiveness of 6-MP as maintenance therapy for ALL.

By extension, other quantiles of the distribution of survival times can be obtained from the Kaplan–Meier estimate $\hat{S}(t)$. The p th quantile is estimated as the earliest time at which the Kaplan–Meier curve drops below $1 - p$. For instance, the lower quartile (i.e., the 0.25 quantile) is the earliest time at which the curve drops below $1 - 0.25 = 0.75$. The lower quartiles for the 6-MP and placebo groups are 13 and 4 weeks, respectively. However, a limitation of the Kaplan–Meier estimate is that when the curve does not reach $1 - p$, the p th percentile cannot be estimated. For example, Fig. 3.2 makes it clear that for the 6-MP group, quantiles of the distribution of remission times larger than the 0.6th cannot be estimated using the Kaplan–Meier method.

Note that while we can estimate the median and other quantiles of the distribution of survival times using the Kaplan–Meier results, we are unable to estimate the mean of the distribution in the typical case, as in the 6-MP group, where the longest follow-up time is censored (Problem 3.7).

A final note: graphs are useful for giving overall impressions of the survival function, but it is difficult to read quantities from them (e.g., median survival time or $\hat{S}(t)$ for some particular t). To obtain precise values, the results in Tables 3.13 and 3.14 can be printed in Stata using the `sts list` and `stsci` commands.

3.5.5 Cumulative Event Function

Another useful summary of survival data is the probability of having experienced the outcome event by time t . In terms of our leukemia example, this would mean estimating the probability of having relapsed by the end of each week of the study.

Definition: The *cumulative event function* at time t , denoted $F(t)$, is the probability that the event has occurred by time t , or equivalently, the probability that the survival time is less than or equal to t . Note that $F(t) = 1 - S(t)$.

The cumulative event function is estimated by the complement of the Kaplan–Meier estimate of the survival function: that is, $\hat{F}(t) = 1 - \hat{S}(t)$. If t has the same value τ for all study participants, then $F(\tau)$ is interpretable as the outcome risk discussed in Sect. 3.4 on contingency table methods for binary outcomes. The cumulative event plots shown in Fig. 3.3 are also easily obtained in Stata by specifying the `failure` option.

Note that parametric methods can also be used to estimate survival distributions, as well as quantities that are not immediately available from the Kaplan–Meier approach (e.g., the mean and specified quantiles). However, because they rest on explicit assumptions about the form of these distributions, they are somewhat less robust than the methods presented here. For example, the mean can be poorly estimated in situations where a large proportion of the data are censored, with the result that the right tail of the survival function is only “known” by extrapolation.

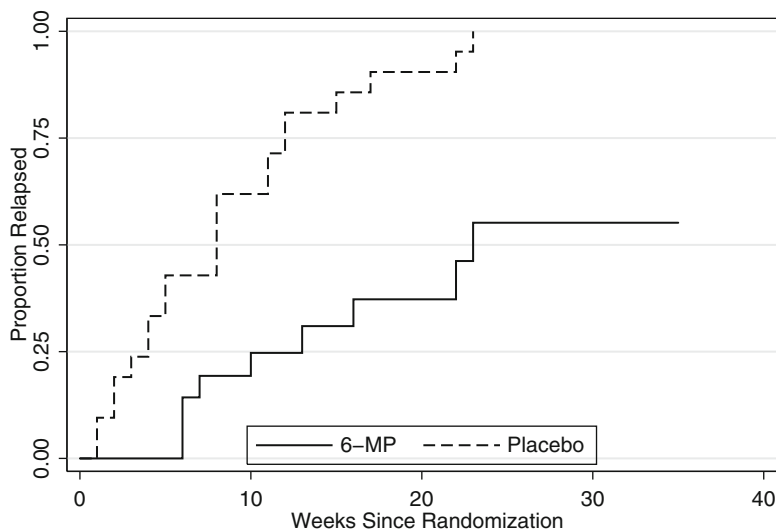


Fig. 3.3 Cumulative event curves by treatment for leukemia patients

3.5.6 Comparing Groups Using the Logrank Test

The Kaplan–Meier estimator provides an interpretable description of the survival experience of two treatment groups in the study of 6-MP as maintenance therapy for ALL. With those descriptions in hand, how do we go on to formally test for differences in relapse between the treatments?

The primary tool for the comparison of the survival experience of two or more groups is the *logrank test*. The null hypothesis for this test is that the survival distributions being compared are equal at all follow-up times. In the leukemia example, this implies that the population survival curves for 6-MP and placebo coincide. The alternative hypothesis is that the two survival curves differ at one or more points in time. Like the Kaplan–Meier estimator, the logrank test accommodates right-censoring. It works by comparing observed numbers of events in each group to the number expected if the survival functions were the same. The comparison accounts for differences in length of follow-up in calculating the expected numbers of events. Results are shown in Table 3.15.

There are a total of 30 events in the sample, 21 in the placebo group and 9 in the 6-MP group. The column labeled *Events expected* gives the expected number of events in the two groups under the null hypothesis of equal survival functions. In the leukemia data, average follow-up was considerably shorter in the placebo group and hence fewer events would be expected in that group. Clearly there were many more events than expected among placebo participants, and many fewer than expected in the 6-MP group. The resulting χ^2 statistic of 16.8 is statistically significant ($P < 0.00005$), in accord with our earlier impression that 6-MP is effective maintenance therapy for patients with ALL.

Table 3.15 Logrank test for leukemia example

Logrank test for equality of survival functions

group	Events observed	Events expected
6 MP	9	19.25
Placebo	21	10.75
Total	30	30.00

chi2(1) =	16.79
Pr>chi2 =	0.0000

The logrank test is easily generalized to the comparison of more than two groups. The logrank test statistic for $K > 2$ groups follows an approximate χ^2 distribution with $K - 1$ degrees of freedom. In this more general case, the null hypothesis is

$$H_0 : S_1(t) = \dots = S_K(t) \quad \text{for all } t \quad (3.16)$$

where $S_k(t)$ is the survival function for the k th group at time t . In analogy to the F -test discussed in Sect. 4.3.3, the alternative hypothesis is that some or all of the survival curves differ at one or more points in time.

When the null hypothesis is rejected, visual inspection of the Kaplan–Meier plots can help to determine where the important differences arise. Another common procedure for understanding group differences is to conduct pairwise logrank tests. This requires cautious interpretation; see Sect. 4.3.4 for approaches to handling potential difficulties with multiple comparisons.

If there are more than two groups which are defined by ordered categories (e.g., disease stage) or categories based on a numerical variable (e.g., number of positive nodes), then a trend test based on the logrank is available. In Stata, this is obtained by using the `trend` option for the command `sts test`.

Like some other nonparametric methods reviewed earlier in this chapter, and as its name implies, the logrank test only uses information about the *ranks* of the survival times rather than their actual values. The semi-parametric Cox proportional hazards model covered in Chap. 6 also works this way. In every instance, the nonparametric approach reduces the need for making restrictive and sometimes hard-to-verify assumptions, with a view toward making estimates more robust.

There is an extensive literature on testing differences in survival between groups. These tests have varying levels of similarity to the logrank test. The most popular are extensions of the Wilcoxon test for censored data; these tests can be viewed as a weighted versions of the logrank test. Such weighting can make sense, for example, if early events are judged to be particularly important. However, in the absence of compelling and prespecified reasons, we recommend the logrank test as a default test.

Chapter 6 covers censoring and other types of missing data in greater depth, and also presents more comprehensive methods of analysis for survival data, including the multipredictor Cox proportional hazards regression model.

3.6 Bootstrap Confidence Intervals

Bootstrapping is a widely applicable method for obtaining standard errors and CIs in cases where approximate methods for computing valid CIs have been developed but not conveniently implemented in statistical packages; other situations where development of such methods has turned out to be intractable; and datasets where the assumptions underlying the established methods are badly enough violated that the resulting CIs would be unreliable.

In general, standard errors and CIs reflect the sampling distribution of statistics of interest, such as regression coefficient estimates: that is, their relative frequency if we repeatedly drew independent samples of the same size from the source population, and recalculated the statistics in each new sample. In standard problems such as linear regression, the sampling distribution of the regression coefficient estimates is well known on theoretical grounds, provided the data meet underlying assumptions.

Bootstrap procedures approximate the sampling distribution of statistics of interest by a *resampling* procedure. Specifically, the actual sample is treated as if it were the source population, and bootstrap samples are repeatedly drawn from it. Bootstrap samples of the same size as the actual sample—a key determinant of precision—are obtained by resampling *with replacement*, so that in a given bootstrap sample some observations appear more than once, some once, and some not at all. We use the sample to represent the population and hence resampling from the actual data mimics drawing repeated samples from the source population. Then, from each of a large number of bootstrap samples, the statistics of interest are computed. For example, if our focus was on the difference between the coefficient estimates for a predictor of interest before and after adjustment for a covariate, the two models would be estimated in each bootstrap sample, and the difference between the two coefficient estimates tabulated across samples. The result would be the bootstrap distribution of the difference, which can in turn be regarded as an estimate of its actual sampling distribution. CIs for the statistic of interest would then be computed from the bootstrap distribution. Stata calculates bootstrap CIs using three procedures:

- *Normal approximation*: If the bootstrap distribution of the statistic of interest is reasonably normal, it may be enough to compute its standard deviation, then compute a conventional CI centered on the observed statistic, simply substituting the bootstrap SD for the usual model-based standard error of the statistic. The bootstrap SD is a relatively stable estimate of the standard error, since it is based on the complete set of bootstrap samples, so a relatively small number of bootstrap samples may suffice. However, we often resort to the bootstrap

Table 3.16 Bootstrap confidence interval for association of age with SBP

<code>. reg SBP age</code>							
Source	SS	df	MS	Number of obs = 276			
Model	2179.70702	1	2179.70702	F(1, 274) = 5.58			
Residual	106991.347	274	390.47937	Prob > F = 0.0188			
Total	109171.054	275	396.985652	R-squared = 0.0200			
				Adj R-squared = 0.0164			
				Root MSE = 19.761			
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
age	.4405286	.186455	2.36	0.019	.0734621	.8075952	
_cons	105.713	12.40238	8.52	0.000	81.2969	130.129	
<code>. bootstrap "reg SBP age" _b, reps(1000)</code>							
command: reg SBP age							
statistics: b_age = _b[age]							
Bootstrap statistics				Number of obs = 276			
				Replications = 1000			
Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]		
b_age	1000	.4405287	-.0078003	.1744795	.0981403	.782917	(N)
					.0655767	.7631486	(P)
					.0840077	.7690148	(BC)
Note: N = normal							
P = percentile							
BC = bias-corrected							

precisely because the sampling distribution of the statistic of interest is unlikely to be normal, particularly in the tails. Thus this method is less reliable for constructing CIs than for estimating the standard error of the statistic.

- *Percentile Method*: The CI for the statistic of interest is constructed from the relevant quantiles of the bootstrap distribution. Because the extreme percentiles of a sample are very noisy estimates of the corresponding percentiles of a population distribution, a much larger number of bootstrap samples is required. If 1,000 samples were used, then a 95% CI for the statistic of interest would span the 25th to 975th largest bootstrap estimates.
- *Bias-Corrected Percentile Method*: The percentile-based confidence interval is shifted to account for bias, as evidenced by a difference between the observed statistic and the median of the bootstrap estimates. Again, a relatively large number of bootstrap samples is required.

Table 3.16 shows Stata output for the simple linear regression model for SBP shown earlier in Table 3.4, now with a bootstrap CI. In this instance, all three bootstrap results are fairly consistent with the parametric 95% CI (0.73–0.81 mmHg). See Sects. 4.5.4, 5.5.1, 6.6.1, and 7.9.1 for other examples where bootstrap CIs are computed.

3.7 Interpretation of Negative Findings

Confidence intervals obtained either by standard parametric methods or by the bootstrap play a particularly important role when the data do not enable us to reject a null hypothesis of interest. It is easy to overstate such negative findings. Recall that $P > 0.05$ does not prove the null hypothesis; it only indicates that the observed result could have arisen by chance, not that it necessarily did. A negative result worth discussing is best interpreted in terms of the point estimate and CI. In the following example, we can distinguish four possible cases, in increasing order of the strength of the negative finding. Suppose that a 20% reduction risk of recurrent heart attacks would justify the risks and costs of a possible new treatment, but that a risk reduction of only 5% would not meet this standard. The four cases are:

- The estimated risk reduction was large enough to be substantively important, but the CI spanned the null value and was thus too wide to provide strong evidence for effectiveness. Example: treatment reduced recurrence risk an estimated 20% (95% CI -1% to 37%). In this case, we might conclude that the study gives inconclusive evidence *for* the potential importance of the treatment; but it would be also important to note that the CI includes effects too small to be worthwhile.
- The estimated risk reduction was too small to be important, but the CI extended to values that could be important. Example: treatment reduced recurrence risk an estimated 5% (95% CI -15% to 22%). In this case the point estimate provides little support for the importance of the treatment, but the CI does not clearly rule out a potentially important effect.
- The estimated risk reduction was too small to be important, and while the CI did not include the null (i.e., $P < 0.05$), it did exclude values that could be important. Example: treatment reduced recurrence risk an estimated 3% (95% CI: 1% to 5%). In this case, we can definitively say that the treatment does not have a clinically important benefit, even though we can also rule out no effect.
- The estimated risk reduction was too small to be important, and the CI both included the null and excluded values that could be important. Example: treatment reduced recurrence risk an estimated 1% (95% CI -2% to 4%). Again, we can definitively say that the treatment does not have a clinically important benefit.

This approach using the point estimate and CI is preferable to interpretations based on *ex post facto* power calculations, which are driven by assumptions about the true effect size, and often inappropriately based on treating the observed effect size as if it were the true population value (Hoenig and Heisey 2001). A variant of this approach is to suggest that with a larger sample, the observed effect would have been statistically significant. But of course the CI for most negative findings tells us that the true effect size may well be nil or worse, which a larger sample might also firmly establish. In contrast to these problematic interpretations, the point estimate and CI can together be used to summarize what the data at hand have to tell us about the strength of the association and the precision of our information about it.

3.8 Further Notes and References

Among the best introductory statistics books are [Freedman et al. \(1991\)](#), [Devore and Peck \(1986\)](#), and [Pagano and Gavreau \(1993\)](#). Consult these for more complete coverage of basic statistical inference, ANOVA, and linear regression. Good references on methods for the analysis of contingency tables include [Fleiss et al. \(2003\)](#) and [Jewell \(2004\)](#). Two applied survival analysis texts with a biomedical orientation are [Miller et al. \(1981\)](#) and [Marubini and Valsecchi \(1995\)](#). Finally, for a review of bootstrap methods, see [Efron and Tibshirani \(1986, 1993\)](#).

3.9 Problems

Problem 3.1. An alternative to OLS is least absolute deviation (LAD) regression, in which the regression line is selected to minimize the sum of the absolute vertical differences (rather than squared differences) between the line and the data. Explain how this might reduce sensitivity to outliers.

Problem 3.2. To create a new age variable age_{10} in units of ten years, we would divide the original variable age (in years) by ten, so that a woman of age 67 would have $\text{age}_{10} = 6.7$. Similarly, the standard deviation of age_{10} is changed by the same factor: that is, the SD of age is 6.38, so the SD of age_{10} is 0.638. Suppose we want to estimate the effect of age in SD units, as is commonly done. How do we compute the new variable and what is *its* SD?

Problem 3.3. Using (3.12) and a statistical analysis program, demonstrate with your own data that the slope coefficient in a univariate linear model with continuous predictor and outcome is a rescaled transformation of the sample correlation between predictor and outcome.

Problem 3.4. The correlation coefficient is a measure of *linear* association. Suppose x takes on values evenly over the range from -10 to 10 , and that $E[y|x] = x^2$. In this case, the correlation of x and y is zero, even though there is clearly a systematic relationship. What does this suggest about the need to test model assumptions? Using a statistical package, generate a random sample of 100 values of x uniformly distributed on $[-10, 10]$, compute $E[y|x]$ for each value of x , add randomly generated standard normal errors to get the 100 values of y , and check the sample correlation of x and y .

Problem 3.5. Verify the estimates for the excess risk, relative risk, and odds ratio for the HIV example presented in Table 3.6.

Problem 3.6. The data presented below are from a case-control study of esophageal cancer. (The study and data are described in more detail in Sect. 5.3.)

```
. tabulate case ditob
```

Case status (1=case, 0=control)	tobacco		Total
	0-9 g/day	10+ g/day	
0	255	520	775
1	9	191	200
Total	264	711	975

The rows (labeled according to Case status) represent 200 cancer cases and 775 cancer-free controls selected from the same population as the cases. The columns represent a binary indicator of reported consumption of more than ten grams of tobacco per day.

Compute the odds ratio comparing the risk of cancer in individuals who report consuming more than ten grams of tobacco per day with the corresponding risk in the group reporting less or no consumption. Next, compute the odds ratio comparing the proportion of individuals reporting higher levels of consumption among cases with that among the controls. Comment.

Problem 3.7. Suppose we could estimate the value of the survival function $S(t)$ for every possible survival time from $t = 0$ onward. Clearly $S(t) \rightarrow 0$ as t becomes large. It can be shown that the mean survival time is equal to the area under this “complete” survival curve. Why are we unable to estimate mean survival from the Kaplan–Meier result when the largest follow-up time is censored? To gain insight, contrast the survival curves for the 6-MP and placebo groups in Fig. 3.2.

Problem 3.8. In the leukemia study, the probability of being relapse-free at 20 weeks, conditional on being relapse-free at 10 weeks, can be estimated by the Kaplan–Meier estimate for 20 weeks, divided by the corresponding estimate for 10 weeks. In the placebo group, those estimates are 0.38 and 0.10, respectively. Verify that the estimated conditional probability of remission at week 20, conditional on being in remission at week 10, is 0.25. In the 6-MP group, estimated probabilities of remaining in remission are 0.81, 0.63, and 0.45 at 10, 20, and 30 weeks, respectively. Use these values to estimate the probabilities of remaining in remission at 20 and 30 weeks, conditional on being in remission at 10 weeks.

3.10 Learning Objectives

- (1) Be familiar with the t -test (including versions for paired and unequal-variance data), one-way ANOVA, the correlation coefficient r , and some nonparametric alternatives.

- (2) Describe the assumptions and mechanics of the simple linear model for continuous outcomes, and interpret the results.
- (3) Define the basic measures of association (i.e., excess risk, relative risk, and odds ratio) for binary outcomes.
- (4) Be familiar with standard contingency table approaches to evaluating associations between binary outcomes and categorical predictors, including the χ^2 test and the Mantel–Haenszel approach to estimating odds ratios adjusted for the confounding influence of additional predictors.
- (5) Define right-censoring.
- (6) Interpret Kaplan–Meier survival and cumulative event curves.
- (7) Calculate median survival from an estimated survival curve.
- (8) Interpret the results of a logrank test.

Chapter 4

Linear Regression

Post-menopausal women who exercise less tend to have lower bone mineral density (BMD), putting them at increased risk for fractures. But they also tend to be older, frailer, and heavier, which may explain the association between exercise and BMD. People whose diet is high fat on average have higher low-density lipoprotein (LDL) cholesterol, a risk factor for CHD. But they are also more likely to smoke and be overweight, factors which are also strongly associated with CHD risk. Increasing body mass index (BMI) predicts higher levels of hemoglobin HbA_{1c} , a marker for poor control of glucose levels; however, older age and ethnic background also predict higher HbA_{1c} .

These are all examples of potentially complex relationships in observational data where a continuous outcome of interest, such as BMD, SBP, and HbA_{1c} , is related to a risk factor in analyses that do not take account of other factors. But in each case the risk factor of interest is associated with a number of other factors, or potential *confounders*, which also predict the outcome. So the simple association we observe between the factor of interest and the outcome may be explained by the other factors.

Similarly, in experiments, including clinical trials, factors other than treatment may need to be taken into account. If the randomization is properly implemented, treatment assignment is on average not associated with any prognostic variable, so confounding is usually not an issue. However, in stratified and other complex study designs, multipredictor analysis is used to ensure that CIs, hypothesis tests, and P -values are valid. For example, it is now standard practice to account for clinical center in the analysis of multisite clinical trials, often using the random effects methodology to be introduced in Chap. 7. And with continuous outcomes, stratifying on a strong predictor in both design and analysis can account for a substantial proportion of outcome variability, increasing the efficiency of the study. Multipredictor analysis may also be used when baseline differences are apparent between the randomized groups, to account for potential confounding of treatment assignment.

Another way the predictor–outcome relationship can depend on other factors is that an association may not be the same in all parts of the population. For example, hormone therapy (HT) has a smaller beneficial effect on LDL levels among postmenopausal women who are also taking statins, and its effect on BMD may be greater in younger postmenopausal women. These are examples of *interaction*, where the association of a factor of primary interest with an outcome is modified by another factor.

The problem of sorting out complex relationships is not restricted to continuous outcomes; the same issues arise with the binary outcomes covered in Chap. 5, survival times in Chap. 6, and repeated measures in Chap. 7. A general statistical approach to these problems is needed.

The topic of this chapter is the multipredictor linear regression model, a flexible and widely used tool for assessing the joint relationships of multiple predictors with a continuous outcome variable. We begin by illustrating some basic ideas in a simple example (Sect. 4.1). Then in Sect. 4.2, we present the assumptions of the multipredictor linear regression model and show how the simple linear model reviewed in Chap. 3 is extended to accommodate multiple predictors. Section 4.3 shows how categorical predictors with multiple levels are coded and interpreted. Sections 4.4–4.6 describe how multipredictor regression models can be used to deal with confounding, mediation, and interaction, respectively. Section 4.7 introduces some simple methods for assessing the fit of the model to the data and how well the data conform to the underlying assumptions of the model. Section 4.8 introduces sample size, power, and minimum detectable effect calculations for the multiple linear model. In Chap. 9, we use a *potential outcomes* view of *causal effects* to show how and under what conditions multipredictor regression models might be used to estimate them, and in Chap. 10 we discuss the difficult problem of which variables and how many to include in a multipredictor model.

4.1 Example: Exercise and Glucose

Glucose levels above 125 mg/dL are diagnostic of diabetes, while levels in the range from 100 to 125 mg/dL signal increased risk of progressing to this serious and increasingly widespread condition. So it is of interest to determine whether exercise, a modifiable lifestyle factor, would help people reduce their glucose levels and thus avoid diabetes.

To answer this question definitively would require a randomized clinical trial, a difficult and expensive undertaking. As a result, research questions like this are often initially looked at using observational data. But this is complicated by the fact that people who exercise differ in many ways from those who do not, and some of the other differences might explain any unadjusted association between exercise and glucose level.

Table 4.1 shows a simple linear model using a measure of exercise to predict baseline glucose levels among 2,032 participants without diabetes in the HERS

Table 4.1 Unadjusted regression of glucose on exercise

```
. regress glucose exercise if diabetes == 0
```

Source	SS	df	MS
Model	1412.50418	1	1412.50418
Residual	191605.195	2030	94.3867954
Total	193017.699	2031	95.0357946

Number of obs = 2032

F(1, 2030) = 14.97

Prob > F = 0.0001

R-squared = 0.0073

Adj R-squared = 0.0068

Root MSE = 9.7153

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exercise	-1.692789	.4375862	-3.87	0.000	-2.550954 - .8346243
_cons	97.36104	.2815138	345.85	0.000	96.80896 97.91313

clinical trial of hormone therapy (HT) (Hulley et al. 1998). Women with diabetes are excluded because the research question is whether exercise might help to prevent progression to diabetes among women at risk, and because the causal determinants of glucose may be different in that group. Furthermore, glucose levels are far more variable among diabetics, a violation of the assumption of homoscedasticity, as we show in Sect. 4.7.3 below. The coefficient estimate (Coef.) for exercise shows that average baseline glucose levels were about 1.7 mg/dL lower among women who exercised at least three times a week than among women who exercised less. This difference is statistically significant ($t = -3.87$, $P < 0.0005$).

However, women who exercise are slightly younger, a little more likely to use alcohol, and in particular have lower average BMI, all factors associated with glucose levels. This implies that the lower average glucose we observe among women who exercise could be due at least in part to differences in these other predictors. Under these conditions, it is important that our estimate of the difference in average glucose levels associated with exercise be “adjusted” for the effects of these potential confounders of the unadjusted association. Ideally, adjustment using a multipredictor regression model provides an estimate of the causal effect of exercise on average glucose levels, by *holding the other variables constant*. In Chap. 9, the rationale for estimation of causal effects using multipredictor regression models is explained in more detail.

From Table 4.2, we see that in a multiple regression model that also includes—that is, adjusts for—age, alcohol use (drinkany), and BMI, average glucose is estimated to be only about 1 mg/dL lower among women who exercise (95% CI 0.1–1.8, $P = 0.027$), holding the other three factors constant. The multipredictor model also shows that average glucose levels are about 0.7 mg/dL higher among alcohol users than among nonusers. Average levels also increase by about 0.5 mg/dL per unit increase in BMI, and by 0.06 mg/dL for each additional year of age. Each of these associations is statistically significant after adjustment for the other predictors in the model. Furthermore, the association of each of the four predictors with glucose levels is adjusted for the effects of the other three, in the sense of taking account of its correlation with the other predictors and their adjusted associations with glucose

Table 4.2 Adjusted regression of glucose on exercise

```
. regress glucose exercise age drinkany BMI if diabetes == 0
```

Source	SS	df	MS	Number of obs =	2028
Model	13828.8486	4	3457.21214	F(4, 2023) =	39.22
Residual	178319.973	2023	88.1463042	Prob > F =	0.0000
Total	192148.822	2027	94.7946828	R-squared =	0.0720
				Adj R-squared =	0.0701
				Root MSE =	9.3886

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exercise	-.950441	.42873	-2.22	0.027	-1.791239 - .1096426
age	.0635495	.0313911	2.02	0.043	.0019872 .1251118
drinkany	.6802641	.4219569	1.61	0.107	-.1472513 1.50778
BMI	.489242	.0415528	11.77	0.000	.4077512 .5707328
_cons	78.96239	2.592844	30.45	0.000	73.87747 84.04732

levels. In summary, the multipredictor model for glucose levels shows that the unadjusted association between exercise and glucose is partly but not completely explained by BMI, age, and alcohol use, and that exercise remains a statistically significant predictor of glucose levels after adjustment for these three other factors—that is, when they are held constant by the multipredictor regression model.

Still, we have been careful to retain the language of association rather than cause and effect, and in Chaps. 9 and 10 will suggest that adjustment for additional potential confounders would be needed before we could consider a causal interpretation of the result.

4.2 Multiple Linear Regression Model

Confounding thus motivates models in which the average value of the outcome is allowed to depend on multiple predictors instead of just one. Many basic elements of the multiple linear model carry over from the simple linear model, which was reviewed in Sect. 3.3. In Sect. 9.1, we show how this model is potentially suited to estimating causal relationships between predictors and outcomes.

4.2.1 Systematic Part of the Model

For the simple linear model with a single predictor, the regression line is defined by

$$\begin{aligned} E[y|x] &= \text{average value of outcome } y \text{ given predictor value } x \\ &= \beta_0 + \beta_1 x. \end{aligned}$$

(4.1)

In the multiple regression model, this generalizes to

$$E[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad (4.2)$$

where $\mathbf{x}$ represents the collection of p predictors $x_1, x_2, \dots, x_p$ in the model, and $\beta_1, \beta_2, \dots, \beta_p$ are the corresponding regression coefficients.

The right-hand side of model (4.2) has a relatively simple form, a *linear combination* of the predictors and coefficients. Analogous linear combinations of predictors and coefficients, often referred to as the *linear predictor*, are used in all the other regression models covered in this book. Despite the simple form of (4.2), the multipredictor linear regression model is a flexible tool, and with the elaborations to be introduced later in this chapter, usually allows us to represent with considerable realism how the average value of the outcome varies systematically with the predictors. In Sect. 4.7, we will consider methods for examining the adequacy of this part of the model and for improving it.

4.2.1.1 Interpretation of Adjusted Regression Coefficients

In (4.2), the coefficient β_j , $j = 1, \dots, p$ gives the change in $E[y|\mathbf{x}]$ for an increase of one unit in predictor x_j , holding other factors in the model constant; each of the estimates is adjusted for the effects of all the other predictors. As in the simple linear model, the intercept β_0 gives the value of $E[y|\mathbf{x}]$ when all the predictors are equal to zero; “centering” of the continuous predictors can make the intercept interpretable. If confounding has been persuasively ruled out, we may be willing to interpret the adjusted coefficient estimates as representing causal effects.

4.2.2 Random Part of the Model

As before, individual observations of the outcome y_i are modeled as varying by an error term ε_i about an average determined by their predictor values $\mathbf{x}_i$:

$$\begin{aligned} y_i &= E[y_i | \mathbf{x}_i] + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \end{aligned} \quad (4.3)$$

where x_{ji} is the value of predictor variable x_j for observation i . We again assume that $\varepsilon_i \sim \text{i.i.d } \mathcal{N}(0, \sigma_\varepsilon^2)$; that is, ε is normally distributed with mean zero and the same standard deviation σ_ε at every value of $\mathbf{x}$, and that its values are statistically independent.

4.2.2.1 Fitted Values, Sums of Squares, and Variance Estimators

From (4.2), it is clear that the fitted values $\hat{y}_i$, defined for the simple linear model in (3.4), now depend on all p predictors and the corresponding regression coefficient estimates, rather than just one predictor and two coefficients. The resulting sums of squares and variance estimators introduced in Sect. 3.3 are otherwise unchanged in the multipredictor model.

In the glucose example, the residual standard deviation, shown as `Root MSE`, declines from 9.7 in the unadjusted model (Table 4.1) to 9.4 in the model adjusting for age, alcohol use, and BMI (Table 4.2).

4.2.2.2 Variance of Adjusted Regression Coefficients

Including multiple predictors does affect the variance of $\hat{\beta}_j$, which now depends on an additional factor r_j , the multiple correlation of x_j with the other predictors in the model. Specifically,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma_{y|x}^2}{(n-1)\sigma_{x_j}^2(1-r_j^2)}, \quad (4.4)$$

where, as before, $\sigma_{y|x}^2$ is the residual variance of the outcome and $\sigma_{x_j}^2$ is the variance of x_j ; r_j is equivalent to $r = \sqrt{R^2}$ from a multiple linear model in which x_j is regressed on all the other predictors. The term $1/(1-r_j^2)$ is known as the *variance inflation factor*, since $\text{Var}(\hat{\beta}_j)$ is increased to the extent that x_j is correlated with other predictors in the model.

However, inclusion of other predictors, especially powerful ones, also tends to decrease $\sigma_{y|x}^2$, the residual or unexplained variance of the outcome. Thus, the overall impact of including other predictors on $\text{Var}(\hat{\beta}_j)$ depends on both the correlation of x_j with the other predictors and how much additional variability they explain. In the glucose example, the standard error of the coefficient estimate for exercise declines slightly, from 0.44 to 0.43, after adjustment for age, alcohol use, and BMI. This reflects the reduction in residual standard deviation previously described, as well as a variance inflation factor in the adjusted model of only 1.03.

4.2.2.3 t -Tests and Confidence Intervals

The t -tests of the null hypothesis $H_0: \beta_j = 0$ and CIs for β_j carry over almost unchanged for each of the β s estimated by the model, only using (4.4) rather than (3.11) to compute the standard error of the regression coefficient, and comparing the t -statistic to a t -distribution with $n - (p + 1)$ degrees of freedom (p is the number of predictors in the model, and an extra degree of freedom is used in estimation of the intercept β_0).

However, there is a substantial difference in interpretation, since the results are now adjusted for other predictors. Thus in rejecting the null hypothesis $H_0: \beta_j = 0$ we would be making the stronger claim that, in the population, x_j predicts y , holding the other factors in the model constant. Similarly, the CI for β_j refers to the parameter which takes account of the other $p - 1$ predictors in the model.

We have just seen that $\text{Var}(\hat{\beta}_j)$ may not be increased by adjustment. However, in Sect. 4.4 we will see that including other predictors in order to control confounding commonly has the effect of attenuating the unadjusted estimate of the association of x_j with y . This reflects the fact that the population parameter being estimated in the adjusted model is often closer to zero than the parameter estimated in the unadjusted model, since some of the unadjusted association is explained by other predictors. If this is the case, then even if $\text{Var}(\hat{\beta}_j)$ is unchanged, it may be more difficult to reject $H_0: \beta_j = 0$ in the adjusted model. In the glucose example, the adjusted coefficient estimate for exercise is considerably smaller than the unadjusted estimate. As a result the t -statistic is reduced from -3.87 to -2.22 —still statistically significant, but less highly so.

4.2.3 Generalization of R^2 and r

The coefficient of determination $R^2 = \text{MSS} / \text{TSS}$ retains its interpretation as the proportion of the total variability of the outcome that can be accounted for by the predictor variables. Under the model, the fitted values summarize all the information that the predictors supply about the outcome. Thus, the multiple correlation coefficient $r = \sqrt{R^2}$ now represents the correlation between the outcome y and the fitted values $\hat{y}$. It is easy to confirm this identity by extracting the fitted values from a regression model and computing their correlation with the outcome (Problem 4.3). In the glucose example, R^2 increases from less than 1% in the unadjusted model to 7% after inclusion of age, alcohol use, and BMI, a substantial increase in relative if not absolute terms.

4.2.4 Standardized Regression Coefficients

In Sect. 3.3.9, we saw that the slope coefficient β_1 in a simple linear model is systematically related to the Pearson correlation coefficient (3.12); specifically, $r = \beta_1 \sigma_x / \sigma_y$, where σ_x and σ_y are the standard deviations of the predictor and outcome. Moreover, we pointed out that the scale-free correlation coefficient makes it easier to compare the strength of association between the outcome and various predictors across single-predictor models. In the context of a multipredictor model, *standardized regression coefficients* play this role. Obtained using the `beta` option

to the `regress` command in Stata, the standardized regression coefficient β_j^s for predictor x_j is defined in analogy to (3.12) as

$$\beta_j^s = \beta_j \sigma_{x_j} / \sigma_y, \quad (4.5)$$

where σ_{x_j} and σ_y are the standard deviations of predictor x_j and the outcome y . These standardized coefficient estimates are what would be obtained from the regression if the outcome and all the predictors were first rescaled to have standard deviation 1. Thus, they give the change in standard deviation units in the average value of y per standard deviation increase in the predictor. Standardized coefficients make it easy to compare the strength of association of different continuous predictors with the outcome within the same model.

For binary predictors, however, the unstandardized regression coefficients may be more directly interpretable than the standardized estimates, since the unstandardized coefficients for such predictors simply estimate the differences in the average value of the outcome between the two groups defined by the predictor, holding the other predictors in the model constant.

4.3 Categorical Predictors

In Chap. 3, the simple regression model was introduced with a single continuous predictor. However, predictors in both simple and multipredictor regression models can be binary, categorical, or discrete numeric, as well as continuous numeric.

4.3.1 Binary Predictors

The exercise variable in the model for LDL levels shown in Table 4.1 is an example of a binary predictor. A good way to code such a variable is as an *indicator* or *dummy* variable, taking the value 1 for the group with the characteristic of interest, and 0 for the group without the characteristic. With this coding, the regression coefficient corresponding to this variable has a straightforward interpretation as the increase or decrease in average outcome levels in the group with the characteristic, with respect to the reference group.

To see this, consider the simple regression model for average glucose values:

$$E[\text{glucose}|x] = \beta_0 + \beta_1 \text{exercise}. \quad (4.6)$$

With the indicator coding of `exercise` (1 = yes, 0 = no), the average value of glucose is $\beta_0 + \beta_1$ among women who do exercise, and β_0 among the rest. It follows

directly that β_1 is the difference in average glucose levels between the two groups. This is consistent with our more general definition of β_j as the change in $E[y|\mathbf{x}]$ for a one-unit increase in x_j . Furthermore, the t -test of the null hypothesis $H_0: \beta_1 = 0$ is a test of whether the between-group difference in average glucose levels differs from zero. In fact, this unadjusted model is equivalent to a t -test comparing glucose levels in women who do and do not exercise. A final point: when coded this way, the average value of the exercise variable gives the proportion of women who exercise.

A commonly used alternative coding for binary variables is (1 = yes, 2 = no). With this coding, the coefficient β_1 retains its interpretation as the between-group difference in average glucose levels, but now among women who do not exercise as compared to those who do, a less intuitive way to think of the difference. Furthermore, with this coding the coefficient β_0 has no straightforward interpretation, and the average value of the binary variable is not equal to the proportion of the sample in either group. However, overall model fit, including fitted values of the outcome, standard errors, and P -values, are the same with either coding (Problem 4.1).

4.3.2 Multilevel Categorical Predictors

The 2,763 women in the HERS cohort also responded to a question about how physically active they considered themselves compared to other women their age. The five-level response variable `physact` ranged from “much less active” to “much more active,” and was coded in order from 1 to 5. This is an example of an *ordinal* variable, as described in Chap. 2, with categories that are meaningfully ordered, but separated by increments that may not be accurately reflected in the numerical codes used to represent them. For example, responses “much less active” and “somewhat less active” may represent a larger difference in physical activity than “somewhat less active” and “about as active.”

Multilevel categorical variables can also be *nominal*, in the sense that there is no intrinsic ordering in the categories. Examples include ethnicity, marital status, occupation, and geographic region. With nominal variables, it is even clearer that the numeric codes often used to represent the variable in the database cannot be treated like the values of a numeric variable such as glucose.

Categories are usually set up to be mutually exclusive and exhaustive, so that every member of the population falls into one and only one category. In that case, both ordinal and nominal categories define subgroups of the population.

Both types of categorical variables are easily accommodated in multipredictor linear and other regression models, using indicator or dummy variables. As with binary variables, where two categories are represented in the model by a single indicator variable, categorical variables with $K \geq 2$ levels are represented by $K - 1$ indicators, one for each of level of the variable except a baseline or reference level. Suppose level 1 is chosen as the baseline level. Then, for $k = 2, 3, \dots, K$, indicator variable k has value 1 for observations belonging to the category k , and 0 for observations belonging to any of the other categories. Note that for $K = 2$, this

Table 4.3 Coding of indicators for a multilevel categorical variable

physact	Indicator variables			
	2.physact	3.physact	4.physact	5.physact
Much less active	0	0	0	0
Somewhat less active	1	0	0	0
About as active	0	1	0	0
Somewhat more active	0	0	1	0
Much more active	0	0	0	1

also describes the binary case, in which the “no” response defines the baseline or reference group and the indicator variable takes on value 1 only for the “yes” group.

Stata automatically defines indicator variables using `i.` variable prefix. By default, it uses the level with the lowest value as the reference group, although this is easily modified using a variable prefix of the form `ibk`, where k is the code of the alternative baseline category. Following the Stata convention for the naming of the four indicator variables, Table 4.3 shows the values of the four indicator variables corresponding to the five response levels of `physact`. Each level of `physact` is defined by a unique pattern in the four indicator variables.

Furthermore, the corresponding β s have a straightforward interpretation. For the moment, consider a simple regression model in which the five levels of `physact` are the only predictors. Then,

$$E[\text{glucose}|\mathbf{x}] = \beta_0 + \beta_2 \cdot \text{physact} + \cdots + \beta_5 \cdot \text{physact}. \quad (4.7)$$

For clarity, the β s in (4.7) are indexed in accord with the levels of `physact`, so β_1 does not appear in the model. Letting the four indicators take on values of 0 or 1 as appropriate for the five groups defined by `physact`, we obtain

$$E[\text{glucose}|\mathbf{x}] = \begin{cases} \beta_0 & \text{physact} = 1 \\ \beta_0 + \beta_2 & \text{physact} = 2 \\ \beta_0 + \beta_3 & \text{physact} = 3 \\ \beta_0 + \beta_4 & \text{physact} = 4 \\ \beta_0 + \beta_5 & \text{physact} = 5. \end{cases} \quad (4.8)$$

From (4.8), it is clear that the intercept β_0 gives the value of $E[\text{glucose}|\mathbf{x}]$ in the reference or much less active group (`physact` = 1). Then it is just a matter of subtracting the first line of (4.8) from the second to see that β_2 gives the difference in the average glucose in the somewhat less active group (`physact` = 2) as compared to the much less active group. Accordingly, the t -test of $H_0: \beta_2 = 0$ is a test of whether average glucose levels are the same in the much less and somewhat less active groups (`physact` = 1 and 2). And similarly for β_3 , β_4 , and β_5 .

Four other points are to be made from (4.8).

- Without other predictors, or covariates, the model is equivalent to a one-way ANOVA (Problem 4.9). Also, the model is said to be *saturated* and the population

group means would be estimated under model (4.8) by the sample averages. With covariates, the estimated means for each group would be adjusted for between-group differences in the covariates included in the model.

- The parameters of the model can be manipulated to give the estimated mean in any group, using (4.8), or to give the estimated differences between any two groups. For instance, the difference in average outcome levels between the much more and somewhat more active groups is equal to $\beta_5 - \beta_4$ (why?). All regression packages make it straightforward to estimate and test hypotheses about these *contrasts*. This implies that choice of reference group is in some sense arbitrary. While a particular choice may be best for ease of presentation, possibly because contrasts with the selected reference group are of primary interest, alternative reference groups result in essentially the same model.
- The five estimated group means can take on almost any pattern with respect to each other, in either the adjusted or unadjusted model. In contrast, if `physact` were treated as a score with integer values 1 through 5, the estimated means would be constrained to lie on a straight regression line.

Table 4.4 shows results for the model with `physact` treated as a categorical variable, again using data for women without diabetes in HERS. In the regression output, $\hat{\beta}_0$ is found in the column and row labeled `Coef.` and `_cons`; we see that average glucose in the much less active group is approximately 98.4 mg/dL. The differences between the reference group and the two most active groups are statistically significant; for instance, the average glucose level in the much more active group (`5.physact`) is 3.3 mg/dL lower than in the much less active group ($t = -2.92$, $P = 0.003$).

Using (4.8), the first `lincom` command after the regression computes the estimated mean in the somewhat less active group, equal to the sum of $\hat{\beta}_0$ (`_cons`) and $\hat{\beta}_2$ (`2.physact`), or 97.6 mg/dL (95% CI 96.5–98.6 mg/dL). The `margins` command is then used to estimate the mean level in all five groups.

We can also use the `lincom` command to assess pairwise differences between two groups when neither is the referent. For example, the second `lincom` result in Table 4.4 shows that average glucose is 2.1 mg/dL lower in among women in the much more active (`physact = 5`) group as compared to those who are about as active (`physact = 3`), and that this difference is statistically significant ($t = -2.86$, $P = 0.004$).

The newer command `contrast{physact 0 0 -1 0 1}` is also used to compare groups 3 and 5. The contrast coefficients correspond in order to the five levels of `physact`. The two nonzero coefficients, -1 for group 3 and 1 for group 5, directly reflect the `lincom` command, and the three zeroes correspond to the omitted groups. The `effects` option is needed to obtain the estimated between-group difference and 95% confidence interval supplied by default by the `lincom` command. We explain *contrasts* in more detail in Sect. 4.3.5 below.

Table 4.4 Regression of physical activity on glucose

```
. regress glucose i.physact if diabetes == 0
```

Source	SS	df	MS
Model	1673.09022	4	418.272554
Residual	191344.609	2027	94.3979322
Total	193017.699	2031	95.0357946

Number of obs = 2032

F(4, 2027) = 4.43

Prob > F = 0.0014

R-squared = 0.0087

Adj R-squared = 0.0067

Root MSE = 9.7159

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
physact					
2	-.8584489	1.084152	-0.79	0.429	-2.984617 1.267719
3	-1.226199	1.011079	-1.21	0.225	-3.20906 .7566629
4	-2.433855	1.010772	-2.41	0.016	-4.416114 -.451595
5	-3.277704	1.121079	-2.92	0.003	-5.476291 -1.079116
_cons	98.42056	.9392676	104.78	0.000	96.57853 100.2626

```
. lincom _cons + 2.physact
```

```
( 1) 2.physact + _cons = 0
```

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	97.56211	.5414437	180.19	0.000	96.50027 98.62396

```
. margins physact
```

Adjusted predictions

Model VCE : OLS

Expression : Linear prediction, predict()

Number of obs = 2032

	Margin	Std. Err.	z	P> z	[95% Conf. Interval]
physact					
1	98.42056	.9392676	104.78	0.000	96.57963 100.2615
2	97.56211	.5414437	180.19	0.000	96.5009 98.62332
3	97.19436	.3742409	259.71	0.000	96.46086 97.92786
4	95.98671	.3734108	257.05	0.000	95.25483 96.71858
5	95.14286	.6120416	155.45	0.000	93.94328 96.34244

```
. lincom 5.physact - 3.physact
```

```
( 1) - 3.physact + 5.physact = 0
```

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	-2.051505	.717392	-2.86	0.004	-3.458407 -.6446024

(continued)

Table 4.4 (continued)

```
. contrast {physact 0 0 -1 0 1}, effects
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
physact	1	8.18	0.0043

	Contrast	Std. Err.	t	P> t	[95% Conf. Interval]
physact					
(1)	-2.051505	.717392	-2.86	0.004	-3.458407 - .6446024

Table 4.5 Overall physical activity effects on glucose

```
. quietly regress glucose i.physact if diabetes == 0

. testparm i.physact
      F( 4, 2027) =    4.43
      Prob > F =    0.0014
```

```
. contrast physact
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
physact	4	4.43	0.0014

4.3.3 The F-Test

Although every pairwise contrast between levels of a categorical predictor is readily available, the *t*-tests for these multiple comparisons provide no overall evaluation of the importance of the categorical variable, or more precisely a single test of the null hypothesis that the mean level of the outcome is the same at all levels of this predictor. In the example, this is equivalent to a test of whether any of the four coefficients corresponding to `physact` differ from zero. The `testparm` result in Table 4.5 ($F(4, 2027) = 4.43, P = 0.0014$) shows that glucose levels clearly differ among the groups defined by `physact`. The same result is also obtained using the `contrast` command.

4.3.4 Multiple Pairwise Comparisons Between Categories

When the focus is on the difference between a single prespecified pair of subgroups, the overall *F*-test is of limited interest and the *t*-test for the single contrast between

those subgroups can be used without inflation of the type-I error rate. All levels of the categorical predictor should still be retained in the analysis, however, because residual variance can be reduced, sometimes substantially, by splitting out the remaining groups. Furthermore, this avoids combining the remaining subgroups with either of the prespecified groups, focusing the contrast on the comparison of interest.

However, it is frequently of interest to examine multiple pairwise differences between levels of a categorical predictor, especially when the overall F -test is statistically significant, and in some cases even when it is not. Examples include comparisons between treatments in a clinical trial with more than one active treatment arm, or in longitudinal data, to be discussed in Chap. 7, when between-treatment differences are evaluated at multiple points in time. We also discuss the implications of multiple comparisons for model selection in Sect. 10.3.2, and more broadly in Sect. 13.4.1.

For this case, various methods are available for controlling the familywise error rate (FER) for the wider set of comparisons being made. These methods differ in the trade-off made between power and the breadth of the circumstances under which the type-I error rate is protected. One of the most straightforward is Fisher's *least significant difference* (LSD) procedure, in which the pairwise comparisons are carried out using t -tests at the nominal type-I error rate, but only if the overall F -test is statistically significant; otherwise the null hypothesis is accepted for all the pairwise comparisons. This protects the FER under the *complete null hypothesis* that all the group-specific population means are the same. However, it is subject to inflation of the FER under *partial null hypotheses*—that is, when there are some real population differences between subgroups.

More conservative procedures that protect the FER under partial null hypotheses include setting the level of the pairwise tests required to declare statistical significance equal to α/k (Bonferroni) or $1-(1-\alpha)^{1/k}$ (Sidak), where α is the desired FER and k is the number of preplanned comparisons to be made. The Sidak correction is slightly more liberal for small values of k , but otherwise equivalent. The Scheffé method is another, although very conservative, method in which differences can be declared statistically significant only when the overall F -test is also statistically significant. The Tukey *honestly significant difference* (HSD) and Tukey–Kramer methods are more powerful than the Bonferroni, Sidak, or Scheffé approaches and also perform well under partial null hypotheses.

As noted in Sect. 3.1.5, the Bonferroni, Sidak, and Scheffé procedures are available with the oneway ANOVA in Stata. In addition, beginning with Version 12, the `contrast` and `margins` postestimation commands implement analogous pairwise comparisons for all regression models discussed in this book, with control of FER using the Bonferroni, Sidak, and Scheffé procedures available via the `mcompare` option. These new commands have extensive capabilities for postestimation hypothesis testing, a few of which are illustrated below, and many others beyond the scope of this book. In Table 4.5, we obtained Bonferroni-corrected comparisons with the reference level of `physact` using the command `contrast`

Table 4.6 Bonferroni-corrected physical activity effects

```
. regress glucose i.physact if diabetes == 0
```

Source	SS	df	MS	Number of obs = 2032		
Model	1673.09022	4	418.272554	F(4, 2027)	=	4.43
Residual	191344.609	2027	94.3979322	Prob > F	=	0.0014
Total	193017.699	2031	95.0357946	R-squared	=	0.0087
				Adj R-squared	=	0.0067
				Root MSE	=	9.7159
glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
physact						
2	-.8584489	1.084152	-0.79	0.429	-2.984617	1.267719
3	-1.226199	1.011079	-1.21	0.225	-3.20906	.7566629
4	-2.433855	1.010772	-2.41	0.016	-4.416114	-.451595
5	-3.277704	1.121079	-2.92	0.003	-5.476291	-1.079116
_cons	98.42056	.9392676	104.78	0.000	96.57853	100.2626

```
. contrast physact, mcompare(bonferroni) effects
```

Contrasts of marginal linear predictions

Margins : asbalanced

	df	F	P>F
physact	4	4.43	0.0014

Note: Bonferroni-adjusted p-values are reported for tests on individual contrasts only.

	Number of Comparisons
physact	4

	Contrast	Std. Err.	Bonferroni t P> t		Bonferroni [95% Conf. Interval]	
physact						
(2 vs base)	-.8584489	1.084152	-0.79	1.000	-3.56876	1.851862
(3 vs base)	-1.226199	1.011079	-1.21	0.901	-3.753832	1.301434
(4 vs base)	-2.433855	1.010772	-2.41	0.065	-4.96072	.093011
(5 vs base)	-3.277704	1.121079	-2.92	0.014	-6.080331	-.4750759

`physact, compare(bonferroni)`. Note that while the estimates and overall F -test are unchanged, the P -values for the pairwise comparisons are larger and the CIs wider than in the regression output (Table 4.6).

A special case arises when only comparisons with a single reference group are of interest, as might arise in a clinical trial with multiple treatments and a single placebo control. In this situation, Dunnett's test achieves better power than

alternatives designed for all pairwise comparisons, while still protecting the FER under partial null hypotheses. It also illustrates the general principle that controlling the FER for a smaller number of contrasts is less costly in terms of power, so that it makes sense to control only for the contrasts of interest. Compare this approach to Scheffé's, which controls the FER for all possible contrasts but at a considerable expense in power.

The previous alternatives provide simultaneous inference on all the pairwise comparisons considered. Various *step-down* and *step-up* multiple-stage testing procedures attempt to improve power using testing of cleverly sequenced hypotheses that only continues as long as the test results are statistically significant. The Duncan and Student-Newman-Keuls procedures fall in this class. However, neither protects the FER under partial null hypotheses.

4.3.5 Testing for Trend Across Categories

The coefficient estimates for the categories of `physact` shown in Table 4.4 decrease in order, suggesting that mean glucose levels are characterized by a linear trend across the levels of `physact`. Tests for linear trend are best performed using a *contrast* in the coefficients corresponding to the various levels of the categorical predictor.

Definition: A *contrast* is a weighted sum of the regression coefficients of the form $a_1\beta_1 + a_2\beta_2 + \cdots + a_p\beta_p$ in which the weights, or *contrast coefficients*, sum to zero: that is, $a_1 + a_2 + \cdots + a_p = 0$.

The contrasts used to test for trend can be motivated as linear regressions of the adjusted means for each category on the categorical variable, treated as a continuous predictor, after centering and possibly rescaling the numeric codes used for each category. The resulting contrast coefficients used to test for linear trend have a simple pattern: they are

- Integer-valued
- Evenly spaced
- Symmetric about zero

Using integers is just a convenience. Underlying the even spacing is the assumption that the “distances” between adjacent categories are all the same; below, we briefly outline how this assumption can be relaxed. Symmetry about zero implies that they also sum to zero, as required.

To make this specific, the contrast coefficients that we would use to test for trend across the five levels of `physact` are $-2, -1, 0, 1, \text{ and } 2$. More generally, when the number of levels is odd, the contrast coefficients are sequential integers (spacing of one), and by symmetry, the middle category has coefficient 0 and drops out. Thus for three categories, the coefficients are $-1, 0, \text{ and } 1$, and for seven, follow in order from -3 to 3 . When the number of levels is even, a spacing of two is the smallest that gives integer-valued contrast coefficients, and none of the categories

Table 4.7 Trend test in a model omitting the intercept

```
. regress glucose ibn.physact if diabetes == 0, noconstant
```

Source	SS	df	MS	Number of obs = 2032		
Model	18987135.4	5	3797427.08	F(5, 2027) =40227.86		
Residual	191344.609	2027	94.3979322	Prob > F = 0.0000		
Total	19178480	2032	9438.22835	R-squared = 0.9900		
				Adj R-squared = 0.9900		
				Root MSE = 9.7159		

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
physact						
1	98.42056	.9392676	104.78	0.000	96.57853	100.2626
2	97.56211	.5414437	180.19	0.000	96.50027	98.62396
3	97.19436	.3742409	259.71	0.000	96.46043	97.9283
4	95.98671	.3734108	257.05	0.000	95.2544	96.71902
5	95.14286	.6120416	155.45	0.000	93.94256	96.34315


```
. * Tests for linear trend
. test -2*1.physact - 2.physact + 4.physact + 2*5.physact = 0
( 1) - 2*1bn.physact - 2.physact + 4.physact + 2*5.physact = 0
      F( 1, 2027) = 12.11
      Prob > F = 0.0005

. contrast {physact -2 -1 0 1 2}, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
physact	1	12.11	0.0005


```
. contrast q(1).physact, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
physact	1	12.11	0.0005

are omitted. Thus with four categories, the contrast coefficients are $-3, -1, 1$, and 3 , and with six, they are $-5, -3, -1, 1, 3$, and 5 . So it is easy to figure out the contrast coefficients for any number of categories.

Table 4.7 shows a linear regression of glucose levels on physical activity, omitting the intercept, which we obtain by specifying `ibn.physact` in the `regress` command, in combination with the option `noconstant`. In this model, the group means for levels 1–5 of `physact` are given by $\beta_1, \beta_2, \beta_3, \beta_4$ and β_5 , rather than by (4.8). The `test` command calculates the contrast using the contrast coefficients $-2, -1, 0, 1$, and 2 , then compares it to the null value of zero; again, β_3 , corresponding to the middle category, drops out. The result ($F(1, 2027) = 12.11, P = 0.0005$) leaves little doubt that there is a declining trend in mean glucose with increasing levels of physical activity.

Table 4.7 also shows two other methods for obtaining the test for linear trend. The first, using the command `contrast{physact -2 -1 0 1 2}`, incorporates the contrast coefficients for the five categories directly, in the same order as the levels of `physact`; this approach was also used to contrast levels 3 and 5 of `physact` in Table 4.4.

The second method uses the so-called *contrast operator* `q(1)`. Including `(1)` as part of the operator specifies the test for linear trend; the default is to provide additional tests for quadratic, cubic, and quartic trends, plus a joint test for all four patterns. In both commands, the `noeffects` option prevents Stata from printing the numeric values of the contrasts, which are uninterpretable in this case.

The `q`. contrast operator treats the ordered categories as evenly spaced, regardless of the coding of the categorical variable. This assumption can be relaxed using the `p`. operator instead, in combination with a coding for the categorical variable that reflects the hypothesized spacing. For example, if we hypothesized spacings of 2, 1, 1, and 2 units between the categories of the physical activity variable, coding the levels as 1, 3, 4, 5, and 7, then testing for linear trend using the command `contrast p(1).physact, noeffects` would obtain the appropriate test.

Of course, the default in Stata and other statistical packages is to include the intercept in almost all regression models; in the Cox model, introduced in Chap. 6, the baseline hazard plays this role. When an intercept is included in the model, one level of the categorical variable must generally serve as the reference category and be omitted from the model. This default form of the model was laid out Table 4.3 and (4.8), and is obtained simply by specifying `i.physact` in the `regress` command.

Fortunately, we can easily adapt the integer-valued, evenly-spaced, symmetric, zero-sum contrast coefficients to the default form of the model with an intercept, simply by dropping the coefficient corresponding to the omitted reference category. To see why this works, and why the intercept does not figure in the contrast, we evaluate the contrast in the regression coefficients specifying the means for each level of `physact`, as shown in (4.8):

$$\begin{aligned} 0 &= -2\beta_0 - (\beta_0 + \beta_2) + (\beta_0 + \beta_4) + 2(\beta_0 + \beta_5) \\ &= -\beta_2 + \beta_4 + 2\beta_5 \end{aligned} \tag{4.9}$$

In (4.9), the mean for level three of `physact`, $\beta_0 + \beta_3$, is omitted because the contrast coefficient $a_3 = 0$, and β_0 disappears because the contrast coefficients sum to zero. Table 4.8 summarizes the resulting contrasts used to test for trend when the categorical variable has 3–6 levels and the lowest category is treated as the reference.

Table 4.9 shows the test for trend in glucose levels across the levels of `physact`, based on the default form of the model including an intercept. The trend test result is exactly the same as in Table 4.7, whether we use `test` or either version of the `contrast` command to obtain it.

Table 4.8 Trend contrasts for models with an intercept

Number of categories	Linear contrast
3	$\beta_3 = 0$
4	$-\beta_2 + \beta_3 + 3\beta_4 = 0$
5	$-\beta_2 + \beta_4 + 2\beta_5 = 0$
6	$-3\beta_2 - \beta_3 + \beta_4 + 3\beta_5 + 5\beta_6 = 0$

Table 4.9 Trend test in a model including the intercept

```
. regress glucose i.physact if diabetes == 0
```

Source	SS	df	MS	Number of obs =	2032
Model	1673.09022	4	418.272554	F(4, 2027) =	4.43
Residual	191344.609	2027	94.3979322	Prob > F =	0.0014
Total	193017.699	2031	95.0357946	R-squared =	0.0087
				Adj R-squared =	0.0067
				Root MSE =	9.7159

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
physact					
2	-.8584489	1.084152	-0.79	0.429	-2.984617 1.267719
3	-1.226199	1.011079	-1.21	0.225	-3.20906 .7566629
4	-2.433855	1.010772	-2.41	0.016	-4.416114 -.451595
5	-3.277704	1.121079	-2.92	0.003	-5.476291 -1.079116
_cons	98.42056	.9392676	104.78	0.000	96.57853 100.2626

```
. * Tests for linear trend
. test -2.physact + 4.physact + 2*5.physact = 0
( 1) - 2.physact + 4.physact + 2*5.physact = 0
      F( 1, 2027) = 12.11
      Prob > F = 0.0005

. contrast {physact -2 -1 0 1 2}, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
physact	1	12.11	0.0005

```
. contrast q(1).physact, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
physact	1	12.11	0.0005

A few more details about these trend tests are worth noting:

- In (4.9), we showed why β_0 does not figure in the contrasts in Table 4.8. By extension, the effects of any adjustment variables held constant by the model would also drop out.

- If a different reference category is used, we simply drop that component of the contrast rather than the first. For example, suppose we specified level two as the reference category for `physact` using `ib2.physact` in the `regress` command. Then the appropriate contrast would be $-2\beta_1 + \beta_4 + 2\beta_5 = 0$. If we specified level three as the reference category, using `ib3.physact`, the contrast would be $-2\beta_1 - \beta_2 + \beta_4 + 2\beta_5 = 0$. The trend test results are unaffected by changing the reference category.
- As compared to a simpler approach in which the categorical variable is treated as a continuous predictor, using the categorical version of the model in conjunction with contrasts to test for trend can be more efficient when there is *both* trend and departure from it, a problem we examine next. This occurs because the model captures the departures from linear trend, reducing the residual variance, and thus making regression effects easier to detect.
- These contrasts are valid for the other models in this book, including logistic, survival, repeated measures, and GLMs. In GLMs and Cox models, treating a multilevel predictor as categorical rather than continuous achieves no efficiency gain of the kind sometimes seen in linear models. Nonetheless, in such cases, treating the predictor as categorical rather than continuous should achieve at least somewhat better fit.
- Similar contrasts are available for assessing quadratic, cubic, and quartic trends across categories, now easily accessible using the `contrast` command with the `q.` and `p.` contrast operators.

4.3.5.1 Departures from Linear Trend

The pattern in average glucose across the levels of a categorical variable could be characterized by both a linear trend and a departure from trend. After demonstrating a statistically significant trend as in Table 4.7 or 4.9, it is easy to test for such a departure. One method for doing this uses a model in which the categorical variable is treated both as continuous and categorical. In this set-up, the continuous version accounts for the trend, while the categorical version captures departure from it. Thus, in Table 4.10 the F -test for the overall effect of `physact` as a categorical variable ($F(3, 2027) = 0.26, P = 0.85$) shows that there is little evidence for departures from a linear trend in this case.

Table 4.10 Testing for departure from linear trend

```
. quietly regress glucose physact i.physact if diabetes == 0
note: 5.physact omitted because of collinearity

. testparm i.physact
      F(   3,   2027) =    0.26
      Prob > F =    0.8511
```

Table 4.11 Testing for departure from linear trend

```
. quietly regress glucose i.physact if diabetes == 0

. contrast q(2/4).physact, noeffects
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
physact			
(quadratic)	1	0.11	0.7411
(cubic)	1	0.01	0.9415
(quartic)	1	0.49	0.4859
Joint	3	0.26	0.8511
Residual	2027		

Two additional comments about the model in Table 4.10:

- The omission of an additional category of `physcat` is expected, in fact necessary for the test for departure from trend to work. For this to occur, `physact` must precede `i.physact` in the regression command; with the reverse ordering, Stata would omit `physact` as continuous instead.
- *This model is only useful for testing from departure from trend.* Neither the coefficient nor the *t*-test for the effect of `physact` as continuous is interpretable. Estimation of the effects of the categorical variable as well as the test for trend must be carried out as in Table 4.7 or 4.9, using a model including the categorical version of the predictor only.

We can obtain exactly the same result from the original model including `physact` only as a categorical variable, using the contrast operator `q(2/4) . physact`. This assesses evidence for quadratic, cubic, and quartic trends, as well as evidence for all three jointly. Because we omitted the test for linear trend, the 3 degree-of-freedom joint test is equivalent to the first approach using `physact` as both continuous and categorical. Note that the specific form of the contrast operator depends on the number of levels: for example, we would need to use `contrast q(2/3) .physact` if `physact` had four levels, and `contrast q(2/5) .physact` if it had six (Table 4.11).

4.4 Confounding

In Table 4.1, the unadjusted coefficient for `exercise` estimates the difference in mean glucose levels between two subgroups of the population of women with heart disease. But this comparison ignores other ways in which those subgroups may differ. In other words, the analysis does not take account of confounding of the association we see. Although the unadjusted coefficient may be useful for describing differences between subgroups, it would be risky to infer any causal connection

between exercise and glucose on this basis. In contrast, the adjusted coefficient for exercise in Table 4.2 takes account of the fact that women who exercise also have lower BMI and are slightly younger and more likely to report alcohol use, all factors which are associated with differences in glucose levels.

While this adjusted model is clearly rudimentary, the underlying premise of multipredictor regression analysis of observational data is that with a sufficiently refined model (and good enough data), we can estimate causal effects, free or almost free of confounding. In Chap. 9, we use the concept of *potential outcomes* to define causal effects more precisely, and to show when multipredictor models can be used to estimate them in the presence of confounding, and when they cannot.

To summarize briefly, the overall point of Chap. 9 is that to assess confounding we first need a hypothesized causal framework. In particular, the potential confounder should be plausible as a cause of both the predictor of interest and the outcome, or as a proxy for such a cause. Within this hypothesized framework, the data provide support for confounding if we find that:

- The potential confounder is associated with the predictor of interest, and also independently associated with the outcome.
- The coefficient for the effect of the primary predictor on the outcome changes when we add the potential confounder to the model. Note, however, that analogous changes are also seen in logistic, Cox, and some other models, discussed in Chaps. 5, 6, and 8, when nonconfounders associated with the outcome but not the predictor of interest are added to the model.

4.4.1 Range of Confounding Patterns

Confounders often explain some of the association of a predictor of interest with the outcome, so that the adjusted effect, which may have a causal interpretation, is often weaker than the unadjusted effect. We saw this pattern in the estimate for the effect of exercise on glucose levels after adjustment for age, alcohol use, and BMI. However, qualitatively different patterns can arise. We now consider a small hypothetical example where $\mathcal{E}$, the exposure of primary interest, is binary and coded 0 and 1, and the potential confounder, $\mathcal{C}$, is continuous. At one extreme, the effect of a factor of interest may be completely confounded by a second variable. In the upper left panel of Fig. 4.1, $\mathcal{E}$ is shown to be strongly associated with y in unadjusted analysis, as represented in the scatterplot. However, the upper right panel shows that the unadjusted difference in y can be entirely explained by the continuous covariate $\mathcal{C}$. The regression lines for $\mathcal{C}$ are the same for both groups defined by $\mathcal{E}$; in other words, there is no association with $\mathcal{E}$ after adjustment for $\mathcal{C}$.

At the other extreme, we may find little or no association in unadjusted analysis, because it is *masked* or *negatively confounded* by another predictor. The lower panels of Fig. 4.1 show this pattern. On the left, there is clearly no association between the binary predictor $\mathcal{E}$ and y , but on the right the regression lines for $\mathcal{C}$

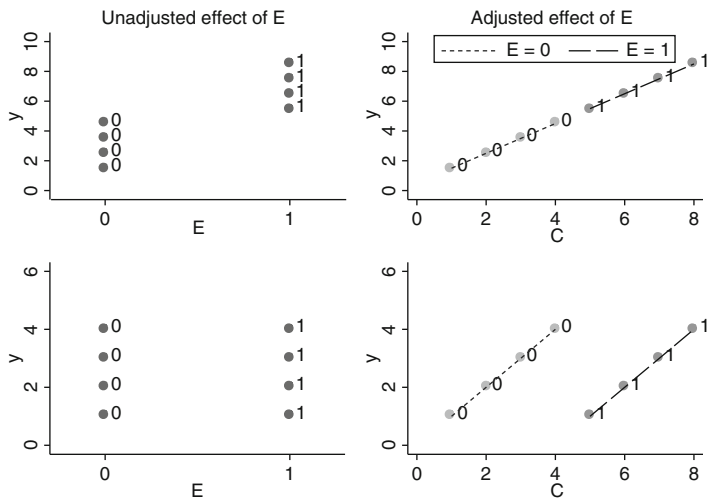


Fig. 4.1 Complete and negative confounding patterns

are very distinct for the groups defined by $\mathcal{E}$. In short, the association between $\mathcal{E}$ and y is unmasked by adjustment for $\mathcal{C}$. Negative confounding can occur under the following circumstances:

- The predictors are inversely correlated, but have regression coefficients with the same sign.
- The two predictors are positively correlated, but have regression coefficients with the opposite sign.

The example shown in the lower panels of Fig. 4.1 is of the second kind.

4.4.2 Confounding Is Difficult to Rule Out

The problem of confounding can be more resistant to multipredictor regression modeling than the example in Table 4.12 might suggest. We assumed in that example that the model included all confounders of the effect of BMI on LDL. Of course, the multipredictor linear model (4.2) can (within limits imposed by sample size) include more than a few predictors, giving us considerable freedom to model the effects of other causal factors. Nonetheless, for the multipredictor linear model to control confounding successfully and estimate causal effects without bias, all potential confounders must have been:

- Recognized and assessed by design in the study
- Measured without error
- Accurately represented in the systematic part of the model

Table 4.12 Unadjusted and adjusted regressions of LDL on BMI

```
. regress LDL bmi
```

Source	SS	df	MS	Number of obs =	2747
Model	14446.0223	1	14446.0223	F(1, 2745) =	10.14
Residual	3910928.63	2745	1424.74631	Prob > F =	0.0015
Total	3925374.66	2746	1429.48822	R-squared =	0.0037
				Adj R-squared =	0.0033
				Root MSE =	37.746

LDL	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
BMI	.4151123	.1303648	3.18	0.001	.1594894 .6707353
_cons	133.1913	3.7939	35.11	0.000	125.7521 140.6305

```
. regress LDL bmi age nonwhite smoking drinkany
```

Source	SS	df	MS	Number of obs =	2745
Model	42279.1877	5	8455.83753	F(5, 2739) =	5.97
Residual	3881903.3	2739	1417.27028	Prob > F =	0.0000
Total	3924182.49	2744	1430.09566	R-squared =	0.0108
				Adj R-squared =	0.0090
				Root MSE =	37.647

LDL	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
BMI	.3591038	.1341047	2.68	0.007	.0961472 .6220605
age	-.1897166	.1130776	-1.68	0.094	-.4114426 .0320095
nonwhite	5.219436	2.323673	2.25	0.025	.6631081 9.775764
smoking	4.750738	2.210391	2.15	0.032	.4165363 9.08494
drinkany	-2.722354	1.498854	-1.82	0.069	-5.661351 .2166444
_cons	147.3153	9.256449	15.91	0.000	129.165 165.4656

Logically, of course, it is not possible to show that all confounders have been measured, and in some cases it may be clear that they have not. Furthermore, the hypothetical causal framework may be uncertain, especially in the early stages of an investigating a research question. Also, measurement error in predictors is common; this may arise in some cases because the study has only measured proxies for the causal variables which actually confound a predictor of interest. Finally, Sect. 4.7 will show that accurate modeling of systematic relationships cannot be taken for granted.

4.4.3 Adjusted Versus Unadjusted $\hat{\beta}$ s

Uncontrolled confounding induces bias in unadjusted (or inadequately adjusted) estimates of the causal effects that are commonly the focus of our attention. This suggests that unadjusted parameter estimates are always biased and adjusted

estimates less so. But there is a sense in which this is misleading. In fact the two estimate different population quantities. The observed difference in average glucose levels between women who do and do not exercise is clearly interpretable, although it almost surely does not have a causal interpretation. Thus, it should not be expected to have the same value as the causal parameter.

4.4.4 Example: BMI and LDL

We turn to a relatively simple example, again using data from the HERS cohort. BMI and LDL cholesterol are both established heart disease risk factors. It is reasonable to hypothesize that higher BMI leads to higher LDL in some causal sense, to be made more specific in Chap. 9. An unadjusted model for BMI and LDL is shown in Table 4.12. The unadjusted estimate shows that average LDL increases .42 mg/dL per unit increase in BMI (95% CI: 0.16–0.67 mg/dL, $P = 0.001$). However, age, ethnicity (nonwhite), smoking, and alcohol use (drinkany) may confound this unadjusted association. These covariates may either represent determinants of LDL or be proxies for such determinants, and are correlated with but almost surely not caused by BMI, and so may confound the BMI–LDL relationship. After adjustment for these four demographic and lifestyle factors, the estimated increase in average LDL is 0.36 mg/dL per unit increase in BMI, an association that remains highly statistically significant ($P = 0.007$). In addition, average LDL is estimated to be 5.2 mg/dL higher among nonwhite women, after adjustment for between-group differences in BMI, age, smoking, and alcohol use. The association of smoking with higher LDL is also statistically significant, and there is some evidence for lower LDL among older women and those who use alcohol.

In this example, smoking is a negative confounder, because women with higher BMI are less likely to smoke, but both are associated with higher LDL. Negative confounding is further evidenced by the fact that the adjusted coefficient for BMI is *larger* (0.36 versus 0.32 mg/dL) in the fully adjusted model shown in Table 4.12 than in a model adjusted for age, nonwhite, and drinkany but not for smoking (reduced model not shown).

The covariates in the adjusted model shown in Table 4.12 can all be shown to meet sample diagnostic criteria for potential confounding of the effect of BMI. For example, LDL is 5.2 mg/dL higher and average BMI 1.7 kg/m² higher among nonwhite women, and the adjusted effect of BMI is 13% smaller than the unadjusted estimate. Note that while the associations of ethnicity with both BMI and LDL are statistically significant in this example, ethnicity might still meaningfully confound BMI even if the differences were not nominally significant. Evidence for this would still be provided by the substantial ($\geq 10\%$) change in the coefficient for BMI after adjustment for ethnicity, according to a useful (albeit ultimately arbitrary) rule of thumb (Greenland 1989). Recommendations for inclusion of potential confounders in multipredictor regression models are given in Chap. 10.

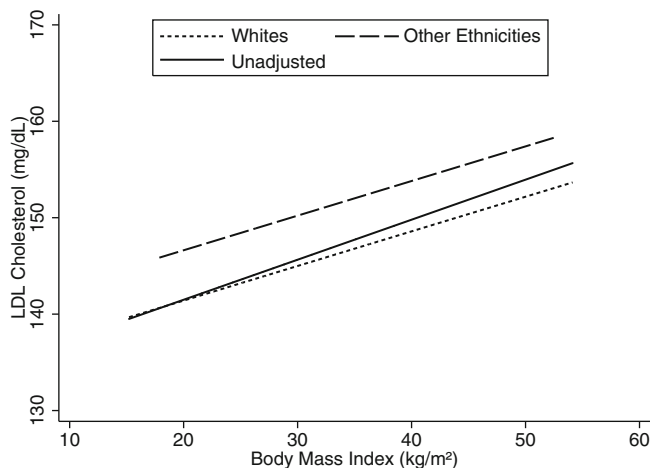


Fig. 4.2 Unadjusted and adjusted regression lines

Figure 4.2 shows the unadjusted regression line for LDL and BMI, together with the adjusted lines specific to the white and nonwhite women, holding the other variables constant at their respective means. Two comments about Fig. 4.2:

- Some of the upward slope of the unadjusted regression line reflects the fact that women with higher BMI are more likely to be nonwhite, younger, and not to use alcohol—all factors associated with higher LDL. Despite the negative confounding by smoking, when these all these effects are accounted for using the multipredictor regression model, the slope for BMI is attenuated.
- The adjusted regression lines for white and nonwhite women are parallel, both with the same slope of 0.36 mg/dL per unit increase in BMI. Similar patterns are assumed to hold for adjusted regression lines specific to subgroups defined by smoking and alcohol use. Accordingly, the lines are separated by a vertical distance of 5.2 mg/dL at every value of BMI—the adjusted difference in average LDL by ethnicity. This pattern reflects the fact that the model does not allow for interaction between BMI and ethnicity. We assume that the slope for BMI is the same in both ethnic groups, and, equivalently, that the difference in LDL due to ethnicity is the same at every value of BMI. Testing the no-interaction assumption will be examined in Sect. 4.6 below.

4.5 Mediation

In the adjusted model for LDL shown in Table 4.12, we assumed that age, race/ethnicity, smoking, and alcohol use might confound the effect of BMI, because they affect both BMI and LDL levels, or are proxies for factors that do. However,

if the primary predictor is a cause of one of the covariates, which in turn affects the outcome, this would be an instance of *mediation*. For example, statin drugs reduce low-density LDL cholesterol levels, which in turn appear to reduce risk of heart attack; in this model, reductions in LDL mediate the protective effect of statins.

Thus a potential mediator, like a potential confounder, must make sense in terms of a hypothetical causal framework. In particular, it should be plausible as an *effect* of the predictor of interest and as a *cause* of the outcome, or as a proxy for the true intermediary factor. Within this framework, the data support mediation if we find that:

- The potential mediator is associated with the predictor of interest *and* with the outcome, controlling for the predictor of interest.
- The coefficient for the effect of the primary predictor on the outcome changes when we add the potential mediator to the model. However, as with confounders, analogous changes are also seen in logistic, Cox, and some other models when nonmediators associated with the outcome but not the predictor of interest are added to the model.

Thus mediators behave like confounders in regression models, and can only be distinguished by the hypothesized causal framework—the data have little to tell us about the direction of the causal effects.

4.5.1 Indirect Effects via the Mediator

The effect of the primary predictor on the mediator, and of the mediator on the outcome, together comprise the hypothesized *indirect causal pathway* via the mediator. If the models used to estimate these effects adequately control confounding of both relationships, then the two effects may together have a causal interpretation as the *indirect effect* of the primary predictor; additional assumptions underlying this interpretation are discussed in Sect. 9.6. Accordingly, primary evidence for the indirect effect via the mediator is given by a test of the effect of the primary predictor on the mediator, in combination with a second test of the effect of the mediator on the outcome. The overall null hypothesis of no indirect effect is rejected only if *both* underlying null hypotheses are rejected at the nominal α level, preventing inflation of the type-I error rate.

4.5.2 Overall and Direct Effects

If the indirect pathway exists, and confounding has been controlled, then the coefficient for the primary predictor before adjustment for the mediator has a causal interpretation as the *overall effect* of the primary predictor on the outcome. The coefficient adjusted for the mediator is interpretable as the so-called *direct effect*

of the primary predictor via other pathways that do not involve the mediator. Finally, the *difference* between overall and direct effects of the primary predictor is interpretable as the indirect effect.

Tests for the difference between the overall and direct effects can also be used to assess mediation. However, these tests are complicated by the need to compare coefficient estimates for the primary predictor from two different models, but estimated using the same data. As a result, the two estimates are correlated, which must be taken into account. Surprisingly, these tests are less powerful in some cases than the joint test of the indirect pathway just discussed.

It is important to note that these interpretations may hold only under additional conditions in the generalized linear and Cox models discussed in Chaps. 5, 6, and 8. In particular, tests for the difference between the overall and direct effects can give false-positive results, because the *collapsibility* issue first introduced in Sect. 3.4.5. As we have already pointed out, in these models the coefficient for the primary predictor will generally change if a powerful predictor is added to the model. This holds even if the new covariate is not associated with the primary predictor, implying that it plays no mediating role.

4.5.3 Percent Explained

The *relative* difference between the overall and direct effects is sometimes referred to as the *percent explained* (PE) and used as an additional summary measure of the indirect effect. Direct estimation of PE rests on the assumption that the primary predictor and mediator do not interact (Robins and Greenland 1992; Freedman et al. 1992). This assumption can be checked using methods explained in Sect. 4.6, and possibly relaxed (Li et al. 2001; Vansteelandt 2009; VanderWeele 2009) as discussed briefly in Sect. 9.6. Testing and CI estimation for PE are even more complicated and problematic than for the difference between the overall and direct effects of the primary predictor.

4.5.4 Example: BMI, Exercise, and Glucose

We examined the extent to which the effects of BMI on glucose levels might be mediated through its effects on likelihood of exercise. Although exercise may in some cases affect BMI, in HERS exercise was weakly associated ($P = 0.06$) with a small *increase* in BMI over the first year of the study. As a result, we would argue that in this population of older women with established heart disease, BMI mainly affects likelihood of exercise, with very little feedback. Thus, mediation of the effects of BMI by exercise makes sense in terms of a hypothesized causal framework. We recognize that our simple models might not completely control confounding of the relationships among BMI, exercise, and glucose, and could be improved with expert input.

To assess mediation of the effects of BMI by exercise, we assessed both links in the hypothesized indirect pathway. Specifically, we first used a logistic model (Chap. 5) to assess the independent effects of BMI on likelihood of exercise, adjusting for age, race/ethnicity, smoking, alcohol use, and poor or fair self-reported health. Results in Table 4.13 show that each kg/m^2 increase in BMI is associated with an 8% decrease in the odds of exercise (95% CI 4–10%, $P < 0.0005$). In addition, the linear model for glucose levels establishes the second link in the indirect pathway, showing that exercise is independently associated with a decrease in average glucose of about 1 mg/dL (95% CI 0.1–1.9, $P = 0.027$). So the proposed mediator is associated with both the primary predictor and independently with the outcome. Since both null hypotheses are rejected at the nominal 2-sided 5% level, there is evidence for the indirect causal pathway via exercise.

On the other hand, the coefficient for BMI is only slightly attenuated when exercise is added to the model, from 0.50 to 0.49 mg/dL per kg/m^2 increase in BMI. We manipulated regression results stored as so-called `scalars` to calculate PE as $(0.5025557 - 0.4859684)/0.5025557 \times 100 = 3.3\%$. Thus, while our joint test of the indirect pathway shows that we can rule out chance at the nominal 5% level, only a very small part of the effect of BMI on glucose levels appears to be mediated by its effects on likelihood of exercising.

4.5.5 Pitfalls in Evaluating Mediation

Evaluating mediation, in particular estimating direct effects and PE, has many potential difficulties. In particular, bias can arise from uncontrolled confounding of the association between the mediator and the outcome (Robins and Greenland 1992; Cole and Hernán 2002)—even in clinical trials where the primary predictor is randomized treatment assignment. In observational data, we obviously need to control confounding of the effects of the primary predictor as well. Additional difficulties arise if a confounder of the mediator/outcome relationship is affected by treatment, and thus a causal intermediate (Petersen et al. 2006). We briefly cover these issues in Sect. 9.6.

4.5.5.1 Temporality

In addition, it is often difficult to infer causal direction in cross-sectional data. Longitudinal data may provide stronger support for the hypothesized indirect pathway by showing that changes or differences in the predictor of interest are associated with subsequent changes in the mediator, which in turn predict the outcome still later in time. However, if these changes all occur more or less simultaneously, and between sequential longitudinal observations, the temporal ordering can easily be obscured. Furthermore, as discussed in Sect. 6.3.1, longitudinal analyses set up to

Table 4.13 Indirect pathway from BMI to glucose levels via exercise

```

. * Overall effect of BMI on glucose, adjusting for age and alcohol use
. regress glucose BMI age10 nonwhite smoking drinkany poorfair if diabetes == 0

```

Source	SS	df	MS	Number of obs = 2025		
Model	13529.786	6	2254.96434	F(6, 2018) = 25.48		
Residual	178590.143	2018	88.4985842	Prob > F = 0.0000		
				R-squared = 0.0704		
				Adj R-squared = 0.0677		
Total	192119.929	2024	94.9209135	Root MSE = 9.4074		

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BMI	.5025557	.0414832	12.11	0.000	.4212013	.5839102
age10	.7093964	.3259568	2.18	0.030	.0701494	1.348643
nonwhite	.8801519	.7610825	1.16	0.248	-.6124377	2.372741
smoking	.1812593	.6135155	0.30	0.768	-1.021931	1.384449
drinkany	.7137293	.4305044	1.66	0.097	-.1305502	1.558009
poorfair	-.2052528	.5394217	-0.38	0.704	-1.263134	.8526288
_cons	77.63278	2.687214	28.89	0.000	72.36278	82.90279


```

. * Store coefficient for BMI as estimate of overall effect
. scalar overall = _b[BMI]

. * First link: logistic model for BMI effect on exercise
. logistic exercise BMI age10 nonwhite smoking drinkany poorfair if diabetes == 0

```

Logistic regression

Number of obs = 2025
LR chi2(6) = 158.56
Prob > chi2 = 0.0000
Log likelihood = -1294.4669
Pseudo R2 = 0.0577

exercise	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
BMI	.9235428	.0093154	-7.89	0.000	.9054643	.9419822
age10	.8171735	.0600467	-2.75	0.006	.7075662	.9437597
nonwhite	.8012592	.1416865	-1.25	0.210	.5665721	1.133159
smoking	.3012331	.0470011	-7.69	0.000	.2218658	.4089921
drinkany	.9159856	.0883199	-0.91	0.363	.758255	1.106527
poorfair	.523097	.0671846	-5.05	0.000	.406684	.6728331


```

. * Second link: fully adjusted model for effect of exercise on glucose levels
. regress glucose BMI age10 nonwhite smoking drinkany poorfair exercise ///
  if diabetes == 0

```

Source	SS	df	MS	Number of obs = 2025		
Model	13964.2063	7	1994.88661	F(7, 2017) = 22.59		
Residual	178155.723	2017	88.3270811	Prob > F = 0.0000		
				R-squared = 0.0727		
				Adj R-squared = 0.0695		
Total	192119.929	2024	94.9209135	Root MSE = 9.3982		

(continued)

Table 4.13 (continued)

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BMI	.4859684	.0421125	11.54	0.000	.4033798	.568557
age10	.6655835	.3262395	2.04	0.041	.0257819	1.305385
nonwhite	.8315359	.7606607	1.09	0.274	-.6602267	2.323299
smoking	-.0612536	.6225991	-0.10	0.922	-1.282258	1.159751
drinkany	.6954023	.4301665	1.62	0.106	-.1482147	1.539019
poorfair	-.3387946	.5422525	-0.62	0.532	-1.402228	.724639
exercise	-.9762492	.4402026	-2.22	0.027	-1.839548	-.1129499
_cons	78.86342	2.74136	28.77	0.000	73.48723	84.23961

```
. * Store coefficient for BMI as estimate of direct effect, and calculate PE
. scalar direct = _b[BMI]
. scalar PE = round((overall-direct)/overall*100, 0.1)
. scalar list PE
      PE =      3.3
```

examine such temporal patterns can be misleading if the mediator also potentially confounds the association between the primary predictor and outcome ([Hernán et al. 2001](#)).

4.5.5.2 Problems with PE

Finally, while PE is a popular and relatively interpretable measure of mediation, CIs for this measure can be wide and unreliable if the overall effect of the primary predictor is weak or noisily estimated. In addition, while PE is nominally a percentage, values outside the interval from 0% to 100% are possible. In particular, this occurs if the direct and indirect effects of the primary predictor are in opposite directions—for instance, if a treatment has both beneficial and adverse effects on the outcome, via different pathways. Even when PE is between 0% and 100%, confidence bounds are commonly outside this range. In addition, [Molenberghs et al. \(2002\)](#) show that estimates of PE are also influenced by the precision of measurements of both the mediator and outcome, potentially leading to highly misleading results.

4.6 Interaction

In Sect. 4.4, we gave examples in which a multipredictor linear model might be used to reduce or eliminate confounding of the effects of a primary predictor. So far, we have made the assumption that causal effect of the primary predictor was the same within strata defined by the covariates. However, this may not hold. In this section, we show how to use regression to model the resulting *interaction*, so that we can estimate causal effects that differ according to the level of a covariate. Interaction is also referred to as *effect modification* or *moderation*, and must be distinguished from both confounding and mediation ([Baron and Kenny 1986](#)).

Table 4.14 Model for interaction of HT and statins

Group	HT	statins	HT#statins	$E[\text{LDL} \mathbf{x}]$
1	0	0	0	β_0
2	1	0	0	$\beta_0 + \beta_1$
3	0	1	0	$\beta_0 + \beta_2$
4	1	1	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

4.6.1 Example: Hormone Therapy and Statin Use

As an example of interaction, we examine whether the effect of HT on LDL cholesterol differs according to baseline statin use, using data from HERS. To do this, a constructed interaction variable is useful. Suppose both assignment to HT and use of statins at baseline are coded using indicator variables. Then, the product of these two variables is also an indicator, equal to one only for the subgroup of women who reported using statins at baseline and were randomly assigned to HT, and zero for everyone else. Now, consider the regression model

$$E[\text{LDL}|\mathbf{x}] = \beta_0 + \beta_1\text{HT} + \beta_2\text{statins} + \beta_3\text{HT}\#\text{statins}, \quad (4.10)$$

where HT is the indicator of assignment to HT, statins the indicator of baseline statin use, and HT#statins the interaction term, which Stata calculates automatically.

Table 4.14 shows the values of (4.10) for each of the four groups of women defined by HT and statins. The difference in $E[y|\mathbf{x}]$ between groups 1 and 2 is β_1 , the effect of HT among women not using statins. Similarly, the difference in $E[y|\mathbf{x}]$ between groups 3 and 4 is $\beta_1 + \beta_3$, the effect of HT among statin users. So the interaction term β_3 gives the difference in treatment effects in these two groups. Accordingly, a t -test of $H_0: \beta_3 = 0$ is a test for the equality of the effects of HT among statin users as compared to nonusers. Note that both overall and within the strata defined by baseline statin use, we can assume that the groups randomly assigned to HT and placebo are comparable.

Taking analogous differences between groups 1 and 3 or 2 and 4 would show that β_2 gives the difference in average LDL among statin users as compared to nonusers among women assigned to placebo, while $\beta_2 + \beta_3$ gives the analogous difference among women assigned to HT. However, women were not randomized to statin use, so unbiased estimation of the causal effects of statin use would require careful adjustment for *confounding by indication*—that is, for the prognostic factors that lead physicians to prescribe this treatment.

Table 4.15 shows that there is some evidence for a smaller effect of HT on LDL among women reporting statin use at study baseline. The command `i.HT##i.statins` instructs Stata to include both so-called main effects, shown as `1.HT` and `1.statins` in the output, as well as the interaction term `HT#statins`, which it calculates only for the purposes of running the regression and does not retain in the data.

Table 4.15 Interaction of hormone therapy and statin use

```
. reg LDL1 i.HT#i.statins
```

Source	SS	df	MS	Number of obs = 2608		
Model	227141.021	3	75713.6735	F(3, 2604) = 52.68		
Residual	3742707.78	2604	1437.29177	Prob > F = 0.0000		
Total	3969848.8	2607	1522.76517	R-squared = 0.0572		
				Adj R-squared = 0.0561		
				Root MSE = 37.912		

LDL1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.HT	-17.72836	1.870629	-9.48	0.000	-21.39643	-14.06029
1.statins	-13.80912	2.15213	-6.42	0.000	-18.02918	-9.589065
HT#statins						
1 1	6.244416	3.076489	2.03	0.042	.2118042	12.27703
_cons	145.1567	1.325549	109.51	0.000	142.5575	147.756


```
. lincom 1.HT + 1.HT#1.statins
( 1) 1.HT + 1.HT#1.statins = 0
```

LDL1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	-11.48394	2.442444	-4.70	0.000	-16.27327	-6.694615

The coefficient for HT, or $\hat{\beta}_1$, shows that among women who did not report statin use at baseline, average cholesterol at the first annual HERS visit was almost 18 mg/dL lower in the HT arm than in placebo, a statistically significant subgroup treatment effect.

To obtain the estimate of the effect of HT among baseline statin users, we sum the coefficients for HT and HT#statins (that is, $\hat{\beta}_1 + \hat{\beta}_3$) using the `lincom` command. Note that in contrast to the `regress` command itself, where we used ## to obtain both main effects and interaction term, in the `lincom` command we used a single # to specify the interaction term only. The result shows that the treatment effect among baseline statin users was only -11.5 mg/dL, although this was also statistically significant. The difference ($\hat{\beta}_3$) of 6.2 mg/dL between the two treatment effects was also statistically significant ($t = 2.03$, $P = 0.042$). Finally, the results for variable `statins` indicate that among women assigned to placebo, baseline statin use is a statistically significant predictor of LDL levels at the first annual visit.

Finally, we note that in the `lincom` command shown in Table 4.15, we have to specify the values of each variable—in this case, 1 and 1—to which the interaction term applies. If either of the two main effects is a multicategory predictor, then the interaction would also have more than one level. For example, if we wanted to assess interaction between HT and level of physical activity, we would use the commands shown in Table 4.16. The `testparm` command is used to obtain a global test of the

Table 4.16 Interaction of hormone therapy and physical activity

```
. regress LDL1 i.HT##i.physact
```

Source	SS	df	MS
Model	160857.353	9	17873.0393
Residual	3808991.44	2598	1466.1245
Total	3969848.8	2607	1522.76517

Number of obs = 2608

F(9, 2598) = 12.19

Prob > F = 0.0000

R-squared = 0.0405

Adj R-squared = 0.0372

Root MSE = 38.29

LDL1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.HT	-4.973552	5.810288	-0.86	0.392	-16.36681	6.419711
physact						
2	4.386916	4.612377	0.95	0.342	-4.65739	13.43122
3	6.96232	4.338071	1.60	0.109	-1.544106	15.46875
4	8.797315	4.378699	2.01	0.045	.2112231	17.38341
5	6.793914	5.040489	1.35	0.178	-3.089867	16.67769
HT#physact						
1 2	-6.714054	6.799605	-0.99	0.324	-20.04725	6.619138
1 3	-10.71075	6.367042	-1.68	0.093	-23.19573	1.774244
1 4	-13.15391	6.411071	-2.05	0.040	-25.72523	-.5825811
1 5	-12.96408	7.314865	-1.77	0.076	-27.30763	1.379473
_cons	133.4211	3.928472	33.96	0.000	125.7178	141.1243

```
. testparm i.HT#i.physact
```

F(4, 2598) = 1.42

Prob > F = 0.2258

```
. contrast HT#physact
```

Contrasts of marginal linear predictions

Margins : asbalanced

	df	F	P>F
HT#physact	4	1.42	0.2258

interaction, which is not statistically significant, despite nearly significant P -values for the interaction terms for HT and levels 3, 4, and 5 of physical activity. The `contrast` command gives an equivalent result.

4.6.2 Example: BMI and Statin Use

Similar approaches can be used to assess modification of the effects of continuous predictors. For example, the association between BMI and baseline LDL cholesterol levels was shown in Sect. 4.4.4 to be statistically significant after adjustment for

demographics and lifestyle factors. However, treatment with statins may modify this association, possibly by interrupting the causal pathway between higher BMI and increased LDL. This would imply that BMI is less strongly associated with increased average LDL among statin users than among nonusers.

In examining this interaction, centering BMI about its mean value of 28.6 kg/m² makes the parameter estimate for statin use more interpretable, as shown below. Then, to implement the analysis, we would first compute `BMIC`, the new centered BMI variable. Note that because `statins` is an indicator variable coded 1 for users and 0 for nonusers, the interaction variable `statins#c.BMIC` automatically made by Stata is by definition equal to `BMIC` in statin users, but equal to zero for nonusers. We then fit a multipredictor regression model including all these three predictors, as well as the potential confounders adjusted for previously. The resulting model for baseline LDL is

$$E[LDL|\mathbf{x}] = \beta_0 + \beta_1 \text{statins} + \beta_2 \text{BMIC} + \beta_3 \text{statins}\#c.\text{BMIC} \\ + \beta_4 \text{age} + \beta_5 \text{nonwhite} + \beta_6 \text{smoking} + \beta_7 \text{drinkany}. \quad (4.11)$$

Thus, among women who do not use statins,

$$E[LDL|\mathbf{x}] = \beta_0 + \beta_2 \text{BMIC} \\ + \beta_4 \text{age} + \beta_5 \text{nonwhite} + \beta_6 \text{smoking} + \beta_7 \text{drinkany}, \quad (4.12)$$

and the slope associated with `BMIC` in this group is β_2 . In contrast, among statin users

$$E[LDL|\mathbf{x}] = \beta_0 + \beta_1 \text{statins} + \beta_2 \text{BMIC} + \beta_3 \text{statins}\#c.\text{BMIC} \\ + \beta_4 \text{age} + \beta_5 \text{nonwhite} + \beta_6 \text{smoking} + \beta_7 \text{drinkany} \\ = \beta_0 + \beta_1 + (\beta_2 + \beta_3) \text{BMIC} \\ + \beta_4 \text{age} + \beta_5 \text{nonwhite} + \beta_6 \text{smoking} + \beta_7 \text{drinkany}. \quad (4.13)$$

In this group, the slope associated with BMI is $\beta_2 + \beta_3$; so clearly the interaction parameter β_3 gives the difference between the two slopes. The model also posits that the difference in average LDL between statin users and nonusers depends on BMI. Subtracting (4.12) from (4.13), the difference in average LDL in statin users as compared to nonusers is $\beta_1 + \beta_3 \text{BMIC}$.

Table 4.17 shows the results of the interaction model for statin use and BMI. The estimated coefficients have the following interpretations:

- `statins`: Among women with `BMIC` = 0, or equivalently, with BMI = 28.6 kg/m², statin use was associated with LDL levels that were more than 16 mg/dL lower on average. Note that if we had not first centered BMI, this coefficient would be an estimate of the statin effect in women with BMI = 0.

Table 4.17 Interaction model for BMI and statin use

```
. regress LDL i.statins#c.BMIc age nonwhite smoking drinkany
```

Source	SS	df	MS
Model	216681.484	7	30954.4978
Residual	3707501	2737	1354.58568
Total	3924182.49	2744	1430.09566

Number of obs = 2745

F(7, 2737) = 22.85

Prob > F = 0.0000

R-squared = 0.0552

Adj R-squared = 0.0528

Root MSE = 36.805

LDL	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
1.statins	-16.25301	1.468788	-11.07	0.000	-19.13305 -13.37296
BMIc	.5821275	.160095	3.64	0.000	.2682082 .8960468
statins# c.BMIc					
1	-.701947	.2693752	-2.61	0.009	-1.230146 -.1737478
age	-.1728526	.1105696	-1.56	0.118	-.3896608 .0439556
nonwhite	4.072767	2.275126	1.79	0.074	-.3883704 8.533903
smoking	3.109819	2.16704	1.44	0.151	-1.139381 7.359019
drinkany	-2.075282	1.466581	-1.42	0.157	-4.950999 .8004355
_cons	162.4052	7.583312	21.42	0.000	147.5356 177.2748

```
. lincom BMIc + 1.statins#c.BMIc
```

```
( 1) BMIc + 1.statins#c.BMIc = 0
```

LDL	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	-.1198195	.2206807	-0.54	0.587	-.5525371 .3128981

- BMIc: Among women who do not use statins, the increase in average LDL is 0.58 mg/dL per unit increase in BMI. The association is statistically significant ($t=3.64$, $P < 0.0005$).
- statins#c.BMIc: The slopes for the average change in LDL per unit increase in BMI differ by approximately -0.70 mg/dL according to baseline statin use. That is, the increase in average LDL associated with increases in BMI is much less rapid among women who use statins. Moreover, the interaction is statistically significant ($t = -2.61$, $P = 0.009$).
- lincom is used to estimate the slope for BMI among statin users, equal to the sum of the slope among nonusers plus the estimated difference in slopes. The estimate of -0.12 mg/dL per unit increase in BMI is not statistically significant ($t = -0.54$, $P = 0.59$), but the 95% CI (-0.55 to 0.31 mg/dL per unit increase in BMI) is fairly wide.

Figure 4.3 shows the estimated regression lines in the two groups, demonstrating that the parallel lines assumption is no longer constrained to hold in the interaction model. In summary, the analysis suggests that the adverse effect of higher BMI on LDL may be blocked by statin use.

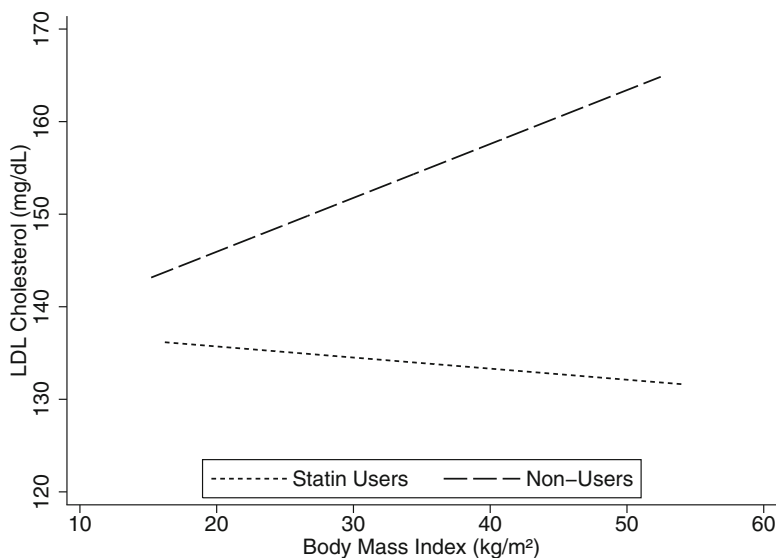


Fig. 4.3 Stratum-specific regression lines

4.6.3 Interaction and Scale

Interaction models are often distinguished from simpler *additive* models which do not include interaction terms. Moreover, the simpler additive model is generally treated as the default in predictor selection, with an interaction term being added only if there is more-or-less persuasive evidence that it is needed. It is important to recognize, however, that the need for interaction terms is dependent on the scale on which the outcome is measured (or, in the models discussed in later chapters, the scale on which its mean is modeled).

In Sects. 4.7.2 and 4.7.3 below we examine changes of the scale on which the outcome is measured to address violations of the linear model assumptions of normality and constant variance. Log transformation of the outcome, among the most commonly used changes of scale, effectively means modeling the average value of the outcome on a relative rather than absolute scale, as we show in Sect. 4.7.5 below. Similarly, in the analysis of before-and-after measurements of a response to treatment, we have the option of modeling percent rather than absolute change from baseline.

The issue of the dependence of interaction on scale arises in a similar but subtly different way with the other models discussed later in this book. For example, in logistic regression (Chap. 5) the *logit* transformation of $E[Y | \mathbf{x}]$ is modeled, while in some generalized linear models (GLMs; Chap. 8), including the widely used Poisson model, the log of $E[Y | \mathbf{x}]$ is modeled. Note that modeling $E[\log(Y) | \mathbf{x}]$, as we might do in a linear model, is different from modeling $\log(E[Y | \mathbf{x}])$ in the Poisson model.

Table 4.18 Interaction model for HT effects on absolute change in LDL

```
. regress LDLch HT#c.LDL0
```

Source	SS	df	MS	Number of obs = 2597		
Model	721218.969	3	240406.323	F(3, 2593) = 258.81		
Residual	2408575.51	2593	928.876015	Prob > F = 0.0000		
Total	3129794.48	2596	1205.62191	R-squared = 0.2304		
				Adj R-squared = 0.2295		
				Root MSE = 30.477		

LDLch	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.HT	-15.47703	1.196246	-12.94	0.000	-17.82273	-13.13134
cLDL0	-.3477064	.0225169	-15.44	0.000	-.3918593	-.3035534
HT#c.LDL0						
1	-.0786871	.0316365	-2.49	0.013	-.1407226	-.0166517
_cons	-4.888737	.8408392	-5.81	0.000	-6.537522	-3.239953

The need to model interaction depends on outcome scale because the simpler additive model can only hold exactly on one such scale, and may be an acceptable approximation on some scales but not others. This is in contrast to confounding; if C confounds $\mathcal{E}$, then it does so on every outcome scale. In the case of the linear model, the dependence of interaction on scale means that transformation of the outcome will sometimes succeed in eliminating an interaction.

4.6.4 Example: Hormone Therapy and Baseline LDL

The effect of HT on LDL cholesterol in the HERS trial was dependent on baseline values of LDL, with larger reductions seen among women with higher baseline values. An interaction model for absolute change in LDL from baseline to the first annual visit is shown in Table 4.18. Note that baseline LDL is centered in this model in order to make the coefficient for hormone therapy (HT) easier to interpret.

The coefficients in the model have the following interpretations:

- HT: Among women with the average baseline LDL level of 135 mg/dL, the effect of HT is to lower LDL an average of 15.5 mg/dL over the first year of the study.
- cLDL0: Among women assigned to placebo, each mg/dL increase in baseline LDL is associated with a 0.35 mg/dL greater decrease in LDL over the first year. That is, women with higher baseline LDL experience greater decreases in the absence of treatment; this is in part due to regression to the mean and in part to greater likelihood of starting use of statins.
- HT#c.LDL0: The effect of HT is to lower LDL an additional 0.08 mg/dL for each additional mg/dL in baseline LDL. In short, larger treatment effects are seen among women with higher baseline values. The interaction is statistically significant ($P = 0.013$).

Table 4.19 Interaction model for HT effects on percent change in LDL

```
. regress LDLpctch HT#c.LDL0
```

Source	SS	df	MS	Number of obs = 2597		
Model	233394.163	3	77798.0542	F(3, 2593) = 165.33		
Residual	1220171.82	2593	470.563756	Prob > F = 0.0000		
Total	1453565.98	2596	559.925263	R-squared = 0.1606		
				Adj R-squared = 0.1596		
				Root MSE = 21.692		

LDLpctch	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.HT	-10.79035	.8514335	-12.67	0.000	-12.45991	-9.120789
cLDL0	-.2162436	.0160265	-13.49	0.000	-.2476697	-.1848176
HT#c.LDL0						
1	.0218767	.0225175	0.97	0.331	-.0222773	.0660307
_cons	-1.284976	.5984713	-2.15	0.032	-2.458506	-.1114456

Inasmuch as the reduction in LDL caused by HT appears to be greater in proportion to baseline LDL, it is reasonable to ask whether the HT effect on *percent change* in LDL might be constant across baseline LDL levels. In that case, modeling an interaction between HT and the baseline value would not be necessary. This turns out to be the case, as shown in Table 4.19. In particular, the interaction term HT#c.LDL0 is no longer statistically significantly ($P = 0.331$) and could be dropped from the model. Note that the coefficient for HT now estimates the average *percent* change in LDL due to treatment, among women at the average baseline level. In summary, analyzing percent rather than absolute change in LDL eliminates the interaction between HT and baseline LDL.

4.6.5 Details

There are several other more general points to be made about dealing with interaction in multipredictor regression models.

- Interactions between two multilevel categorical predictors require extra care in coding and interpretation. Simple computation of interaction terms involving a categorical predictor will almost always give mistaken results. In contrast, the `i.` and `##` operators in Stata will handle this situation. However, suppose one of the predictors has four levels and the other three levels. Then the interaction is modeled using an extra $(4-1)(3-1) = 6$ indicator variables. Many different patterns are subsumed by the alternative hypothesis of interaction, only a few of which may be of interest or biologically plausible; moreover, the F -test for interaction may have low power.

- Interactions between two continuous variables are also tricky, especially if the two predictors are highly correlated. Both main effects in this case are hard to interpret. “Centering” of both variables on their respective sample means (Problem 4.6) resolves the interpretative problem only in part, since the coefficient for each predictor still refers only to the case where the value of other predictor is at its sample mean. Both the linearity of the interaction effect and the need for higher order interactions would need to be checked.
- In examining interactions, it is not enough to show that the predictor of primary interest has a statistically significant association with the outcome in a subgroup, especially when it is not a statistically significant predictor overall. So-called subgroup analysis of this kind can severely inflate the type-I error rate, and has a justifiably bad reputation in the analysis of clinical trials. Showing that the subgroup-specific regression coefficients are statistically different by testing for interaction sets the bar higher, is less prone to type-I error, and thus more persuasive (Brookes et al. 2001).
- Methods have been developed (Gail and Simon 1985) for assessing *qualitative interaction*, in which the sign of the coefficient for the predictor of interest differs across subgroups. This was nearly the case in the interaction of BMI and statin use. A more specific alternative of this kind is often easier to detect.
- Interaction can be hard to detect if the interacting variables are highly correlated. For example, it would be difficult to assess the interaction between two types of exposure if they occurred together either little or most of the time. This was not the case in the second HERS example, because statin use was reported by 36% of the cohort at baseline, and was uncorrelated with assignment to HT by virtue of randomization. However, in an observational cohort it might be much less common for women to report use of both medications. In that case, oversampling of dual users might be used if the interaction were of sufficient interest.

4.7 Checking Model Assumptions and Fit

In the simple linear model (4.1) as well as the multipredictor linear model (4.2), it has been assumed so far that $E[y|\mathbf{x}]$ changes linearly with each continuous predictor, and that the error term ε has a normal distribution with mean zero and constant variance for every value of the predictors. We have also implicitly assumed that model results are not unduly driven by any small subset of observations. Violations of these assumptions have the potential to bias regression coefficient estimates and undermine the validity of CIs and P -values.

In this section, we show how to assess the validity of the linearity assumption for continuous predictors and suggest modifications to the model which can make it more reasonable. We also discuss assessments of normality, how to transform the outcome in order to make this assumption approximately hold, and discuss conditions under which it may be relaxed. We then discuss departures from the

assumption of constant variance and methods for addressing them. Many of these procedures rely heavily on the transformations of both predictor and outcome that were introduced in Chap. 2. Finally, we show how to deal with *influential points*. Throughout, we emphasize the *severity* of departures, since model assumptions rarely hold exactly, and small departures are often benign, especially in large data sets. Nonetheless, careful attention to meeting model assumptions can prevent us from being seriously misled, and sometimes increase the efficiency of our analysis into the bargain.

4.7.1 Linearity

In modeling the effect of BMI on LDL, we have assumed that the regression is a straight line. However, this may not be an adequate representation of the true relationship. For example, we might find that average LDL stops increasing, or increases more slowly, among women with BMI in the upper reaches of its range—a *ceiling effect*. Analogously, the inverse relationship between BMI and HDL (“good”) cholesterol may depart from linearity, with floor effects among very heavy women.

4.7.1.1 Component-Plus-Residual Plots

In unadjusted analysis, checks for departures from linearity could be carried out using LOWESS, the nonparametric scatterplot smoother introduced in Chap. 2. This smoother approximates the regression line under the weaker assumption that it is smooth but not necessarily linear, with the degree of smoothness under our control, via the bandwidth. If the linear fit were satisfactory, the LOWESS curve would be close to the model regression line; that is, the nonparametric estimate found under the weaker assumption of smoothness would agree with the estimate found when linearity is assumed.

However, the direct approach of adding a LOWESS smooth to a scatterplot of predictor versus outcome is only effective for simple linear models with a single continuous predictor. For multipredictor regression models, the analogous plot would have to accommodate $p + 1$ dimensions, where p is the number of predictors in the model—hard to imagine even for $p = 2$. Moreover, nonparametric smoothers work less well in higher dimensions.

Fortunately, the residuals from a regression model make it possible to examine the linearity of the adjusted association between a given predictor and the outcome, after taking account of the other predictors in the model. The basic idea is to plot the residuals versus each continuous predictor in the model; then a nonparametric smoother is used to detect departures from a linear trend in the average value of the residuals across the values of the predictor. This is a *residual versus predictor* (RVP) plot, obtained in Stata using the `rvpplot` command.

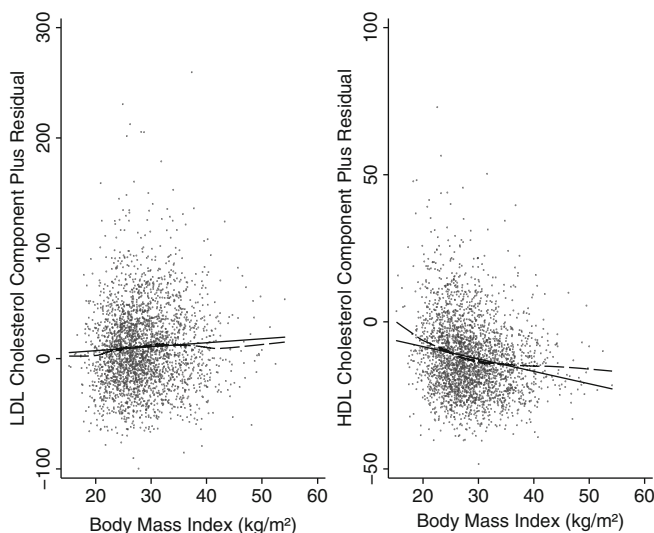


Fig. 4.4 CPR plots for multiple regressions of LDL and HDL on BMI

However, for doing this check in Stata, we recommend the closely related *component plus residual* (CPR) plot, mainly because the `cprplot` command allows LOWESS smooths, which we find more informative and easier to control than the smooths available with `rvpplot`. Rather than the residuals of the RVP plot, the residuals plus the component of the fitted values due to BMI are plotted and smoothed against BMI.

Figure 4.4 shows CPR plots for multipredictor regression models for LDL and HDL, each adjusting the estimated effect of BMI for age, ethnicity, smoking, and alcohol use, with solid lines representing the linear fits for BMI, and the dashed lines the LOWESS smooths of the plotted component-plus-residuals (CPRs) against BMI. If the linear fits for BMI were satisfactory, then there would be no nonlinear pattern across values of BMI in the CPRs. For LDL, shown on the left, the linear and LOWESS fits agree quite well, but for HDL, there is a substantial divergence. Thus the linearity assumption is rather clearly met by BMI in the model for LDL, but not in the model for HDL.

The curvature in the relationship between BMI and HDL can be approximated by adding a quadratic term in BMI to the multipredictor linear model. The fitted model is shown in Table 4.20.

For interpretability, we centered the linear term BMI_C on the sample mean of 28.6 kg/m^2 before calculating the quadratic term, BMI_C^2 , and also centered age. The linear and quadratic terms in centered BMI are both clearly needed ($P < 0.0005$). In this model, the intercept 47.6 estimates expected HDL for a 67-year old, white nonsmoking abstainer with $BMI = 28.6 \text{ kg/m}^2$. The BMI_C coefficient

Table 4.20 Linear plus quadratic model for effect of BMI on HDL

```
. regress HDL BMIC BMIC2 agec nonwhite smoking drinkany
```

Source	SS	df	MS	Number of obs = 2745		
Model	38474.0925	6	6412.34874	F(6, 2738) = 39.99		
Residual	439006.42	2738	160.338356	Prob > F = 0.0000		
Total	477480.512	2744	174.008933	R-squared = 0.0806		
				Adj R-squared = 0.0786		
				Root MSE = 12.662		

HDL	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
BMIC	-.5272063	.0507626	-10.39	0.000	-.6267432	-.4276693
BMIC2	.0242527	.0053231	4.56	0.000	.013815	.0346904
agec	.1893209	.0380347	4.98	0.000	.1147414	.2639005
nonwhite	2.494766	.7815733	3.19	0.001	.9622325	4.027299
smoking	-2.070298	.7449086	-2.78	0.005	-3.530938	-.6096584
drinkany	4.345096	.5041409	8.62	0.000	3.356561	5.333631
_cons	47.86615	.3794279	126.15	0.000	47.12215	48.61014

estimate of -0.53 estimates the decrease in average HDL per unit increase in BMI, *at the point where BMI = 28.6 kg/m²*, while the coefficient for BMIC2 captures the (upward) curvature of the regression line.

A CPR plot for the relationship between BMI and HDL in this model is shown in Fig. 4.5. Except at the extremes of the range of BMI, where the LOWESS smooth would usually be unreliable, the quadratic fit is clearly an improvement on the simpler model.

4.7.1.2 Smooth Transformations of the Predictors

In the example of HDL and BMI, the departure from linearity was approximately addressed by adding a quadratic term in BMI to the model. This solution is often useful when the regression line estimated by the LOWESS smooth is convex or concave, and especially if the line becomes steeper at either side of the CPR plot.

However, other transformations of the predictor may sometimes be more successful and should be considered. Figure 4.6 shows some of the predictor transformations commonly used to linearize the association between the predictor and the outcome. The upper left panel shows the typical curvature captured by adding a quadratic term in the predictor to the model. On the upper right, both quadratic and cubic terms have been included; in general, such higher order polynomial transformations are useful for S-shapes. A drawback is that these lines often fit badly in the tails of the predictor distribution, especially if the data there are sparse. As in the HDL example in Table 4.20, lower order terms are generally retained in polynomial models: specifically, we would include the linear term along with the quadratic term in the upper left panel, as well as with the quadratic plus cubic terms on the upper right.

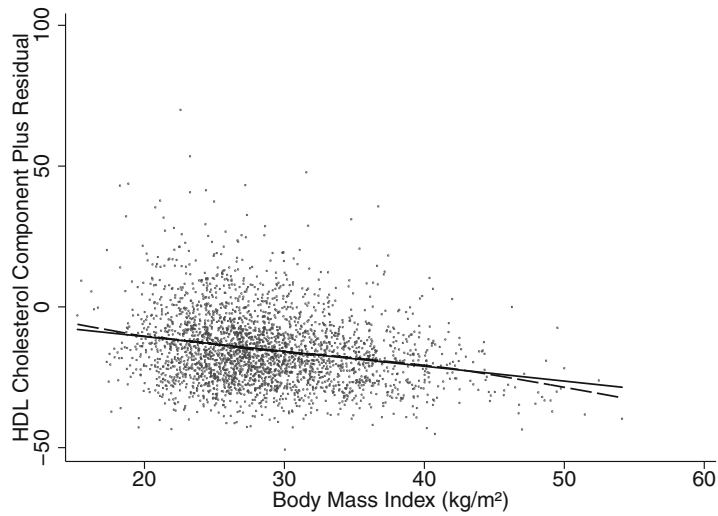


Fig. 4.5 CPR plot for HDL model with linear and quadratic terms in BMI

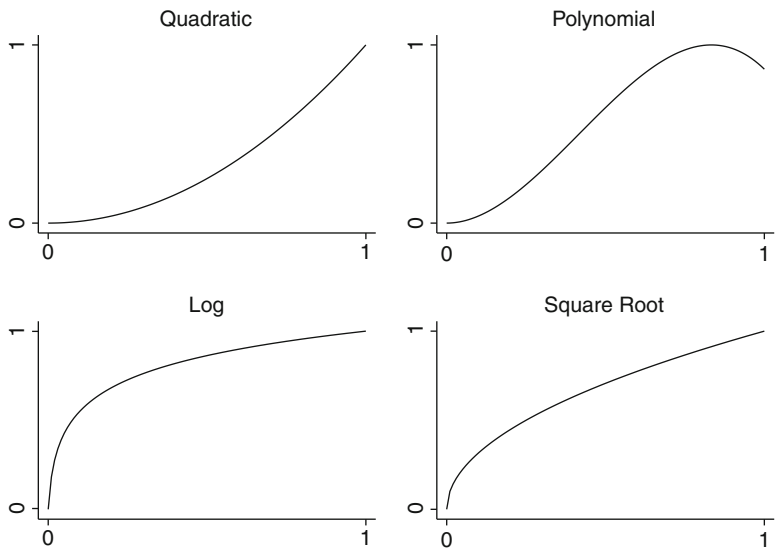


Fig. 4.6 Linearizing predictor transformations

The lower panels of Fig. 4.6 show the log and square root transformations, which are useful in situations where the regression line increases more slowly with increasing values of the predictor, as we might expect in cases of floor or ceiling effects, and more generally where the slope becomes less steep. Each of

these transformations would work just as well for modeling the mirror image of the nonlinear shape, reversed top-to-bottom. In Sect. 4.7.5 below, we discuss interpretation of the regression coefficients for a log-transformed predictor.

Comparison of the LOWESS smooth in CPR plots with the transformations in Fig. 4.6 can help identify the best candidate transformations. After the revised model is estimated, repeating the diagnostic using a new CPR plot then provides an initial check on the adequacy of the transformation: there should be no remaining pattern in the residuals, and the smooth should be close to the linear fit.

In cases where a quadratic or quadratic plus cubic term is added to the model, we can use t - or F -tests to evaluate the statistical significance of the addition to the model. This works because the original model is “nested” in the final model, in the sense that the predictors in the smaller model are a subset of those in the larger model. In other cases, for example, when we substitute the log-transformed for the untransformed predictor, the original and final models are not nested, so this testing procedure does not apply, although alternatives are available (Vuong 1989). In both cases, however, we can check whether R^2 improves substantially with the transformation.

4.7.1.3 Restricted Cubic Splines

Improving on the flexibility of polynomial transformations but with better behavior in the tails, *restricted cubic splines* are now implemented in Stata and other packages. This transformation requires selecting a small number of *knots*, or cutpoints, usually placed at symmetric percentiles of the predictor distribution. If there are k knots, the predictor is represented in the model by $k - 1$ spline variables. The effect of the predictor on the mean of the outcome is then modeled as cubic polynomials in the intervals between knots (achieving flexibility), is smooth at each knot (avoiding unrealistic sharp bends), but is constrained to be linear beyond the extreme knots (improving behavior in the tails). Suppose that in the model for the effect of BMI on HDL, we represent BMI by a restricted cubic spline with the default five knots. The results are shown in Table 4.21.

A primary advantage of restricted cubic splines is that the first of the $k - 1$ spline variables is just the untransformed predictor, so that all nonlinearity is captured by the other $k - 2$ variables. This affords a straightforward statistical test for departure from linearity, analogous to the tests for the contribution of quadratic and cubic terms in a polynomial model. The first F -test in Table 4.21 for the joint effect of the nonlinear components $\text{BMI}_{\text{sp}2}$, $\text{BMI}_{\text{sp}3}$, and $\text{BMI}_{\text{sp}4}$ confirms that the departure from linearity is important, despite the large t -test P -values. The second F -test confirms the overall importance of BMI for predicting HDL.

Another big advantage of restricted cubic splines is that graphical diagnostics for nonlinearity are considerably more difficult with the logistic, Cox, repeated measures, and GLMs presented in later chapters. However, departures from linearity can be conveniently assessed and modeled using restricted cubic splines in all of these settings.

Table 4.21 Restricted cubic spline model for effect of BMI on HDL

```
. mkspline BMIspl = BMI, cubic
. regress HDL BMIspl1 BMIspl2 BMIspl3 BMIspl4 age10 nonwhite smoking drinkany
```

Source	SS	df	MS		Number of obs =	2745
Model	38913.5934	8	4864.19917		F(8, 2736) =	30.35
Residual	438566.919	2736	160.294926		Prob > F =	0.0000
					R-squared =	0.0815
					Adj R-squared =	0.0788
Total	477480.512	2744	174.008933		Root MSE =	12.661

HDL	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
BMIspl1	-1.008258	.2823244	-3.57	0.000	-1.561849 - .4546676
BMIspl2	1.139488	2.424866	0.47	0.638	-3.615266 5.894242
BMIspl3	-.4761041	9.557886	-0.05	0.960	-19.21751 18.2653
BMIspl4	-1.757718	11.21143	-0.16	0.875	-23.74145 20.22601
age10	1.882574	.3807256	4.94	0.000	1.136035 2.629113
nonwhite	2.469817	.7823079	3.16	0.002	.9358431 4.003791
smoking	-2.097091	.7452066	-2.81	0.005	-3.558315 -.6358663
drinkany	4.376239	.5041816	8.68	0.000	3.387624 5.364854
_cons	62.2474	6.939817	8.97	0.000	48.63959 75.85521


```
. * test for departure from linearity
. test BMIspl2 BMIspl3 BMIspl4
      F( 3, 2736) = 7.84
      Prob > F = 0.0000

. * test for overall effect of BMI
. test BMIspl1 BMIspl2 BMIspl3 BMIspl4
      F( 4, 2736) = 27.67
      Prob > F = 0.0000
```

The primary disadvantage of restricted cubic splines is that the numeric results for BMIspl1, BMIspl2, BMIspl3, and BMIspl4 in Table 4.21 are uninterpretable. The resulting fit can only be adequately represented graphically, as in Fig. 4.7. The `adjustrcspline` command, part of the downloadable `postrcspline` package, can also be used to plot restricted cubic spline fits with CIs, for logistic and GLMs as well as standard linear models.

In addition, spline fits can be sensitive to the number of knots (Stone 1986). The flexibility of the fit increases with the number and placement of the knots, just as LOWESS smooths become more flexible with smaller bandwidths. In Stata, the default number is 5, but with datasets with fewer than 100 observations, 4 or 3 knots may work better. More than 5 knots are seldom necessary in large datasets unless the response to the predictor is unusually complicated. Plotting the fitted regression line is useful for judging the plausibility of the fit.

4.7.1.4 Categorizing the Predictor

Another transformation useful in exploratory analysis is to categorize the continuous predictor, either at cutpoints selected a priori or at percentiles that ensure adequate

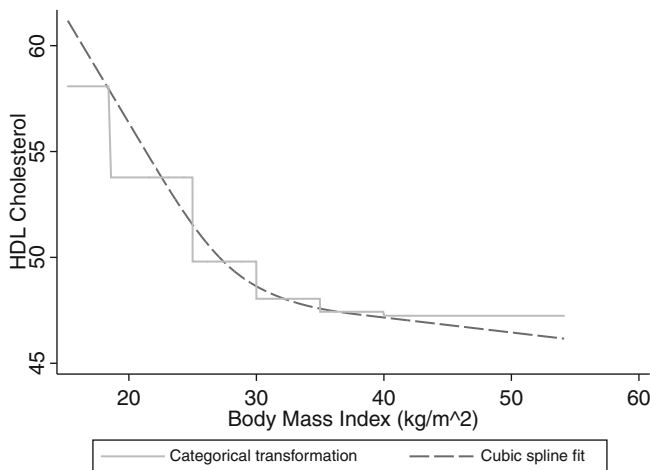


Fig. 4.7 HDL model with restricted cubic spline and categorical transformations of BMI

representation in each category. Then the model is estimated using indicators for all but the reference category of the transformed predictor, as in the `physact` example in Sect. 4.3. This method models the association between the ordinal categories and the outcome as a *step function*, also shown in Fig. 4.7. Although this approach is unrealistic in not providing a smooth estimate of the regression line, and also less efficient, it has the advantage of flexibility, in that each step can be of any height. Such transformations are also easy to understand, especially when the categories are defined by familiar clinical cutpoints. In contrast, smooth transformations, including polynomials and restricted cubic splines, are harder to motivate, present, and interpret.

4.7.1.5 Nonlinearity, Interaction, and Covariate Overlap

Apparent nonlinearity can sometimes mask interactions. For example, suppose that both the average value of a continuous predictor and its effect on the outcome differ across subgroups defined by a binary covariate. If we fail to model the interaction, the effect of the continuous predictor will appear nonlinear, even if its effects are completely linear *within* each subgroup. Furthermore, we show in Sect. 9.2.3 that unless there is considerable overlap in the values of the continuous predictor in the two subgroups—Fig. 9.1 is an extreme example—it can be difficult to distinguish non-linearity from effect modification by the covariate. This illustrates the difficulty of identifying a reasonably accurate model, especially if the sample size is small-to-moderate.

4.7.2 Normality

In Sect. 4.1, we stated that in the multipredictor linear model, the error term ε is assumed to have a normal distribution. Confidence intervals for regression coefficients and related hypothesis tests are based on the assumption that the coefficient estimates have a normal distribution. If ε has a normal distribution, and other assumptions of the multipredictor linear model are met, then ordinary least squares estimates of the regression coefficients can be shown to have a normal distribution, as required.

However, it can be shown that the regression coefficients are approximately normal in larger samples even if ε does not have a normal distribution. In that case, characterizing the distribution of the residuals is helpful for assessing whether the sample is large enough to trust the confidence intervals and hypothesis tests, since larger samples are required for this approximation to hold when departures from the normality of the errors are relatively serious. As with the t -test reviewed in Sect. 3.1, outliers are the principal worry with such departures, with the potential to erode the power of the model to detect real effects.

4.7.2.1 Residual Plots

Various graphical methods introduced in Chap. 2 are useful for assessing the normality of ε . In using these tools, it is important to distinguish between the distribution of the outcome y and the distribution of the residuals, which are the sample analogue of ε . The point here is that the residuals may be normally distributed when y is not, and conversely. Since our assumptions concern the distribution of ε , it is important to apply the diagnostic tools to the residuals rather than to the outcome variable itself.

Figure 4.8 shows four useful graphical tools for assessing the normality of the residuals, in this case from our multipredictor regression model for LDL. In the upper panels, the histogram and boxplot both suggest a somewhat long tail on the right. The lower left panel presents a nonparametric estimate of the distribution of the residuals obtained using the `kdensity`, `normal` command in Stata. For comparison, the dashed line in that panel shows the normal distribution with the same mean and standard deviation. Comparing these two curves suggests some skewing to the right, with a long right and short left tail; but overall the shapes are quite close. Finally, as explained in Chap. 2, the upward curvature of the normal Q–Q plot on the lower right is also diagnostic of right skewness.

Interpretation of the results shown in Fig. 4.8 depends on the sample size. With 2,763 observations, there is little reason for concern about the moderate right skewness. Given such a large data set, the distribution of the parameter estimates is likely to be well approximated by the normal, despite the mild departure from normality in the residuals. However, in a small data set, with 50 or fewer observations, the long right tail might be reason for concern, in part because it could make parameter estimates less precise and tests less powerful.

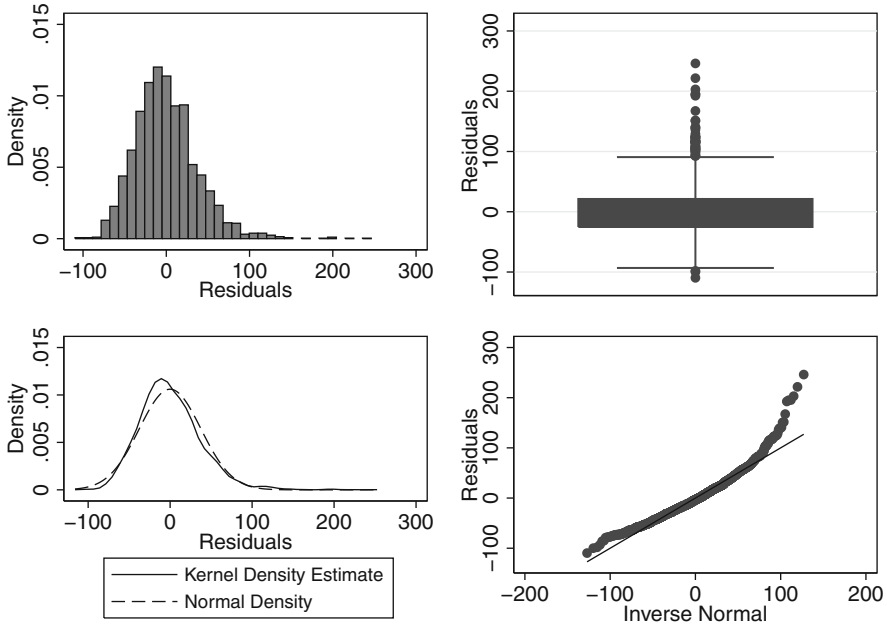


Fig. 4.8 Residuals with untransformed LDL

4.7.2.2 Testing for Departures from Normality

Various statistical tests are available for assessing the normality of the residuals, but have the drawback of being sensitive to sample size, often failing to reject the null hypothesis of normality in small samples where meeting this assumption is most important, and conversely rejecting it even for small violations in large data sets where inferences are relatively robust to departures from normality. For this reason, we do not recommend use of these tests; instead, the graphical methods just described should be used to judge the potential seriousness of the violation in the light of the sample size.

4.7.2.3 Normalizing Transformations of the Outcome

Transforming the outcome is often successful for reducing the skewness of residuals. The rationale is that the more extreme values of the outcome are usually the ones with large residuals (defined as $r_i = y_i - \hat{y}_i$); if we can “pull in” the outcome values in the tail of the distribution toward the center, then the corresponding residuals are likely to be smaller too.

One such transformation is to replace the outcome y with $\log(y)$. A constant can be added to an outcome variable with negative or zero values, so that all values are

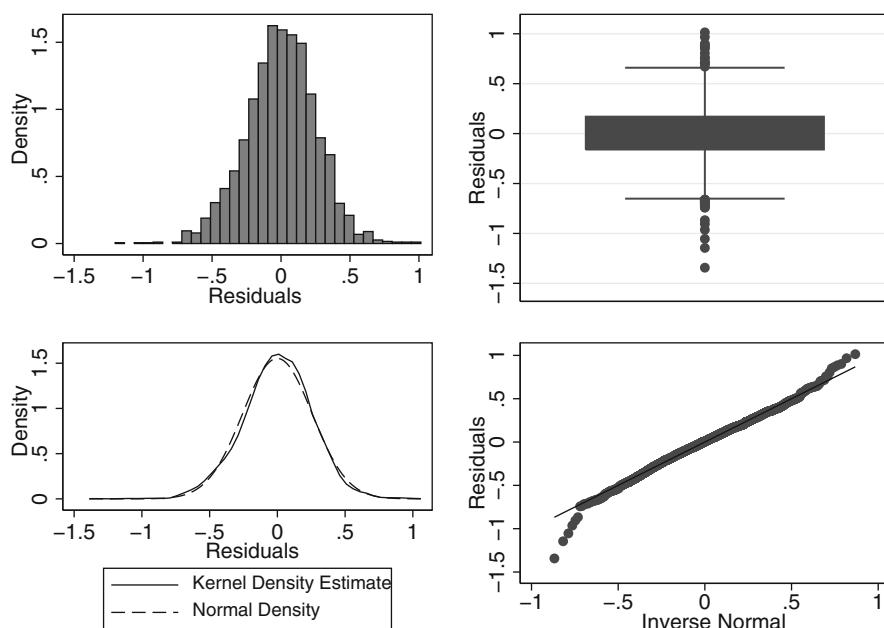


Fig. 4.9 Residuals with log-transformed LDL

positive, although this may complicate interpretation. The log transformation is now conventionally used to analyze viral load in studies of HIV and hepatitis infections, triglyceride levels in studies of cardiovascular disease, and in many other contexts. Figure 4.9 shows that after log transformation of LDL, there is no more evidence of right skewness; in fact, there is slight evidence of too long a tail on the left. It should also be noted that there is no qualitative change in inferences for BMI. In Sect. 4.7.5 below, we discuss interpretation of regression coefficients in models where the outcome is log transformed.

Power transformations are a flexible alternative to the log transformation. In this case, y is replaced by y^k . Smaller values of k “pull in” the right tail more strongly. As an example, square ($k = 1/2$) and cube ($k = 1/3$) root transformations were commonly used in analyzing CD4 lymphocyte counts in studies of HIV infection, since the distribution is very long tailed on the right. Adding a constant so that all values of the outcome are nonnegative will sometimes be necessary in this case too. The `ladder` command in Stata systematically searches for the power transformation of the outcome which is closest to normality, providing Q–Q plots for each candidate.

A more difficult problem arises if both tails of the distribution of the residuals are too long, since neither log nor fractional power transformations will fix both tails. In this case one solution is the rank transformation, in which each outcome is replaced by its rank in the ordering of all the outcomes, as in the computation

of the Spearman correlation coefficient (Sect. 3.2); this does not achieve normality but may reduce the loss of power. Another possibility is trimming the tails; for example, “Winsorizing” the outcome involves replacing outcome values more than 2 or 3 standard deviations from the average by that limiting value.

4.7.2.4 Alternatives to Transformation: Bootstrap and GLMs

Some outcome variables cannot be satisfactorily normalized by transformation, or there may be compelling reasons to analyze them on the original scale. Bootstrap CIs, as introduced in Sects. 3.6 and 4.5.4, are a useful alternative, implemented for most Stata procedures. We recommend use of percentile-based intervals, obtained using the `estat bootstrap` postestimation command, preferably based on 500 or more bootstrap samples, rather than the default of 50. These should be more reliable than the default intervals provided by the `vce(bootstrap)` option, which are based on the assumption that the coefficient estimate is normally distributed and use only the bootstrap estimate of the standard error.

Another good alternative is provided by the GLMs discussed in Chap. 8, in particular the gamma model, suitable for some badly skewed variables. Second-line options include dichotomizing the outcome, with analysis using logistic models, or categorizing the outcome into at least 3 ordered categories, then using proportional-odds or continuation-ratio models (Ananth and Kleinbaum 1997; Greenland 1994), as briefly described in Chap. 5.

4.7.3 Constant Variance

An additional assumption concerning ε is *homoscedasticity*, meaning that its variance σ_ε^2 is constant across observations. When this assumption is violated, the validity of CIs and *P*-values can be affected. In particular, between-group contrasts can be misleading if σ_ε^2 differs substantially across the subgroups being compared, and the subgroups differ in size. Furthermore, in contrast to violations of the assumption that the residuals are normally distributed, heteroscedasticity is no less a problem in large samples than in small ones. Finally, while violations do not make the coefficient estimates biased, some precision can be lost.

4.7.3.1 Residual Plots

Diagnostics for violations of the constant variance assumption also use the RVP plots used to check linearity of response to continuous predictors, as well as analogously defined residual versus fitted (RVF) plots. If the constant variance assumption is met, then the vertical spread of the residuals should be similar across the ranges of the predictors and fitted values; in contrast, heteroscedasticity is

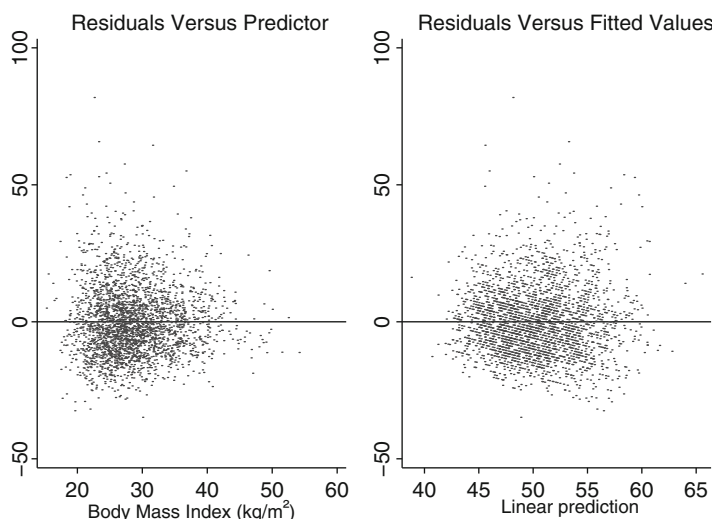


Fig. 4.10 Checking for constant residual variance

signaled by horizontal funnel shapes. Since the residuals of the LDL analysis gave no evidence of trouble, we examined the residuals from the companion model for HDL, which was shown in Sect. 4.7.1 to need a quadratic term in BMI to meet the linearity assumption.

Figure 4.10 shows scatterplots of the residuals of the regression of HDL on BMI and its square, as well as age, ethnicity, smoking, and alcohol use. The plot against BMI shows somewhat wider range on the left, although this may partly be due to the fact that there are more observations on the left, and so more likely a few large residuals purely by chance. This evidence for nonconstant variance is mirrored in the slightly wider spread on the right in the facing plot of the residuals against the fitted values.

4.7.3.2 Subsample Variances

Constancy of variance across levels of categorical predictor can be checked by comparing the sample variance of the residuals for each category. In this example, the variance was essentially identical across groups defined by ethnicity, smoking, and alcohol use. In contrast, in our analysis of the influence of exercise on glucose levels in Sect. 4.1, violation of the assumption of constant variance was one of several motivations for excluding women with diabetes. If they had been included, the variance of the residuals would have varied between this group of 734 women and the remainder of the HERS cohort by a factor of 26 (2,332 versus 90). Even after log transformation of glucose, the variance would still have differed by a factor of

10 (0.097 versus 0.0094). This pattern reflects the fact that diabetes is characterized by loss of control over glucose levels, and also variation in the use of medications that control them. These large differentials in residual variance would call into question inferences drawn from comparisons between women with and without diabetes.

4.7.3.3 Testing for Departures from Constant Variance

Statistical methods available for testing the assumption of homoscedasticity share the sensitivity to sample size described earlier for tests of normality. The resulting potential for giving false reassurance in small samples leads us to recommend against the use of these formal tests. Instead, we need to examine the severity of the violation.

4.7.3.4 When Departures May Cause Trouble

Violations of the assumption of constant variance should be addressed in cases where the variance of the residuals:

- Changes by a factor of 2 or more across the range of the fitted values of a continuous predictor, judging from the LOWESS smooth of the squared residuals.
- Differs by a factor of 2 or more between subgroups that differ in size by a factor of 2 or more.
- Differs by a factor of 3 or more between subgroups that differ in size by a factor of less than 2.

Note that smaller differences in the *standard deviation* of the residuals would give reason for transformation.

4.7.3.5 Variance-Stabilizing Outcome Transformations

In simple cases where multiple predictors do not need to be taken into account, we could use *t*-tests with the `unequal` option to compare subgroups, allowing for the unequal variances. However, multipredictor modeling is often crucial; furthermore, use of a *t*-test with unequal variances would not address smooth dependence of σ_ϵ^2 either on $E[y|\mathbf{x}]$ or on a continuous predictor. In that case, nonconstant variance can sometimes be addressed using a *variance-stabilizing* transformation of the outcome, including the log and square root transformations. As shown in Fig. 4.11, log transformation of HDL reduces, though it does not completely eliminate, the evidence for nonconstant variance we found in Fig. 4.10. However, in this case our qualitative conclusions would be unchanged by log transformation of HDL.

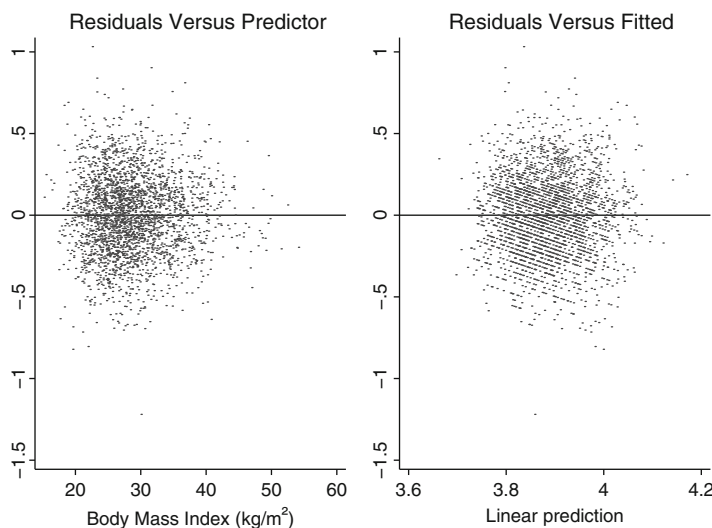


Fig. 4.11 Rechecking constant variance after log-transforming HDL

4.7.3.6 Robust Standard Errors

So-called robust or “sandwich” standard errors ([Huber 1967](#)), available with many Stata regression procedures using the option `vce(robust)`, are another convenient means of dealing with nonconstant residual variance. This method will provide more reliable inferences when the constant-variance assumption is violated, provided the model for $E[y|x]$ is approximately correct. However, some caution is warranted in using these standard errors in smaller samples. In extensive simulations, [Long and Ervin \(2000\)](#) show that robust standard errors can be too small in samples as large as 250 observations. They find that a more conservative alternative developed by [MacKinnon and White \(1985\)](#) has the best properties; this can be specified using the option `vce(hc3)` with the `regress` command. Table 4.22 shows linear models for glucose levels, successively estimated using model-based, robust, and HC3 standard errors. While the very large difference in glucose levels according to diabetes status is unambiguous, even in this small sample, the robust standard errors are considerably larger. Moreover, evidence for the adverse effect of BMI appears considerably weaker with the more conservative robust SEs.

4.7.3.7 GLMs

GLMs are another important alternative when transformation of the outcome fails to rectify substantial violations of the assumption of constant variance. For example,

Table 4.22 Models with conventional, robust, and HC3 standard errors

<code>. regress glucose diabetes BMI age drinkany</code>					
Source	SS	df	MS	Number of obs = 137	
Model	84874.7167	4	21218.6792	F(4, 132) = 37.43	
Residual	74823.7504	132	566.846594	Prob > F = 0.0000	
				R-squared = 0.5315	
				Adj R-squared = 0.5173	
Total	159698.467	136	1174.25343	Root MSE = 23.809	

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
diabetes	50.64445	4.585857	11.04	0.000	41.57318	59.71573
BMI	1.033281	.3662364	2.82	0.006	.3088297	1.757733
.....						

<code>. regress glucose diabetes BMI age drinkany, vce(robust)</code>					
Linear regression				Number of obs = 137	
				F(4, 132) = 19.32	
				Prob > F = 0.0000	
				R-squared = 0.5315	
				Root MSE = 23.809	

glucose	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
diabetes	50.64445	6.527487	7.76	0.000	37.73244	63.55647
BMI	1.033281	.4967385	2.08	0.039	.0506837	2.015879
.....						

<code>. regress glucose diabetes BMI age drinkany, vce(hc3)</code>					
Linear regression				Number of obs = 137	
				F(4, 132) = 17.96	
				Prob > F = 0.0000	
				R-squared = 0.5315	
				Root MSE = 23.809	

glucose	Coef.	Robust HC3 Std. Err.	t	P> t	[95% Conf. Interval]	
diabetes	50.64445	6.715182	7.54	0.000	37.36116	63.92775
BMI	1.033281	.5244014	1.97	0.051	-.0040363	2.070599
.....						

Poisson and negative binomial models have now mostly taken the place of linear models for count outcomes using the variance-stabilizing square root transformation. In GLMs, including the logistic model (Chap. 5), the variance of the outcome is modeled as a function of its mean (Table 8.8); in the Poisson model, for example, the variance is assumed *equal* to the mean. Furthermore, the mean-variance assumption can be relaxed using variants of these models allowing for so-called *overdispersion*, or using robust standard errors, as just described.

4.7.4 Outlying, High Leverage, and Influential Points

We have already pointed out that outlying observations with relatively large residuals can cause trouble, in part by inflating the variance of coefficient estimates, making it harder to detect statistically significant effects. In this section, we consider *high-leverage* points, which could be described as x -outliers, since they tend to have extreme values of one or more predictors, or represent an unusual combination of predictor values. The importance of high-leverage points is that they are also potentially *influential*, in the sense that one or more of the coefficient estimates would change by an unduly large amount if the influential points were omitted from the data set. This can happen when a high-leverage point also has a large residual.

Definition: *High leverage points* are x -outliers with the potential to exert undue influence on regression coefficient estimates. *Influential points* are points that have exerted undue influence on the regression coefficient estimates.

Ultimately, our concern is that changes in coefficient estimates resulting from the omission of one or a few influential points could qualitatively affect the conclusions drawn from the analysis. This could arise if associations that were clearly statistically significant become clearly nonsignificant, or vice versa, including interaction and quadratic terms, or if associations change substantially in magnitude or direction. We would have good reason to mistrust substantive conclusions that were dependent on a few observations in this way. Similarly, in regression models oriented to prediction of future outcomes (Sect. 10.1), prediction error might be substantially affected.

Outlying, high leverage, and influential points are illustrated in Fig. 4.12. In all three of these small samples ($n = 26$), a problematic data point, marked with an X, is included. The solid and dashed lines in each plot show the regression lines estimated with and without the point, as a graphical measure of influence. The sample shown on the upper left includes an outlier with a very large positive residual. However, the leverage of the outlier is minimal, because it is in the center of the distribution of x . Accordingly, the slope estimate is unaffected by omission of this data point. Note that the point is influential for the intercept estimate, but this parameter may be of less direct interest.

In the upper right panel, the point at the extreme right has high leverage, but because this data point is fairly consistent with the prediction based on the other 25 data points, its influence is limited, and the estimated slope and its statistical significance are almost unchanged by omission of the high-leverage point. Certainly our qualitative interpretation of the slope would be unaffected.

In contrast, the point at the extreme right in the lower left panel has the same leverage as the point in the upper right panel, but in this case its influence is very strong, moving the slope estimate by more than 2 standard errors. The slope remains positive and statistically significant in this instance, so our qualitative interpretation would be similar, but in some circumstances omission of such a data point could

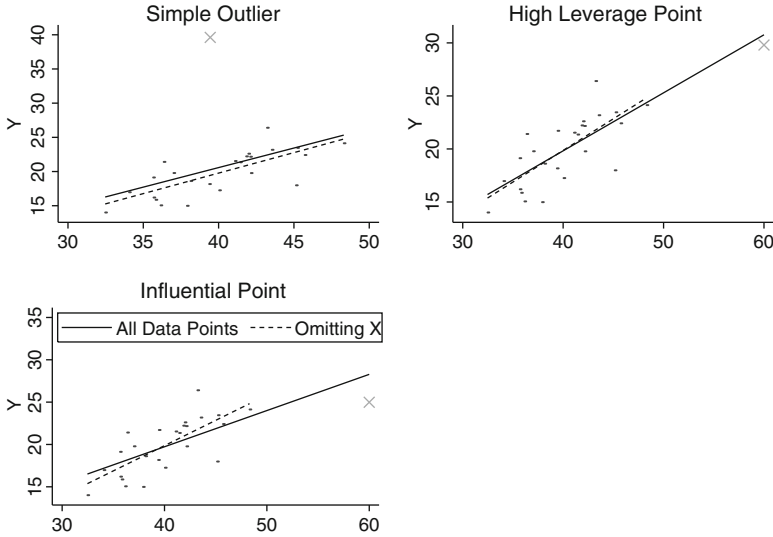


Fig. 4.12 Outlying, high-leverage, and influential points

make a nonsignificant result highly statistically significant, or vice versa. In part, this reflects the small sample size, since a high leverage point has a better chance of outweighing a relatively small number of other observations.

4.7.4.1 DFBETAs

To check for sensitivity of the conclusions of an analysis to a small number of high-leverage observations, we first need to identify potentially influential points. Of the various statistics for quantifying influence that have been defined, we recommend using DFBETA statistics, which quantify how much each of the coefficients would change if each observation were omitted from the data set. In linear regression, these statistics are exact; for logistic and Cox models, accurate approximations are available. DFBETA statistics are in standard error units—effectively on the same scale as the t -statistic, which is equal to $\hat{\beta}$ divided by its standard error. If the analysis is focused on one predictor of primary interest, then clearly the DFBETAs for that predictor are of central concern.

Boxplots are convenient for identifying a small set of extreme outliers among the DFBETA values for each predictor. DFBETAs often have a very small interquartile range, so that a substantial set of observations may lie beyond the whiskers of the plot. Thus, we need to look for a small number of extreme values that are set off from the rest. Figure 4.13 shows boxplots of the DFBETA statistics for the single predictor in the three data sets shown in Fig. 4.12. These plots clearly indicate the single influential point.

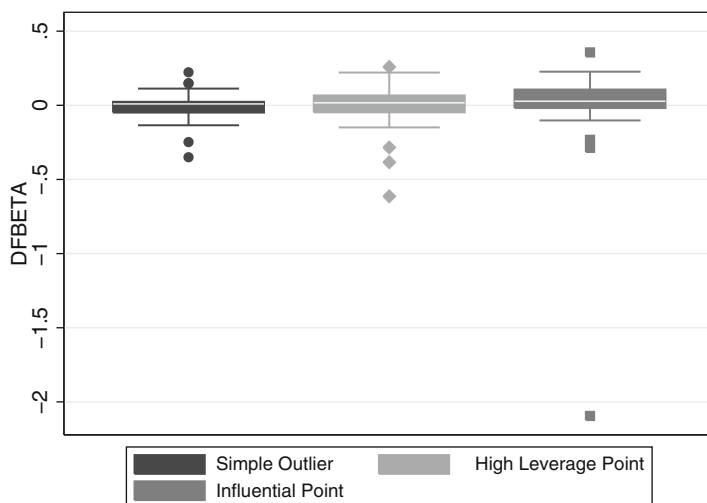


Fig. 4.13 DFBETAs for data sets shown in Fig. 4.12

If a small set of observations meeting diagnostic criteria for undue influence is identified, the accuracy of those data points should first be checked and clearly erroneous observations corrected, or if this is impossible, deleted. Then if any of the apparently influential points are retained, a final step is sensitivity analyses in which the final model is rerun omitting some or all of the retained influential points. For example, suppose we have identified ten influential points that are not due to data errors, and that these include two observations with absolute DFBETAs greater than 2, three observations with values between 1 and 2, and five more with values between 0.5 and 1. Then, a convenient ad hoc procedure would be to delete the two worst observations, then the worst five, and finally all ten potentially influential points. In each model, we would check whether the important conclusions of the analysis were affected. In prediction models, sensitivity would be assessed in terms of estimated prediction error (Sect. 10.1). In summary, we emphasize the underlying theme of sensitivity to the omission of a *small* number of points, relative to sample size; if we omit 10% or 20% of the data and the conclusions change, this would probably not indicate undue sensitivity.

Figure 4.14 above shows boxplots of DFBETAs for the multiple regression of LDL on BMI, age, ethnicity, smoking, and alcohol use. As compared to the clearly influential point shown in Fig. 4.13, the largest DFBETAs are much less extreme. Examination of the four observations with DFBETAs > 0.2 identified women with high LDL values between 346 and 393 mg/dL.

The sensitivity of model results to the omission of these four points is summarized in Table 4.23. The changes are mostly minor, in particular, for BMI, the predictor of primary interest. The P -values for ethnicity and smoking shift from nominally statistically significant to borderline significant, but these are not variables of primary interest and in any case our conclusions should not be unduly influenced by small shifts of this kind.

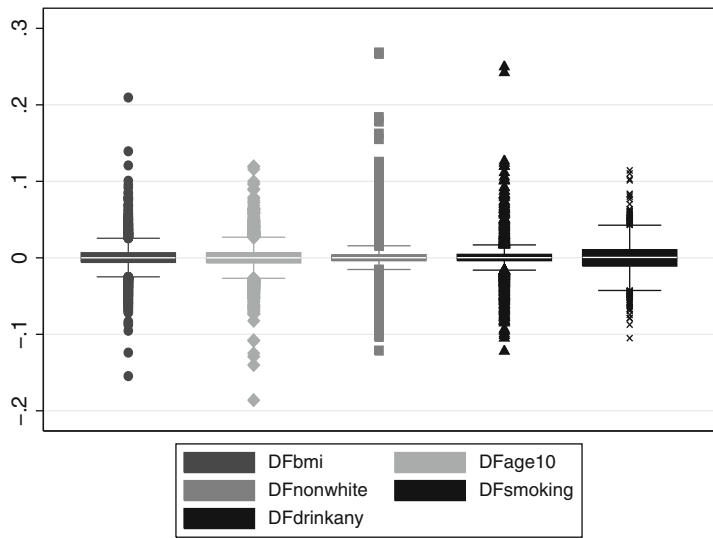


Fig. 4.14 DFBETAs for LDL model

Table 4.23 Sensitivity of LDL model to omission of four most influential points

Predictor variable	All observations			Omitting four observations		
	$\hat{\beta}$	95% CI	<i>P</i> -Value	$\hat{\beta}$	95% CI	<i>P</i> -Value
BMI	0.36	0.10, 0.62	0.007	0.34	0.08, 0.60	0.010
Age	−1.89	−4.11, 0.32	0.090	−1.86	−4.03, 0.31	0.090
Nonwhite	5.22	0.66, 9.78	0.025	4.19	−0.27, 8.66	0.066
Smoking	4.75	0.42, 9.08	0.032	3.78	−0.47, 8.03	0.081
Alcohol use	−2.72	−5.66, 0.22	0.069	−2.64	−5.51, 0.23	0.072

A weakness of these procedures is that DFBETAs capture the influence of omitting one observation at a time, but do not tell us how the omission of various *sets* of points, some of which may have small DFBETAs, will affect our conclusions. Unfortunately, user-friendly diagnostics for checking sensitivity to omission of sets of observations have not been developed, in part because the computational burden is too great.

4.7.4.2 Addressing Influential Points

If substantive conclusions are qualitatively affected by omission of influential points in the sensitivity analysis, *this should be reported*. In addition, it is often worthwhile to consider in substantive terms why these points have high leverage and are influential. For example, the western collaborative group study (WCGS) data include an influential point with an extreme but accurately recorded cholesterol level

of 645 mg/dL, which resulted from familial hypercholesterolemia, a rare condition. For research questions concerning the effects of cholesterol levels in the usual range determined by common risk factors, it would be reasonable to delete this point. But in many circumstances, deletion of influential points is hard to justify.

In that case, it may also be worth considering a more complex model that better accommodates the influential points. In Fig. 4.12, for example, a quadratic term would almost certainly reduce the influence of the observation causing trouble. Alternatively, interaction terms might accommodate influential data points characterized by an unusual combination of two predictor values. Nonetheless, changing the model in such a substantial way to accommodate one or a few data points should be undertaken with caution, with attention to the plausibility of the modified model, and the results clearly presented as data driven, sensitive to influential points, and hypothesis generating.

4.7.5 Interpretation of Results for Log Transformed Variables

In Sect. 4.7, we discussed log-transforming predictors to achieve linearity, and proposed log transformation of the outcome as a means of normalizing the residuals or stabilizing their variance. Even if substantive interpretation and P -values are often not much changed, these transformations have a substantial effect on the estimated regression coefficients and their literal interpretation.

For both predictors and outcomes, log transformation changes the focus from absolute to relative or percentage change. Recall that for a predictor and outcome on their measured scale, the regression coefficient is interpretable as the change in the average value of the outcome for every unit increase in the predictor; for both predictor and outcome, we mean change on the measured, or absolute, scale.

4.7.5.1 Log Transformation of the Predictor

First consider log transformation of the predictor. In this case, the regression coefficient multiplied by $\log(1.01)$ can be interpreted as the change in the average value of the outcome for every 1% increase in the predictor. This is valid whether we use the natural log or logarithms with other bases. In a linear model using the natural log ($\ln$) transformation of weight to predict SBP, the estimated coefficient for $\ln$ weight is 3.004517. Thus, we estimate that average SBP increases $3.004517 \times \ln(1.01) \approx 0.03$ mmHg for each 1% increase in weight. Similarly, if we multiply $\hat{\beta}$ by $\ln(1.05)$ or $\ln(1.1)$ we obtain the estimates that average SBP increases 0.15 mmHg for each 5% increase in weight and 0.29 mmHg for each 10% increase.

Within limits, we can approximate these results without using a calculator. Specifically, if the predictor is natural log-transformed, we can estimate the increase in the average value of the outcome per 1% increase in the predictor simply

by $\hat{\beta}/100$. This follows because $\ln(1.01) \approx 0.01$. But this shortcut is not valid for logarithms with other bases, and analogous calculations for larger percentage increases in the predictor get progressively less accurate and should not be attempted by this means.

4.7.5.2 Log Transformation of the Outcome

Similarly, with natural log transformation of the outcome, $100(e^{\hat{\beta}} - 1)$ is interpretable as the *percentage* increase in the average value of the outcome per unit increase in the predictor. If base-10 logs were used to transform the outcome, then $100(10^{\hat{\beta}} - 1)$ has this interpretation. The coefficient for BMI in a linear model for the natural log transformation of triglyceride (TGL) is 0.0133487, so the model predicts a $100(e^{0.0133487} - 1) = 1.34\%$ increase in TGL per unit increase in BMI.

Again, we can approximate these results without a calculator under some circumstances. When the outcome is natural log transformed, we can approximate the percentage change in the average value of the outcome per unit increase in the predictor by $100\hat{\beta}$. But this is acceptably accurate only if $\hat{\beta}$ is smaller than 0.1 in absolute value, and is again not valid using log transformations with other bases.

4.7.5.3 Log Transformation of Both Predictor and Outcome

If both predictor and outcome are transformed using natural logs, then $100(e^{\hat{\beta} \ln(1.01)} - 1)$ can be interpreted as the percentage increase in the average value of the outcome per 1% increase in the predictor. With the $\log_{10}$ transformation, $100(10^{\hat{\beta} \log_{10}(1.01)} - 1)$ has this interpretation. In this case, the back-of-the-envelope approximation for the percent increase in outcome for each 1% increase in the predictor is simply $\hat{\beta}$; this is accurate if both predictor and outcome are natural log transformed and $\hat{\beta}$ is smaller than 0.1 in absolute value.

4.7.6 When to Use Transformations

Our graphical diagnostics for linearity, normality, and constant variance do not provide clearcut decision rules analogous to $P < 0.05$, and we do not recommend formal statistical tests in this context. Furthermore, addressing these violations will in many cases involve using transformations of predictors or outcomes that may make the results harder to interpret. A natural criterion for assessing the necessity for transformation is whether important substantive results differ qualitatively before and after transformation. If not, it may be reasonable not to use the transformations. Our example using BMI and diabetes to predict HDL is probably a case in point: while log transformation of HDL corrected departures from both normality and

constant variance, the conclusions were unchanged. But if substantial differences do arise, then using transformed variables to meet model assumptions more closely helps us to avoid misleading results.

4.8 Sample Size, Power, and Detectable Effects

Section 4.2.2 presented the t -test of the null hypothesis $\beta_j = 0$, in which we compare $\hat{\beta}_j/\text{SE}(\hat{\beta}_j)$ to the t -distribution with $n - (p + 1)$ degrees of freedom. This test leads directly to methods for estimating sample size and power for analyses using the linear model. Suppose we would like to calculate the sample size that would provide power of γ to reject $\beta_j = 0$ in a two-sided test with type-I error rate α , under the alternative hypothesis $\beta_j = \beta_j^a$, assuming for now that $\beta_j^a > 0$. We begin with an expression for power, relying on the large-sample equivalence of the t and standard normal Z -distributions:

$$\begin{aligned}
 \gamma &= P\left[|\hat{\beta}_j|/\text{SE}(\hat{\beta}_j) > z_{1-\alpha/2}\right] \\
 &\approx P\left[\hat{\beta}_j/\text{SE}(\hat{\beta}_j) > z_{1-\alpha/2}\right] \\
 &= P\left[(\hat{\beta}_j - \beta_j^a)/\text{SE}(\hat{\beta}_j) > z_{1-\alpha/2} - \beta_j^a/\text{SE}(\hat{\beta}_j)\right] \\
 &= 1 - \Phi\left[z_{1-\alpha/2} - \beta_j^a/\text{SE}(\hat{\beta}_j)\right] \\
 &= \Phi\left[\beta_j^a/\text{SE}(\hat{\beta}_j) - z_{1-\alpha/2}\right].
 \end{aligned} \tag{4.14}$$

In (4.14), $|\cdot|$ denotes absolute value; $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution (1.96 for a two-sided test with type-I error rate of 5%); and $\Phi(\cdot)$ is the cumulative distribution function for a standard normal variate Z , so that $\Phi(z_{1-\alpha/2}) = P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$. The first approximation in (4.14) holds because if β_j is positive, $P(\hat{\beta}_j/\text{SE}(\hat{\beta}_j) < z_{\alpha/2}) \approx 0$. The second step is simple algebra. The third follows because $(\hat{\beta}_j - \beta_j^a)/\text{SE}(\hat{\beta}_j)$ has an approximate Z -distribution in large samples, and the fourth because of the symmetry of the Z -distribution about zero. Using (4.4) (with n in place of $n - 1$) to evaluate $\text{SE}(\hat{\beta}_j)$, then applying the inverse transformation Φ^{-1} to both sides of (4.14), and solving for n gives

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \sigma_{y|x}^2}{(\beta_j^a \sigma_{x_j})^2 (1 - \rho_j^2)}. \tag{4.15}$$

In (4.15), z_γ is the quantile of the standard normal distribution for power (0.84 for 80% power, 1.28 for 90%), $\sigma_{y|x}^2$ is the residual variance of the outcome, σ_{x_j} is the standard deviation of X_j , and ρ_j is its multiple correlation with the other covariates. The variance inflation factor $1/(1 - \rho_j^2)$ in (4.15) accounts for the potential loss of precision due to the inclusion of other predictors in the model (Hsieh et al. 1998).

In some problems, including secondary analyses of existing data, n is fixed. In that case, (4.15) can be solved to calculate power, if we specify β_j^a :

$$\gamma = 1 - \Phi \left[z_{1-\alpha/2} - |\beta_j^a| \sigma_{x_j} \sqrt{n(1 - \rho_j^2)} / \sigma_{y|x} \right]. \quad (4.16)$$

Similarly, we can calculate the *minimum detectable effect*—that is, the smallest value of β_j^a for which a sample of size n would provide power of γ to reject the null hypothesis $\beta_j = 0$ in a two-sided test with type-I error of α . The minimum detectable effect is

$$\pm \beta_j^a = \frac{(z_{1-\alpha/2} + z_\gamma) \sigma_{y|x}}{\sigma_{x_j} \sqrt{n(1 - \rho_j^2)}}. \quad (4.17)$$

Some additional points:

- When X_j is binary with prevalence f_j , $\sigma_{x_j} = \sqrt{f_j(1 - f_j)}$ in (4.15)–(4.17).
- When X_j is continuous with standard deviation σ_{x_j} , it is important to recognize that sample size, power, and minimum detectable effects do not depend in any real way on the units in which X_j is measured. This is most clearly seen in (4.17). Suppose X_j is usually measured in grams. Changing the unit to milligrams increases σ_{x_j} by a factor of 1,000, and shrinks β_j^a by the same factor. But of course the effect on the outcome of a 1-milligram increase in the predictor is 1,000 times smaller than the effect of a 1-gram increase. One way to avoid confusion is to consider the minimum detectable effect size for a one standard deviation change in X_j , which is often a reasonable-sized change to consider. That effect size is obtained by setting $\sigma_{x_j} = 1$ in (4.17).
- If $\beta_j^a < 0$ under the alternative, we have to use $|\beta_j^a|$ in (4.16) to get the correct result. It follows that the negative of the value given by (4.17) is also a valid solution for the minimum detectable effect.
- Because they are based on the standard normal distribution, (4.15)–(4.17) are only approximate. Exact solutions involve the noncentral t -distribution and iterative calculations. Numerous packages supply these estimates for small as well as large sample sizes; the `sampsi` and `sampsi_reg` commands in Stata work for binary and continuous predictors respectively. An approximate correction is to add 2 to the estimate of n provided by (4.15) for tests with α of 5%, and add 4 with α of 1% (Snedecor and Cochran 1989, page 104). The correction can be important when $n < 50$ and especially when $n < 25$.
- Sample size (4.15) and minimum detectable effect (4.17) calculations simplify considerably when we specify $\alpha = 0.05$ and $\gamma = 0.8$, β_j^a is the effect of a one standard deviation increase in continuous x_j , and we do not need to penalize for covariate adjustment. In that standard case,

$$n = 7.849 \times \sigma_{y|x}^2 / (\beta_j^a)^2. \quad (4.18)$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = 2.802 \times \sigma_{y|x} / \sqrt{n}. \quad (4.19)$$

For 90% power, substitute 10.51 for 7.849 and 3.242 for 2.802.

- Similarly, for a 2-arm clinical trial with equal allocation to arms, so that β_j^a is the between-group difference in means and $s_{x_j}^2 = 0.25$, we can calculate

$$n = 4 \times 7.849 \times \sigma_{y|x}^2 / (\beta_j^a)^2. \quad (4.20)$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = 2 \times 2.802 \times \sigma_{y|x} / \sqrt{n}. \quad (4.21)$$

- Power calculations using (4.16) simplify analogously, but still require a statistical calculator or computer package to evaluate the normal cumulative distribution function $\Phi(\cdot)$.
- The Stata commands `sampsi` and `sampsi_reg` can also be used to compute power, but not minimum detectable effects.
- In using sample size calculators that do not allow for covariate adjustment, including the `sampsi` and `sampsi_reg` commands, the unadjusted sample size estimate should be inflated by $1/(1-\rho_j^2)$; similarly, the minimum detectable effect estimate should be inflated by $\sqrt{1/(1-\rho_j^2)}$. To calculate power, use $n(1-\rho_j^2)$ in place of n as an input.
- For the linear model, the proposed adjustment may be conservative, since adjustment for covariates will also reduce the residual variance $\sigma_{y|x}^2$, to some extent offsetting the loss of precision due to the correlation ρ_j between X_j and the other covariates. This is particularly relevant in calculations for stratified randomized trials with continuous outcomes, since the stratification factor may account for a large proportion of the variance of the outcome, but is in expectation uncorrelated with treatment assignment.

To illustrate these calculations, suppose we are planning a randomized trial with equal allocation to active treatment and control ($f = 0.5$) to assess the effect of a new lipid-lowering agent on LDL levels. From pilot data, the residual standard deviation $\sigma_{y|x}$ for LDL is expected to be ≈ 38 mg/dL, and we hypothesize that the agent will lower average LDL levels about 40 mg/dL. Because this is a clinical trial, it is unlikely that we will need to adjust for covariates, so we can assume $\rho_j = 0$. The sample size must provide 80% power in a two-sided test with α of 5%.

We first calculate the sample size using the `sampsi` command in Stata, then using its capacity as a calculator to evaluate (4.15). Table 4.24 shows the results. In using `sampsi`, any values of the means for populations 1 and 2 that differ by 40 mg/dL would give the same answer, so for convenience we used 0 and 40. With the Snedecor and Cochran correction, using Stata to evaluate (4.15) gives about the same result as `sampsi`.

Table 4.24 Sample size calculations for a small clinical trial

```
. sampsi 0 40, sdi(38) alpha(0.05) power(0.8)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
               and m2 is the mean in population 2

Assumptions:

      alpha =    0.0500   (two-sided)
      power =    0.8000
      m1 =           0
      m2 =          40
      sd1 =          38
      sd2 =          38
      n2/n1 =       1.00

Estimated required sample sizes:

      n1 =          15
      n2 =          15

. * solution using Snedecor and Cochran correction
. display (invnormal(.975)+invnormal(.8))^2*38^2/(40^2*0.5*(1-0.5))+2
30.334456
```

When the predictor of interest is continuous, we can use the downloadable `sampsi_reg` command in Stata. Suppose, for example, that we would like to estimate the power of a study with 485 participants to detect an effect of higher BMI on SBP, controlling for age, race/ethnicity, smoking, alcohol use, and physical activity levels. From pilot data, we estimate that $\sigma_{y|x} \approx 18.5$ mmHg, $\sigma_x \approx 5.5$ kg/m², and $\rho_j \approx 0.33$. We hypothesize that average SBP increases 0.5 mmHg for every kg/m² increase in BMI—that is, $\beta_j^a = 0.5$. What is the power of the study to detect this effect of BMI on SBP in a two-sided test with α of 5%?

Table 4.25 shows results of the computation using `sampsi_reg` in Stata, as well as a direct implementation of (4.16). Since `sampsi_reg` does not allow for the adjustment based on the variance inflation factor, we first deflate the available sample size by $1 - \rho_j^2$. The two estimates of power are in close agreement.

4.8.1 Calculations Using Standard Errors Based on Published Data

Equations (4.15)–(4.17) depend on $\sigma_{y|x}$, σ_{x_j} , and ρ_j , for which it may be hard to obtain estimates. However, the derivation using (4.4) suggests a solution. Suppose an estimate $\tilde{SE}(\hat{\beta}_j)$ for the standard error of $\hat{\beta}_j$ is available, based on a multiple linear regression model with appropriate covariates and estimated using $\tilde{n}$ observations. For example, we could compute $\tilde{SE}(\hat{\beta}_j)$ from a published article as

Table 4.25 Power calculation for independent effect of BMI on SBP

```

. display 485*(1-.33^2)
432.1835
. sampsi_reg, alt(0.5) nl(432.1835) s(power) sx(5.5) sd1(18.5)

Estimate power for linear regression
Test Ho: Alt. Slope = Null Slope, usually Null Slope is 0

Assumptions:
      Alpha =      0.0500   (two-sided)
      N =      432.1835
      Null Slope =      0.0000
      Alt Slope =      0.5000
      Residual sd =      18.5000
      SD of X's =      5.5000

Estimated power:
      Power = .86934271

. display 1-normal(invnormal(0.975)-0.5*5.5*sqrt(485*(1-.33^2)))/18.5)
.8708243

```

the width of the 95% CI for $\hat{\beta}_j$, divided by $2z_{.975} \approx 3.92$. Care must be taken to ensure that the hypothesized value of β_j^a corresponds to the same measurement scale for X_j as in the source article. Then, (4.15) can be simplified as

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \tilde{n} [\tilde{SE}(\hat{\beta}_j)]^2}{(\beta_j^a)^2}. \quad (4.22)$$

Similarly, power in a new sample of size n is given by

$$\gamma = 1 - \Phi \left[z_{1-\alpha/2} - |\beta_j^a| / [\sqrt{\tilde{n}/n} \tilde{SE}(\hat{\beta}_j)] \right]. \quad (4.23)$$

Finally, the minimum detectable effect in a new sample of size n can be obtained as

$$\pm \beta_j^a = (z_{1-\alpha/2} + z_\gamma) \sqrt{\tilde{n}/n} \tilde{SE}(\hat{\beta}_j). \quad (4.24)$$

As an example, we could use the multiple linear model in Table 4.2 to obtain sample size, power, and minimum detectable effect estimates for a new study of the effect of BMI on glucose levels in nondiabetic women. Based on the HERS data with $\tilde{n} = 2028$, $\tilde{SE}(\hat{\beta}_j) = (0.5707328 - 0.4077512)/3.92 \approx 0.0415528$. Suppose we hypothesize that glucose levels increase 0.5 mg/dL for each kg/m² increase in BMI, so $\beta_j^a = 0.5$.

In Table 4.26, we first use (4.22) to estimate that a new sample of 147 participants would provide 90% power in a 2-sided test with α of 5% to detect the hypothesized increase in glucose of 0.5 mg/dL for each kg/m² increase in BMI. Then, using (4.23), we find that a sample of 200 participants would provide almost 97% power to detect

Table 4.26 Calculations based on regression output

```

. * sample size for a new study providing 90% power
. display (invnormal(.975)+invnormal(.9))^2*2028*0.0415528^2/0.5^2
147.17185

. * power in a new study with 200 participants
. display 1-normal(invnormal(0.975)-0.5/(sqrt(2028/200)*0.0415528))
.96552967

. * minimum effect detectable with 80% power in a new study with 100 participants
. display (invnormal(.975)+invnormal(.8))*sqrt(2028/100)*0.0415528
.5242496

```

the hypothesized effect. Finally, using (4.24) suggests that a smaller sample of 100 participants would provide 80% power to detect a minimum effect of 0.52 mg/dL for each kg/m² increase in BMI.

4.9 Summary

The multipredictor linear model is a straightforward extension of the simple linear model for continuous outcomes. Inclusion of multiple predictors in the model makes it possible to adjust for confounding variables, examine mediation, check for and model interactions, and increase efficiency, especially in experiments, by accounting for design factors. To avoid misleading conclusions, it is important to check assumptions, including normality of the residuals, especially in small samples; transformations of the outcome, bootstrapping, and GLMs can be used to address violations. Nonconstant variance of the residuals is a potentially serious concern even in large samples, but can be resolved using robust standard errors. As with the models discussed in later chapters, nonlinear effects of continuous predictors can be accommodated using predictor transformations, including restricted cubic splines, and interactions modeled using product terms. Finally, it is important to recognize outcomes for which linear regression is not appropriate; these include binary, time-to-event, count, and repeated measures or clustered outcomes, and are addressed in subsequent chapters.

4.10 Further Notes and References

For more detailed information on the linear regression model, first-rate books include [Weisberg \(1985\)](#) and [Draper and Smith \(1981\)](#). A standard book on regression diagnostics is [Belsey et al. \(1980\)](#), while [Cleveland \(1985\)](#) covers graphical methods for model checking in detail. See [Breiman \(2001\)](#) for a skeptical view of the sensitivity of the methods presented here for detecting lack of fit.

4.10.1 Generalized Additive Models

Methods have also been developed for fitting linear as well as logistic (Chap. 5) and other GLMs (Chap. 8) in which the adjusted response to each predictor can be flexibly modeled as a smooth (piecewise cubic rather than piecewise linear) spline, or alternatively using a LOWESS curve. In both cases, the degree of smoothness is under the control of the analyst. Known as *generalized additive models* (Hastie and Tibshirani 1986, 1999), implementations in the R statistical package make it easy to model and test the statistical significance of departures from linearity. Implementations in R of smooth spline transformations of predictors are also available for the Cox model, discussed in Chap. 6.

4.11 Problems

Problem 4.1. Using the WCGS data for middle-aged men at risk for heart disease, fit a multipredictor model for total cholesterol (`chol`) that includes the binary predictor `arcus`, which is coded 1 for the group with *arcus senilis*, a milky ring in the iris associated with high cholesterol levels, and 0 for the reference group. Save the fitted values. Now refit the model with the code for the reference group changed to 2. Compare the coefficients, standard errors, P -values, and fitted values from the two models. The WCGS data are available at <http://www.biostat.ucsf.edu/vgsm>.

Problem 4.2. Using (4.2), show that β_j gives the difference in $E[y|\mathbf{x}]$ for a one-unit increase in x_j , no matter what the values of x_j or the other predictors. *Hint:* Write the value of (4.2) for $x_j = x$ and then for $x_j = x + 1$, for arbitrary (unspecified) values of the other predictors, all of which are held fixed, and subtract the first value from the second.

Problem 4.3. Using the WCGS data referenced in Problem 4.1, extract the fitted values from the multipredictor linear regression model for cholesterol and show that the square of the sample correlation between the fitted values and the outcome variable is equal to R^2 . In Stata, the following code saves the predicted values from the regression model in Table 4.2 to a new variable `yhat`:

```
. regress glucose exercise BMI smoking drinkany
. predict yhat
```

Then use the `pwcorr` and `display` commands to get the correlation between `yhat` and the predictor and square it.

Problem 4.4. Use the `test` command in Stata or an equivalent command in another statistical package to show that $F = t^2$ for a pairwise contrast between any other level of a categorical predictor and the reference group used in the model.

Problem 4.5. In the model including an interaction between BMI and statin use, define a second new BMI variable so that estimates for BMI specific to women who do and do not use statins can be obtained directly from the regression coefficients, rather than having to compute sums of the coefficients for one of these groups. Define the values of the new BMI variable in the two groups, and then write down the regression equations analogous to (4.11)–(4.13). Explain why the statin use variable needs to be included in this model.

Problem 4.6. If we “center” age—that is, replace it with a new variable defined as the deviation in age from the sample mean, what would be the interpretation of the intercept in the model for SBP (3.2)? If BMI had *not* been centered, how would the interpretation of the statin use variable change in the model in Sect. 4.6.2 allowing for interaction in predicting LDL?

Problem 4.7. Consider the associations between exercise and glucose levels among women without diabetes. What are the interpretations of the coefficient for exercise:

- In a simple linear model for glucose levels.
- In a multipredictor linear regression model for glucose adjusting for all known confounders of the exercise association.

Suppose factor X had been identified as a mediator of the exercise/glucose association. What would be the interpretation of the exercise coefficient in a multipredictor regression model that also adjusted for factor X, supposing that the exercise coefficient remained statistically significantly different from zero?

Problem 4.8. Suppose that in a clinical trial of the effects of a new treatment on glucose levels, the randomization is stratified on diabetes, an important predictor of this outcome. By virtue of randomization, the treatment is uncorrelated with diabetes. Using (4.4), explain why including diabetes in the analysis should provide a more efficient estimate of the treatment effect. Would it be a good idea to check for interaction between treatment and diabetes in this analysis? Why?

Problem 4.9. Using Stata (or another statistical package) and the WCGS data set referenced above in Problem 4.1 (or your own data set), verify that you get equivalent results from:

- A *t*-test and a simple linear model with one binary predictor.
- One-way ANOVA and a linear model with one multilevel categorical predictor.

Problem 4.10. What is the difference between showing that an interaction is statistically significant and showing that an association is statistically significant in one group but not in the other? Describe a pattern where the second condition holds but there would clearly be no interaction. Is that pattern of substantive interest?

Problem 4.11. Consider a predictor of interest for an important outcome in your field of expertise. Are there other predictors that might be hypothesized a priori to interact with the predictor of interest? Why?

Problem 4.12. Suppose you have used a restricted cubic spline to model a non-linear response to your predictor of primary interest, similar to one of the models for HDL in Fig. 4.7. Figure out how to use the spline basis variables, which in Stata would be made by the `mk spline` command, and corresponding regression coefficients to plot the shape of the response estimated by the regression model.

Problem 4.13. Consider a right-skewed outcome variable that could be adequately normalized using an unfamiliar fractional power transformation (say, the cube root). A simpler alternative is just to dichotomize the variable. Why would you expect this to be a costly choice in terms of efficiency? Now consider birth weights. Why might analysis of an indicator of low birth weight be worth the loss of efficiency in this case?

Problem 4.14. Suppose you fit a model with an influential point. With the point, the association of interest is just statistically significant, and without it, it is clearly not. What would you do?

4.12 Learning Objectives

- (1) Describe situations in which multipredictor analysis is needed. Given an analysis situation, decide if linear regression is appropriate.
- (2) Translate research questions appropriate for a regression model into specific questions about the coefficients of the model.
- (3) Use linear regression models to test hypotheses about relationships between variables, including confounding, mediation, and interaction.
- (4) Describe the linear regression model, its key assumptions, and their implications.
- (5) Explain why the estimates are called least squares estimates.
- (6) Define regression line, fitted value, residual, and influence.
- (7) State the relationships between:
 - Correlation and regression coefficients
 - The two-sample t -test and a regression model with one binary predictor
 - ANOVA and a regression model with categorical predictors
- (8) Know how a statistical package is used to estimate the parameters in a regression model and make diagnostic plots to assess how well model assumptions are met.
- (9) Interpret regression model output including regression coefficient estimates, hypothesis tests, CIs, and statistics which quantify the fit of the model.
- (10) Interpret regression coefficients when the predictor, outcome, or both are log transformed.

Chapter 5

Logistic Regression

Patients testing positive for a sexually transmitted disease at a clinic are compared to patients with negative tests to investigate the effectiveness of a new barrier contraceptive. One-month mortality following coronary artery bypass graft surgery is compared in groups of patients receiving different dosages of beta blockers. Many clinical and epidemiological studies generate outcomes which take on one of two possible values, reflecting presence/absence of a condition or characteristic at a particular time, or indicating whether a response occurred within a defined period of observation. In addition to evaluating a predictor of primary interest, it is important to investigate the importance of additional variables that may influence the observed association and therefore alter our inferences about the nature of the relationship. In evaluating the effect of contraceptive use in the first example, it would be clearly important to control for age in addition to behaviors potentially linked to infection risk. In the second example, a number of demographic and clinical variables may be related to both the mortality outcome and treatment regime. Both of these examples are characterized by binary outcomes and multiple predictors, some of which are continuous.

Methods for investigating associations involving binary outcomes using contingency table methods were briefly covered in Sect. 3.4. Although these techniques are useful for exploratory investigations, and in situations where the number of predictor variables of interest is limited, they can be cumbersome when multiple predictors are being considered. Further, they are not well suited to situations where predictor variables may take on a large number of possible values (e.g., continuous measurements). Similar to the way linear regression techniques expanded our arsenal of tools to investigate continuous outcomes, the logistic regression model generalizes contingency table methods for binary outcomes. In this chapter, we cover the use of the logistic model to analyze data arising in clinical and epidemiological studies. Because the basic structure of the logistic model mirrors that of the linear regression model, many of the techniques for model construction, interpretation, and assessment will be familiar from Chap. 4.

5.1 Single Predictor Models

Recall the example in Sect. 3.4 investigating the association between CHD and age for the WCGS. Table 5.1 summarizes the observed proportions (P) of CHD diagnoses for five categories of age, along with the estimated risk difference (RD), relative risk (RR), and odds ratio (OR). The last three measures are computed according to procedures described in Sect. 3.4, using the youngest age group as the baseline category. The estimates show a tendency for increased risk of CHD with increasing age. Although this information provides a useful summary of the relationship between CHD risk and age, the choice of five-year categories for age is arbitrary. A regression representation of the relationship would provide an attractive alternative and obviate the need to choose categories of age.

Recall that in standard linear regression, we modeled the average of a continuous outcome variable y as a function of a single continuous predictor x using a linear relationship of the form

$$E[y|x] = \beta_0 + \beta_1 x.$$

We might be tempted to use the same model for a binary outcome variable. First, note that if we follow convention and code the values of a binary outcome as one for those experiencing the outcome and zero for everyone else, the observed proportion of outcomes among individuals characterized by a particular value of x is simply the mean (or “expected value”) of the binary outcome in this group. In the notation introduced in Sect. 3.4, we symbolize this quantity by $P(x)$. The linear model for our binary outcome might then be expressed as

$$P(x) = E[y|x] = \beta_0 + \beta_1 x. \quad (5.1)$$

This has exactly the same form as the linear regression model; the expected value of the outcome is modeled as a linear function of the predictor. Further, changes in the outcome associated with specified changes in the predictor x have a risk difference interpretation: For example, if x is a binary predictor taking on the values 0 or 1, the effect of increasing x one unit is to add an increment β_1 to the outcome. From (5.1),

$$P(1) - P(0) = \beta_1.$$

Referring back to Definition (3.14) in Sect. 3.4, we see that this is the risk difference associated with a unit increase in x . Models with this property are often referred to as *additive risk models* (Clayton and Hills 1993).

Table 5.1 CHD for five age categories in the WCGS sample

Age group	P	$1 - P$	RD	RR	OR
35–40	0.057	0.943	0.000	1.000	1.000
41–45	0.050	0.950	−0.007	0.883	0.877
46–50	0.093	0.907	0.036	1.635	1.700
51–55	0.123	0.877	0.066	2.156	2.319
56–60	0.149	0.851	0.092	2.606	2.886

There are several limitations with the linear model (5.1) as a basis for regression analysis of binary outcomes. First, the statistical machinery which allowed us to use this linear model to make inferences about the strength of relationship in Chap. 4 required that the outcome variable follow an approximate normal distribution. For a binary outcome, this assumption is clearly incorrect. Second, the outcome in the above model represents a probability or risk. Thus, any estimates of the regression coefficients must constrain the estimated probability to lie between zero and one for the model to make sense. The first of these problems is statistical, and addressing it would require generalizing the linear model to accommodate a distribution appropriate for binary outcomes. The second problem is numerical. To ensure sensible estimates, our estimation procedure would have to satisfy the constraints mentioned.

Another issue is that in many settings, it seems implausible that outcome risk would change in a strictly linear fashion for the entire range of possible values of a continuous predictor x . Consider a study examining the likelihood of a toxicity response to varying levels of a treatment. We would not expect the relationship between likelihood of toxicity and dose to be strictly linear throughout the range of possible doses. In particular, the likelihood of toxicity should be zero in the absence of treatment and increase to a maximum level, possibly corresponding to the proportion of the sample susceptible to the toxic effect, with increasing dose.

Figure 5.1 presents four hypothetical models linking the probability $P(x)$ of a binary outcome to a continuous predictor x . In addition to the linear model (a), there is the exponential model (b) that constrains risk to increase exponentially with x , the “step function” model (c) that allows irregular (but piecewise-constant) change in risk with increasing values of x , and the smooth S-shaped curve in (d) known as the *logistic* model. The exponential model is also known as *log linear* because it specifies that the logarithm of the outcome risk is linear in x . It presents a problem similar to that noted for the linear model above: Namely, that risk is not obviously constrained to be less than one for large values of $\beta_0 + \beta_1 x$. The outcome probabilities for model (c) simply represent the estimated proportion of positive outcomes in each group specified by the categories of x , and has the desirable properties that risks are clearly constrained to fall in the interval $[0, 1]$, and that the nature of the increase in the interval can be flexibly represented by different “step” heights. However, it lacks smoothness, a property that is biologically plausible in many instances. In addition, the choice of break points delineating the changes in risk is subjective. By contrast, the logistic model allows for a smooth change in risk throughout the range of x , and has the property that risk increases slowly up to a “threshold” range of x , followed by a more rapid increase and a subsequent leveling off of risk. This shape is consistent with many dose-response relationships (illustrated by the toxicity example from the previous paragraph). As we will see later in this chapter, all of these models represent valid alternatives for assessing how risk of a binary outcome changes with the value of a continuous predictor. However, most of our focus will be on the logistic model.

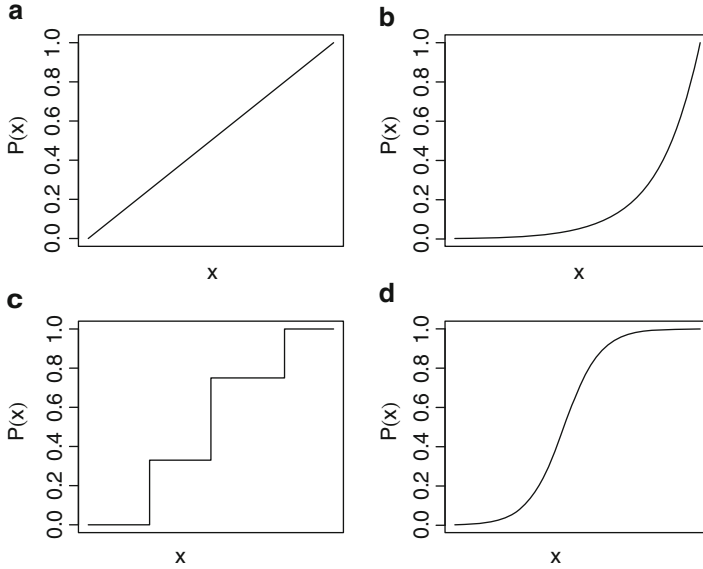


Fig. 5.1 Risk models for a binary outcome and continuous predictor (a) Linear (b) Exponential (c) Step function (d) Logistic

In addition to a certain degree of biological plausibility, the logistic model does not pose the numerical difficulties associated with the linear and log-linear models, and has a number of other appealing properties that will be described in more detail below. For these reasons, it is by far the most widely used model for binary outcomes in clinical and epidemiological applications, and forms the basis of logistic regression modeling. However, adoption of the logistic model still implies strong assumptions about the relationship between outcome risk and the predictor. In fact, expressed on a transformed scale, the model prescribes a linear relationship between the logarithm of the odds of the outcome and the predictor.

The logistic model plotted in Fig. 5.1d is defined by the equation

$$P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \quad (5.2)$$

In terms of the odds of the outcome associated with the predictor x , the model can also be expressed as

$$\frac{P(x)}{1 - P(x)} = \exp(\beta_0 + \beta_1 x). \quad (5.3)$$

Consider again the simple case where x takes on the values 0 or 1. From the last equation, the ratio of the odds for these two values of x are

$$\frac{P(1)/[1 - P(1)]}{P(0)/[1 - P(0)]} = \exp(\beta_1). \quad (5.4)$$

Expressed in this form, we see that the logistic model specifies that the ratio of the odds associated with these two values of x is given by the factor $\exp(\beta_1)$. Equivalently, the odds for $x = 1$ are obtained by multiplying the odds for $x = 0$ by this factor. Because of this property, the logistic model is an example of a *multiplicative risk model* (Clayton and Hills 1993). (Note that the log-linear model is also multiplicative in this sense, but is based on the outcome risks rather than the odds.)

Although not easily interpretable in the form given in (5.2) and (5.3), expressed as the logarithm of the outcome odds (as given in (5.3)), the model becomes linear in the predictor

$$\log \left[\frac{P(x)}{1 - P(x)} \right] = \beta_0 + \beta_1 x. \quad (5.5)$$

This model states that the log odds of the outcome is linearly related to x , with intercept coefficient β_0 and slope coefficient β_1 (i.e., the logistic model is an additive model when expressed on the log odds scale). The logarithm of the outcome odds is also frequently referred to as the *logit* transformation of the outcome probability.

In the language introduced in Chaps. 3 and 4, (5.2), (5.3), and (5.5) define the systematic part of the logistic regression model, linking the average $P(x)$ of the outcome variable y to the predictor x . The random part of the model specifies the distribution of the outcome variable y_i , conditional on the observed value x_i of the predictor (where the subscript i denotes the value for a particular subject). For binary outcomes, this distribution is called the *binomial* distribution and is completely specified by the mean of y_i conditional on the value x_i . To summarize, the logistic model makes the following assumptions about the outcome y_i :

- (1) y_i follows a Binomial distribution.
- (2) The mean $E[y|x] = P(x)$ is given by the logistic function (5.2).
- (3) Values of the outcome are statistically independent.

These assumptions closely parallel those associated with the linear regression (in Sect. 3.3), the primary difference being the use of the binomial distribution for the outcome y . Note that the assumption of constant variance of y across different values of x is not required for the logistic model. Another difference is that the random aspect of the logistic model is not included as an additive term in the regression equation. However, it is still an integral part of estimation and inference regarding model coefficients. (This is discussed further in Sect. 5.6.)

As we will see in the rest of this chapter, both of the alternative expressions (5.2) and (5.5) for the logistic model are useful: the linear logistic form (5.5) is the basis

Table 5.2 Logistic model for the relationship between CHD and age

```
. logistic chd69 age, coef
```

Logit estimates

Log likelihood = -869.17806

Number of obs = 3154

LR chi2(1) = 42.89

Prob > chi2 = 0.0000

Pseudo R2 = 0.0241

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0744226	.0113024	6.58	0.000	.0522703	.0965748
_cons	-5.939516	.549322	-10.81	0.000	-7.016167	-4.862865

for regression modeling, while the (nonlinear) logistic form (5.2) is useful when we want to express the outcome on its original scale (e.g., to estimate outcome risk associated with a particular value of x).

One of the most significant benefits of the linear logistic formulation (5.5) is that the regression coefficients are interpreted as log odds ratios. These can be expressed as odds ratios via simple exponentiation (as demonstrated above in (5.4)), providing a direct generalization of odds ratio methods for frequency tables to the regression setting. This property follows directly from the definition of the model, and is demonstrated in the next section. Finally, we note that there are a number of alternative regression models for binary outcomes that share similar properties to the logistic model. Although none of these comes close to the logistic model in terms of popularity, they offer useful alternatives in some situations. Some of these will be discussed in Sect. 5.5.

5.1.1 Interpretation of Regression Coefficients

Table 5.2 shows the fit of the logistic model (5.5) for the relationship between CHD risk and age in the WCGS study. The coefficient labeled `_cons` in the table is the intercept (β_0), and the coefficient labeled `age` is the slope (β_1) of the fitted logistic model. Since the outcome for the model is the log odds of CHD risk, and the relationship with `age` is linear, the slope coefficient β_1 gives the change in the log odds of `chd69` associated with a one-year increase in `age`. We can verify this by using the formula for the model (5.5) and the estimated coefficients to calculate the difference in risk between a 56- and a 55-year-old individual:

$$\log \left[\frac{P(56)}{1 - P(56)} \right] - \log \left[\frac{P(55)}{1 - P(55)} \right]$$
$$= (-5.940 + 0.074 \times 56) - (-5.940 + 0.074 \times 55) = 0.074.$$

This is just the coefficient β_1 as expected; performing the same calculation on an arbitrary one-year age increase would produce the same result (as shown at the end of this section). The corresponding odds ratio for any one-year increase in age can then be computed by simple exponentiation:

$$\exp(0.074) = 1.077.$$

This odds ratio indicates a small (approximately 8%) but statistically significant increase in the odds of CHD for each one-year age increase. We can estimate the (clinically more relevant) odds ratio associated with a ten-year increase in age the same way, yielding:

$$\exp(0.074 \times 10) = 2.105.$$

Following the same approach we can use (5.5) to calculate the log odds ratio and odds ratio for an arbitrary Δ unit increase in a predictor x as follows:

$$\log \left[\frac{\frac{P(x+\Delta)}{1-P(x+\Delta)}}{\frac{P(x)}{1-P(x)}} \right] = \beta_1 \Delta, \quad \frac{\frac{P(x+\Delta)}{1-P(x+\Delta)}}{\frac{P(x)}{1-P(x)}} = \exp(\beta_1 \Delta). \quad (5.6)$$

In addition to computing odds ratios, the estimated coefficients can be used in the logistic function representation of (5.2) to estimate the probability of having CHD during study follow-up for a individual with any specified age. For a 55-year-old individual:

$$P(55) = \frac{\exp(-5.940 + 0.074 \times 55)}{1 + \exp(-5.940 + 0.074 \times 55)}.$$

Of course, such an estimate only makes sense for ages near the values used in fitting the model.

The output in Table 5.2 also gives standard errors and 95% CIs for the model coefficients. The interpretation of these is the same as for the linear regression model. The fact that the interval for the coefficient of age excludes zero indicates statistically significant evidence that the true coefficient is different than zero. Similar to linear regression, the ratio of the coefficients to their standard errors forms the Wald (z) test statistic for the hypothesis that the true coefficients are different than zero. This statistic is assumed to approximately follow a normal distribution, and the associated P -value and 95% confidence intervals rely on this assumption. As introduced in Sect. 3.6, bootstrap confidence intervals are useful when the accuracy of this approximation is questionable. The logarithm of the likelihood for the fitted model along with a likelihood ratio (LR) statistic `LR_chi2(1)` and associated P -value (`Prob > chi2`) are also provided. Maximum likelihood is the standard method of estimating parameters from logistic regression models, and is based on finding the estimates which maximize the joint probability (or *likelihood*—see Sect. 5.6) for the observed data under the chosen model.

Table 5.3 Effects of age differences of 1 and 10 years, by reference age

Age (x)	$P(x)$	$P(x + 1)$	odds(x)	odds($x + 1$)	OR	RR	ER
40	0.049	0.053	0.052	0.056	1.077	1.073	0.004
50	0.098	0.105	0.109	0.117	1.077	1.069	0.007
60	0.186	0.198	0.229	0.247	1.077	1.062	0.012
Age (x)	$P(x)$	$P(x + 10)$	odds(x)	odds($x + 10$)	OR	RR	ER
40	0.049	0.098	0.052	0.109	2.105	1.996	0.049
50	0.098	0.186	0.109	0.229	2.105	1.899	0.088
60	0.186	0.325	0.229	0.482	2.105	1.746	0.139

The LR statistic given in the table compares the likelihood from the fitted model with the corresponding model excluding age, and addresses the hypothesis that there is no (linear) relationship between age and the log odds of CHD occurrence. The associated P -value is obtained from the χ^2 distribution with one degree of freedom (corresponding to the single predictor used in the model). LR tests are covered in more detail in Sect. 5.2.1. Note that the $Pseudo\ R^2$ value in the table is intended to provide a measure paralleling that used in linear regression models, and is related to the LR statistic.

As an additional illustration of the properties of the logistic model, Table 5.3 presents a number of quantities calculated directly from the coefficients in Table 5.2 and (5.2) and (5.5). For the ages 40, 50, and 60, the table gives the estimated response probabilities and odds. These are also calculated for one- and ten-year age increases so that corresponding odds ratios can be computed. As prescribed by the model, the odds ratios associated with a fixed increment change in age remain constant across the age range. Estimates of RR and ER are also computed for one- and ten-year age increments to illustrate that the fitted logistic model can be used to estimate a wide variety of quantities in addition to odds ratios. Note that the estimated values of ER and RR are not constant with increasing age (because the model does not restrict them to be so). Note also that although measures such as ER and RR can be computed from the logistic model, the resulting estimates will not in general correspond to those obtained from a regression model defined on a scale on which ER or RR is assumed constant. We will return to this topic when we consider alternative binary regression approaches in Sect. 5.5, and again in Sect. 9.3, where we consider use of the logistic model to estimate response probabilities for binary predictors representing contrasting exposure scenarios in the context of causal inference.

5.1.2 Categorical Predictors

Similar to the conventional linear regression model, the logistic model (5.5) is equally valid for categorical risk factors. For example, we can use it to look again at the relationship between CHD risk and the binary predictor *arcus senilis* as

Table 5.4 Logistic model for CHD and arcus senilis

```
. logistic chd69 i.arcus

Logistic regression

Log likelihood = -879.10783
```

Number of obs

=

3152

LR chi2(1)

=

12.98

Prob > chi2

=

0.0003

Pseudo R2

=

0.0073

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
1.arcus	1.63528	.2195035	3.66	0.000	1.257	2.127399

shown in Table 5.4. The regression output in Table 5.4 summarizes the model fit in terms of the odds ratio for the included predictor, and does not include estimates of the regression coefficients. In particular, the model intercept is omitted. This is the default option in many statistical packages such as Stata. Specifying the `coef` option as illustrated in Table 5.2 provides coefficient estimates, including the intercept. Note also that the estimated odds ratio, P -value for the Wald test that the true value the odds ratio is one (or, equivalently that the coefficient is zero), and corresponding 95% CI are virtually the same as the results obtained in Table 3.5. Because `arcus` is a binary predictor (coded as one for individuals with the condition and zero otherwise), entering it directly into the model as if it were a continuous measurement produces the desired result: the coefficient represents the log odds ratio associated with a one-unit increase in the predictor. (In this case, only one, single unit increase is possible by definition.) For two-level categorical variables with levels coded other than zero or one, care must be taken so that they are appropriately treated as categories (and not continuous measurements) by the model-fitting software.

Categorical risk factors with multiple levels are treated similarly to the procedure introduced in Sect. 4.3 for linear regression. In this way, we can repeat the analysis in Table 5.1, dividing study participants into five age groups and taking the youngest group as the reference. In order to estimate odds ratios for each of the four older age groups compared to the youngest group, we need to construct four indicator variables corresponding to the levels of the categorical variable encoding the age groups. Stata does this automatically via the `i.` prefix for the categorical predictor `agec`, as shown in Table 5.5. This variable is constructed with categories corresponding to the age divisions shown in Table 5.1.

Note that the estimated odds ratios appear to be identical to those in the table. In fact, because we are estimating a parameter for each age category except the youngest (reference) group, we are not imposing any restrictions on the parameters (i.e., the logistic assumption does not come into play as it does for continuous predictors). Thus, we would expect the estimated odds ratios to be identical to those estimated using the contingency table approach.

The LR test for this model compares the likelihood for the model with four indicator variables for age with that from the corresponding model with no

predictors. In contrast to the individual Wald tests provided for each level of age, the LR test examines the overall effect of age represented as a five-level predictor. The results indicate that inclusion of age affords a statistically significant improvement in the fit of the model.

The table also includes output from the Stata `testparm` and `contrast` commands, used here to test the global hypothesis that the coefficients for the four older age categories are all equal to zero. This hypothesis is identical to the one addressed by the LR test in this case, and the resulting Wald `chi2` test statistic is quite similar to the LR statistic. The correspondence between these two tests is also discussed in Sects. 5.2.1 and 10.4.2.

We note that caution should be exercised in interpretation of significance results for individual Wald tests for categorical predictors with multiple levels, especially in cases where the overall hypothesis test is not statistically significant. As discussed in Sect. 4.3.4, the `mcompare` option allows for control of the familywise Type-I error rate (FER) in making multiple pairwise comparisons, using Bonferroni, Sidak, and Scheffé procedures. In this case, we used the `contrast` command with option `mcompare(sidak)` to obtain more conservative *P*-values and CIs for the age effects (the odds-ratios are unchanged).

An additional test of interest in this example is evaluation of the presence of linear trend in the log odds of CHD with increasing age category. This test is implemented exactly as described for linear regression models in Sect. 4.3.5, using the contrast coefficients given in Table 4.8; the test is also obtained using both `contrast` commands introduced in Table 4.9. The result shown in Table 5.5 is quite significant, indicating evidence for a linear trend in the log odds of disease with increasing category of age, and confirming our impression of a regular increase in odds ratios with increasing age. The methods presented there for evaluating departure from linearity are also directly applicable to the logistic model.

Estimating regression coefficients for levels of a categorical predictor often involves specification of an appropriate reference category, especially for nominal categorical predictors. For the example in Table 5.5, this was chosen automatically by Stata as the age category with the smallest numerical label. (A similar procedure is followed by most major statistical packages.) Since age can be considered as ordinal, it makes sense in this case to preserve the ordering of the categories, especially if assessing trends in outcome odds with increasing age is of interest. However, in cases where a reference group different from the default is of interest, most statistics packages (including Stata and SAS) have methods for changing the default. For example, using `ib2.agec` rather than `i.agec` in the `logistic` command in Table 5.5 will result in the second age category being used as the reference. Alternatively, the model can be re-fit using a recoded version of the predictor. Note that it is also possible to compute odds ratios comparing arbitrary groups from the coefficients obtained using the default reference group. For example, the odds ratio comparing the fourth age group in Table 5.5 to the third can be shown to be $\frac{2.88}{2.32} = 1.24$. (This calculation is left as an exercise.)

Another important consideration in selecting a reference group for a categorical predictor are the sample sizes in each category. As a general rule, when individuals

are unevenly distributed across categories it is desirable to avoid making the smallest group the reference category. This is because standard errors of coefficients for other categories will be inflated due to the small sample size in the reference group.

A final issue that arises in fitting models with ordinal categorical predictors formed based on an underlying continuous measurement is the choice of how many categories, and how these should be defined. In the example in Table 5.5, the choice of five-year age groups was somewhat arbitrary. In many cases, categories will correspond to pre-existing hypotheses or be suggested by convention (e.g., ten-year age categories in summaries of cancer rates). In the absence of such information, a good practice is to choose categories of equal size based on quantiles of the distribution of the underlying measure.

How many categories a given model will support depends on the overall sample size as well as the distribution of outcomes in the resulting groups. In the WCGS sample, a logistic model including a coefficient for each unique age (assigning the youngest age as the reference group) yields reasonable estimates and standard errors. There are 266 individuals in the smallest group. (A much simpler model that fits the data adequately can also be constructed using the methods discussed in Sect. 5.4.1.) Care must be taken in defining categories to ensure that there are adequate numbers in the subgroups (possibly by collapsing categories). In general, avoid categorizations that result in categories that are homogeneous with respect to the outcome or that contain fewer than ten observations. Problems that arise when this is not the case are discussed in Sect. 5.4.4.

5.2 Multipredictor Models

Clinical and epidemiological studies of binary outcomes typically focus on the potential effects of multiple predictors. When these are categorical and few in number, contingency table techniques suffice for data analyses. However, for larger numbers of potential predictors and/or when some are continuous measurements, regression methods have a number of advantages. For example, the WCGS study measured a number of potential predictors of CHD, including total serum cholesterol, diastolic and SBP, smoking, age, body size, and behavior pattern. The investigators recognized that these variables all may contribute to outcome risk in addition to being potentially associated with each other, and that in assessment of the influence of a selected predictor, it might be important to control for the potential confounding influence of others. Because there are a number of candidate predictors, some of which can be viewed as continuous measurements, multiple regression techniques are very appealing in analyzing such data.

The logistic regression model for multiple predictor variables is a direct generalization of the version for a single predictor introduced above (5.5). For a binary

Table 5.6 Multiple logistic model for CHD risk

. logistic chd69 age chol bmi sbp i.smoke if chol<645, coef						
Logistic regression			Number of obs		= 3141	
			LR chi2(5)		= 159.80	
			Prob > chi2		= 0.0000	
Log likelihood = -807.19249			Pseudo R2		= 0.0901	
chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0644476	.0119073	5.41	0.000	.0411097	.0877855
chol	.0107413	.0015172	7.08	0.000	.0077675	.013715
bmi	.0574361	.0263549	2.18	0.029	.0057814	.1090907
sbp	.0192938	.0040909	4.72	0.000	.0112759	.0273117
1. smoke	.6344778	.1401836	4.53	0.000	.3597231	.9092325
_cons	-12.31099	.977256	-12.60	0.000	-14.22638	-10.3956

outcome y , and p predictors $x_1, x_2, \dots, x_p$, the systematic part of the model is defined as follows:

$$\log \left[\frac{P(x_1, x_2, \dots, x_p)}{1 - P(x_1, x_2, \dots, x_p)} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p. \quad (5.7)$$

This can be re-expressed in terms of the outcome probability as follows:

$$P(x_1, x_2, \dots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}. \quad (5.8)$$

As with standard multiple linear regression, the predictors may include continuous and categorical variables. The multiple-predictor version of the logistic model is based on the same assumptions underlying the single predictor version. (These are presented in Sect. 5.1.) In addition, it assumes that multiple predictors are related to the outcome in an additive fashion on the log odds scale. The interpretation of the regression coefficients is a direct generalization of that for the simple logistic model:

- For a given predictor x_j , the coefficient β_j gives the change in log odds of the outcome associated with a unit increase in x_j , for arbitrary fixed values for the predictors $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p$.
- The exponentiated regression coefficient $\exp(\beta_j)$ represents the odds ratio associated with a one unit change in x_j .

Table 5.6 presents the results of fitting a logistic regression model examining the impact on CHD risk of age, cholesterol (mg/dL), SBP (mmHg), BMI (computed as weight in kilograms divided by the square of height in meters), and a binary indicator of whether or not the participant smokes cigarettes, using data from the WCGS sample. This model is of interest because it addresses the question of whether a select group of established risk factors for CHD are independent predictors for the WCGS study.

Twelve observations were dropped from the analysis in Table 5.6 because of missing cholesterol values. An additional observation was dropped (via the `if` statement in the `regress` command) because of an unusually high cholesterol value (645 mg/dL) that is clearly an outlier. Note that all predictors are entered as continuous measurements in the model. The coefficient for any one of these (e.g., `chol`) gives the log odds ratio (change in the log odds) of CHD for a unit increase in the predictor, adjusted for the presence of the others. The small size of the coefficients for these measures reflects the fact that a unit increase on the measurement scale is a very small change, and does not translate to a substantial change in the log odds.

Log odds ratios associated with larger increases are easily computed as described in Sect. 5.1. The 95% CIs for coefficients of all included predictors exclude zero, indicating that each is a statistically significant independent predictor of outcome risk (as measured by the log odds). Of course, additional assessment of this model would be required before it is adopted as a “final” representation of outcome risk for this study. In particular, we would want to evaluate whether the linearity assumption is met for continuous predictors, evaluate whether additional confounding variables should be adjusted for, and check for possible interactions. These topics are discussed in more detail below.

As an example of an application of the fitted model in Table 5.6, consider calculating the log odds of developing CHD within ten years for a 60-year-old smoker, with 253 mg/dL of total cholesterol, SBP of 136 mmHg, and a BMI of 25. Applying (5.7) with the estimated coefficients from Table 5.6,

$$\begin{aligned}\log \left[\frac{P(60, 253, 136, 25, 1)}{1 - P(60, 253, 136, 25, 1)} \right] &= -12.311 + .0644 \times 60 + .0107 \times 253 \\ &\quad + .0193 \times 136 + .0574 \times 25 + .6345 \times 1 \\ &= -1.046.\end{aligned}$$

A similar calculation gives the corresponding log odds for a similar individual of age 50:

$$\begin{aligned}\log \left[\frac{P(50, 253, 136, 25, 1)}{1 - P(50, 253, 136, 25, 1)} \right] &= -12.311 + .0644 \times 50 + .0107 \times 253 \\ &\quad + .0193 \times 136 + .0574 \times 25 + .6345 \times 1 \\ &= -1.690.\end{aligned}$$

Finally, the difference between these gives the log odds ratio for CHD associated with a ten year increase in age for individuals with the specified values of all of the included predictors:

$$-1.046 - (-1.690) = 0.644.$$

Table 5.7 Multiple logistic model with rescaled predictors

```
. logistic chd69 age_10 chol_50 bmi_10 sbp_50 i.smoke if chol<645
```

Logistic regression

Number of obs = 3141

LR chi2(5) = 159.80

Prob > chi2 = 0.0000

Pseudo R2 = 0.0901

Log likelihood = -807.19249

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age_10	1.904989	.2268333	5.41	0.000	1.508471	2.405735
chol_50	1.710974	.1297977	7.08	0.000	1.474584	1.985259
bmi_10	1.775995	.4680613	2.18	0.029	1.059518	2.976973
sbp_50	2.623972	.5367142	4.72	0.000	1.757326	3.918016
1.smoke	1.886037	.2643914	4.53	0.000	1.432933	2.482417

Closer inspection reveals that this result is just ten times the coefficient for age in Table 5.6. In addition, we see that we could repeat the above calculations for any ten-year increase in age, and for any fixed values of the other predictors and obtain the same result. Thus, the formula (5.6) for computing log odds ratios for arbitrary increases in a single predictor applies here as well. The odds ratio for a ten-year increase in age (adjusted for the other included predictors) is given simply by

$$\exp(0.0644 \times 10) = \exp(.644) = 1.90.$$

Interpretation of regression coefficients for categorical predictors also follow that given for single predictor logistic models. For example, the coefficient (0.634) for the binary predictor variable smoke in Table 5.6 is the log odds ratio comparing smokers to nonsmokers for fixed values of age, chol, sbp, and bmi. The corresponding odds ratio

$$\exp(0.634) = 1.89$$

measures the proportionate increase in the odds of developing CHD for smokers compared to nonsmokers adjusted for age, cholesterol, SBP, and BMI.

The estimated coefficients for the first four predictors in Table 5.6 are all very close to zero, reflecting the continuous nature of these variables and the fact that a unit change in any one of them does not translate to a large increase in the estimated log odds of CHD. As shown above, we can easily calculate odds ratios associated with clinically more meaningful increases in these predictors. An easier approach is to decide on the degree of change that we would like the estimates to reflect and fit a model based on predictors rescaled to reflect these decisions. For example, if we would like the model to produce odds ratios for ten-year increases in age, we should represent age as the rescaled predictor $\text{age}_{.10} = \text{age}/10$. Table 5.7 shows the estimated odds ratios from the model including rescaled versions of the first four predictors in Table 5.6. (The numbers after the underscores in the variable names indicate the magnitude of the scaling.) We also “centered” these predictors

before scaling them by subtracting of the mean value for each. (Centering predictors is discussed in Sects. 3.3.1 and 4.6.) Note that the log-likelihood and Wald test statistics for this model are identical to their counterparts in Table 5.6.

5.2.1 Likelihood Ratio Tests

In Sect. 5.1, we briefly introduced the concept of the likelihood, and the LR test for logistic models. The likelihood for a given model is interpreted as the joint probability of the observed outcomes expressed as a function of the chosen regression model. The model coefficients are unknown quantities and are estimated by maximizing this probability (hence the name maximum-likelihood estimation). For numerical reasons, maximum-likelihood estimation in statistical software is usually based on the logarithm of the likelihood. An important property of likelihoods from nested models (i.e., models in which predictors from one are a subset of those contained in the other) is that the maximized value of the likelihood from the larger model will always be at least as large as that for the smaller model.

Although the numerical value of the likelihood (or log-likelihood) for a single model does not have a particularly useful interpretation, the LR statistic assessing the difference in likelihoods from two nested models is a valuable tool in model assessment (analogous to the F tests introduced in Sect. 4.3.3). It is especially useful when investigating the contribution of more than one predictor, or for predictors with multiple levels.

For example, consider assessment of the contribution of self-reported behavior pattern to the model summarized in Table 5.7. In the WCGS study, investigators were interested in “type A” behavior as an independent risk factor for CHD. Behavior was classified as either type A or type B, with each type subdivided into two further levels A_1 , A_2 , B_3 , and B_4 (coded as 1, 2, 3, and 4, respectively). The expanded model addresses the question of whether behavior pattern contributes to CHD risk when other established risk factors are accounted for.

Table 5.8 displays the results of including the four-level categorical variable `behpat` in the model from Table 5.7. The natural coding of the variable results in type A_1 behavior being taken as the reference level. Examination of the coefficients and associated 95% CIs for the remaining indicators reveals that although the second category of type A behavior appears not to differ from the reference level, both categories of type B behavior do display statistically significant differences, and are associated with lower outcome risk.

The LR statistic is computed as twice the difference between log likelihoods from the two models, and can be referred to the χ^2 distribution for significance testing. Because the likelihood for the larger model must be larger than the likelihood for the smaller (nested) model, the difference will always be positive. Twice the difference between the log likelihood for the model including `behpat` (Table 5.8) and that for the model excluding this variable (Table 5.6) is

$$2 \times [-794.81 - (-807.19)] = 24.76.$$

Table 5.8 Logistic model for WCGS behavior pattern

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 i.smoke i.behpat if chol<645
```

Logistic regression

Log likelihood = -794.81

Number of obs = 3141

LR chi2(8) = 184.57

Prob > chi2 = 0.0000

Pseudo R2 = 0.1040

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age_10	1.83375	.2198681	5.06	0.000	1.449707	2.319529
chol_50	1.704097	.1301391	6.98	0.000	1.467201	1.979243
sbp_50	2.463504	.5086518	4.37	0.000	1.643621	3.692369
bmi_10	1.739415	.4620341	2.08	0.037	1.033479	2.927551
1.smoke	1.830672	.2583097	4.29	0.000	1.38837	2.413882
behpat						
2	1.068257	.2363271	0.30	0.765	.6924157	1.648103
3	.5141593	.1245593	-2.75	0.006	.3198064	.8266243
4	.572071	.1826117	-1.75	0.080	.3060107	1.069457

```
. estimates store mod1
```

Table 5.9 Likelihood ratio test for four-level WCGS behavior pattern

```
. lrtest mod1
```

likelihood-ratio test

(Assumption: . nested in mod1)

LR chi2(3) = 24.76

Prob > chi2 = 0.0000

This value follows a χ^2 distribution, with degrees of freedom equal to the number of additional variables present in the larger model (three in this case). Statistical packages like Stata can often be used to compute the LR test directly by first fitting the larger model (in Table 5.8), and saving the likelihood in the user-defined variable (in this case, in the variable `mod1` created in the last line of the table). Next, the reduced model eliminating `behpat` is fit, followed by a command to evaluate the LR test as displayed in the Table 5.9. (See Table 5.6 for the full regression output for this model.) The result agrees with the calculation above, and the associated P -value indicates that collectively, the four-level categorical representation of behavior pattern makes a statistically significant independent contribution to the model.

The similarity between the two odds ratios for type A (the reference level and the second indicator for type A_2 behavior) and type B (the indicators representing types B_3 and B_4 behavior) in Table 5.8 suggests that a single binary indicator distinguishing the A and B patterns might suffice. Note that the logistic model that represents behavior pattern as a two-level indicator (with type B behavior as the reference category) is actually nested within the model in Table 5.8. (The model including the two-level representation is a special case of the four-level version when the coefficients for the two levels of type B and type A behavior, respectively, are identical.) Table 5.10 displays the fitted model and LR test results for this reduced model including the two-level binary indicator `dibpat`. The fact that the difference between the likelihoods for the two models is not statistically significant

Table 5.10 Likelihood ratio test for two-level WCGS behavior pattern

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 i.smoke i.dibpat if chol<645
```

Logistic regression

Log likelihood = -794.92603

Number of obs = 3141

LR chi2(6) = 184.34

Prob > chi2 = 0.0000

Pseudo R2 = 0.1039

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age_10	1.830252	.2190623	5.05	0.000	1.44754	2.314147
chol_50	1.702406	.1299562	6.97	0.000	1.465835	1.977157
sbp_50	2.467919	.5084377	4.38	0.000	1.648039	3.695681
bmi_10	1.732349	.4596114	2.07	0.038	1.029917	2.913859
i.smoke	1.829163	.2580698	4.28	0.000	1.387265	2.411822
i.dibpat	2.006855	.2897341	4.82	0.000	1.512259	2.663212

```
. lrtest mod1
```

likelihood-ratio test

(Assumption: . nested in mod1)

LR chi2(2) = 0.23

Prob > chi2 = 0.8904

confirms our suspicion that modeling the effect of behavior pattern as a two-level predictor is sufficient to capture the contribution of this variable.

As demonstrated above, the LR test is a very useful tool in comparing nested logistic regression models. Note that alternate tests based on Wald statistics can also be used, as illustrated in Tables 4.4 and 5.5. In moderate to large samples, the results from the LR and Wald tests for the effects of single predictors will agree quite closely. However, in smaller samples the results of these two tests may differ substantially. In general, the LR test is more reliable than the Wald test, and is preferred when both are available. Finally, note that because the likelihood is computed based on the observations used to fit the model, it is important to ensure that the same observations are included in each candidate model considered in LR testing. This was accomplished in the examples by insuring that the fitted models excluded 12 observations with missing values for cholesterol, and another with an outlying value of 645. Likelihoods from models fit on differing sets of observations are not comparable. A more complete discussion of the concepts of likelihood and maximum-likelihood estimation is given in Sect. 5.6.

5.2.2 *Confounding*

A common goal of multiple logistic regression modeling is to investigate the association between a primary predictor and the outcome, accounting for the possible mediating or confounding influence of additional measured predictors. For example, in evaluating the observed association between behavior pattern (considered in the previous section) and CHD risk, it is important to consider the effects of additional variables that might be related to both behavior and

Table 5.11 Logistic model for type A behavior pattern and selected predictors

```
. logistic dibpat age_10 chol_50 sbp_50 bmi_10 i.smoke
```

Logistic regression

Log likelihood = -2150.1739

Number of obs = 3141

LR chi2(5) = 53.80

Prob > chi2 = 0.0000

Pseudo R2 = 0.0124

dibpat	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age_10	1.324032	.0881552	4.22	0.000	1.16205	1.508594
chol_50	1.084241	.0464136	1.89	0.059	.9969839	1.179135
sbp_50	1.461247	.1876433	2.95	0.003	1.136104	1.879442
bmi_10	1.123846	.1672474	0.78	0.433	.8395252	1.504459
1.smoke	1.26933	.0930786	3.25	0.001	1.099403	1.465522

CHD occurrence. Recall from Chap. 4 that regression models can account for potential confounding or mediation influences of such variables by considering the adjusted and unadjusted associations between the outcome and predictor of primary interest. In this section, we briefly review these issues in the logistic regression context.

Consider again the assessment of behavior pattern as a predictor of CHD in the WCGS example considered in the previous section. In the analysis summarized in Table 5.10, we concluded that a two-level indicator (`dibpat`) distinguishing type A and B behaviors adequately captures the effects of this variable on CHD (in place of a more complex, four-level summary of behavior). The discussion in Chap. 9 will suggest that we should consider the possible causal relationships of the additional variables in the model with both the outcome and behavior pattern before making any conclusions about the possible causal connection between behavior type and the outcome.

Recall the discussion of confounding and mediation presented in Sects. 4.4 and 4.5. To be a confounder of an association of primary interest, a variable must be associated with both the outcome and the primary predictor. From Table 5.10, all of the predictors in addition to `dibpat` are independently associated with the CHD outcome. Since `dibpat` is a binary indicator, we can examine its association with these predictors via logistic regression as well. Table 5.11 presents the resulting model. With the exception of BMI (`bmi_10`), all appear to be associated with behavior pattern. In deciding which variables to adjust for in summarizing the CHD-behavior pattern association, it is worth considering the possible causal relationships to help identify or distinguish variables with confounding influence from those that could be potential mediators or effect modifiers.

Causal connections are likely to be very complex. For example, age can be considered as a possible confounder of the relationship between behavior type and CHD. However, BMI, cholesterol, SBP (hypertension), and smoking could either exert a confounding influence or be viewed as mediating variables in the pathway between behavior and CHD. The unadjusted odds ratio (95% CI) for the association

between type A behavior and CHD is 2.36 (1.79, 3.10). By contrast, the adjusted odds ratio in Table 5.10 is 2.01 (95% CI 1.51, 2.66). Note that dropping any of the additional predictors from the model singly results in little change to the estimated OR for type A behavior (less than 5%). Thus if any of these variables acts as a mediator, the influence appears to be weak. This suggests that the influence of type A behavior on CHD may act partially through another unmeasured pathway. (Or that this characterization of behavior is itself mediated through other unmeasured behavioral characteristics.) In this case, adjustment for the other variables is appropriate if they are considered as confounders. However, if they (with the possible exception of age) are regarded as mediators, then the effect assessed on the adjusted model can be viewed as an estimate of the direct effect of behavior not mediated through the pathways mediated by these variables. See Sects. 9.6 and 10.2 for further discussion of these issues. Of course, before concluding that we have adequately modeled the relationship between behavior pattern and CHD we need to account for possible interactions between included predictors (Sect. 5.2.4), and conduct diagnostic assessments of the model fit including nonlinearity in relationships with continuous predictors (Sect. 5.4).

5.2.3 Mediation

As an example of assessment of mediation in the context of a binary outcome, we consider an example from the FIT study, a randomized trial investigating the effect of a treatment for reducing spinal fracture risk in postmenopausal women with prior history of fracture due to osteoporosis (Black et al. 1996b). We are interested in evaluating possible mediation of treatment effects through changes in bone mineral density (BMD). A finding that much of the beneficial effect of treatment operated through this pathway would be of practical interest in development of future treatments.

Table 5.12 presents two logistic regression models for the effect of randomized treatment assignment on a binary indicator of spinal fracture occurrence. The first model gives the marginal effect of assignment to treatment in the entire sample of 5,470 women. Assuming that randomization was effective, the unadjusted odds ratio for treatment in this model represents an *intention to treat* estimate of the effectiveness of treatment assignment. The second model in the table includes predictors for change in BMD (in standard deviation units) between follow-up and baseline, baseline level of BMD (also in standard deviation units), baseline smoking status (former and current smokers compared to nonsmokers as the reference category), and a binary indicator of a history of previous spinal fracture (`frac_base`). Age (in years) is also included as a restricted cubic spline with three knots. Note that since the follow-up level of BMD reflects changes that occurred postrandomization, these baseline measures represent potential confounders of the association between change in BMD and new fracture occurrence. As discussed in Sect. 4.5, interpretation of the apparent attenuation of the effect of treatment in this model

Table 5.12 Logistic regression estimation of marginal and direct effect of treatment assignment on new fracture risk in the FIT study example

```
*** Marginal treatment effect ***
. logistic frac_new i.treat
```

Logistic regression	Number of obs	=	5470
	LR chi2(1)	=	32.05
	Prob > chi2	=	0.0000
Log likelihood = -1163.5889	Pseudo R2	=	0.0136

frac_new	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
1.treat	.5052736	.0624452	-5.52	0.000	.3965785 .64376


```
*** Direct treatment effect not mediated by change in BMD ***
. logistic frac_new i.treat bmd_diff bmd_base i.frac_base i.smoking age_spl*
```

Logistic regression	Number of obs	=	5339
	LR chi2(8)	=	311.04
	Prob > chi2	=	0.0000
Log likelihood = -982.6019	Pseudo R2	=	0.1366

frac_new	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
1.treat	.5966412	.0829632	-3.71	0.000	.4543112 .7835616
bmd_diff	.7062953	.0505978	-4.85	0.000	.6137729 .8127648
bmd_base	.6885569	.0412505	-6.23	0.000	.6122735 .7743444
1.frac_base	3.428229	.4569538	9.24	0.000	2.640048 4.451719
smoking					
1	1.141699	.1555701	0.97	0.331	.8741083 1.491207
2	1.379136	.2722494	1.63	0.103	.9366451 2.030669
age_spl1	1.123983	.0413332	3.18	0.001	1.045822 1.207986
age_spl2	.9476609	.0329655	-1.55	0.122	.8852031 1.014526

relative to the first (unadjusted) model requires assumptions about the causal nature of the relationships represented. In this example, a plausible interpretation is that treatment effects are mediated through treatment-induced changes in BMD.

Following the approach introduced in Sect. 4.5, we can assess whether the conditions for mediation are met by fitting two models: the first, a linear regression for the dependence of change in BMD on treatment assignment; the second, a logistic regression of the dependence of the outcome on change in BMD. In both cases, we adjust for the possible confounders displayed in Table 5.12. Both models yield highly significant results for the Wald tests of the coefficients representing the key components of the mediating relationships. Further, there is no evidence for interaction between treatment assignment and change in BMD. This, and the observed attenuation in the estimated effect of treatment in the second model in Table 5.12 provides evidence for the possible mediating role of change in BMD.

As also discussed in Sect. 4.5, it may also be of interest to make separate estimates of the direct and indirect components of the overall effect of treatment

assignment on fracture risk, and to estimate the proportion of the treatment effect explained (PTE) by the mediating influence of changes in BMD. Similar to the examples presented in that section, the odds ratio of 0.597 for treatment assignment in the second model shown in Table 5.12 can be interpreted as an estimate of the direct effect of treatment not mediated through effects on BMD.

By contrast to the results presented in Sect. 4.5, decomposing the relationships between outcome, treatment, and a mediator into overall, indirect, and direct effect components poses additional difficulties in the context of logistic regression models. This results from the use of the odds ratio as a measure of association, as discussed in Sect. 3.4.4. (A similar phenomenon occurs for the Cox regression model introduced in the next chapter.) Performing analyses using an alternative binary regression model based on relative risks rather than odds ratios (see Sect. 5.5.3) avoids this difficulty. Chapter 9 presents further discussion of this topic, including an introduction to more general techniques for assessment of mediation based on causal inference methods. In particular, these methods allow estimation of the causal direct effect of treatment, not mediated through the mediating variable. This estimate will generally differ from the regression estimate described here and has a clearer causal interpretation, especially when additional confounding variables play a role.

5.2.4 Interaction

Recall from Chap. 4 that an interaction between two predictors in a regression model means that the degree of association between each predictor and the outcome varies according to levels of the other predictor. The mechanics of fitting logistic regression models including interaction terms is quite similar to standard linear regression (see Sect. 4.6). For example, to fit an interaction between two continuous predictors x_1 and x_2 , we include the product x_1x_2 as an additional predictor in a model containing x_1 and x_2 as shown in (5.9):

$$\log \left[\frac{P(x_1, x_2, x_1 \times x_2)}{1 - P(x_1, x_2, x_1x_2)} \right] = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1 \times x_2. \quad (5.9)$$

Fitting interactions between categorical predictors and between continuous and categorical predictors also follows the procedures outlined in Chap. 4. However, because of the log odds ratio interpretation of regression coefficients in the logistic model, interpreting results of interactions is somewhat different. We review several examples below.

For an illustrative example of a two-way interaction between two binary indicator variables from the WCGS study, consider the regression model presented in Table 5.13. The fitted model includes the indicator `arcus` for arcus senilis (defined in Sect. 3.4), a binary indicator `bage_50` for participants over the age of 50, and the product between them, `bage_50#arcus`, made automatically by the `##` operator in the `logistic` command. The research question addressed is whether the

Table 5.13 Logistic model for interaction between arcus and age as a categorical predictor

<code>. logistic chd69 i.bage_50##i.arcus, coef</code>						
Logistic regression			Number of obs		= 3152	
			LR chi2(3)		= 40.33	
			Prob > chi2		= 0.0000	
Log likelihood = -865.43251			Pseudo R2		= 0.0228	
chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
1.bage_50	.8932677	.1721239	5.19	0.000	.5559111	1.230624
1.arcus	.6479628	.1788637	3.62	0.000	.2973964	.9985293
bage_50#arcus						
1 1	-.5920552	.2722269	-2.17	0.030	-1.12561	-.0585002
_cons	-2.882853	.1089261	-26.47	0.000	-3.096344	-2.669362

association between arcus and CHD is age dependent. The statistically significant result of the Wald test for the coefficient associated with the product of the indicators for age and arcus indicates that an interaction is present. This means that we cannot interpret the coefficient for arcus as a log odds ratio without specifying whether or not the participant is older than 50. (A similar result holds for the interpretation of bage_50.)

The procedure for obtaining the component odds ratios is similar to the methods for obtaining main and interaction effects for linear regression models, and is straightforward using the regression model. If we represent `1.arcus` and `1.bage_50` as x_1 and x_2 in (5.9), we can compute the log odds for any combination of values of these predictors using coefficients from Table 5.13. For example, the log odds of CHD occurrence for an individual over 50 years old without arcus is given by

$$\begin{aligned}\log \left[\frac{P(0, 1, 0)}{1 - P(0, 1, 0)} \right] &= \beta_0 + \beta_2 \\ &= -2.883 + 0.893 = -1.990.\end{aligned}$$

Similarly, the log odds for an individual between 39 and 49 years old without arcus is

$$\log \left[\frac{P(0, 0, 0)}{1 - P(0, 0, 0)} \right] = \beta_0.$$

With these results, we see that the five expressions below define the component log odds ratios in the example:

$$\begin{aligned}\log \left[\frac{P(1, 0, 0)}{1 - P(1, 0, 0)} \right] - \log \left[\frac{P(0, 0, 0)}{1 - P(0, 0, 0)} \right] &= \beta_1 = 0.648 \\ \log \left[\frac{P(1, 1, 1)}{1 - P(1, 1, 1)} \right] - \log \left[\frac{P(0, 1, 0)}{1 - P(0, 1, 0)} \right] &= \beta_1 + \beta_3 = 0.056\end{aligned}$$

Table 5.14 Component odds ratios for arcus-age interaction model

Odds ratio	Groups compared
$\exp(\beta_1) = 1.91$	Arcus vs. no arcus, age 39–49
$\exp(\beta_1 + \beta_3) = 1.06$	Arcus vs. no arcus, age 50–59
$\exp(\beta_2) = 2.44$	Age 50–59 vs. age 39–49, no arcus
$\exp(\beta_2 + \beta_3) = 1.35$	Age 50–59 vs. age 39–49, arcus
$\exp(\beta_1 + \beta_2 + \beta_3) = 2.58$	Arcus and age 50–59 vs. no arcus and ages 39–49

Table 5.15 Example odds ratio for arcus-age interaction model

```
. lincom 1.bage_50 + 1.bage_50#1.arcus
```

(1) [chd69]1.bage_50 + [chd69]1.bage_50#1.arcus = 0

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
(1)	1.351497	.2850372	1.43	0.153	.8939071 2.043325

$$\begin{aligned} \log \left[\frac{P(0, 1, 0)}{1 - P(0, 1, 0)} \right] - \log \left[\frac{P(0, 0, 0)}{1 - P(0, 0, 0)} \right] &= \beta_2 = 0.893 \\ \log \left[\frac{P(1, 1, 1)}{1 - P(1, 1, 1)} \right] - \log \left[\frac{P(1, 0, 0)}{1 - P(1, 0, 0)} \right] &= \beta_2 + \beta_3 = 0.301 \\ \log \left[\frac{P(1, 1, 1)}{1 - P(1, 1, 1)} \right] - \log \left[\frac{P(0, 0, 0)}{1 - P(0, 0, 0)} \right] &= \beta_1 + \beta_2 + \beta_3 = 0.949. \end{aligned} \quad (5.10)$$

The corresponding odds ratios are then easily calculated by exponentiation, as shown in Table 5.14.

Referring back to Table 5.13, we see that all of the component odds ratios aren’t immediately obvious from standard regression output. However, the log odds ratio and associated 95% CIs for *arcus* among individuals in the younger age group and for older individuals among those without *arcus* can be read directly. This is because when we set either variable to zero (the reference level), the interaction term evaluates to zero and is eliminated. Estimated log odds ratios corresponding to the nonreference levels of these variables involve the interaction term, and differ from their counterparts by the value of its coefficient (−0.592). Standard errors and 95% CIs for these estimates require additional calculations that cannot be completed without further information about the fitted model. Fortunately, many statistical packages have facilities that greatly simplify these calculations. Table 5.15 illustrates the use of the `lincom` command in Stata to compute the odds ratio comparing the odds of CHD in individuals of age 50 and over with the odds among those under 50, among individuals with *arcus*.

By specifying the correct combination of coefficients (corresponding to those in Table 5.14), the output in the Table 5.15 provides the desired odds ratio estimate along with the 95% CI. Results of the accompanying hypothesis test that the underlying log odds ratio is zero are also provided.

Table 5.16 Logistic model for interaction between arcus and age as continuous

```
. logistic chd69 i.arcus#c.age, coef
```

Logistic regression

Log likelihood = -858.93362

Number of obs = 3152

LR chi2(3) = 53.33

Prob > chi2 = 0.0000

Pseudo R2 = 0.0301

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
1.arcus	2.754185	1.140118	2.42	0.016	.5195952	4.988774
age	.089647	.0148904	6.02	0.000	.0604623	.1188317
arcus#c.age						
1	-.0498298	.0233431	-2.13	0.033	-.0955814	-.0040782
_cons	-6.788086	.7179977	-9.45	0.000	-8.195335	-5.380836

Interactions between a continuous and categorical variable are handled in a similar fashion to those involving binary predictors. In the previous example, the categorization of age was somewhat arbitrary. In fact, because age was represented by two categories, essentially the same results could have been obtained using frequency table techniques (as illustrated in Table 3.9). A more complete assessment of the interaction can be obtained by considering age as a continuous variable (previously considered in Table 5.2). For example, this would allow us to investigate whether increase in CHD risk with increasing age differs in individuals with and without arcus. The logistic model addressing this question is displayed in Table 5.16.

Note the use of the ## operator in Stata, introduced in Sect. 4.6, which instructs the program to include an interaction term between the two variables. This is accomplished by inclusion of the product of arcus and age (arcus#c.age) as well as the individual predictors age and 1.arcus. For a fixed age (e.g., 55), the log odds ratio associated with having arcus is calculated as follows, using the estimated coefficients from Table 5.16:

$$\begin{aligned} &\log \left[\frac{P(1, 55, 55)}{1 - P(1, 55, 55)} \right] - \log \left[\frac{P(0, 55, 0)}{1 - P(0, 55, 0)} \right] \\ &= (-6.788 + 2.754 + (0.090 - 0.050) \times 55) - (-6.788 + 0.090 \times 55) \\ &= (2.754 - 0.050 \times 55) = 0.014. \end{aligned}$$

We see that this corresponds to an odds ratio of $\exp(0.014) = 1.01$, which is similar to that calculated for the corresponding age group in Table 5.14. We can obtain this estimate and its 95% CI directly as shown in Table 5.17.

Note that because age is represented as a continuous variable, its value must be specified in interpreting the effect of arcus on the log odds of CHD risk. Similarly, among individuals with arcus, log odds ratios can be computed for any specified increase in age. Figure 5.2 displays the estimated log odds as a function of age,

Table 5.17 Logistic model for interaction between arcus and age as a continuous predictor

```
. lincom 1.arcus + 55*1.arcus#c.age
```

```
( 1)  [chd69]1.arcus + 55*[chd69]1.arcus#c.age = 0
```

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
(1)	1.013637	.2062336	0.07	0.947	.6802954	1.510313

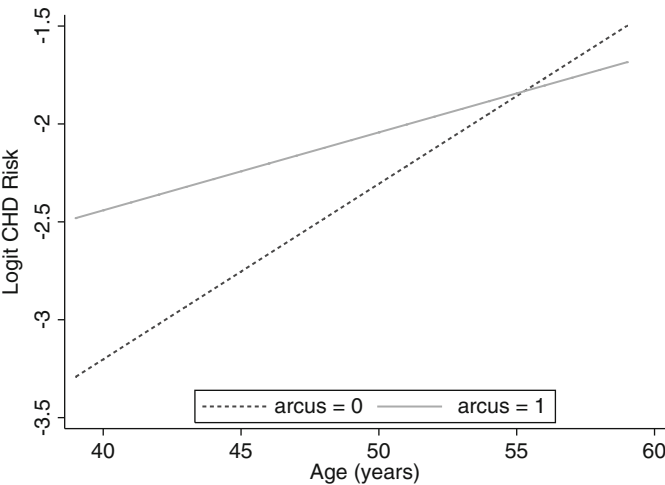


Fig. 5.2 Log odds of CHD and age for individuals with and without arcus senilis

separately for individuals with and without arcus. The equations for these two lines can be obtained directly from the coefficients in Table 5.16 and are printed below for individuals with and without arcus, respectively:

$$\begin{aligned}\log \left[\frac{P(\text{age})}{1 - P(\text{age})} \right] &= (-6.788 + 2.754) + (0.090 - 0.050) \times \text{age} \\ &= -4.034 + 0.040 \times \text{age}.\end{aligned}$$

and

$$\log \left[\frac{P(\text{age})}{1 - P(\text{age})} \right] = -6.788 + 0.0896 \times \text{age}.$$

Figure 5.2 displays the results obtained above, indicating that CHD risk is higher for younger participants with arcus. However, older participants with arcus seem to be at somewhat lower risk than those without arcus. Of course, further interpretation

of these equations should be preceded by thorough checking of the linearity of the relationship between age and the log odds of the outcome, including whether more complicated, higher-order interaction terms are needed.

Recall the discussion in Sect. 5.1 where we motivated the logistic model as an example of a multiplicative risk model (see (5.4)). By contrast, the risk difference model (introduced in (5.1) and discussed further in Sect. 5.5.3) is an example of an additive risk model. In addition to defining two distinct ways in which a predictor can act to modify outcome risk, this distinction turns out to be very important in the context of interaction: For a specified outcome and predictor pair, it is possible to have interaction under the multiplicative model and not under the additive model, and vice versa.

For example, if we fit the additive risk model to the data from the age/arcus example in Table 5.16, the Wald test P -value for inclusion of the product term (age_50arcus) is 0.15. (The corresponding value from the logistic model was 0.03.) The implications of this are that we should not necessarily regard interaction as mirroring a biological mechanism, but rather as a property of the data and model being fit. In the example, we would want to account for the interaction if we were using the logistic model but not necessarily if we were analyzing the WCGS data using the additive model. The additive regression model is described further in Sect. 5.5.3. Also, see Clayton and Hills (1993) and Jewell (2004) for more detailed discussions of the distinction between multiplicative and additive interaction.

5.2.5 Prediction

Frequently, the goal of fitting a logistic model is to predict risk of the binary outcome given a set of risk factors. Recall that in Sect. 5.2.1, we fit a logistic model for the CHD outcome in the WCGS sample, using age, cholesterol level, systolic blood pressure, BMI, a binary indicator of current cigarette smoking (with nonsmokers composing the reference group), and an indicator of type A behavior as predictors. Table 5.10 summarizes the results. Table 5.18 presents an expanded version of this model that includes two additional predictors `bmichol` and `bmisbp` for the interactions between BMI and serum cholesterol level and BMI and SBP (both centered and scaled as described in Sect. 5.2). These were both found to make statistically significant contributions to the model in further analyses investigating two way interactions between the original predictors in Table 5.10.

As shown in Sect. 5.2, the estimated coefficients from the model in Table 5.18 can be used directly in the logistic formula (5.8) to compute the log odds (or the corresponding probability) of CHD for an arbitrary individual by specifying the desired values for the predictors. Table 5.19 displays a few such predictions (labeled `prchd`) for five individuals in the WCGS sample (obtained using the `predict` command in Stata).

Table 5.18 Expanded logistic model for CHD events

```
. logistic chd69 age_10 chol_50 sbp_50 bmi_10 smoke dibpat bmichol bmisbp,
coef

Logistic regression

Log likelihood = -788.01957
```

Number of obs = 3141

LR chi2(8) = 198.15

Prob > chi2 = 0.0000

Pseudo R2 = 0.1117

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age_10	.5949713	.1201092	4.95	0.000	.3595615	.830381
chol_50	.5757131	.07779	7.40	0.000	.4232474	.7281787
sbp_50	1.019647	.2066014	4.94	0.000	.6147159	1.424579
bmi_10	1.048839	.2998176	3.50	0.000	.4612074	1.636471
smoke	.6061929	.1410533	4.30	0.000	.3297335	.8826523
dibpat	.7234267	.1448996	4.99	0.000	.4394288	1.007425
bmichol	-.8896932	.2746471	-3.24	0.001	-1.427992	-.3513948
bmisbp	-1.503455	.631815	-2.38	0.017	-2.74179	-.2651208
_cons	-3.416061	.1504717	-22.70	0.000	-3.71098	-3.121142

Table 5.19 Sample predictions from the logistic model in Table 5.18

	chd69	age	chol	sbp	bmi	smoke	dibpat	prchd
1.	no	49	225	110	19.78795	smoker	A1,A2	.0433952
2.	no	42	177	154	22.9551	smoker	A1,A2	.0708145
3.	no	42	181	110	23.62529	nonsmoker	B3,B4	.0082533
4.	no	41	132	124	23.109	smoker	B3,B4	.0089318
5.	yes	59	255	144	21.52041	smoker	B3,B4	.1926046

5.2.6 Prediction Accuracy

In some applications, we may be interested in using a logistic regression model as a tool to classify outcomes of newly observed individuals based on values of measured predictors. For the WCGS example just considered, this may involve deciding on treatment strategy based on prognosis as measured by the predicted probability from the logistic model in Table 5.18. Similar to the goals of developing diagnostic tests for detecting diseases, this approach requires us to choose a cut-off or threshold value of the predicted outcome probability above which treatment would be initiated. A fundamental consideration in choosing this threshold is in evaluating the degree of misclassification of outcomes incurred by the choice. For a binary outcome, misclassification can be quantified by calculating the proportion of individuals incorrectly classified as either having the outcome or not. These are known as the *false-positive* and *false-negative* rates, respectively, and are standard measures of prediction error in the logistic regression context. Rather than state prediction performance in terms of misclassification, the following complementary measures are frequently used in assessment of prediction rules for binary outcomes:

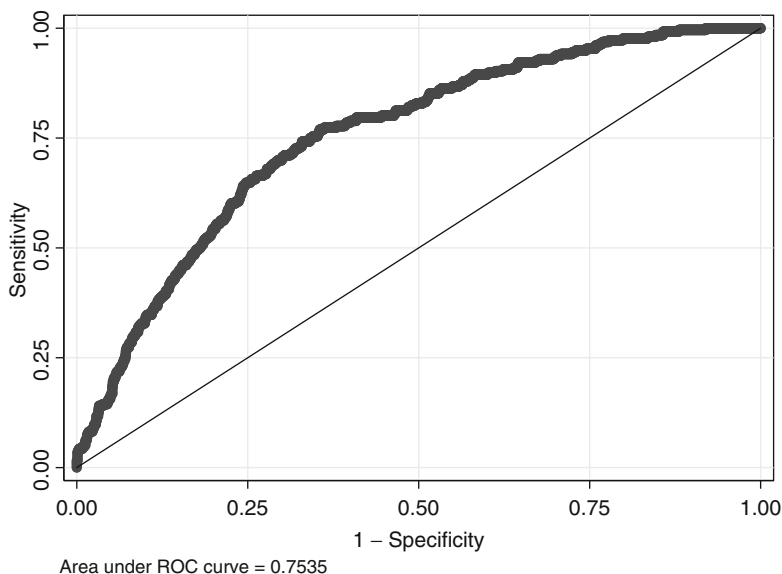


Fig. 5.3 ROC curve for logistic prediction of CHD events

Sensitivity The proportion of individuals with the outcome that are correctly classified, calculated as the complement of the false-negative rate.

Specificity The proportion of individuals without the outcome that are correctly classified, calculated as the complement of the false-positive rate.

As the threshold value of a prediction rule varies between zero and one, these quantities can be calculated and compared to evaluate overall performance. A *receiver operating characteristic* (ROC) curve plots the sensitivity against the false-positive rate (i.e., one minus the specificity) for a range of thresholds to help visualize test performance. Figure 5.3 shows the ROC curve for the current example (obtained using the `lrroc` command in Stata), along with a diagonal reference line, usually interpreted as representing the ROC curve for a test that is no better than the flip of a coin.

ROC curves for tests with overall good performance (i.e., low misclassification rates for both positive and negative outcomes) will lie close to the left and topmost margins of the plot. In Fig. 5.3, a test with a sensitivity of around 75% is close to optimal in this sense. (The threshold value corresponding to a sensitivity of 0.75 and a specificity of 0.64 in Fig. 5.3 is about 0.07.) Note that in most practical situations, assessment of test performance has a subjective component: The cost of misclassifying an individual as positive may be deemed more serious than the alternative situation, or vice versa. These considerations weigh into evaluation of test results. The area under an ROC curve (also known as the *C-statistic*) provides an

overall measure of classification accuracy, with the value of one representing perfect accuracy. In the present case, the value of 0.754 does not indicate very impressive performance.

A clear limitation with the example above is that the individuals used to evaluate the performance are the same as those used to fit the model on which the classification rule is based. Alternative techniques that do not share this limitation include cross-validation and learning set/test set validation (both described in Sect. 10.1). Finally, note that although logistic regression is a valid approach for development of prediction tools, alternative techniques are available. Classification trees are an example of a larger class of tree-based methods, and involve fewer modeling assumptions than the logistic approach. See [Goldman et al. \(1996\)](#) for an example of their application in a clinical context. Prediction is discussed in greater detail in Sect. 10.1.

5.3 Case-Control Studies

In situations where binary outcomes are rare or difficult to observe, it is not always feasible to collect a large enough sample to investigate the relationship between the outcome and predictors of interest. Consider the problem of evaluating dietary risk factors for stomach cancer. Because this disease is relatively rare (accounting for approximately 2% of annual cancer deaths in the United States), only a very large cross-sectional or prospective sample would include sufficient numbers of cases to evaluate associations with predictors of interest. Case-control studies address this problem by recruiting a fixed number of individuals with the outcome of interest (the cases) and a number of comparable control individuals free of the outcome. Retrospective histories of predictor variables of interest are then collected via questionnaire after recruitment.

A well-known example of a case-control study is the Ille-et-Vilaine study of cancer conducted in France between 1972 and 1974. It includes 200 cases and 775 comparable controls, and was designed to investigate alcohol, diet, and tobacco consumption as risk factors for esophageal cancer in men. This is known as an *unmatched* study since cases and controls were sampled separately in predetermined numbers. An alternative type of case-control study is based on *matching* a fixed number of controls to each sampled case based on selected characteristics. Methods for matched studies are different and will be covered briefly below in Sect. 5.3.1.

Because the overall proportion of individuals is fixed by design in a case-control study (e.g., 200/995, or approximately five controls per case for Ille-et-Vilaine), it is not meaningful to make direct comparisons of outcome risk (estimated as the proportion of individuals with the outcome) between groups defined by predictor variables, as is conventional in studies where participants are not sampled based on their outcome status. Rather, analyses are based on the distribution of predictors variables compared across case/control status. At first glance, this approach does not seem to address the fundamental question of whether or not the predictor is associated with increased risk of developing the outcome. For example, observing that

Table 5.20 Odds ratio for smoking and esophageal cancer

```
. tabodds case ditob, or
```

ditob	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
0-9 g/day	1.000000	.	.	.	.
10+ g/day	10.407051	64.89	0.0000	5.119049	21.157585

```
. tabodds ditob case, or
```

case	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
0	1.000000	.	.	.	.
1	10.407051	64.89	0.0000	5.119049	21.157585

self-reported alcohol consumption differed between cases and controls in Ille-et-Vilaine does not seemingly translate into a clear statement about esophageal cancer risk associated with alcohol use. Further, application of conventional measures of association to settings where the role of the outcome and predictor are reversed seemingly leads to unintuitive results. For example, observing that individuals with esophageal cancer risk are twice as likely (in terms of the relative risk) as cancer-free individuals to report a specified degree of alcohol consumption does not state the association in a way that makes the possible causal connection clear.

Recall that our definitions of the relative risk, risk difference, and odds ratios in Chap. 3 were stated in terms of the outcome probabilities. This limits their usefulness in retrospective settings such as case-control studies. However, it is a unique property of the odds ratio that it retains its validity as a measure of outcome risk, even for case-control sampling. To demonstrate this for a simple example, Table 5.20 presents odds ratios for the Ille-et-Vilaine study estimated using the `tabodds` procedure in Stata. The first part of the table gives the odds of the binary case-control status indicator `case` compared in two groups defined by the binary indicator `ditob` of moderate to heavy level of smoking (10+ grams/day of tobacco smoked), and the second part gives the corresponding odds ratio comparing moderate-to-heavy level of smoking between cases and controls. The estimated odds ratios are identical. This property does not hold for the risk difference and relative risk.

We can also demonstrate this property directly using the definition of the odds ratio. Table 5.21 presents a hypothetical 2×2 table for a binary outcome and predictor in terms of the frequencies of n individuals in the four possible cross-categorizations (labeled a , b , c , and d). We estimate the outcome probability among individuals with and without the predictor with the proportions $a/(a + c)$ and $b/(b + d)$, respectively, and the corresponding odds of the outcome as

$$\frac{a/(a + c)}{c/(a + c)} \text{ and } \frac{b/(b + d)}{d/(b + d)}. \quad (5.11)$$

The resulting odds ratio is then ad/bc .

Table 5.21 Outcome by predictor status for a case-control study

Outcome	Predictor		
	Yes	No	Total
Yes	a	b	$a + b$
No	c	d	$c + d$
Total	$a + c$	$b + d$	n

Similarly, we can estimate the exposure probability among individuals with and without the outcome as $a/(a + b)$ and $c/(c + d)$, and the corresponding odds as above. It is easy to verify that the odds ratio based on these is also ad/bc . This property of the odds ratios is central to the wide use of case-control studies, and suggests that logistic regression may be applicable as well. The additional fact that the odds ratio approximates the relative risk for rare outcomes (e.g., many forms of cancer) increases its appeal.

Recall that in the logistic regression model, the intercept coefficient β_0 is interpreted as the “baseline” log odds of outcome risk obtained when no predictors are included in the model (or, equivalently, when all predictors take on the value zero). As we have stated above, this quantity cannot be meaningfully estimated from case-control studies. As a result, the intercept coefficient in logistic regression models for case-control data can not be interpreted as providing an estimate of baseline risk in the population from which the sample was drawn. It is a remarkable fact that the logistic model is nonetheless directly applicable to data from case-control studies, and that estimated regression coefficients for included predictors provide valid estimates of log odds ratios, sharing the interpretation from other study types. Note that the logistic is the only binary regression model with this property.

A primary hypothesis underlying the Ille-et-Vilaine study was that alcohol consumption was related to esophageal cancer. Alcohol consumption was measured in average total daily consumption in grams, estimated directly from questionnaire responses on a number of different types of alcoholic beverages. The investigators recognized that age and smoking were potential confounding influences, and should be accounted for in assessing the association between alcohol consumption and cancer risk. (Dietary factors were also considered, but are not discussed here.)

Table 5.22 presents the results of a logistic regression model fit to these data, including a four-level categorization `alcgp` of average daily alcohol consumption and controlling for the dichotomous indicator `ditob` of moderate-to-heavy smoking (introduced above) and `age` (in years) as a continuous predictor. The lowest level of alcohol consumption (0–39 g/day) is taken as the reference category, and the three included indicators represent 40–79, 80–119, and 120+ g/day, respectively. The results indicate a clear increase in cancer risk with increasing alcohol consumption, and that this effect is evident when age and smoking are accounted for.

Estimated odds ratios in Table 5.22 are larger than 1.0, and the associated 95% CIs exclude 1.0, indicating that each of the predictors is associated with statistically

Table 5.22 Logistic model for alcohol consumption and esophageal cancer

```
. logistic case i.alcgp i.ditob age

Logistic regression

Log likelihood = -354.34556
```

Number of obs = 975

LR chi2(5) = 280.80

Prob > chi2 = 0.0000

Pseudo R2 = 0.2838

case	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
alcgp					
2	4.063502	1.024363	5.56	0.000	2.47926 6.66007
3	7.526931	2.138602	7.10	0.000	4.312895 13.13612
4	32.07349	11.58611	9.60	0.000	15.80015 65.10752
1.ditob	7.375744	2.732364	5.39	0.000	3.56842 15.24529
age	1.068417	.0087666	8.07	0.000	1.051372 1.085738

significant increases in risk of esophageal cancer. Further, since esophageal cancer is relatively rare in the general population on which this study was conducted, interpreting the odds ratios as estimated relative risks is approximately correct.

A single summary of the contribution of alcohol consumption to a model including age and smoking can be obtained by fitting the same model excluding the indicators for alcohol, and performing a likelihood ratio test, as shown in Table 5.23. This procedure assumes that the full model including alcohol in Table 5.22 is fit first, and the model log likelihood is stored for future reference as `mod1` (in the second line of the output in Table 5.23). The results indicate a substantial contribution of the categorical summary `alcgp` of alcohol consumption to the overall fit of the model as summarized by the large log LR statistic (128.7). Further analyses might investigate the relationship between alcohol, smoking, and the log odds of cancer risk in more detail, possibly including these variables as continuous measures. We would naturally want to evaluate the linearity assumption implicit in including the variables (and age) in this form as well.

5.3.1 Matched Case-Control Studies

Consider the issues that would arise in designing a case-control study investigating esophageal cancer in a different population than Ille-et-Vilaine, possibly focusing on exposures other than alcohol as potential risk factors: We certainly would like to take into account known confounding factors such as those considered above as part of our design. If there are many such variables, we may be concerned that they will not be well represented in our chosen sample, and/or that analyses accounting for their influence may be overly complex. If we could recruit study subjects accounting for their profiles for these suspected confounders, we might be able to avoid some of these difficulties. This is the rationale for *matching*. We can

Table 5.23 Likelihood ratio test for contribution of `alcgrp`

```

. quietly logistic case i.alcgrp i.ditob age
. est store mod1
. logistic case i.ditob age

Logistic regression                                Number of obs   =          975
                                                    LR chi2(2)      =        152.11
                                                    Prob > chi2     =         0.0000
Log likelihood = -418.68894                        Pseudo R2       =         0.1537

-----+-----
      case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      1.ditob |   9.463852   3.362354    6.33   0.000    4.716825    18.9883
      age     |   1.055568   .0073642    7.75   0.000    1.041232    1.0701
-----+-----

. lrtest mod1

likelihood-ratio test                                LR chi2(3)      =        128.69
(Assumption: . nested in mod1)                     Prob > chi2     =         0.0000

```

build in control for confounding by incorporating knowledge of known confounders into the design of the study. By matching cases with controls that have the same values of these variables, we ensure control for confounding by comparing cases and controls within strata defined by the matching factors. In one of the simplest matched designs, disease cases are paired with controls into *matched sets* having similar values of the matching variables.

Because cases and controls within matched sets are sampled together based on shared values of the matching variables, the structure of the overall sample differs from that of an unmatched study. If we were to try to account for the sampling design via a standard logistic model that accounted for the matched sets with indicator variables, the number of parameters would frequently be too large for reliable estimation. For example, in a matched pair study with 200 matched pairs, as many as 199 parameters would be needed to account for the matching criteria. Clearly another regression approach is called for.

Regression modeling for matched data is based on a modification of the maximum-likelihood estimation approach used for the conventional logistic model (and described in more detail in Sect. 5.6). The *conditional logistic regression model* avoids estimating parameters accounting for the matching via *conditioning*. The parameters for predictors in this model have the log odds ratio interpretation familiar from the standard logistic model. The result is that we can conduct regression analyses exactly as before. However, the variables used in matching are controlled for automatically and not used directly in modeling. The `clogit` command in Stata provides a very convenient way to fit conditional logistic regression models. Most major statistical packages have similar facilities.

Matching is not always a good idea and should never be undertaken lightly. Effective matching (in cases where matching variables are strong confounders) can yield more precise estimates of the disease/exposure relationship. However, in cases where the matching variables do not actually confound the relationship between

the exposure of interest and the outcome, the matching can lead to estimates with decreased precision relative to those obtained from an unmatched study. Further, satisfying matching criteria can be difficult and may result in a loss of cases. Good basic references for statistical analysis of data from matched case-control studies include [Breslow and Day \(1984\)](#) and [Jewell \(2004\)](#).

5.4 Checking Model Assumptions and Fit

Section 4.7 presented a number of techniques for assessing model fit and assumptions for linear regression models. Here, we cover many of the same topics for logistic models. Fortunately, many of the issues and techniques are similar and the methods from linear models apply more or less directly. One simplification of model assessment for binary outcomes is that no checks of distributional assumptions analogous to normally distributed residuals and constant variance are required. This is because the probability distribution for binary outcomes has a simple form that does not include a separate variance parameter. All required parameters are included in the model for the relationship between the log odds of the outcome and the predictors as described in Sects. 5.1 and 5.2. By contrast, construction and interpretation of graphical methods of assessment are more complex because of the nature of residuals from logistic models. We focus here on issues that differ from the approaches discussed in Sect. 4.7. We also note that additional issues arise in assessment of models for repeated or longitudinal binary outcomes such as those introduced in Chap. 7, due to the nature of the assumed dependence between outcomes.

5.4.1 *Linearity*

In Table 5.2, we fit a simple logistic regression model relating CHD risk and age for the WCGS data. In addition to providing a simple description of the relationship, the model makes it easy to compute the log odds associated with an arbitrary value of age. However, as in simple linear regression (Sect. 4.7), the uncritical adoption of the assumption that variables are linearly related to the outcome can lead to biased estimates and incorrect inferences. LOWESS scatterplot smoothing methods (introduced in Chap. 2) offer an exploratory approach to assessing the form of relationship between the log odds of the outcome and age that obviates the need to impose a particular parametric form. In the case of binary outcomes, these average the outcome proportions (or the corresponding log odds) over groups whose size is specified the bandwidth of the selected smoothing method. Figure 5.4 displays the log odds estimated by LOWESS (obtained using the `lowess` command in Stata with the `logit` option) along with the linear logistic fit. The latter is represented by the dashed line, obtained by simply plotting the log odds estimated by the model

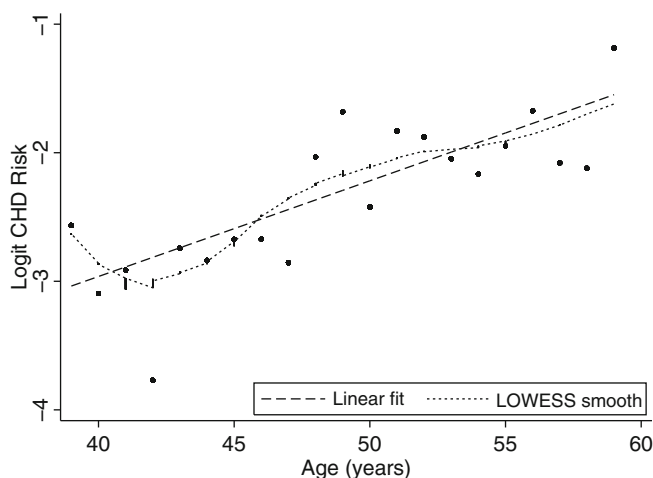


Fig. 5.4 Assessing linearity in the relationship between CHD risk and age

for all the (3,154) individuals in the sample. The smoothed estimate is given by the dotted line. The plotted points are the empirical log odds of the outcome for each of the unique values of age observed in the sample.

Although not conclusive, the results indicate that the linear logistic model fits the data reasonably well. However, the smoothed estimate suggests an initial decrease in the log odds of CHD risk for ages less than 42, followed by a fairly regular increase. The decrease might be due to elevated CHD risk among younger participants. In fact, 7% of the 39-year-olds ($n = 266$) in the study had CHD compared to 4% of the 40-year-old participants. The initial decline in the smoothed estimate is clearly influenced by the observed 2% rate of CHD among the 42-year-olds as well. A reasonable approach to evaluating this further would be to test for particular departures from linearity by adding polynomial terms in age or using restricted cubic splines (similar to the approach described in Sect. 4.10). Table 5.24 displays results from a model including a quadratic term in age (centered to reduce possible collinearity with the linear term). The Wald test statistic clearly indicates that the addition of this term does not afford a statistically significant improvement in the fit over the linear model. We can conclude that the linear model is adequate.

If the role of age in modeling is primarily as an adjustment factor, we would also want to examine whether the assumption of linearity impacts inferences about other predictors. Adoption of the linear form is acceptable if no impacts are seen, but predictions of outcome risk based on the linear model may yield biased results for ages not well represented in the data. Diagnostics for checking linearity in the context of multiple predictor models are somewhat less well developed for logistic models than for linear models. For example, tools like the component plus residual (CPR) plots presented in Sect. 4.7 are not generally available. However, the techniques presented here in combination with LR comparisons of models are

Table 5.24 Logistic model incorporating a quadratic effect of age

```
. logistic chd69 age agesq, coef
```

Logistic regression

Log likelihood = -869.14333

Number of obs = 3154

LR chi2(2) = 42.96

Prob > chi2 = 0.0000

Pseudo R2 = 0.0241

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0769963	.0150015	5.13	0.000	.0475938	.1063987
agesq	-.0005543	.0021066	-0.26	0.792	-.0046831	.0035745
_cons	-6.04301	.678737	-8.90	0.000	-7.37331	-4.71271

usually sufficient to diagnose and correct nonlinearity problems. The increased availability of nonparametric regression approaches for binary regression (discussed briefly in Sect. 5.5) is rapidly expanding the arsenal of available tools in this area.

5.4.2 Outlying and Influential Points

Similar to the definition of residuals for linear regression (in Sect. 4.7), *standardized Pearson residuals* for logistic regression models are based on comparing observed values of the outcome variable with predictions from a fitted model. However, because outcomes in logistic models are binary, the values of these residuals cluster in two groups corresponding to the two values of the outcome. This makes graphical displays of residuals more difficult to interpret than in the linear regression case. An exception occurs when there are relatively few unique covariate patterns in the data (e.g., when predictors are categorical) and residuals and predictions can be grouped.

Figure 5.5 shows standardized Pearson residuals for the model in Table 5.18, plotted against the ordered observation number for the individual subjects. This *index plot* allows observations with unusually large residuals relative to other observations to be identified and investigated as potential outliers. The grouping of residuals based on outcome status is evident from the plot. In this case, although a number of observations have fairly large residuals (i.e., greater than two), none appear to be indicative of outlying observations. A number of other plots based on residuals are possible. In our experience, these are less useful in general than the investigation of influential points discussed in the next paragraph.

Diagnostic techniques for identifying influential observations in logistic regression models are also quite similar in definition and interpretation to their counterparts for linear regression. Most statistical packages that feature logistic regression allow computation of influence statistics that measure how much the estimated coefficients for a fitted model would change if the observation were deleted. Figure 5.6 shows influence statistics (often called DFBETA values) for the model in Table 5.18, plotted against the estimated outcome probabilities.

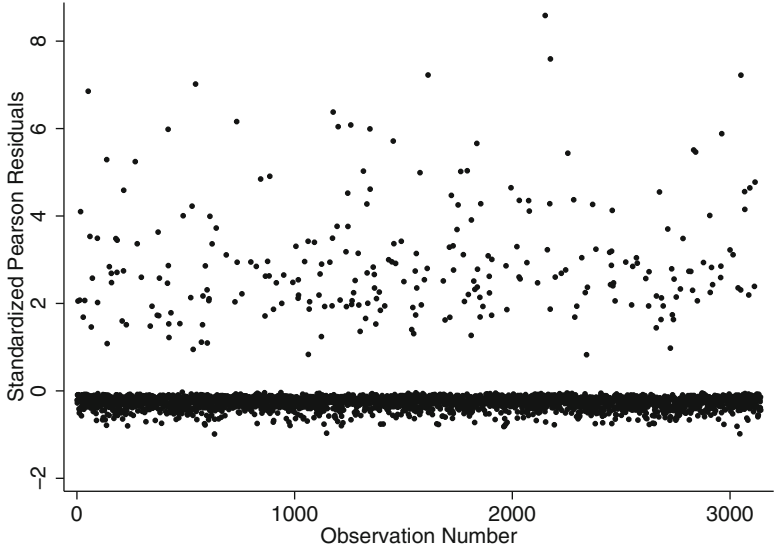


Fig. 5.5 Standardized pearson residuals for logistic model in Table 5.18

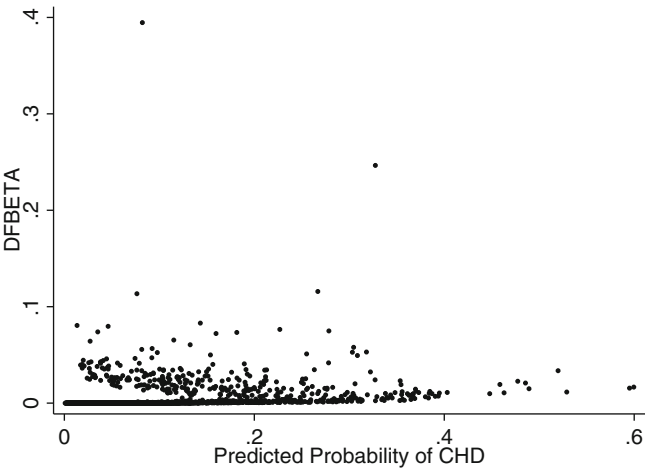


Fig. 5.6 Influence statistics for logistic model in Table 5.18

Two observations appear to have more influence than the rest. The most extreme observation is for an individual who is a nonsmoker with CHD, characterized by below average cholesterol (188) and a very high BMI value (39). Deletion of either observation (or both) resulted in no noticeable changes to model coefficients. Since there is no reason to suspect that any of the data are incorrect, both observations were retained.

Table 5.25 Link test for logistic model in Table 5.18

```
. linktest
```


Logit estimates

Log likelihood = -786.89258

Number of obs = 3141

LR chi2(2) = 200.40

Prob > chi2 = 0.0000

Pseudo R2 = 0.1130

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_hat	.5646788	.306056	1.85	0.065	-.0351799	1.164538
_hatsq	-.1002356	.0688901	-1.46	0.146	-.2352576	.0347865
_cons	-.3983753	.3230497	-1.23	0.218	-1.031541	.2347904

5.4.3 Model Adequacy

The techniques discussed above address potential nonlinearity in the relationship between the log odds of the outcome and the predictor, but implicitly assume that the logistic model is correct. Recall from Sect. 4.7 that transformations of the outcome variable can be used to ensure that the distribution of the errors in a regression model are normally distributed. In a similar way, we can investigate the adequacy of the logistic model.

5.4.3.1 Specification Tests

A simple (and rather crude) approach to evaluating whether a given logistic model provides an adequate description of the data is through the use of a *specification test*. The `linktest` procedure in Stata is an example. Table 5.25 presents the results of applying `linktest` immediately after fitting the model in Table 5.18. This test involves fitting a second model, using the estimated right-hand side (i.e., the linear predictor) from the previously fitted model as a predictor. We would expect that the Wald test result for this predictor (labeled `_hat`) to be statistically significant if the original model provided a reasonable fit. The model fit by `linktest` also includes the square of this predictor (labeled `_hatsq`). The Wald test for inclusion of the latter variable is used to evaluate the hypothesis that the model is adequate; that is, the inclusion of the squared linear predictor should not improve prediction if the original model was adequate. Rejection indicates that the model is inadequate, and that an alternative binary regression model should be considered. Inadequacy may reflect the fact that even though important predictors are included and modeled correctly, the logistic model is not an appropriate representation of the relationship between outcome and predictors. It may also indicate that important predictors have been omitted, or are represented incorrectly in the model. The test can not distinguish between these two alternative explanations. It also does not suggest what alternate model form might be preferable.

In the example, the P -value for the Wald test for the predictor `_hat_sq` does not provide strong evidence of inadequacy of the logistic model. However, the fact that the P -value for the predictor `_hat` in Table 5.25 is also not very small provides some indication that the overall fit may not be very good. (This is consistent with the large residuals noted in Sect. 5.4.2.)

Possible alternatives to the logistic model were discussed in Sect. 5.1, and will be covered in more detail in Sect. 5.5. Because these typically involve the use of specialized methods of estimation and result in coefficients with different interpretations, they are rarely used in practice. Fortunately, differences between results from alternative models are often small, and the logistic model applies in a very wide range of problems involving binary outcomes. Problems with fit can frequently be addressed using judicious selection and appropriate transformations of predictors.

5.4.3.2 Goodness of Fit Tests

Another approach to assessing model adequacy is provided by *goodness of fit* tests. The *Hosmer–Lemeshow* test is an example of this approach applicable to binary regression models such as the logistic. The test works by forming groups of the ordered, estimated outcome probabilities (e.g., ten equal-size groups based on deciles of the distribution of the outcome probabilities) and evaluating the concordance of the expected outcome frequencies in these groups with their empirical counterparts. The underlying hypothesis is that the estimated and observed frequencies agree. Thus, a statistically significant finding (i.e., rejection) indicates lack of fit. A nonsignificant finding rules out gross lack of fit.

Table 5.26 displays results of the Hosmer–Lemeshow test for the regression model fitted in Table 5.18. The `table` option requests that the observed and expected frequencies of the binary outcome (ones and zeros) for the requested groups be printed as well. The nonsignificant results do not indicate evidence for gross lack of fit. Increasing the number of groups to 20 yields a larger P -value (0.35), illustrating the sensitivity of the test to the number of groups chosen, and raising the possibility that judicious choice of group size may allow an investigator to choose the number of groups resulting in the most favorable P -value. To avoid this subjectivity, ten groups are generally recommended.

The Hosmer–Lemeshow test has a number of serious limitations. First, it is not sensitive to a number of sources of lack of fit such as misspecification of the model, and lacks power in these situations as a consequence. Further, the results of the test depend on the number of groups specified as well as the distribution of predictor values within these groups. Finally, the test can be very sensitive to fairly small fit discrepancies in large samples. Thus, a significant result may not signal a serious fit problem in such cases. Similarly, failure to find a statistically significant result does not necessarily mean that the model fits the data well. This test is most useful as a very crude way to screen for fit problems, and should not be taken as a definitive diagnostic of a “good” fit. Use in conjunction with a specification test (such as

Table 5.26 Hosmer–Lemeshow goodness of fit test

. lfit, group(10) table

Logistic model for chd69, goodness of fit test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0160	1	3.3	314	311.7	315
2	0.0251	6	6.5	308	307.5	314
3	0.0344	11	9.3	303	304.7	314
4	0.0450	12	12.5	302	301.5	314
5	0.0575	18	16.0	296	298.0	314
6	0.0728	10	20.4	304	293.6	314
7	0.0963	28	26.5	286	287.5	314
8	0.1268	44	34.7	270	279.3	314
9	0.1791	50	46.7	264	267.3	314
10	0.5996	76	80.3	238	233.7	314

number of observations = 3141
number of groups = 10
Hosmer--Lemeshow chi2(8) = 11.36
Prob > chi2 = 0.1824

the one described above) may provide a bit broader screen to detect problems. However, results of either approach should not be relied on to guarantee model fit in the absence of supplementary investigations, including diagnostic assessment of residuals and influential observations.

5.4.4 Technical Issues in Logistic Model Fitting

In some cases, measures of association for binary outcomes such as odds ratios and relative risks take on the value zero, or are infinite. This happens when subgroups formed by the predictors are homogeneous with respect to outcome status. This translates to estimation problems in regression models, where parameters are typically represented as the logarithm of the underlying association measures.

Table 5.27 presents an example from the WCGS study using a four-level categorization of cholesterol level (0–150, 151–200, 201–250, and 251+) as a predictor of CHD outcome. Note the missing odds ratio estimates and the note explaining that “0.cholc dropped and 89 obs not used.” Examination of the data reveals that there are no observed CHD cases among the 89 individuals with cholesterol in the default reference category (0–150 mg/dL). Because the odds of CHD are zero for this group, it is not possible to estimate valid odds ratios for the other categories. Choosing an alternate reference group allows valid estimates to be made. However, the odds ratio of zero for the lowest category still causes a fitting issue: the log odds ratio is infinite, and the parameter can not be estimated.

The problem raised in this example can be addressed by choosing a different categorization of cholesterol. However, this approach changes the interpretation of

Table 5.27 Logistic model for CHD and categorized cholesterol level

```
. logistic chd69 i.cholc
note: 0.cholc != 0 predicts failure perfectly
      0.cholc dropped and 89 obs not used

note: 3.cholc omitted because of collinearity

Logistic regression
Log likelihood = -855.50635
```

Number of obs = 3053

LR chi2(2) = 52.77

Prob > chi2 = 0.0000

Pseudo R2 = 0.0299

chd69	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
cholc						
0	(empty)					
1	.2884527	.0574556	-6.24	0.000	.1952214	.4262082
2	.4514053	.0642673	-5.59	0.000	.3414914	.5966966
3	(omitted)					

the categorized variable, and will not work in all cases. In small samples, frequently no amount of regrouping or recategorizing will eliminate these issues. In these situations, exact logistic regression methods (discussed in Sect. 5.5.4) should be considered. When exact methods are not computationally feasible, the penalized maximum likelihood approach proposed by Firth (1993) provides another possible alternative. This is available for Stata in a downloadable, user-defined module entitled *firthlogit*. We recommend that a statistician be consulted to diagnose the exact nature of the problem and suggest appropriate solutions.

Another issue to consider in fitting logistic regression models with multiple predictors is deciding how many predictors is appropriate. Fitting too many predictors can result in biased estimates and incorrect inferences. The severity of these problems is directly related to the sample size, the number of observed outcomes, and the distribution of outcomes over the included predictors. Chapter 10 discusses model building strategies in detail, and Sect. 10.4.2 provides guidelines for selection of appropriate number of predictors.

5.5 Alternative Strategies for Binary Outcomes

A review of current clinical and epidemiological research studies involving binary outcomes will reveal that the overwhelming majority of regression analyses are based on the logistic model. In some instances, specific knowledge about a disease-exposure relationship may suggest a different model. Alternatively, it may be desirable to summarize observed associations using measures such as the relative risk or risk difference in preference to the odds ratio. Because the logistic model yields only the latter, there are situations where alternative regression approaches may be preferred. Finally, diagnostic evaluations may lead to the conclusion that

the logistic model is simply not right for a particular data set. In this section, we review some examples of alternative approaches to binary regression. We also briefly discuss models for categorical outcomes with more than two levels.

5.5.1 Infectious Disease Transmission Models

Recall the CDC transmission study data discussed in Sect. 3.4 (O'Brien et al. 1994). The goal of this study was to investigate risk factors for sexual transmission of HIV in susceptible female partners of previously infected males. Although the outcomes were restricted to prevalent HIV serostatus measured at enrollment, the infection dates of the male partners were approximately known from transfusion records. In addition, self-reported information on number of unprotected sexual contacts was also collected. These data pertain to contacts that occurred between the time of infection of the male partner and the time of enrollment. (Note that monogamy was an eligibility criterion, to reduce the possibility of infection from other sources.)

Unlike many chronic diseases, the mechanism of acquisition of many infectious diseases is well understood. In these cases, simple probabilistic *transmission models* linking outcomes with exposures are frequently used to quantify infection risk. One of the most basic such models links the cumulative probability of escaping infection following a series of exposed contacts. The model assumes that each contact carries an identical risk λ of infection, and that outcomes of successive contacts are independent. Under these assumptions, the chance of escaping infection following k contacts is

$$(1 - \lambda)^k,$$

with the complementary probability of being infected following k contacts given by

$$P(k) = 1 - (1 - \lambda)^k.$$

This model corresponds well to the observed data from the CDC study: each female partner can be characterized by the binary infection status and the reported number of exposed contacts k (the predictor), with the outcome probability given above. This suggests that a binary regression approach linking these two variables would be ideal for estimating the per-contact transmission probability λ . Unfortunately, the logistic model does not provide a direct estimate. By contrast, an alternative transformation of $P(k)$, known as the complementary log–log, provides a model with a more appealing structure:

$$\log\{-\log[1 - P(k)]\} = \log[-\log(1 - \lambda)] + \log(k). \quad (5.12)$$

This model is similar to the familiar linear model

$$\log\{-\log[1 - P(x)]\} = \beta_0 + \beta_1 x, \quad (5.13)$$

Table 5.28 Complementary log–log regression model for per-contact risk

```
. glm hivp, family(binomial) link(cloglog) offset(logcontacts)
```

Generalized linear models		No. of obs	=	31
Optimization	: ML: Newton-Raphson	Residual df	=	30
		Scale parameter	=	1
Deviance	= 40.8340195	(1/df) Deviance	=	1.361134
Pearson	= 84.90572493	(1/df) Pearson	=	2.830191
Variance function: V(u) = u*(1-u)		[Bernoulli]		
Link function : g(u) = ln(-ln(1-u))		[Complementary log-log]		
Standard errors : OIM				
Log likelihood = -20.41700975		AIC	=	1.381743
BIC = -62.18559663				

hivp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_cons	-7.033126	.3803284	-18.49	0.000	-7.778556	-6.287696
logcontacts	(offset)					

```
. bootstrap "glm hivp, family(binomial) link(cloglog) offset(logcontacts)"
_b_se, reps(1000)
```

```
command:      glm hivp, family(binomial) link(cloglog) offset(logcontacts)
statistics:    b_cons      = [hivp]_b[_cons]
```

Bootstrap statistics		Number of obs	=	31
		Replications	=	1000

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]	
b_cons	1000	-7.033126	-.0629388	1.163788	-8.216878	-6.296359

where the intercept coefficient $\beta_0 = \log[-\log(1-\lambda)]$, but includes the predictor $x = \log(k)$ as a fixed *offset*, with corresponding coefficient $\beta_1 = 1$ as specified by model (5.12). Predictors with fixed coefficients are referred to as *offsets*, and can be easily accommodated by standard statistical software packages. (Part of the model evaluation procedure in this case may include checking whether this is reasonable in terms of fit.) Similar to the logistic model, an inverse transformation allows us to represent this model on the probability scale as follows:

$$P(x) = 1 - \exp[-\exp(\beta_0 + \beta_1 x)],$$

(5.14)

Table 5.28 shows the results of fitting model (5.12) using the generalized linear model estimation program `glm` in Stata, which we explain in greater detail in Chap. 8. Note that the logarithm of the number of contacts `logcontacts` appears as an offset, and no coefficient for this predictor was estimated.

An additional calculation inverting the complementary log–log transform of the intercept `_cons` provides the estimate of λ :

$$\lambda = 1 - \exp[-\exp(-7.033)] = 0.0009.$$

The approximate 95% CI (0.0004, 0.0019) can be obtained via a similar calculation applied to confidence limits given in the regression output. Because of the small sample size ($n = 31$), the approximate CIs may not be reliable. For comparison, Table 5.28 also gives bias-corrected 95% bootstrap CIs (calculated using 1,000 bootstrap samples) for the same model. The bias-corrected CI (0.0003, 0.0018) for the parameter λ can be obtained from the interval for the intercept coefficient β_0 (represented by `b_cons` in the table) via the calculation used for the approximate interval. The lower bound of this interval is only slightly more conservative than the approximate interval, but otherwise they are remarkably similar. The bootstrap interval should still be considered a better summary of uncertainty about λ .

Clearly, model (5.12) is very simple, and a number of the underlying assumptions are questionable (e.g., that the per-contact risk λ is constant). However, it is a useful “null” model to which more complex alternatives may be compared. Further, the parameter λ is an important ingredient in more complex mathematical epidemic models. This model is also interesting because it is an example of a *proportional hazards model*. These arise frequently in studies where controlling for duration of follow-up is an important consideration in data analyses, and are the subject of the next chapter. Finally, model (5.13) and the conventional logistic model are examples of the family of GLMs that includes most of the regression models considered in this book.

5.5.2 Pooled Logistic Regression

The MIRA study was a randomized trial designed to investigate the effectiveness of diaphragms as a means of prevention of sexual transmission of HIV in women in sub-Saharan Africa (Padian et al. 2007). Here, we consider data on 1,000 randomly selected individuals participating in a substudy investigating risk factors for infection with herpes simplex virus type 2 (HSV-2), conducted among women testing negative for infection at enrollment (de Bruyn et al. 2011). The study design is characterized by visits at three month intervals following enrollment, with infection outcome and predictor information collected at each. HSV-2 infection can occur at most once, so individual outcomes can be summarized by a binary indicator of whether or not infection has occurred during follow-up. Also, the interval of infection occurrence is informative about the possible time of infection. For example, individuals at higher risk for infection at any time during follow-up may also tend to be infected earlier. Direct application of the logistic model described in previous sections would not account for this, or the fact that multiple observations of both predictors and outcomes are available. Methods for regression analysis of survival outcomes (as discussed in Chap. 6) based on precisely measured times of outcome occurrence also don’t apply unless we make an assumption about the actual occurrence times of infections (e.g., the midpoint of the interval between visits). *Pooled logistic regression* provides a hybrid approach that avoids such assumptions, and also allows the information on outcome occurrence collected in multiple study visits to be used appropriately.

Table 5.29 Example data from MIRA study

```
. list id mos hsv2 age stihx newparts if id==2 | id==54
```

	id	mos	hsv2	agecat	stihx	newparts
4.	2	3	0	1	0	0
5.	2	6	0	1	0	0
6.	2	9	0	1	0	0
7.	2	12	0	1	0	1
8.	2	15	0	1	0	0
9.	2	18	0	1	0	0
10.	2	21	0	1	0	0
11.	2	24	0	1	0	0
409.	54	3	0	2	0	1
410.	54	6	0	2	0	1
411.	54	9	1	2	0	0

Table 5.29 illustrates observations of key study variables for two selected participants from the MIRA study. In addition to indicators of outcome occurrence `hsv2`, each individual contributes observations of predictors that are fixed (`agecat`, a categorical representation of age at enrollment; `stihx`, a binary indicator of self-reported history of prior sexually transmitted infections) or time varying (`newparts`, a binary indicator of self report of recent new sexual partners). In addition to outcome and predictor values, the follow-up duration (`mos`) is recorded as the number of months elapsed since enrollment. The first individual remained uninfected for all study visits, and provides measures of the HSV-2 outcome and fixed and time-varying predictors for each interval. The actual time of infection is said to be *right censored* at the time of the last visit. The second individual was first observed to be infected at the fifth visit (12 months), and was removed from observation for treatment thereafter. The time of infection in this case is censored into the interval between the fourth and fifth visits. This data structure is typical for application of the pooled logistic model, and also shares features with survival data that are the subject of the next chapter. We would like the analysis to assess the association between predictors and outcome occurrence, and also account for the duration of follow-up.

Table 5.30 displays the results of fitting a logistic regression model to the data partially shown in Table 5.29. Because we want to make as few assumptions as possible about how infection risk varies with duration of follow-up, the model uses a restricted cubic spline (discussed in Chap. 4) with three knots to account for the effects of time. The estimated odds ratios for the spline predictors (`sp11` and `sp12`), together with the intercept odds (not shown) can be regarded as the “baseline” infection odds that applies to individuals with the additional predictors of interest set to zero. The significant result for the `testparm` command, which evaluates the Wald test for the hypothesis that both spline coefficients equal zero, indicates that accounting for time variation via a spline results in a significantly improved fit to the data relative to a model including only an intercept term. The

Table 5.30 Pooled logistic regression model for MIRA study example

```
. logistic hsv2 spl* i.agecat i.stihx newparts
```

Logistic regression		Number of obs	=	6069
		LR chi2(6)	=	33.69
		Prob > chi2	=	0.0000
Log likelihood = -509.18109		Pseudo R2	=	0.0320

	hsv2	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	spl1	.9256021	.0379368	-1.89	0.059	.8541555 1.003025
	spl2	1.035493	.0556198	0.65	0.516	.9320218 1.15045
agecat						
	2	.6090384	.1369677	-2.20	0.027	.3919372 .946396
	3	.5522162	.2213153	-1.48	0.138	.2517489 1.211297
1.stihx		1.92962	.4938946	2.57	0.010	1.168431 3.186695
newparts		1.826313	.4372661	2.52	0.012	1.142288 2.919945

```
. testparm spl*

( 1) [hsv2]spl1 = 0
( 2) [hsv2]spl2 = 0

      chi2( 2) =    10.59
Prob > chi2 =    0.0050
```

odds ratios for the additional predictors are interpreted similarly to the conventional logistic model studied in previous sections. For example, increased age is associated with decreased odds of infection relative to the youngest age group. Also, report of a recent new partner is associated with an approximate doubling of infection odds. As discussed previously, the form of the model specifies that the predictors act to increase or decrease baseline infection odds by fixed increments. This is an example of a *proportional odds model*. For outcomes that are rare, the odds ratios closely approximate *relative hazards* estimated from the closely related *proportional hazards model* that is discussed in the next chapter.

Although the pooled logistic regression model is widely applicable and easy to fit, it suffers from several disadvantages relative to the regression methods for survival data that are the subject of Chap. 6: First, it requires explicit modeling of the effects of time in the analysis, a feature not shared with the Cox proportional hazards regression model. Second, because event times are not known precisely, the causal links between time-varying predictors and event occurrence are less clear than in settings where such information is completely observed. Finally, in situations with missed and/or irregularly spaced intervals the estimates are susceptible to bias because of the need to make assumptions about the behavior of predictors and outcomes in unobserved periods of follow-up. Despite these issues, this approach is becoming increasingly popular for longitudinal data of the type described here, and, as discussed in Chap. 9, is commonly used in applications of regression to causal inference. More information about the approach, including a comparison to the Cox proportional hazards model is included in [D’Agostino et al. \(1990\)](#).

5.5.3 Regression Models Based on Risk Differences and Relative Risks

A recent study of prevalent human T-cell leukemia/lymphoma virus (HTLV) infection in infants born to mothers in the United Kingdom identified a number of factors associated with infection, including the parent's country of birth and ethnicity of the mother (Ades et al. 2000). The authors found that a regression model based on risk differences provided a better fit to the data than the logistic model, and reported their results accordingly.

Recall the linear regression model defined in (5.1) that relates risk for a binary outcome to a single predictor x :

$$P(x) = \beta_0 + \beta_1 x.$$

As noted in Sect. 5.1, the coefficient β_1 measures the risk difference associated with a unit increase in x . This model is often referred to as the “additive risk model” because the effect of any unit increase in the predictor x is to add an increment β_1 to the outcome risk. This was the model employed in the HTLV example. Although it provides a valid alternative to logistic regression, it is important to keep in mind the potential problems with fitting and interpretation (raised in Sect. 5.1).

As discussed in Sect. 3.4, the odds ratio is known to approximate the relative risk in the rare outcome setting. Consequently, odds ratios are frequently reported as relative risks in research findings. Unfortunately, this practice is not limited to rare outcomes, and has been the subject of considerable debate in the research literature (Holcomb et al. 2001). This has led many investigators to advocate that regression models based on the relative risk be used in preference to the logistic model (other than in case-control designs where standard regression approaches other than the logistic model do not directly apply). This is possible using the following regression model:

$$\log [P(x)] = \beta_0 + \beta_1 x. \quad (5.15)$$

This is the *log linear* model discussed in Sect. 5.1. The regression coefficient β_1 has the interpretation of the logarithm of the relative risk associated with a unit increase in x . Analogous to the procedure for obtaining odds ratios from logistic models, exponentiated coefficients yield relative risk estimates in this case. Although this model can be fitted with many standard software packages, numerical difficulties may arise because of the constraint that the sum of terms on the right-hand side must be no greater than zero for the results to make sense (due to the constraint that the outcome probability $P(x)$ must lie in the interval $[0, 1]$). In such cases, treating the observed binary responses as if they were distributed according to the Poisson distribution, and using estimation methods for GLMs (Chap. 8) generally yields very similar log relative risk estimates. If robust variances are used to estimate variability, the resulting inferences have been shown to yield results very similar to the conventional binomial estimation procedure and to avoid the associated

Table 5.31 Generalized linear models for CHD risk (P) and age (x)

Model	β_1 (95% CI)	Log-likelihood	$P(55)$
$P(x)$	0.005(0.004, 0.007)	−869.96	0.129
$\log[P(x)] - \text{Binomial}$	0.067(0.047, 0.087)	−869.24	0.136
$\log[P(x)] - \text{Poisson}$	0.067(0.048, 0.087)	−881.86	0.136
$\log\{-\log[1 - P(x)]\}$	0.071(0.050, 0.092)	−869.21	0.136
$\log\{P(x)/[1 - P(x)]\}$	0.074(0.052, 0.097)	−869.18	0.136

convergence problems with the latter (Zou 2004; Yelland et al. 2011). The Poisson approach is generally recommended in cases where relative risk estimates are desired and the log binomial model fails to converge.

Alternative approaches for obtaining adjusted relative risks from odds ratios estimated using logistic regression have been proposed in the literature (Zhang and Yu 1998). These are based on simple transformations of the estimated coefficients similar to the illustrative calculations demonstrated in Sect. 5.1.1. Unfortunately, such calculations can produce incorrect estimates for models including multiple predictors and should be avoided in favor of fitting appropriately defined regression models as described above (McNutt et al. 2003).

Table 5.31 presents the results of fitting five alternative GLMs for the relationship between CHD and age using the WCGS data. (Results were obtained with the Stata GLMs procedure `glm`, also applied in Table 5.28.) These correspond to the binomial regression models considered in this section (i.e., (5.1), (5.2), (5.13), and (5.15)) and the alternative Poisson regression approach. Results for the intercept parameter β_0 are similar. Note that the estimated regression coefficients cannot be directly compared because the models are based on different representations of the outcome. However, since all of them are based on the same number of parameters, comparison of the likelihoods provides a cursory look at how well they describe the data in relative terms. Although the likelihood for the logistic model is slightly larger, there is very little overall difference between the models. Similarly, the estimated coefficients for the log, complementary log–log, and logit models are remarkably similar. (The coefficients for the risk difference model differ because the outcome is modeled without transformation.) Finally, the estimated probabilities for a 55-year-old individual ($P(55)$) are also quite similar. Based on these results, there would be no particular reason to prefer any alternatives over the logistic model.

The results in Table 5.31 illustrate that a variety of models other than the logistic may be appropriate for a given problem. We note that additional binary regression models also exist that are useful in other contexts. For example, the *probit model* is used in the context of instrumental variable methods for binary outcomes in Sect. 9.7. However, given the ease of interpretation, wide use, and software availability of the logistic model, it is by far the most common choice in practice. In general, we advocate fitting the logistic model unless another model is preferable on scientific grounds. Lack of fit can often be dealt with via the techniques discussed in

Sect. 4.7, obviating the need to investigate alternative model formulations. Finally, note that the approaches discussed here are not directly applicable to data from case-control studies (Scott and Wild 1997).

5.5.4 Exact Logistic Regression

Recall the HIV transmission example considered in Tables 3.6 and 5.28. The dataset contains binary outcomes for 31 monogamous female sexual partners of males previously infected with HIV. With so few observations, the reliability of statistical inference relying on conventional statistical procedures such as the Wald and χ^2 test is questionable. The Fisher's exact test, discussed in Sect. 3.4 provides a useful alternative for outcome–predictor comparisons addressable using a two-by-two table. However, this approach limits inference to problems involving a single categorical predictor. The *exact logistic regression* model allows exact inferences to be applied in the regression setting, including models with continuous predictors.

In the example presented in Table 3.6, interest focuses on the possible association between presence of an AIDS diagnosis in the male partner and transmission to the female partner. In Table 5.28, we considered a specialized model linking the degree of sexual contact measured by the logarithm number of contacts reported by each partnership to transmission risk. In Table 5.32, we show the results of fitting both standard and exact logistic regression models to these data using Stata, including both the logarithm of the number of contacts and the indicator of AIDS diagnosis as predictors. Although no exact procedure is available to fit the GLM considered in Table 5.32, there is still interest in examining whether the observed effect of AIDS diagnosis from Table 3.6 is influenced by controlling for degree of exposure to infection. The estimated odds ratios from the two models are comparable. However, the degree of precision for the estimated effect of AIDS appears to be overstated in the standard logistic model. Note that in place of the Wald test results, the exact logistic reports columns labeled `Suff.` and `2*Pr(Suff.)`. These are based on *sufficient statistics* for each predictor in the model conditional on the values of the other predictor(s). Exact inference is based directly on these conditional distributions. If interest focuses on a particular predictor in a model (e.g., AIDS) it is possible to restrict inference to that variable, resulting in some computational savings.

Computational procedures for exact logistic regression are intensive, and frequently it will be unfeasible to fit models with multiple continuous covariates, even for datasets as small as considered in the above example. For this reason, the exact approach is recommended for small samples (typically less than 100), especially when *P*-values from standard asymptotic approaches such as the Wald test are in the range of plausible significance. Exact logistic regression can also be useful in situations where standard models fail to yield valid estimates, such as those discussed in Sect. 5.4.4. Finally, note that most of the procedures discussed for model assessment and postestimation inference that are applicable for the standard logistic model are not available for exact logistic regression.

Table 5.32 Conventional and exact logistic regression models for transmission risk in female partner for the CDC example

```
. logistic hivp logcontacts i.aids
```

Logistic regression

Log likelihood = -14.368496

Number of obs = 31

LR chi2(2) = 4.38

Prob > chi2 = 0.1119

Pseudo R2 = 0.1323

hivp	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
logcontacts	1.183255	.3653485	0.54	0.586	.6460355	2.167206
1.aids	8.091292	8.64643	1.96	0.050	.9963591	65.70824

```
. exlogistic hivp logcontacts aids
```

Exact logistic regression

Number of obs = 31

Model score = 4.855402

Pr >= score = 0.0813

hivp	Odds Ratio	Suff.	2*Pr(Suff.)	[95% Conf. Interval]	
logcontacts	1.169055	35.86	0.6319	.6776411	2.229598
aids	8.085458	3	0.1547	.5835864	492.6288

5.5.5 Nonparametric Binary Regression

The examples of alternative techniques for binary regression considered above represent only a small subset of the available possibilities for estimating the relationship between a binary outcome and a predictor variable. The goal of *nonparametric regression* methods is to provide estimates of this relationship based on minimal assumptions about its form.

Recall the assessment of linearity for the logistic model for the relationship between CHD and age in the WCGS data in Sect. 5.4.1. The smoothed LOWESS estimate displayed in Fig. 5.4 is an example of a nonparametric logistic regression model for this relationship. Although the assumption that the predictor is related to the disease outcome in an additive fashion via the log odds is retained, this technique allowed us to relax the assumption that the relationship is linear by assuming only that the change in CHD risk with age has a certain degree of smoothness. This can prove very useful in exploring the form of the relationship between outcome and predictor, but does not yield readily interpretable parameter estimates or generalize easily to models including more than one predictor. The class of *generalized additive models* provide an extension to the LOWESS technique, allowing multiple predictors to be fit simultaneously, each of which can be represented as a smooth function (Hastie and Tibshirani 1999). Although very useful in evaluating outcome–predictor relationships, these models are frequently difficult to fit and interpret.

Methods for significance testing, CIs, and model evaluation are less well developed for nonparametric alternatives than for conventional logistic regression. In addition, decisions about degree of smoothness and interpretation of resulting estimates is often very complex. Finally, practical implementations of nonparametric binary regression that handle multiple predictors are not widely available in standard statistical packages. For these reasons, we recommend that flexible parametric approaches be used in accounting for nonlinearities in the relationship between predictor and outcome, and that nonparametric alternatives be used primarily for exploratory purposes.

Classification trees (Breiman et al. 1984) are another popular approach to nonparametric binary regression. As discussed in Sect. 5.2.5, these lack the linear and additive structure shared by other approaches, and have been used to develop prediction tools for using measured characteristics to correctly distinguish binary outcomes. However, classification trees can also be used to explore complex relationships between multiple predictors and a binary response. Because they do not yield estimates of association parameters, interpretation of the contribution of individual predictors to the outcome risk is complex. However, like the nonparametric regression approaches discussed above, they are very useful tools in exploratory analyses and can be very helpful in discovering and interpreting interaction.

5.5.6 More Than Two Outcome Levels

Research studies frequently yield outcomes that have multiple categories. (See Chap. 2 for definitions of categorical variable types.) Consider the back pain example introduced in Sect. 1.1, where pain intensity was measured on an ordered, ten-point scale. In addition to the *ordinal* categorical outcome just considered, *nominal* categorical outcome measures are also commonplace in clinical research. For example, the outcome in a study of cancer outcomes by cell type is a nominal categorical variable. Both type of outcomes can be investigated using contingency table methods. The limitations of these when multiple predictors are involved are clear. For certain questions, considering a binary representation might also be reasonable. For example, to investigate factors that distinguish patients suffering from severe pain from all others in the pain example. In this case, logistic regression is an appropriate tool to consider. However, there is clearly information lost in reducing ten levels down to two. In the remainder of this section we briefly review regression methods for nominal and ordinal categorical outcomes.

5.5.6.1 Ordinal Categorical Outcomes

The *proportional odds model* is a commonly used generalization of the logistic model that accommodates a multilevel categorical response with ordered categories. Rather than modeling the probability of response in a particular category, this model

is based on the cumulative probability that the response is not greater than a chosen category. The dependence of this response on predictors is identical to the form of the logistic model. For the back pain example, (assuming a ten-level response and a single predictor x), the form of this model for a response probability of severity no greater than 5 is given by

$$\log \left[\frac{\Pr(y \leq 5)}{\Pr(y > 5)} \right] = \alpha_5 - \beta x.$$

A similar expression applies to all ten levels of the response. (We assume that the levels of the response are coded 1, 2, ..., 10.)

Note that the intercept parameter α_5 is unique to this response level, and represents the probability of a response of no more than 5 among individuals with $x = 0$. Because the response is expressed as a cumulative probability, the intercept coefficients are constrained as $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_{10}$. The coefficient β is interpreted as the log odds ratio associated with a unit increase in x , assumed to be constant across response levels. (i.e., response levels are parallel, each with slope β .) This assumption amounts to a strong restriction on the effect of the predictor on the response, and needs to be validated.

Note that there are many alternatives to the proportional odds model, including the *continuation ratio model*. We refer the reader to the references provided below for additional information on these.

5.5.6.2 Nominal Categorical Outcomes

When there is no natural ordering implicit in a categorical response, or when the assumptions implicit in the models above do not apply to an ordinal outcome, the *multinomial logistic* model (also known as the polytomous logistic model) can be used for regression analyses. For a single predictor x , the model specifies that each response level follows a logistic regression model for x , with a selected level specified as the reference. The regression coefficients for each level are unique; so for the pain example the model would include nine intercept and slope coefficients. For level 5, and specifying the first level as the reference category, the model would take the form

$$\log \left[\frac{\Pr(y = 5)}{\Pr(y = 1)} \right] = \alpha_5 + \beta_5 x.$$

Notice that when there are more than two outcome levels, the two levels specified in the model are not binary alternatives. The outcome then represents a log relative risk rather than a log odds. Thus, the coefficient β_5 represents the change in the log relative risk for level 5 (relative to the reference level 1) associated with a unit increase in x . The exponentiated value of this coefficient is interpreted as a *relative risk ratio* rather than an odds ratio. Because this model does not involve the restrictions implicit in the proportional odds model, it is an attractive alternative when the

proportional odds assumption is not satisfied. However, because of the potentially large number of parameters and the flexibility of choice for the reference group, the multinomial logistic model can be challenging to interpret.

The models outlined here represent a few of those available for analyzing categorical responses. For further information on these and other models, including examples and a description of available software resources, see [Ananth and Kleinbaum \(1997\)](#) and [Greenland \(1994\)](#).

5.6 Likelihood

One of the common themes uniting methods presented in this book is the principle of using observed data to estimate unknown quantities of interest. The majority of the methods presented are regression models relating outcome and predictor variables measured on a sample of individuals. The principal unknown quantities in the models are the regression parameters. Once these are estimated, inferences can be made about the true values of these parameters and related quantities of interest such as predicted outcomes. All available information about the parameters is contained in the observed data. A standard approach to estimating parameters in models like the ones covered here is known as *maximum likelihood estimation*. Although not required for applications, a basic understanding of this topic helps in unifying the concepts underlying estimation and inference in most of the regression models covered in this book. Here, we provide a brief discussion of some of the key ideas in the binary regression context.

The *likelihood* associated with a set of independent observations of an outcome is just the product of their respective probabilities of occurrence under the assumed model relating outcomes to predictors. Because this represents the joint probability of observing all of the outcomes in the sample, the likelihood can then be interpreted as a measure of support provided for the model by the data. The maximum-likelihood estimate of the parameter(s) is the value for the parameter(s) that yields the maximum value of the likelihood for the observed data.

To take a very simple example from the binary outcome context, consider the problem of estimating the prevalence of HIV for the sample of 31 female partners of previously infected males from the CDC transmission study considered in the examples presented above and in Sect. 3.4. The assumed model is that the actual prevalence in the target population is represented by a constant that we can symbolize by P (similar to the definition introduced earlier in this chapter). We can think of P as the probability that a randomly sampled individual will test positive. The corresponding probability of observing a negative is $1 - P$. However, P is unknown. The observed data consist of the 31 indicators of HIV status, and the likelihood, as defined above, is just the product of the individual outcome probabilities:

$$P^7 \times (1 - P)^{23}.$$

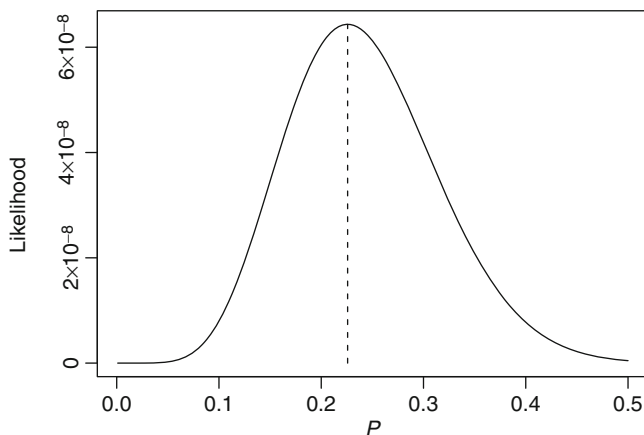


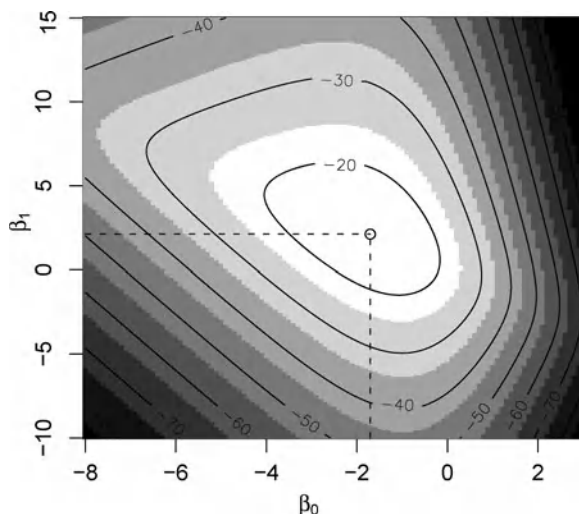
Fig. 5.7 Likelihood function for HIV prevalence

The likelihood is formed as the product of the individual outcome probabilities because these are independent events. It is a function of the unknown constant P , with the observed infection indicators providing the number of positive and negative individuals. Figure 5.7 presents a plot of this function for a range of values for P . The maximum-likelihood estimator of P is just the value of P that maximizes the likelihood function. This value is indicated in the figure. The maximum can be found easily in this example using calculus. Not surprisingly, it corresponds exactly to the intuitive estimate of the actual prevalence of HIV-positive individuals in the sample of 31: Because there are seven such individuals in the sample, the estimated prevalence is 0.226. For more complicated models (e.g., regression models with multiple predictors) computing the maximum typically involves iterative calculations on a computer.

Likelihood functions for binary regression models are defined following the procedure used above, but the outcome probability P for each individual is replaced with the form defined by the logistic model (5.2). To take another example from the CDC study, consider a regression model relating HIV status of the female partners to a binary indicator of presence of an AIDS diagnosis in the male. (This example was already considered in Sect. 3.4.) Following our conventional notation, we will represent the outcome as Y and the predictor as x . The observed data now include both Y and the binary predictor x for each individual in the sample. The likelihood takes exactly the same form as in the last example, except the constant P is replaced with the expression for the logistic model, substituting in each individual's value of the predictor (i.e., x_i for the i_{th} individual):

$$\prod_{i=1}^{31} \left[\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{Y_i} \times \left[1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{1-Y_i}.$$

Fig. 5.8 Likelihood function for a two-parameter logistic model



Since both Y and x (the indicator of AIDS status) are observed, the only unknown quantities are the regression parameters β_0 and β_1 . These are generally estimated using an iterative maximization algorithm. Figure 5.8 presents a plot of the logarithm of this function for a range of values for P . Because the likelihood function depends on two unknown parameters, it has the form of a “surface” when plotted in three dimensions. The two-dimensional figure represents the contours of this surface as seen from above. The maximum value is indicated, and the corresponding maximum-likelihood estimates for β_0 and β_1 are -1.705 and 2.110 , respectively.

Because likelihoods are formed from the product of outcome probabilities for all individuals in a sample, the numerical value of a given likelihood depends on the sample size and is not particularly interpretable by itself. However, comparing likelihoods from nested models is a direct way to evaluate improvements in fit. This is the basis of the LR test.

Finally, we note that although the discussion here is limited to the binary outcome context, estimation methods for most of the regression models presented in this book are likelihood based. For example, least squares estimation and F -testing for comparing nested models in linear regression and analysis of variance models are examples of likelihood methods. Further, likelihood methods are fundamental to the family of GLMs discussed in Chap. 9.

5.7 Sample Size, Power, and Detectable Effects

Section 4.8 provides formulas for calculating sample size, power, and minimum detectable effects for the linear model. Analogous results hold for the logistic model. To compute the sample size that will provide power of γ in two-sided tests with

type-1 error of α to reject the null hypothesis $\beta_j = 0$ for the effect of a predictor X_j , accounting for the loss of precision due to multiple predictors, we can use

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2}{(\beta_j^a \sigma_{x_j})^2 p(1-p)(1-\rho_j^2)}, \quad (5.16)$$

where β_j^a is the hypothesized value of β_j under the alternative, $z_{1-\alpha/2}$ and z_γ are the quantiles of the standard normal distribution corresponding to the specified type-1 error and power, σ_{x_j} is the standard deviation of X_j and ρ_j is its multiple correlation with the other covariates, and p is the marginal prevalence of the outcome (Hsieh et al. 1998; Hsieh and Lavori 2000). For problems with predetermined n , power is given by

$$\gamma = 1 - \Phi \left[z_{1-\alpha/2} - |\beta_j^a| \sigma_{x_j} \sqrt{np(1-p)(1-\rho_j^2)} \right] \quad (5.17)$$

and the minimum detectable effect (on the log-odds scale) by

$$\pm \beta_j^a = \frac{z_{1-\alpha/2} + z_\gamma}{\sigma_{x_j} \sqrt{np(1-p)(1-\rho_j^2)}}. \quad (5.18)$$

Some additional points:

- When X_j is binary with prevalence f_j , $\sigma_{x_j} = \sqrt{f_j(1-f_j)}$ in (5.16)–(5.18).
- When X_j is continuous with standard deviation σ_{x_j} , it is important to recognize that sample size, power, and minimum detectable effects do not depend in any real way on the units in which X_j is measured. This is most clearly seen in (6.26). Suppose X_j is usually measured in grams. Changing the unit to milligrams increases σ_{x_j} by a factor of 1,000, and shrinks β_j^a by the same factor. But of course the effect on the outcome of a 1-mg increase in the predictor is 1,000 times smaller than the effect of a 1-g increase. One way to avoid confusion is to consider the minimum detectable effect size for a one standard deviation change in X_j , which is often a reasonable-sized change to consider. That effect size is obtained by setting $\sigma_{x_j} = 1$ in (5.18).
- Sample size (5.16) and minimum detectable effect (5.18) calculations simplify considerably when we specify $\alpha = 0.05$ and $\gamma = 0.8$, β_j^a is the effect of a one standard deviation increase in continuous x_j , and we do not need to penalize for covariate adjustment. In that standard case,

$$n = \frac{7.849}{(\beta_j^a)^2 p(1-p)}. \quad (5.19)$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = \frac{2.802}{\sqrt{np(1-p)}}. \quad (5.20)$$

For 90% power, substitute 10.51 for 7.849 and 3.242 for 2.802.

- Similarly, for a 2-arm clinical trial with equal allocation to arms, so that β_j^a is the log odds-ratio for treatment and $s_{x_j}^2 = 0.25$, we can calculate

$$n = 4 \times \frac{7.849}{(\beta_j^a)^2 p(1-p)}. \quad (5.21)$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = 2 \times \frac{2.802}{\sqrt{np(1-p)}}. \quad (5.22)$$

- Power calculations using (5.17) simplify analogously, but still require a statistical calculator or computer package to evaluate the normal cumulative distribution function $\Phi(\cdot)$.
- As in calculations for the linear model, we need to use $|\beta_j^a|$ in (5.17) if $\beta_j^a < 0$. It follows that the negative of the value given by (5.18) is also a valid solution for the minimum detectable effect.
- These computations are valid for unmatched case-control studies, in which p , the sample prevalence of the outcome, is controlled by design. However, special methods are required for matched case-control studies.
- Sample size and power (but not minimum detectable effects) can be calculated using the `sampsi` command in Stata as well as many other statistical packages. Alternatively, (5.16)–(5.18) can easily be programmed in Stata, R, or Excel, or evaluated by hand if values of $z_{1-\alpha/2}$, z_γ , and $\Phi(\cdot)$ are available.
- The use of the factor $1 - \rho_j^2$ to account for covariate adjustment carries over from linear to logistic models. However, there is no analog to the reduction in residual variance that can result from including covariates in linear models, so that the adjustment to these calculations using $1 - \rho_j^2$ is less likely to be conservative.
- The `sampsi` command does not incorporate the factor $1 - \rho_j^2$ for covariate adjustment. An unadjusted estimate of n should be inflated by $1/(1 - \rho_j^2)$; similarly the unadjusted minimum detectable effect estimate should be inflated by $\sqrt{1/(1 - \rho_j^2)}$. To calculate power, use $n(1 - \rho_j^2)$ in place of n as an input.
- For logistic models with a continuous predictor, `sampsi` can be made to work by reversing the role of predictor and outcome, as we show in an example below.
- These calculations were derived in Chap. 4 from the Wald test of $\beta_j = 0$. Calculations based on the more reliable LR test (Self and Mauritsen 1992) have been implemented in the Egret statistical package.

- In Sect. 4.8, we showed how the standard error $SE(\hat{\beta}_j)$ plays a central role in sample size, power, and minimum detectable effect calculations for regression problems. $SE(\hat{\beta}_j)$ is a large-sample approximation for the logistic model, and more exact small-sample computations using the noncentral t -distribution do not carry over from the linear model. Simulations of power may be a more reliable guide when the calculated or available sample size is small.
- Equations (5.16)–(5.18) are based on the assumption that the conditional mean of the outcome does not vary strongly across observations, which would hold if X_j is a relatively weak predictor, or equivalently if $|\beta_j^a|$ is small. Methods based on simulation avoid this simplification and perform slightly better in some circumstances (Vittinghoff et al. 2009). However, errors from these sources are usually small compared to errors arising from uncertainty about the required inputs.
- The alternative calculations (4.22)–(4.24) presented in Sect. 4.8, which use an estimate $\tilde{SE}(\hat{\beta}_j)$ based on published results for an appropriately adjusted model using $\tilde{n}$ observations, carry over directly. There we showed that

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \tilde{n} [\tilde{SE}(\hat{\beta}_j)]^2}{(\beta_j^a)^2}. \quad (5.23)$$

Similarly, power in a new sample of size n is given by

$$\gamma = 1 - \Phi \left[z_{1-\alpha/2} - |\beta_j^a| / [\sqrt{\tilde{n}/n} \tilde{SE}(\hat{\beta}_j)] \right]. \quad (5.24)$$

Finally, the minimum detectable effect in a new sample of size n can be obtained as

$$\pm \beta_j^a = (z_{1-\alpha/2} + z_\gamma) \sqrt{\tilde{n}/n} \tilde{SE}(\hat{\beta}_j). \quad (5.25)$$

In implementing these calculations, care must be taken to obtain the SE of the regression coefficient β_j , not the SE of the odds ratio e^{β_j} . Since results are usually available only for the odds ratio, this can be computed as $\tilde{SE}(\hat{\beta}_j) = \log(UL/LL)/3.92$, where UL and LL are the upper and lower 95% confidence bounds for the odds ratio. We must also ensure that X_j is measured on the same scale as in the published results.

To illustrate these methods, we first use the `sampsiz` command to estimate the sample size providing 80% power in two-sided tests with α of 5% for a clinical trial of a new technique hypothesized to reduce the incidence of an adverse postsurgical outcome from 15% to 5%. We specify the proportion with the outcome in each group, which are equivalent to the means of a continuous outcome. By omitting the `sd1` option, we signal that the outcome is binary, with SD determined under the statistical model as $\sqrt{p(1-p)}$. With equal allocation to treatment and control, the `r()` option, which specifies the ratio of the sizes of the groups being compared,

Table 5.33 Sample size calculation for randomized trial

```

. sampsi 0.05 0.15, power(0.8)

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
              and p2 is the proportion in population 2

Assumptions:

      alpha =    0.0500   (two-sided)
      power =    0.8000
      p1     =    0.0500
      p2     =    0.1500
      n2/n1  =    1.00

Estimated required sample sizes:

      n1 =      160
      n2 =      160

. display log((0.05/0.95)/(0.15/0.85))
-1.2098379

. display (invnormal(0.975)+invnormal(.8))^2/((-1.2098379)^2*0.25*0.075*
(1-0.075)) 309.17921

```

can be left at the default value of 1. In addition, we can safely assume that $\rho_j = 0$, so no adjustment for covariates is likely to be necessary in a randomized trial.

Table 5.33 shows the results. The `sampsi` command estimates that we need 160 participants per group. We also used (5.16) to estimate sample size. For that calculation, β_j^a , the hypothesized log-odds ratio for the effect of the new technique, is $\log(0.05/0.95)/(0.15/0.85) \approx -1.2098$. With equal allocation ($f = 0.5$) to treatment and control, $\sigma_x^2 = f(1 - f) = 0.25$, and the marginal prevalence p of the outcome $\approx 7.5\%$. This gives an overall sample size estimate of 309.

Now, suppose we would like to estimate the sample size that will provide 80% power in two-sided tests with α of 5% to detect an independent association of SBP with CHD, adjusting for age, smoking, BMI, cholesterol levels, and behavioral patterns, as suggested by the results in Table 5.10. From pilot data, we estimate that the prevalence of CHD in the new sample of high risk men will be 30%, that SBP will be approximately 5 mmHg higher among the men with CHD, that the within-group SD of SBP will be 15 mmHg, and finally that $\rho_j \approx 0.33$. To do this computation using the `sampsi` command, we reverse the role of the outcome and predictor, so f is now the prevalence of CHD. Pre-calculation of the local variable `ratio` is required because the `r()` option will not allow the fractional input 3/7.

Table 5.34 first shows the calculation using the `sampsi` command, with the adjustment using the variance inflation factor then carried out based on the unadjusted results. In addition, we computed the sample size using (5.16), relying on the fact that $\beta_j \sigma_x$, the log-odds per SD increase in SBP, is approximately equal to the standardized (in SDs) difference in mean SBP between the subgroups with and without CHD. The two sample size estimates are close.

Table 5.34 Sample size calculation for the effect of SBP on risk of CHD

```

. display 3/7

. 42857143

. sampsi 0 5, sd1(15) r(.42857143) power(0.8)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
               and m2 is the mean in population 2

Assumptions:

      alpha =    0.0500   (two-sided)
      power =    0.8000
      m1 =           0
      m2 =           5
      sd1 =          15
      sd2 =          15
      n2/n1 =      0.43

Estimated required sample sizes:

      n1 =          236
      n2 =          102

. display (236+102)/(1-0.33^2) 379.30648

. display (invnormal(.975)+invnormal(0.8))^2/((5/15)^2*.3*.7*(1-.33^2))
377.48913

```

5.8 Summary

The logistic regression model extends frequency table techniques for investigating the association between a binary outcome and categorical predictor to include continuous predictors and allow simultaneous consideration of multiple (continuous and categorical) predictors.

Modeling techniques for logistic regression mirror those for linear regression, allowing many of the concepts and methods learned in Chap. 4 to be applied directly to studies involving binary outcomes. However, interpretation of logistic regression models is slightly more complex due to the model's nonlinear relationship between outcome risk and predictors. In particular, regression coefficients need to be transformed to be interpretable as odds ratios.

Although a powerful and useful tool, there are a number of situations where logistic regression is not the best method for analyzing binary outcome data. As we have seen in several examples, when attention is restricted to one or a few categorical predictors, regression techniques are not needed. In other situations, an alternative binary regression model linked to alternate measures of association such as relative risks or risk differences may be preferred. We refer readers to Chap. 6 for methods appropriate for regression analysis for event time outcomes. Although we have provided a brief illustration in Sect. 5.5.2 of how logistic regression can be used to investigate the effects of predictors on binary outcomes that are duration dependent,

we refer readers to Chap. 9 for a more complete coverage of regression methods for event time outcomes. Finally, we note that when analysis focuses on causal inference about the effect of a particular binary predictor representing a treatment or exposure, the methods covered in Chap. 9 are generally preferred.

5.9 Further Notes and References

There are a number of excellent textbooks on logistic regression, including [Breslow and Day \(1984\)](#), [Hosmer and Lemeshow \(2000\)](#), [Kleinbaum \(2002\)](#), and [Collett \(2003\)](#). All of these provide more details and cover a broader range of topics than provided here. Although we have focused on Stata in our example analyses, most modern statistical software packages provide extensive facilities for fitting and interpretation of logistic models, including R, SAS, S-PLUS, and SPSS. More extensive facilities for exact logistic regression and contingency table methods are available in the programs StatXact and LogXact.

Throughout this chapter, we have concentrated on analysis of data where the outcomes and predictors were measured without substantial error and missing observations were not considered a major problem. In many studies, we cannot assume that this is the case. There is an extensive literature on the impacts of misclassified outcomes and measurement error in predictors in the context of logistic regression ([Carroll et al. 1995](#); [Magder and Hughes, 1997](#)).

Missing data are an issue in most studies involving binary outcomes, and arise through a variety of mechanisms. When relatively few observations are involved, the problem can be handled via the default procedure in most available software programs (i.e., to eliminate any observations with one or more missing values among the predictors). The validity of this approach rests on the assumption that the individuals dropped from the analysis are “missing completely at random.” However, when a substantial fraction of observations involve missing values, more care is required. In addition to the obvious problem of the reduction in power incurred by dropping observations there are substantial concerns that the results based on the remaining complete data may be biased. There are a number of approaches to handling missing observations, including sensitivity analyses, imputation, and modified maximum likelihood estimation methods. (See [Jewell 2004](#) for a more complete discussion.) These tend to be complex to apply and are not generally well represented in standard software.

5.10 Problems

Problem 5.1. Verify that the numerical average (mean) of the following sample of 25 binary outcomes equals the proportion of positive outcomes (ones) in the sample:

(1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0)

Table 5.35 Logistic model for CHD and age

```
. logistic chd69 i.agec, coef
```

Logistic regression

Log likelihood = -868.14866

Number of obs = 3154

LR chi2(4) = 44.95

Prob > chi2 = 0.0000

Pseudo R2 = 0.0252

chd69	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agec						
1	-.1314518	.2309941	-0.57	0.569	-.5841919	.3212882
2	.5307399	.2235341	2.37	0.018	.0926211	.9688586
3	.8409976	.2274986	3.70	0.000	.3951085	1.286887
4	1.05998	.2585408	4.10	0.000	.5532496	1.566711
_cons	-2.804337	.1849627	-15.16	0.000	-3.166858	-2.441817

Problem 5.2. Use the regression coefficients from the logistic model presented in Table 5.2 in the logistic formula (5.2) to estimate the quantities in Table 5.3 for a 65-year-old individual. Use additional calculations to add a new section to Table 5.3 for an age increment of five years.

Problem 5.3. Perform the basic algebra necessary to verify the properties of the logistic regression coefficient β_1 stated in (5.6).

Problem 5.4. The output in the Table 5.35 provides the regression coefficients corresponding to the model fitted in Table 5.5. Use the coefficients and calculations similar to those illustrated in Sect. 5.1.1 to compute the log odds ratio comparing CHD risk in the fourth age category (4 .agec) with the third (3 .agec). Also, compute the odds ratio for this comparison. Comment on how we might obtain an estimated standard error and 95% CI for this quantity.

Problem 5.5. For the fitted logistic regression model in Table 5.6, calculate the log odds for a 60-year-old smoker with cholesterol, SBP, and BMI values of 250 mg/dL, 150 mmHg, and 20, respectively. Now calculate the log odds for an individual with a cholesterol level of 200 mg/dL, holding the values of the other predictors fixed. Use these two calculations to estimate an odds ratio associated with a 50 mg/dL increase in cholesterol. Repeat the above calculations for a 70-year-old individual with identical values of the other predictors. Comment on any differences between the two estimated odds ratios.

Problem 5.6. Use the regression output in Table 5.16 and a calculation similar to that presented in (5.11) to compute the odds ratio comparing the odds of CHD in a 55-year-old individual with arcus to the corresponding odds for a 40-year-old who also has arcus.

Problem 5.7. Use the WCGS data set to fit the regression model presented in Table 5.18. Perform the Hosmer–Lemeshow goodness of fit test for the following number of groups: 10, 15, 20, and 25. Comment on the differences. The data set is available at <http://www.biostat.ucsf.edu/vgsm>.

Problem 5.8. Verify that the odds ratio formed from the two odds presented in (5.11) is given by ad/bc . Verify that the same odds ratio is obtained if the two component odds are computed based on the probability of exposure conditional on outcome status.

Problem 5.9. Compute the approximate 95% CI for the following per-contact infection risk based on the intercept coefficient and associated standard errors given in Table 5.28:

$$1 - \exp[-\exp(-7.033)].$$

5.11 Learning Objectives

- (1) Describe situations in which logistic regression analysis is needed.
- (2) Translate research questions appropriate for a logistic regression model into specific questions about model parameters.
- (3) Use logistic regression models to test hypotheses about relationships between a binary outcome variable and a continuous or categorical predictor.
- (4) Describe the logistic regression model, its key assumptions, and their implications.
- (5) State the relationships between:
 - Odds ratios and logistic regression coefficients.
 - A two $\times$ two table analysis of the association between a binary outcome and single categorical predictor and a logistic regression model for the same variables.
- (6) Know how a statistical package is used to fit a logistic regression model to continuous and categorical predictors.
- (7) Interpret logistic regression model output, including:
 - Regression parameter estimates, hypothesis tests, CIs.
 - Statistics which quantify the fit of the model.

Chapter 6

Survival Analysis

Children receiving a kidney transplant may be followed to identify predictors of mortality. Specifically, is mortality risk lower in recipients of kidneys obtained from a living donor? If so, is this effect explained by the time the transplanted kidney is in transport or how well the donor and recipient match on characteristics that affect immune response? Similarly, HIV-infected subjects may be followed to assess the effects of a new form of therapy on incidence of opportunistic infections. Or patients with liver cirrhosis may be followed to assess whether liver biopsy results predict mortality.

The common interest in these studies is to examine predictors of time to an event. The special feature of the survival analysis methods presented in this chapter is that they take time directly into account: in our examples, time to transplant rejection, incidence of opportunistic infections, or death from liver failure. Basic tools for the analysis of such *time-to-event* data were reviewed in Sect. 3.5. This chapter covers multipredictor regression techniques for the analysis of outcomes of this kind.

6.1 Survival Data

6.1.1 Why Linear and Logistic Regression Would not Work

In Sect. 3.5, we saw that a defining characteristic of survival data is *right-censoring*:

Definition: A survival time is said to be *right-censored* at time t if it is only known to be greater than t .

Because of right-censoring, survival times cannot simply be analyzed as continuous outcomes. But survival data also involve an outcome *event*, so why is logistic regression not applicable? The reason is variable lengths of follow-up. In Chap. 5, the logistic model was used to study CHD events among men in the WCGS (Rosenman et al. 1964). But in that study, the investigators were able to determine

whether each one of the study participants experienced the outcome event at any time in the well-defined 10-year follow-up period; follow-up was constant across participants.

In contrast, follow-up times were quite variable in ACTG 019 (Volberding et al. 1990), a randomized double-blind placebo-controlled clinical trial of zidovudine (ZDV) for prevention of AIDS and death among patients with HIV infection. Between April 1987 and July 1989, 453 patients were randomized to ZDV and 428 to placebo. When the data were analyzed in July 1989, some had been in the study for less than a month, while others had been observed for more than 2 years. Simply applying logistic regression to the binary indicator of mortality in this example would ignore the broad variation between patients in length of follow-up. Regression adjustment for duration of follow-up would address this partially, but impose unnecessary assumptions about the relationship between event risk and duration. Although the pooled logistic regression model introduced in Sect. 5.5.2 addresses some of these concerns, that approach is more appropriate when follow-up and event time information is restricted to intervals corresponding to regular study visits. The concepts and methods introduced in this chapter offer a more complete approach to regression for survival data including observations of actual event times.

6.1.2 Hazard Function

In Sect. 3.5, we introduced the survival function and its complement, the cumulative event function, as useful summaries of the distribution of a survival time.

Definition: The *survival function* at time t , denoted $S(t)$, is the probability of being event-free at t . The *cumulative event function* at time t , denoted $F(t) = 1 - S(t)$, is the complementary probability that the event has occurred by time t .

Another useful summary is the hazard function $h(t)$.

Definition: The *hazard function* $h(t)$ is the short-term event rate for subjects who have not yet experienced the outcome event.

The hazard function is systematically related to both the survival and cumulative event functions.

Table 6.1 shows mortality rates for children who have recently undergone kidney transplantation, on each of the first ten days after surgery, using data from the united network for organ sharing (UNOS). At the beginning of fifth day after surgery, for example, 9,651 children remained alive and in the study, and of these, 3 died during the next 24 h, yielding an estimated death rate of 0.31 deaths per 1,000 subjects per day. From the rightmost column of the table, it appears that the mortality rate declines over the first 10 days, although the estimates spike on days 8 and 10.

Table 6.1 Mortality among pediatric kidney transplant recipients

Days since transplant	No. in follow-up	No. died	No. censored	Death rate per 1,000 subject-days
1	9,750	7	14	$7/9,750 \times 1,000 = 0.72$
2	9,729	5	8	$5/9,729 \times 1,000 = 0.51$
3	9,716	5	12	$5/9,716 \times 1,000 = 0.51$
4	9,699	7	41	$7/9,699 \times 1,000 = 0.72$
5	9,651	3	54	$3/9,651 \times 1,000 = 0.31$
6	9,594	2	57	$2/9,594 \times 1,000 = 0.21$
7	9,535	0	50	$0/9,535 \times 1,000 = 0.00$
8	9,485	4	49	$4/9,485 \times 1,000 = 0.42$
9	9,432	1	49	$1/9,432 \times 1,000 = 0.11$
10	9,382	3	28	$3/9,382 \times 1,000 = 0.32$

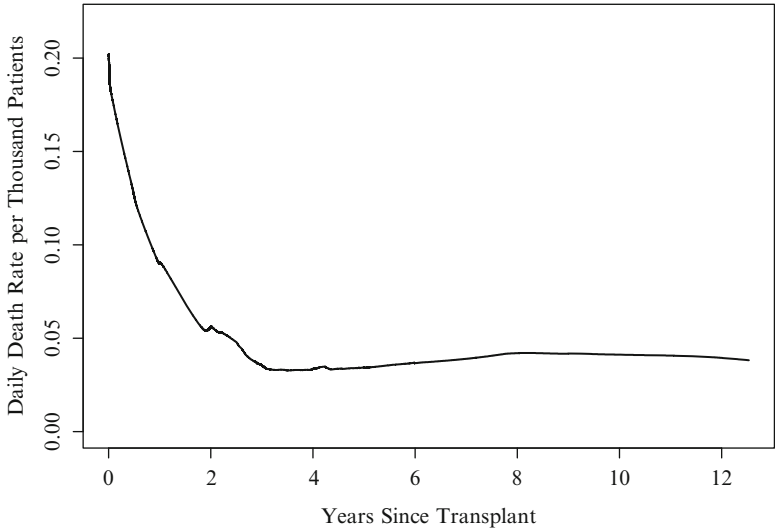


Fig. 6.1 Mortality rate for pediatric kidney transplant recipients

In Fig. 6.1, daily death rates, smoothed by LOWESS, are used to estimate the mortality hazard for a much longer time period, the first 12 years after transplantation. The mortality hazard declines rapidly over the course of the first 2 years, reaching a plateau approximately 3 years after transplantation.

6.1.3 Hazard Ratio

We now compare the hazard functions for children whose transplanted kidney was provided by a living donor, commonly a family member, and those for whom the

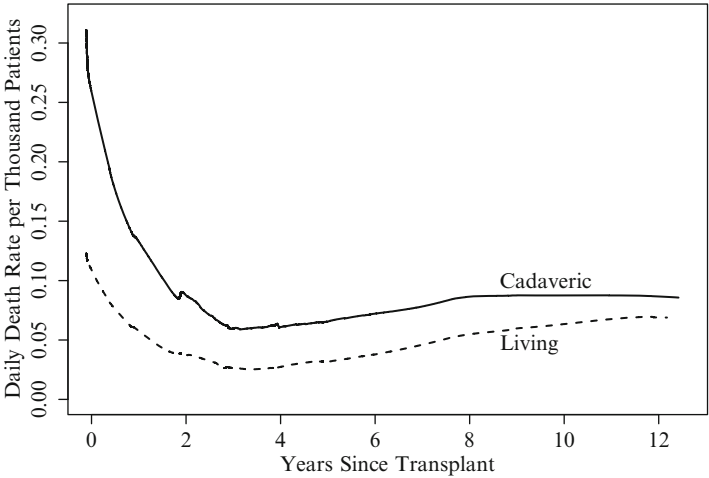


Fig. 6.2 Smoothed mortality rates for recipients by kidney donor type

Table 6.2 Smoothed death rates (per 1,000 days) by donor type

Years since transplantation	Smoothed rates		Death rate ratio
	Cadaveric	Living	
0.25	0.235	0.098	2.40
0.50	0.193	0.082	2.36
1.00	0.138	0.061	2.27
2.00	0.088	0.038	2.30
3.00	0.061	0.027	2.25
4.00	0.063	0.026	2.37
5.00	0.065	0.032	2.03

source was recently deceased. Figure 6.2 shows LOWESS-smoothed death rates for the recipients of kidneys from living and recently deceased donors. The mortality rate is considerably lower among the recipients of kidneys from living donors at all time points, but the curves are similar in shape.

Table 6.2 gives the values of the LOWESS-smoothed death rates shown in Fig. 6.2 for selected time points, which estimate the hazard functions in each group, as well as the death rate ratio, an estimate of the *hazard ratio*. We could write the hazard ratio as

$$\text{HR}(t) = h_c(t)/h_l(t), \tag{6.1}$$

where $h_c(t)$ is the hazard function in the recipients of kidneys from cadaveric donors, and $h_l(t)$ is the corresponding hazard function in the reference group, the recipients of kidneys from living donors.

6.1.4 Proportional Hazards Assumption

The results in Table 6.2 show that while the mortality hazards decline over time in both groups of pediatric kidney transplant recipients, the hazard ratio is roughly constant. In other words, the hazard in the comparison group is a constant proportion of the hazard in the reference group.

Definition: Under the *proportional hazards assumption*, the hazard ratio does not vary with time. That is, $HR(t) \equiv HR$.

Provided the hazards are proportional in this sense, the effect of donor source on post-transplant mortality risk can be summarized by a single number. This simplification is useful, but not necessary for the Cox proportional hazards model described in the next section. We can generalize the model by including an interaction between the predictor and time; this allows the hazard ratio for that predictor to change with time. In Sect. 6.4.2, we show how this strategy can be used to check and model nonproportional hazards with respect to a variable. This is implemented using time dependent covariates (*TDCs*), an extension of the basic Cox model introduced in Sect. 6.3.1.

6.2 Cox Proportional Hazards Model

The Cox proportional hazards regression model is a flexible tool for assessing the relationship of multiple predictors to a right-censored, time-to-event outcome, and has much in common with linear and logistic models. To understand how the Cox model works, we first consider the broader class of proportional hazards models.

6.2.1 Proportional Hazards Models

In the linear model for continuous outcomes, covered in Chaps. 4 and 10, the linear predictor $\beta_1 x_1 + \dots + \beta_p x_p$, which captures the effects of predictors, is linked directly to the conditional mean of the outcome, $E[y|\mathbf{x}]$:

$$E[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \quad (6.2)$$

In the logistic model for binary outcomes, covered in Chap. 5, the linear predictor is linked to the conditional mean through the logit transformation:

$$\log \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \quad (6.3)$$

In (6.3), $p(\mathbf{x}) = E[y|\mathbf{x}]$ is the probability of the outcome event for a observation with predictor values $\mathbf{x} = (x_1, \dots, x_p)$.

In proportional hazards regression models, the linear predictor is linked through the log-transformation to the hazard ratio introduced in Sect. 6.1.3. If the hazard ratio obeys the proportional hazards assumption, and thus does not depend on time, we can write

$$\log [\text{HR}(\mathbf{x})] = \log \frac{h(t|\mathbf{x})}{h_0(t)} = \beta_1 x_1 + \dots + \beta_p x_p. \quad (6.4)$$

In (6.4), $h(t|\mathbf{x})$ is the hazard at time t for an observation with covariate value $\mathbf{x}$, and $h_0(t)$ is the *baseline hazard function*, defined as the hazard at time t for observations with all predictors equal to zero. As with the intercept in linear and logistic regression, this may mean that the baseline hazard does not apply to any possible observation, and argues for centering continuous predictors.

Solving (6.4) for $h(t|\mathbf{x})$ gives

$$\begin{aligned} h(t|\mathbf{x}) &= h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p x_p) \\ &= h_0(t) \text{HR}(\mathbf{x}). \end{aligned} \quad (6.5)$$

Note that exponentiating the linear predictor ensures that $\text{HR}(\mathbf{x})$ cannot be negative, as required. Furthermore, taking the log of both sides of (6.5), we obtain

$$\log[h(t|\mathbf{x})] = \log[h_0(t)] + \beta_1 x_1 + \dots + \beta_p x_p. \quad (6.6)$$

This shows that the log baseline hazard plays the role of the intercept in other regression models, though in this case it can change over time. Furthermore, (6.6) defines a *log-linear* model, which implies that the log of the hazard is assumed to change linearly with any continuous predictors.

Note also that (6.5) defines a *multiplicative* model, in the sense that the predictor effects act to multiply the baseline hazard. This is like the logistic model, where the linear predictor acts multiplicatively on the baseline odds. In contrast, (6.2) shows that in the linear model the predictor effects are *additive* with respect to the intercept β_0 .

6.2.2 Parametric Versus Semi-parametric Models

We have two options in dealing with the baseline hazard $h_0(t)$. One is to model it with a parametric function. For instance, the exponential survival model specifies that the hazard is a constant while the Weibull regression model has a hazard which is a polynomial in time. In both of these models, the baseline hazard $h_0(t)$ is specified by a small number of additional parameters, which are estimated along

Table 6.3 Cox model for type of donor

stcox i.txttype

No. of subjects =	9750	Number of obs =	9750
No. of failures =	461		
Time at risk =	38004.90961		
Log-likelihood =	-3952.3735	LR chi2(1) =	44.82
		Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
1.txttype	1.879674	.1801323	6.59	0.000	1.557795 2.26806

with $\beta_1, \beta_2, \dots, \beta_p$. If the baseline hazard is specified correctly, this approach is efficient, handles right-censoring as well as more complicated censoring schemes with ease, and makes it simple (though still risky) to extrapolate beyond the data. Of course the adequacy of the model for the baseline hazard has to be checked.

In contrast to parametric models, the *Cox model*, or Cox proportional hazards model, does not require us to specify a parametric form for the baseline hazard, $h_0(t)$. Because we still specify (6.4) as the model for the log-hazard ratio, the Cox model is considered semi-parametric. Nonetheless, estimation of the regression parameters $\beta_1, \beta_2, \dots, \beta_p$ is done without having to estimate the baseline hazard function. Note that estimates of this function can be useful in summarizing hazards associated with particular predictor values, and can be obtained once the regression parameters are estimated (Kalbfleisch and Prentice 1980). The Cox model is more robust than parametric proportional hazards models because it is not vulnerable to misspecification of the baseline hazard. Furthermore, the robustness is commonly achieved with little loss of precision in the estimated predictor effects.

6.2.2.1 Proportionality and Multiplicativity

Figure 6.2 and the summary statistics in Table 6.2 showed that the two mortality hazards for pediatric recipients of kidney transplants from living and recently deceased donors were very nearly proportional over time, in the sense that the ratio of the LOWESS-smoothed death rates was approximately constant. So the Cox model appears appropriate for these data, because the proportional hazards assumption appears to be met for this important predictor. Table 6.3 shows the unadjusted Cox model hazard ratio estimate for txttype, a binary indicator identifying the group receiving transplants from recently deceased donors. The estimated hazard ratio of 1.9 (95% CI 1.6–2.3 $P < 0.0005$) is consistent with the estimates shown in Table 6.2, and suggests that receiving a transplant from a recently deceased donor roughly doubles the mortality risk at every point over the 12 years of follow-up.

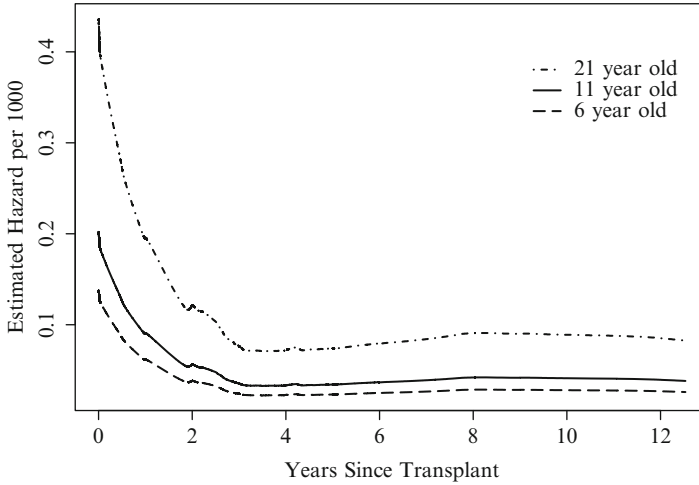


Fig. 6.3 Hazard functions for 6-, 11-, and 21-year-old transplant recipients

Another important determinant of mortality after kidney transplant is the age of the recipient. Using results from a Cox model with age as continuous (results not shown), Fig. 6.3 shows fitted hazards for 6-, 11-, and 21-year-olds. The hazards for the three groups differ proportionally. However, it is important to point out that the perfect proportionality of the hazard functions plotted in Fig. 6.3 is imposed under the fitted model, like the perfectly parallel regression lines for the additive linear model without interaction terms shown in Fig. 4.2. This is in contrast to the apparently proportional relationship between the independently smoothed death rates in Fig. 6.2, which are based only on the data.

While the hazard ratio is assumed to be constant over time in the basic Cox model, under this multiplicative model the between-group *differences* in the hazard can easily be shown to depend on $h_0(t)$ and thus on time. This is reflected in the fact that the hazard functions in Fig. 6.3 are considerably farther apart immediately after transplant when the baseline hazard is higher.

6.2.2.2 DPCA Study of Primary Biliary Cirrhosis

To illustrate interpretation of Cox model results, we consider a cohort of 312 participants in a placebo-controlled clinical trial of D-penicillamine (DPCA) for primary biliary cirrhosis (PBC) (Dickson et al. 1989). PBC destroys bile ducts in the liver, causing bile to accumulate. Tissue damage is progressive and ultimately leads to liver failure. Time from diagnosis to end-stage liver disease ranges from a few months to 20 years. During the approximate 10-year follow-up period, 125 study participants died.

Table 6.4 Cox model for treatment and bilirubin

stcox i.rx bilirubin						
No. of subjects =		312	Number of obs =		312	
No. of failures =		125				
Time at risk =		1713.853528				
Log-likelihood =		-597.08411	LR chi2(2) =		85.79	
			Prob > chi2 =		0.0000	

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	

1.rx	.8181612	.1500579	-1.09	0.274	.5711117	1.172078
bilirubin	1.163459	.0154566	11.40	0.000	1.133556	1.194151

Predicting survival in PBC patients is important for clinical decision making. The investigators collected data on age as well as baseline laboratory values and clinical signs including serum bilirubin levels, enlargement of the liver (hepatomegaly), accumulation of water in the legs (edema), and visible veins in the chest and shoulders (spiders)—all signs of liver damage.

In the sections that follow, we will illustrate use of the Cox model for testing and interpretation. This will present a series of largely unrelated models. The objective will not be to illustrate a model selection strategy.

6.2.3 Hazard Ratios, Risk, and Survival Times

Table 6.4 displays a Cox model for the effects of treatment with DPCA (rx) and bilirubin (bilirubin) on mortality risk in the PBC cohort.

The hazard ratio for treatment, 0.82, means that estimated short-term mortality risk among patients assigned to DPCA was 82% of the risk in the placebo group. This ratio is assumed to be constant over the 10 years of follow-up. Likewise, the hazard ratio for bilirubin levels means that for each mg/dL increase in bilirubin, short-term risk is increased by a factor of 1.16.

More broadly, (6.6) implies that in a model with predictors $x_1, x_2, \dots, x_p$, coefficient β_j is the increase in the log-hazard ratio for a one-unit increase in predictor x_j , holding the values of the other predictors constant. It follows that $\exp(\beta_j)$ is the hazard ratio for a one-unit increase in x_j . Below, we show how this applies to continuous as well as binary and categorical predictors. Furthermore, for predictors with hazard ratios less than 1 ($\beta < 0$), increasing values of the predictors are associated with lower risk and longer survival times. Conversely, when hazard ratios are greater than 1 ($\beta > 0$), increasing values of the predictor are associated with increased risk and shorter survival times. In using the term *risk* in this context, it is important to keep in mind the definition of the hazard as a short-term rate and distinguish risk in this sense from cumulative risk over a defined follow-up period.

Table 6.5 Cox model for treatment and bilirubin showing coefficients

stcox i.rx bilirubin, nohr				LR chi2(2)	=	85.79
Log-likelihood =		-597.08411		Prob > chi2	=	0.0000
_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
1.rx	-.2006959	.1834088	-1.09	0.274	-.5601705	.1587787
bilirubin	.1513976	.0132851	11.40	0.000	.1253594	.1774358

6.2.4 Hypothesis Tests and Confidence Intervals

In the Cox model, as in the logistic model, the estimated coefficients have an approximate normal distribution when there are adequate numbers of events in the sample. The normal approximation is better for the coefficient estimates than for the hazard ratios, so hypothesis tests and confidence intervals are based on calculations involving the coefficients and their standard errors. If there are fewer than 15–25 events, the normal approximation is suspect and bootstrap CIs may work better; see Sect. 6.6.1. Table 6.5 displays the Cox model for the effects of DPCA and bilirubin on mortality risk with results on the coefficient rather than the hazard ratio scale.

For each predictor in the model, Wald Z -tests are the default used by Stata to test the null hypothesis $H_0: \beta = 0$, or equivalently that the hazard ratio equals 1. Under the null, the ratio of the coefficient estimate to its standard error tends to a standard normal, or Z , distribution with mean 0 and standard deviation 1. In Table 6.5, the Z -statistics and associated P -values for `rx` and `bilirubin` appear in the columns headed $|z|$ and $P > |z|$, respectively. The evidence for the efficacy of DPCA is not persuasive ($P = 0.27$), but there is strong evidence that bilirubin levels are associated with mortality risk ($P < 0.0005$). You can verify that the test results in Table 6.4 are identical to those in Table 6.5 and refer to the Z -test involving the actual coefficients and their standard errors, and not to a Z -test involving the ratio of the hazard ratio to its standard error (Problem 6.1).

Since Cox regression is a likelihood-based method, tests for predictors can also be obtained using the LR tests introduced in Sect. 5.2.1 for the logistic regression model. The procedure is the same in this setting, comparing twice the difference in log-likelihoods for nested models to a χ^2 distribution with degrees of freedom equal to the between-model difference in the number of parameters. For instance, to obtain an LR test of the null hypothesis that the hazard ratio for treatment is 1, we would compare the log-likelihood for the model in Table 6.4 to the log-likelihood for a model with `bilirubin` as the only predictor. These log-likelihoods are -597.1 and -597.7 , yielding a LR test statistic of $2[(-597.1) - (-597.7)] = 1.2$, with an associated p -value of 0.27.

In this case, the Wald and LR results are essentially identical. In most situations, these tests give results which are similar but not exactly the same. The results

be will closest when the sample size is large or the estimated hazard ratio is near 1. However, in datasets with few events, the LR test gives more accurate p -values, and so is recommended in that context. As noted in Sect. 10.4.2, qualitative discrepancies between the two test results may indicate that the model includes too many predictors for the number of events.

A 95% CI for each β is obtained by computing $\hat{\beta} \pm 1.96\text{SE}(\hat{\beta})$. Stata and other packages usually make it possible to compute CIs with other significance, or α , levels. In Stata, this can be done by using the `level()` option.

In turn, CIs for the hazard ratios are obtained by exponentiating the upper and lower limits of the CIs for the coefficients, again because the normal approximation is better on the coefficient scale. From Table 6.4, the CI for `rx`, the indicator for treatment with DPCA, shows that the data are consistent with risk reductions as large as 43%, but also with risk increases of 17%. It is also clear that the increase in risk associated with each mg/dL increase in bilirubin is rather precisely estimated (95% CI for the hazard ratio 1.13–1.19).

You can also verify that the CIs in Table 6.4 are *not* equal to the estimated hazard ratio plus or minus 1.96 times *its* standard error (Problem 6.1). For `rx`, that calculation would yield (0.52–1.11) rather than (0.57–1.17). In reasonably large samples like this one, the two intervals are usually very similar. However, since the intervals based on exponentiating the confidence limits for the coefficients are more accurate in small samples, they are the ones used in Stata.

6.2.5 Binary Predictors

Binary predictors can be coded as 1 and 0 and entered as numeric predictors, as opposed to categorical. For example, we could code `rx` as 1 for the DPCA arm and 0 for placebo. Then the exponentiated coefficient gives the hazard ratio for treatment versus placebo (and retains its literal interpretation as the hazard ratio for a one-unit increase in the predictor). Some alternative codings, (e.g., placebo = 1 and treatment = 2) would give the same results in this instance, but would complicate interpretation in the presence of an interaction involving the binary predictor. This would also make the baseline hazard harder to interpret; in the DPCA example, the baseline hazard would not refer to either the placebo or the treatment group. Thus, if binary predictors are treated as numeric, we recommend the 0/1 coding in this context as well (Problem 6.2).

6.2.6 Multilevel Categorical Predictors

Patients in the PBC study underwent a liver biopsy to determine their level of tissue damage. The scores ranged from 1 to 4, with increasing values reflecting

Table 6.6 Categorical fit for histology

```
. stcox i.histol
```

Cox regression -- Breslow method for ties

No. of subjects =	312	Number of obs =	312
No. of failures =	125		
Time at risk =	1713.853528		
Log-likelihood =	-613.62114	LR chi2(3) =	52.72
		Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
histol					
2	4.987976	5.143153	1.56	0.119	.6610611 37.63631
3	8.580321	8.685371	2.12	0.034	1.179996 62.39165
4	21.38031	21.57046	3.04	0.002	2.959663 154.4493

```
testparm i.histol
```

(1) 2.histol = 0
(2) 3.histol = 0
(3) 4.histol = 0

chi2(3) = 43.90
Prob > chi2 = 0.0000

```
lincom -3.histol + 4.histol, hr
```

(1) - 3.histol + 4.histol = 0

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
(1)	2.491785	.4923268	4.62	0.000	1.691727 3.67021

greater damage. When we model a multiple category variable, a series of new variables are created to represent group membership with one group serving as the reference. Results are shown in Table 6.6. By default, Stata has chosen the group with the lowest score as the reference category. Estimated hazard ratios with respect to the reference group are 5.0, 8.6, and 21.4 for the groups with ratings of 2, 3, and 4, respectively, suggesting a steady increase in the hazard with higher ratings.

In addition to the default comparisons with the selected reference group, pairwise comparisons between any two categories can be obtained using the `lincom` command, as shown in Table 6.6 for groups 3 and 4. The hazard in group 4 is 2.5 times higher than in group 3 (95% CI 1.7–3.7, $P < 0.0001$).

6.2.6.1 Categories with No Events

In our example, the default reference category is sensible and does not cause problems. However, categories may sometimes include no events, because the group is small or cumulative risk is low. Hazard ratios with respect to a reference category with no events are infinite, and the accompanying hypothesis tests and CIs are hard to interpret. In this case, selecting an alternative reference group can

correct the problem, although the hazard ratio, Wald test, and CI for the category without events, with respect to the new reference category, will remain difficult to interpret.

6.2.6.2 Global Hypothesis Tests

As in logistic models, global hypothesis tests for the overall effect of a multilevel categorical predictor can be conducted using Wald or likelihood ratio (LR) χ^2 tests, with degrees of freedom equal to the number of categories minus 1. The Wald test result ($\chi^2 = 43.9$, $P < 0.00005$), obtained using the `testparm` command, is displayed in Table 6.6. The LR test result ($\chi^2 = 52.7$, $P < 0.00005$) also appears in the upper right corner of the table. Note that if covariates were included in the model, this default Stata output would refer to a test of the overall effect of *all* covariates in the model, not just `histology`; thus a LR test focused on the overall effect of `histology` would require combining the results of models with and without this predictor. Finally, a logrank test, as in Sect. 3.5.6, is available; this yields a χ^2 of 53.8 ($P < 0.0001$). The tests agree closely and all show that the groups with different histology scores do not have equal survival.

The statistical significance of pairwise comparisons should be interpreted with caution, especially if the global hypothesis test is not statistically significant, as discussed in Sect. 4.3.4. With a large number of categories, multiple comparisons can lead to inflation of the familywise type-I error rate (FER); Bonferonni, Sidak, and Scheffé adjustments are implemented in the `contrast` command, as explained in Sect. 4.3.4. In addition, some comparisons may lack power due to small numbers in either of the categories being compared.

6.2.6.3 Ordinal Predictors and Tests for Trend

The histology score is ordinal, suggesting a more specific question: does the log mortality hazard increase linearly with higher histology ratings? This question can be addressed using tests for trend across categories like those introduced in Sect. 4.3.5. Note that these tests, like other hypothesis tests for the Cox model, are conducted using the coefficients and their standard errors, rather than the relative hazards. Thus for the Cox model, these linear trend tests assess log-linearity of the hazard ratios. From Table 4.8, the trend test for a four-category variable such as `histol` is

$$-\beta_2 + \beta_3 + 3\beta_4 = 0. \quad (6.7)$$

In Stata, the test for linear trend can be obtained using the `test` or `contrast` commands—see Sect. 4.3.5 for an explanation of use of `contrast` command. The three equivalent tests presented in Table 6.7 confirm an increasing linear trend across the four histologic categories ($\chi^2 = 10.23$, $P = 0.0014$).

Table 6.9 Cox model for age in 1-year units

stcox age

Log-likelihood = -629.72592

LR chi2(1) = 20.51

Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age	1.04081	.0091713	4.54	0.000	1.022989 1.058941

6.2.7 Continuous Predictors

Age at enrollment of participants in the PBC study was recorded in years. The Cox model shown in Table 6.9 shows that the hazard ratio for a 1-year increase in age is 1.04 (95% CI 1.02–1.06, $P < 0.0005$). The hazard ratio for continuous predictors is affected by the scale of measurement. In the PBC study, ages range from 26 to 78; thus, a 1-year difference in age is small compared to the range of values. A 5-year increase in age might provide a more clinically interpretable result (Problem 6.5).

Using (6.5), we can write down the ratio of the hazards for any two patients who differ in age by k years—that is, for a patient at age $x + k$ compared with another at age x :

$$\frac{h_0(t) \exp(\beta(x + k))}{h_0(t) \exp(\beta x)} = \frac{\exp(\beta(x + k))}{\exp(\beta x)}$$
$$= \exp(\beta(x + k) - \beta x)$$
$$= \exp(\beta k).$$

(6.8)

Thus a k -unit change in a predictor multiplies the hazard by $\exp(\beta k)$, no matter what reference value x is considered.

Applying (6.8), with $\hat{\beta} = \log(1.04081)$ being the log of the hazard ratio for age from Table 6.9, the hazard ratio for an increase in age of 5 years is $\exp(\hat{\beta}5) = 1.22$. The same transformation can be applied to the confidence limits for age giving a 95% CI for a 5-year increase in age of 1.12–1.33. Equivalently, we could raise the hazard ratio estimate for an increase of one unit to the fifth power, that is, $[\exp(\beta)]^k$, and apply the same operation to the confidence limits (Problem 6.6).

The hazard ratio for a five-unit change can also be obtained by defining a new variable age5 equal to age in years divided by 5. The Cox model for age5 appears in Table 6.10. Note that the Wald and LR test results are identical in Tables 6.9 and 6.10; changes in the scale of a continuous variable do not affect these tests.

Hazard ratios can be interpreted in terms of percent changes in risk. It is easy to see from Table 6.9 that estimated mortality risk among PBC patients increases about 4% for every year increase in age. We could also compute the percent increase risk associated with larger increases in age. A k -unit increase in the predictor implies a

Table 6.10 Cox model for age in 5-year units

stcox age5

Log-likelihood = -629.72592

LR chi2(1) = 20.51

Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age5	1.221397	.0538127	4.54	0.000	1.120352 1.331556

Table 6.11 Unadjusted Cox model for bilirubin

stcox bilirubin

Log-likelihood = -597.6845

LR chi2(1) = 84.59

Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
bilirubin	1.160509	.0151044	11.44	0.000	1.131279 1.190494

$100(\exp \hat{\beta}k - 1)\%$ change in risk. Note that this is the back transformation presented in Sect. 4.7.5 for linear regression models with log-transformed outcomes. Using the log of the hazard ratio estimate from Table 6.9 in place of $\hat{\beta}$, this calculation gives 22% for the increase in mortality risk associated with a 5-year increase in age, a result we could get more directly from Table 6.10.

6.2.8 Confounding

The definition of confounding in Sect. 4.4 is not specific to the linear regression model. The conceptual issues and statistical framework for dealing with confounding are similar across all regression models and discussed in more depth in Chap. 9. To illustrate regression adjustment to control confounding in the Cox model, we examined the association between bilirubin levels and survival among patients in the DPCA trial. We first fit the simple Cox model which appears in Table 6.11. For each one-point increase in baseline bilirubin, the hazard is increased by 16%.

However, patients with higher bilirubin may also be more likely to have hepatomegaly, edema, or spiders—other signs of liver damage which are correlated with elevated bilirubin levels but not mediators of its effects, and all associated with higher mortality risk. Table 6.12 shows the estimated effect of bilirubin on mortality risk adjusted for hepatomegaly, edema, and spiders.

The adjusted hazard ratio for a one-point increase in bilirubin is 1.12 (95% CI 1.09–1.15, $P < 0.0005$). This coefficient represents the effect of a one-unit change in bilirubin while holding edema, hepatomegaly, and spiders constant. The other predictors, which may reflect other aspects of PBC-associated damage to the liver,

Table 6.12 Adjusted Cox model for bilirubin

stcox bilirubin i.edema i.hepatom i.spiders

Log-likelihood = -580.56805

LR chi2(4) = 118.82

Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
bilirubin	1.118276	.0166316	7.52	0.000	1.086149	1.151353
1.edema	2.126428	.4724983	3.40	0.001	1.375661	3.286927
1.hepatom	2.050617	.434457	3.39	0.001	1.353765	3.106173
1.spiders	1.474788	.28727	1.99	0.046	1.00676	2.160393

account for a modest proportion of the unadjusted effect of bilirubin, and clearly contribute independent information about mortality risk. The attenuation of the unadjusted hazard ratio for bilirubin in the adjusted model is typical of confounding.

6.2.9 Mediation

Mediation can also be addressed with the Cox model, using the strategies outlined in Sect. 5.2.3. Here, we use data from the FIT trial [Black et al. 1996b](#), which showed that treatment with alendronate can reduce the risk of fracture in the spine. The relative hazard of fracture of participants on alendronate was 0.52 compared with placebo with a 95% CI from 0.41 to 0.66 ($p < 0.001$). Measures of BMD were also increased by alendronate—the placebo arm showed a 0.8% decrease from baseline while the treated group had a 3.8% increase in BMD from baseline, yielding a net increase in BMD due to alendronate of 4.5% with 95% CI from 4.2% to 4.8%. We can reject a null hypothesis that change in BMD is equal for the two arms ($p < 0.001$). This raises the natural question as to whether the reduction in fracture risk is mediated, or captured by, the observed changes in BMD. Whenever we approach an analysis of mediation, a causal role of the primary predictor is implied. Hence, we should believe that the association between the primary predictor and the possible mediator is a causal one. Here, we have a randomized trial and can comfortably make such an assumption.

As we showed in Sect. 5.2.3, we establish mediation by requiring an association between the predictor of interest (treatment by alendronate in this example) with the mediator (BMD here) *and* the outcome (time to fracture here). The statistical test to establish mediation requires that we test each of these associations at the 0.05 level. Both null hypotheses are rejected with $p < 0.01$, establishing that BMD plays some mediating role in the effect of alendronate on fracture risk.

A fuller picture emerges when we examine the magnitude of the direct effect of alendronate on fracture risk. We can approach this by examining hazard ratios for treatment group in a Cox model which includes an adjustment for BMD. Because

Table 6.13 Cox model for FIT data assessing mediating value of changes in BMD due to alendronate

```
. stcox i.treat i.smoking age bmd_diff bmd_base, strata(frac_base)
```

No. of subjects = 5324
No. of failures = 294
Time at risk = 20494.62287
Log-likelihood = -1911.6879

Number of obs = 5324

LR chi2(6) = 123.14
Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
1.treat	.6237068	.082513	-3.57	0.000	.4812505	.8083319
smoking						
1	1.107652	.1422723	0.80	0.426	.8611343	1.424741
2	1.391522	.254672	1.81	0.071	.9720888	1.991931
age	1.069186	.0116993	6.11	0.000	1.0465	1.092364
bmd_diff	.8274497	.0558578	-2.81	0.005	.724904	.9445018
bmd_base	.004533	.0082887	-2.95	0.003	.0001259	.1632403

Stratified by frac_base

of the possibility of confounding between the outcome and the mediator, we recommend including potential confounders of the outcome/mediator relationship in a Cox model which examines direct effects. Table 6.13 fits a Cox model to the risk of a spinal fracture to examining the mediating value of change in BMD after adjustment for baseline BMD, smoking, age, and history of fractures at baseline. The latter variable used as a stratification variable in the Cox model because direct adjustment yields an infinite hazard ratio. The Cox model shows that there is clearly a statistically and clinically important benefit of treatment even after adjustment for BMD.

There is a temptation to compare the effect of the treatment prior to and after adjustment for the mediator. A model with treatment alone yields a hazard ratio of 0.52 with 95% CI of 0.41 to 0.66. However, it is not straightforward to compare hazard ratios for treatment across the models. Methods that compare these coefficients directly using “proportion of the treatment effect explained” are problematic. For instance, a variable which is strongly associated with the outcome but not a mediator can change the coefficient for the treatment effect in a Cox model. Hence, we do not recommend methods which calculate the “proportion of effects explained” for examining mediation.

6.2.10 Interaction

The concept of interaction presented in Sect. 4.6 is also common to other multipredictor models. To illustrate its application to the Cox model, we examined

Table 6.14 Cox model with interaction

stcox rx##hepatom

Log-likelihood = -619.7079 Prob > chi2 = 0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
1.rx	.8365301	.2778607	-0.54	0.591	.4362622 1.604041
1.hepatom	3.15151	.8380138	4.32	0.000	1.871444 5.30714
rx#hepatom					
1 1	1.099791	.4343044	0.24	0.810	.5071929 2.384775

. lincom 1.rx+1.rx#1.hepatom, hr
(1) 1.rx + 1.rx#1.hepatom = 0

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
(1)	.9200085	.1963396	-0.39	0.696	.6055309 1.397807

Table 6.15 Cox model with interaction

Group	rx	hepatom	$h(t \mathbf{x})$
1	Placebo	No	$h_0(t)$
2	DPCA	No	$h_0(t) \exp(\beta_1)$
3	Placebo	Yes	$h_0(t) \exp(\beta_2)$
4	DPCA	Yes	$h_0(t) \exp(\beta_1 + \beta_2 + \beta_3)$ $= h_0(t) \exp(\beta_1) \exp(\beta_2) \exp(\beta_3)$

interaction between two binary variables in the PBC data, treatment with DPCA (rx), and the presence of liver enlargement or hepatomegaly (hepatom). This analysis examines the hypothesis that the effect of treatment is modified by the presence of hepatomegaly. As in linear and logistic models, interaction is handled by including additional terms in the model. In Stata, interaction terms are created by including the # operator between the two interacting variables. Including the ## operator between the two variables is shorthand for the interaction term and each of the two variables themselves. The interaction model is shown in Table 6.14.

Column 4 of Table 6.15 shows the hazard functions for the four groups defined by treatment and hepatomegaly (Problem 6.7). The coefficients β_1, β_2 , and β_3 correspond to the predictors rx, hepatom, and rx#hepatom, where the latter is the interaction term. We obtain the hazard ratios of interest by dividing the hazard functions for the different rows. Specifically, the comparison of the hazard for group 2 to the hazard for group 1 gives the effect of DPCA in the absence of hepatomegaly. The model specifies that the ratio of these is $\exp(\beta_1)$. In Table 6.14, the estimated hazard ratio for rx is 0.84 (95% CI 0.44–1.60, $P = 0.59$).

Similarly, the ratio of the hazard for group 4 to the hazard for group 3, or $\exp(\beta_1) \exp(\beta_3)$, gives the effect of DPCA in the presence of hepatomegaly. From Table 6.14, the estimated effect is then the product of the estimated hazard ratios for rx and rx#hepatom, or $0.84 \times 1.1 = 0.92$. This estimate, along with a 95% CI (0.61–1.40) and p -value (0.70), can also be obtained using the `lincom` command shown in Table 6.14.

It follows that the interaction hazard ratio $\exp(\beta_3)$ gives the ratio of the DPCA treatment effects among patients with and without hepatomegaly. In Table 6.14, the estimated hazard ratio for rx#hepatom is 1.1 (95% CI 0.51–2.4, $P = 0.81$). The Z -test of $H_0: \beta_3 = 0$ assesses the equality of the effects of DPCA in the two groups.

To interpret these negative findings fully, as discussed in Sect. 3.7, both the point estimates and CIs need to be considered. The stratum-specific treatment effect estimates as well as the interaction are weakly negative, in the sense that the point estimates represent almost no effect or interaction, but the confidence limits include fairly large effects. In view of the weak evidence for interaction, the overall—also negative—finding for treatment with DPCA is the more sensible summary. Similar methods can be used to obtain estimates of the effect of hepatomegaly stratified by treatment assignment: that is, by comparing groups 3 and 1, then 4 and 2.

Interactions involving continuous or multilevel categorical predictors can also be set up using the # and ## operators, but as Sect. 4.6 explains, care must be taken with these more complex cases.

6.2.11 Model Building

Model building with the Cox model is similar to other regression models. Chapter 10 discusses the issues and makes recommendations. To prevent erosion of efficiency as well as bias, models should avoid including too many predictors for the number of observed events. A familiar guideline (Peduzzi et al. 1995, 1996; Concato et al. 1995) prescribes at least ten events per predictor. Vittinghoff and McCulloch (2007) show that as few as five events per predictor may give consistent results in cases where the additional covariates are needed to rule out confounding, but point out that precision in this case may often be poor.

6.2.12 Adjusted Survival Curves for Comparing Groups

Suppose we would like to examine the survival experience of pediatric recipients of kidney from living as compared to recently deceased donors, using the UNOS data. Kaplan–Meier curves, introduced in Sect. 3.5.2, would be a good place to start and are shown in Fig. 6.4.

In accord with the hazard ratio of 2.1 estimated by the unadjusted Cox model shown in Table 6.3, the curves show superior survival in the group with living

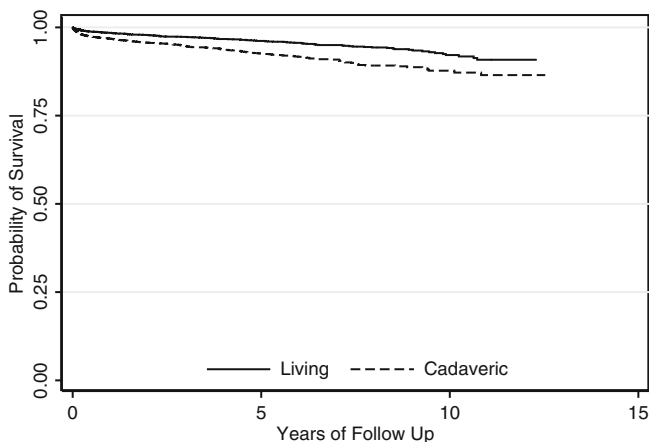


Fig. 6.4 Kaplan-Meier curves for transplant recipients by donor type

donors. However, there are two potentially important confounders of this effect. First, living donors are more likely to be related and thus are closer tissue matches, as reflected in the number of matching human leukocyte antigen (HLA) loci (range 0–6). Second, cold ischemia time (essentially the time spent in transport) is shorter for kidneys obtained from living donors. After adjustment for these two factors, the hazard ratio for donor type is reduced to 1.3 (95% CI 0.9–1.9, $P = 0.19$).

To see how adjusted survival curves might be constructed, first recall that adjustment for these covariates implies that adjusted curves for the two groups should differ only by donor type, with the other covariates being held constant. Curves meeting these criteria can be obtained using the coefficient estimates from the Cox model and an estimate of the baseline survival function, $\hat{S}_0(t)$, based on the Breslow baseline hazard estimate described earlier. Like the baseline hazard, the baseline survival function refers to observations with all predictor values equal to zero. If we assume a proportional hazard model, then a formula which links hazard and survival functions implies, the survival function follows:

$$\left\{ \hat{S}_0(t) \right\}^{\exp(\hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p)}. \quad (6.9)$$

That is, we raise the baseline survival to the $\exp(\hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p)$ power. To evaluate (6.9), we need to specify a value for each of the predictors. In our example with three predictors, we would need to choose and hold constant values for x_2 (cold ischemia time) and x_3 (number of matching HLA loci), then generate the two curves by varying the predictor x_1 (recently deceased versus living donor).

It is conventional to use values for the adjustment variables which are close to the “center” of the data. Thus we centered cold ischemia time at its mean value of 10.8 h and number of matching variable HLA loci at its median, three. With

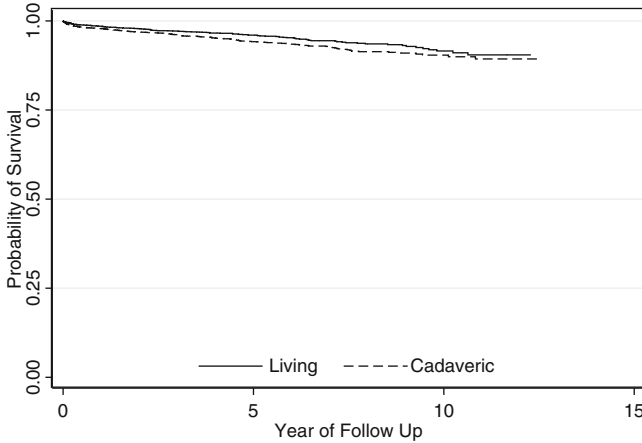


Fig. 6.5 Adjusted survival curves for transplant recipients by donor type

this centering, the baseline hazard and survival functions now refer to observations with cold ischemia time of 10.8h, three matching HLA loci, and a living donor. Then our adjusted estimate of the survival function for the group with living donors, holding the covariates constant at the chosen values, is $\hat{S}_0(t)$, while the corresponding estimate for the group with recently deceased donors is $\{\hat{S}_0(t)\}^{\exp(\hat{\beta}_1)}$. These adjusted curves, obtained in Stata using the `stcurve` command, are shown in Fig. 6.5. The differences between the survival curves are, as expected, narrower after adjustment. Note that the adjusted survival curves could also be estimated using a stratified Cox model, as discussed in Sect. 6.3.2.

6.2.13 Predicted Survival for Specific Covariate Patterns

The estimated survival function (6.9) is also useful for making predictions for specific covariate patterns. For example, consider predicting survival for a PBC patient based on hepatomegaly status and bilirubin level, the two strongest predictors in the model shown in Table 6.12. Figure 6.6 displays the predicted survival curve for a PBC patient with hepatomegaly and a bilirubin level of 4.5 mg/dL. From the curve, the median survival function for this covariate pattern is 6.3 years. Survival probabilities at key time points can likewise be read from the plot: at 5 years, predicted survival for this covariate pattern is below 60%, and by 10 years, it has dropped to less than 20%. However, mean survival cannot be estimated in this case, because the longest follow-up time in the PBCA data is censored (Sect. 3.5).

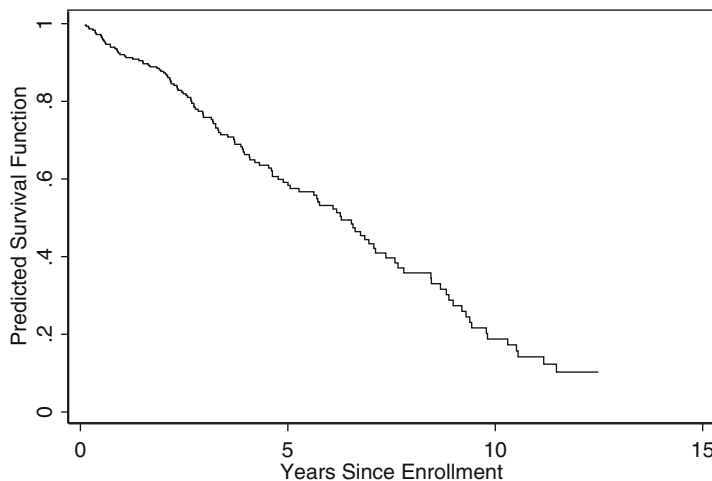


Fig. 6.6 Predicted survival curve for PBC covariate pattern

6.3 Extensions to the Cox Model

6.3.1 Time-Dependent Covariates

So far we have only considered fixed predictors measured at study baseline, such as bilirubin in the DPCA study. However, multiple bilirubin measurements were made over the 10 years of follow-up, and these could provide extra prognostic information. A special feature of the Cox model is that these valuable predictors can be included as TDCs.

Definition: A *time-dependent covariate* in a Cox model is a predictor whose values may vary with time.

In some cases, use of TDCs is critical to obtaining reasonable effect estimates. For example, [Aurora et al. \(1999\)](#) followed 124 patients to study the effect of lung transplantation on survival in children with cystic fibrosis. The natural time origin in this study is the time of *listing* for transplantation, not transplantation itself, because the children are most comparable at that point. However, waiting times for a suitable transplant can be long, and there is considerable mortality among children on the waiting list.

In this context, lung transplantation has to be treated as a TDC. To see this, consider the alternative in which transplantation is modeled as a fixed binary covariate, in effect comparing mortality risk in the group of children who undergo transplantation during the study to risk among those who do not. This method can

make transplantation look more protective than it really is. Here is how the artifact, sometimes called *immortal time bias* (Suisse 2008), comes about:

- Because transplanted patients must survive long enough to undergo transplantation, and waiting times can be long, the survival times measured from listing forward will on average be longer in the transplanted group even if transplantation has no protective effect.
- Because of this, children in the transplanted group are selected for better prognosis. So the randomization assumption discussed in Sect. 9.1.4 does not hold.
- Children are counted as having received a transplant from the time of listing forward, in many cases well before transplantation occurs. As a result they appear to be protected by a procedure that has not yet taken place. This illustrates the general principle that we can get into trouble by using information from the future to estimate current risk.

Treating transplantation as a TDC avoids this artifact. For each child, we define an indicator of transplantation $X(t)$, which takes on value 0 before transplantation and 1 subsequently. For children who are not observed to undergo transplantation, $X(t)$ retains its original value of 0. Thus in an unadjusted model, the hazard at time t can be written as

$$\begin{aligned} h(t|x) &= h_0(t) \exp\{\beta X(t)\} \\ &= \begin{cases} h_0(t) & \text{before transplantation} \\ h_0(t) \exp(\beta) & \text{at or after transplantation.} \end{cases} \end{aligned} \quad (6.10)$$

So now, all children are properly classified at t as having undergone transplantation or not, and we avoid the artifact that comes from treating transplantation as a fixed covariate. Note that Kalbfleisch and Prentice (1980) cite additional conditions concerning the allocation of transplants that must be met for the randomization assumption to hold and an unbiased estimate of the effect of transplantation to be obtained.

The transplantation TDC is relatively simple, because it is binary and cannot change back in value from 1 to 0. In practice, however, use of TDCs in Cox models is often more complicated. Some additional considerations include the following:

- In most prospective studies, predictors like bilirubin will only be measured occasionally, but we need a value at each event time. A commonly used approach is to evaluate $X(t)$ using the most recent measurement before t , but this so-called *last observation carried forward* (LOCF) approach is susceptible to bias; we return to this in Chap. 11. More difficult is a two-stage approach in which we first model the mean trajectory of the TDC for each subject. Then in the second stage we can set $X(t)$ equal to its expected value at t , based on the first-stage model. However, fitting and inference are both complicated in this procedure (Self and Pawitan 1992; DeGruttola and Tu 1994; Wulfsohn and Tsiatis 1997; Tsiatis and Davidian 2004).

- While $X(t)$ cannot legitimately be evaluated using information from the future, it often should be evaluated using all available information up until t . Consider two PBC patients, one with bilirubin values of 0.8 and 3.5 at baseline and year two, and the other with values of 2.5 and 3.5 at those times. In evaluating a TDC for bilirubin at year two, it might not be adequate to account only for the most recent values. A commonly used approach is to include the baseline value as a fixed covariate along with the change since baseline as a TDC. But other combinations of baseline and TDCs summarizing history up to t may be more appropriate.
- Mediation can be evaluated using TDCs, extending the analysis of mediation of the effect of alendronate on fracture by first year changes in BMD, treated as a fixed covariate, as discussed in Sect. 6.2.9. For example, we could examine mediation of the effects of ZDV via its effects on CD4 counts in the ACTG 019 trial by assessing both links in the hypothesized indirect pathway. Specifically, we might use a model for repeated measures, covered in Chap. 7, to assess ZDV effects on CD4 counts over time, and then assess the independent effects of post-randomization CD4 values in a Cox model for AIDS-free survival, controlling for treatment. Finally, we might informally compare the effect estimates for ZDV before and after adjustment for post-randomization CD4 counts.
- Special methods are needed if a TDC both confounds and mediates the effects of a time-dependent exposure or treatment. Suppose we wanted to evaluate the overall effect of highly active anti-retroviral therapy (HAART) on progression to AIDS, using data from an observational cohort. To avoid immortal time bias, HAART would need to be modeled as a TDC. Now suppose we attempt to control confounding by disease severity at treatment initiation by adjusting for time-dependent prognostic measures including CD4 count. The problem is that the effects of HAART on progression to AIDS are also *mediated* via its effects on CD4 count, so this would adjust away some of the protective effect of treatment. As a result, we would not obtain an estimate of the *overall* treatment effect. In Sect. 9.5, we discuss a solution to this problem using *IPW*.
- Ideally TDCs are measured at regularly scheduled visits, so ascertainment does not depend on prognosis. Missing visits can induce bias if the missingness is related to the value of the TDC that would have been obtained. Likewise, ascertainment of TDCs by clinical chart review can be fraught with pitfalls.
- In Stata, accommodating TDCs like the post-randomization CD4 counts in the ACTG 019 example requires a specially constructed dataset with multiple records for each unit. The `stsplit` and `stjoin` commands make this straightforward. In Sect. 6.4.2, we also show how the `stcox` option `tvc` accommodates a different kind of TDC, specifically interactions between a fixed covariate and time, which are useful in dealing with violations of the *proportional hazards assumption*.

6.3.2 Stratified Cox Model

Suppose we want to model the effect of edema (coded 1 for patients with edema and 0 for others) among patients with PBC in the DPCA cohort. Then in an unadjusted model, the hazard for patients with edema is $h(t|x) = h_0(t) \exp(\beta)$, while for other patients it is just $h_0(t)$. So the hazard for patients with edema is modeled as a constant proportion $\exp(\beta)$ of the baseline hazard $h_0(t)$.

However, we will show in Sect. 6.4.2 that the proportional hazards assumption does not hold for edema. We can accommodate the violation by fitting a stratified Cox model in which a separate baseline hazard is used for patients with and without edema. Specifically, we let

$$h(t|\text{edema} = 1) = h_{01}(t) \quad (6.11)$$

for patients with edema, and

$$h(t|\text{edema} = 0) = h_{00}(t) \quad (6.12)$$

for other patients. Now the hazards for the two groups can differ arbitrarily.

Generalizing from edema to a stratification variable with two or more levels, and to a model with covariates $(x_1, \dots, x_p)$, the hazard for an observation in stratum j would have the form

$$h_{0j}(t) \exp(\beta_1 x_1 + \dots + \beta_p x_p). \quad (6.13)$$

Note that in this model we assume that the effect of each of the covariates is the same across strata; later, we examine methods for relaxing this assumption. It is also important to point out that while the stratified, adjusted survival curves presented in Sect. 6.2.12 above can give a clear visual impression of the effect of the stratification variable after adjustment, current methods for the stratified Cox model do not allow us to estimate or test the statistical significance of its effect. Thus stratification could be used in our example to adjust for edema, but might be less useful if edema were a predictor of primary interest. In Sect. 6.4.2, we show how TDCs can be used to obtain valid estimates of the effects of a predictor which violates the proportional hazards assumption. Stratification is also useful in the analysis of stratified randomized trials. We pointed out in Sect. 10.2.6 that we need to take account of the stratification to make valid inferences. But we also need to avoid making an unwarranted assumption of proportional hazards for the stratification variable that could potentially bias the treatment effect estimate. The stratified Cox model is easy to implement in Stata as well as other statistical packages. In ACTG 019, participants were randomized within two strata defined by baseline CD4 count. To conduct the stratified analysis, we defined `strcd4` as an indicator coded 0 for the stratum with baseline CD4 count of 200–499 cells/mm³ and 1 for the stratum with baseline CD4 of less than 200. The stratified model for the

Table 6.16 Cox model for treatment with ZDV, stratified by baseline CD4

stcox i.rx, strata(strcd4)					
Log-likelihood = -276.45001			LR chi2(1)	=	7.36
			Prob > chi2	=	0.0067

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]

1.rx	.4646665	.1362697	-2.61	0.009	.2615261 .8255963

Stratified by strcd4					

effect of ZDV treatment (rx) is shown in Table 6.16. In this instance, the estimated 54% reduction in risk for treatment with ZDV is the same as an estimate reported below in Sect. 6.6.3, which was adjusted for rather than stratified on CD4.

6.3.2.1 Number of Strata

Stratification is a flexible approach to adjustment for a categorical variable even when it has a large number of levels. An example is in a multicenter randomized trial with many centers. For stratification to work well, there do need to be a reasonable number of events (about 5 to 7) in each stratum. When the number of strata gets large, there can be some loss of efficiency in estimation of the treatment or other covariate effects, since the stratified model does not “borrow strength” across strata. Nonetheless, Glidden and Vittinghoff (2004) showed that in this situation, the stratified Cox model performs better than an unstratified model in which the strata are entered into the model as a nominal categorical predictor.

6.3.2.2 Interaction Between Stratum and a Predictor of Interest

In Table 6.16, the model assumes that the ZDV effect is the same in both strata. It is possible, however, that patients with less severe HIV disease, as reflected in higher CD4 counts, may respond better to ZDV. Such an interaction between stratum and treatment can be examined by including a product term between the treatment and stratum indicators. Note that in the stratified model only the product term i.rx#strcd4 and the treatment indicator rx term are entered as predictors. The predictor strcd4 is dropped automatically by Stata, because it has already been incorporated as a stratification factor. In Table 6.17, we find persuasive evidence of an effect of ZDV (rx = 1) in the higher CD4 stratum (strcd4 = 0) with hazard ratio of 0.32 (95% CI 0.14–0.74, P = 0.008). However, from the lincom result, we derive the effect of ZDV in the lower CD4 stratum (strcd4 = 1) where there is weak evidence for a protective effect of ZDV (hazard ratio 0.71, 95% CI 0.32–1.65, P = 0.43). There is the suggestion for interaction (hazard ratio 0.45, 95% CI 0.14–1.48, P = 0.19), given by the product term rx#strcd4, although this is not statistically significant.

Table 6.17 Stratified fit with interaction term

```
. stcox i.rx##i.strcd4, strata(strcd4)
```

No. of subjects =	880	Number of obs =	880
No. of failures =	55		
Time at risk =	354872		
Log-likelihood =	-275.56324	LR chi2(2) =	9.14
		Prob > chi2 =	0.0104

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
1.rx	.3211889	.136976	-2.66	0.008	.1392362	.7409156
1.strcd4	(omitted)					
rx#strcd4						
1 1	2.218026	1.342113	1.32	0.188	.677501	7.261448

```
. lincom 1.rx+1.rx#1.strcd4, hr
```

(1) 1.rx + 1.rx#1.strcd4 = 0

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
(1)	.7124052	.305808	-0.79	0.430	.307142	1.652399

Stratified by strcd4

6.3.2.3 Stratified and Adjusted Survival Curves

In Sect. 6.2.12, we presented adjusted survival curves for pediatric kidney transplant recipients according to donor type, based on an adjusted model in which the effect of donor type was modeled as proportional. We can also obtain adjusted survival curves according to the levels of a stratification factor. We will show in Sect. 6.4.2 that the effects of baseline edema on mortality risk among PBC patients in the DPCA cohort were not proportional. Suppose we would like to compare the survival curves according to edema, adjusting for age. As in the earlier example, we need to specify a value for age in order to estimate the survival curves, and make a similar choice in centering age on its mean of 50. Under the stratified Cox model, the survivor function for a PBC subject with centered `agec` is given by

$$[S_{0j}(t)]^{\exp(\beta_{agec})}.$$

(6.14)

The adjusted survival curves for the edema ($j = 1$) and no edema ($j = 0$) strata, adjusted to age 50 (i.e., `agec` = 0), are therefore $S_{01}(t)$ and $S_{00}(t)$, respectively. Figure 6.7 shows shorter survival in patients with edema at baseline. However, these stratum-specific survival functions also suggest that the multiplicative effect of edema on the mortality hazard is not constant over time. We examine this more carefully in Sect. 6.4.2.

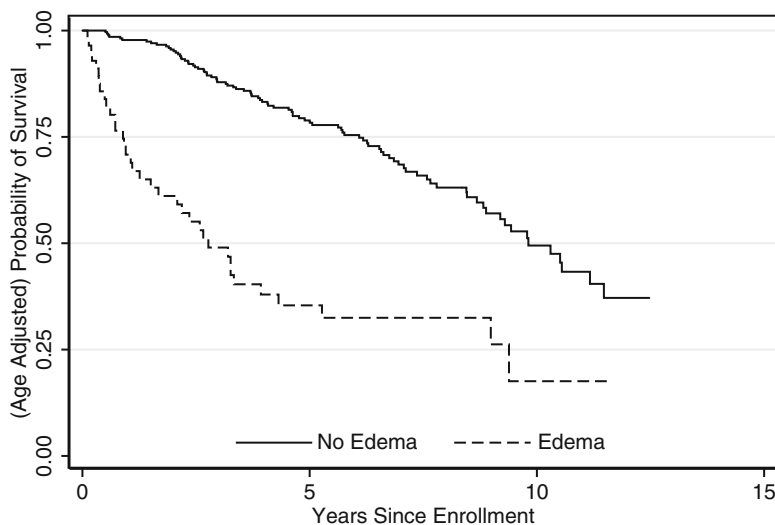


Fig. 6.7 Stratified survival curves for edema adjusted for age

6.4 Checking Model Assumptions and Fit

Two basic assumptions of the Cox model are *log-linearity* and *proportional hazards*. Just as with other regression models, these assumptions can be examined, and extensions of the model can deal with violations and model more complex effects.

6.4.1 Log-Linearity of the Hazard Function

In Sect. 6.2.1, (6.6) specifies that each unit change in a continuous predictor has the same effect on the log of the hazard. This implies that the hazard ratio is log-linear in continuous predictors.

Unlike the linear model, but like the logistic, diagnostics for violations of log-linearity using plots of residuals do not work very well for the Cox model. However, violations of this assumption are easy to detect and accommodate with the tools covered in Sect. 4.7.1 for the linear model. The approach is simple: attempt more general models and examine improvements in fit.

Like other models, the Cox model can be generalized by adding polynomial terms for the predictor in question to the model. Effect sizes and p -values are then checked to determine whether the higher order terms are important; or the predictor can be log-transformed and the log-likelihoods informally compared (Problem 6.4). Alternatively, the continuous predictor can be categorized using well-chosen cut-points; then log-linearity is checked using the methods outlined above in Sect. 6.2.2

for assessing both trend and departures from trend in ordinal predictors. These approaches have limitations: a susceptibility to outliers for polynomial models and sensitivity to the number and placement of the cutpoints for categorizations.

Restricted cubic splines are an alternative approach offering flexibility with relative parsimony. These methods, discussed in Sect. 4.7.1, lay down a series of “knots” along the values of the predictor and fit a polynomial curve between them—allowing for a wide variety of shapes. Consider the relationship between age and hazard of death for the PBC dataset. Using a spline fit, we could detect a nonlinear pattern between age and mortality. Unlike categorization of a continuous predictor, splines are not greatly sensitive to number and placement of knots. Three to five knots provide a great deal of flexibility. The choice of the number of knots is a balance between the sample size and the degree of flexibility desired. A further advantage is the similarity of implementation across diverse regression models. First, a spline *basis* is derived—in Stata this uses the `mkspline` command. This basis comprises $k - 1$ predictors, where k is the number of knots. These predictors then take the place of the continuous variable in the regression model.

Table 6.18 uses the commands and output for splines in a Cox model. Note, the similarity to the application of splines linear model in Sect. 4.7.1. Two test statements appear in Table 6.18. The tests suggest strong support for an overall effect of age on survival (given by the $p = 0.0004$) but find no evidence that the spline model fits the data better than a log-linear term in age (given by the $p = 0.99$). A p -value alone should not be used for model selection; hence, we compare the log-linear and spline fits graphically to examine the magnitude of the differences. Figure 6.8 shows the fits compared with a categorical fit placing cutpoints at the knots. The linear and restricted spline fit agree closely, suggesting a log-linear model fit age reasonable well.

6.4.2 Proportional Hazards

The adjusted Cox model shown in Table 6.12 shows that mortality risk is increased about twofold in PBC patients with edema at baseline. However, Fig. 6.7 suggests that edema may violate the proportional hazards assumption: specifically, the hazard ratio in edema is greatest in the first few years and then diminishes. Thus the effect of edema on the hazard is time-dependent. A transformed version of Fig. 6.7 turns out to be more useful for examining violations of the proportional hazards assumption.

6.4.2.1 Log-Minus-Log Survival Plots

To illustrate the use of transformed survival plots for assessing proportionality for binary or categorical predictors, we consider the treatment indicator (rx) in the DPCA trial. This method exploits the relationship between the survival and hazard functions. If proportional hazards hold for rx , then by (6.9)

Table 6.18 Restricted cubic spline Cox model for the effect of age on mortality

```

. mkspline age_sp = age, cubic

. stcox age_sp1 age_sp2 age_sp3 age_sp4

Cox regression -- Breslow method for ties

No. of subjects =          312                Number of obs   =          312
No. of failures =           125                LR chi2(4)      =          20.59
Time at risk    =  1713.853528                Prob > chi2       =          0.0004

Log-likelihood   =  -629.68657

-----+-----
      _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
    age_sp1 |   1.049751   .0702359    0.73   0.468   .9207355   1.196845
    age_sp2 |   .9849075   .3583548   -0.04   0.967   .482713   2.009564
    age_sp3 |   .9746888   1.345662   -0.02   0.985   .0651166   14.5895
    age_sp4 |   1.17647    2.203381    0.09   0.931   .0299493   46.21414
-----+-----

. * test for departure from linearity

. test age_sp2 age_sp3 age_sp4

( 1)  age_sp2 = 0
( 2)  age_sp3 = 0
( 3)  age_sp4 = 0

      chi2( 3) =    0.08
    Prob > chi2 =    0.9943

. * test for overall effect

. test age_sp1 age_sp2 age_sp3 age_sp4

( 1)  age_sp1 = 0
( 2)  age_sp2 = 0
( 3)  age_sp3 = 0
( 4)  age_sp4 = 0

      chi2( 4) =   20.54
    Prob > chi2 =    0.0004

```

$$S_1(t) = [S_0(t)]^{\exp(\beta)}, \quad (6.15)$$

where $S_0(t)$ is the survival function for placebo patients and $S_1(t)$ is the corresponding survival function for the DPCA-treated patients. Then, the *log-minus-log* transformation of (6.15) gives

$$\log\{-\log[S_1(t)]\} = \beta + \log\{-\log[S_0(t)]\}. \quad (6.16)$$

Thus when proportional hazards hold, the two transformed survival functions will be a constant distance β apart, where β is the log of the hazard ratio for treatment with DPCA. This approach assumes a categorical variable but can be adapted to a continuous variable by, for instance, factoring a continuous variable into quartiles.

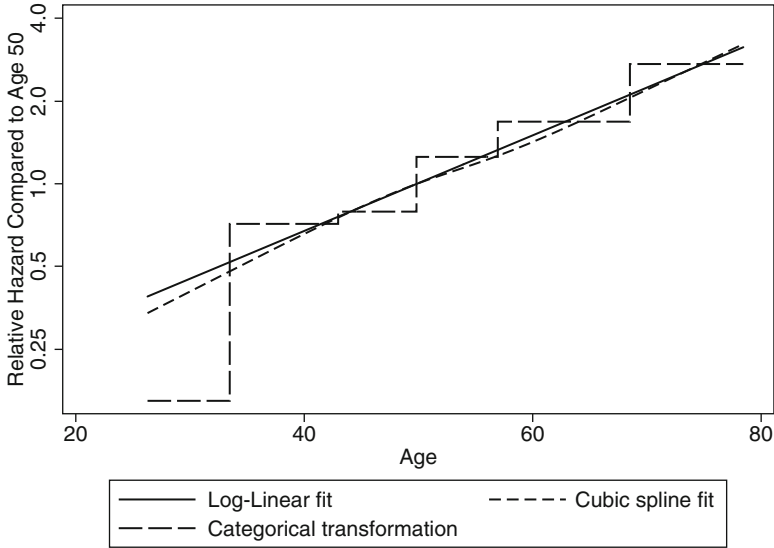


Fig. 6.8 Cox model fit to PBC data using a log-linear fit, restricted cubic spline, and categorical transformation of age

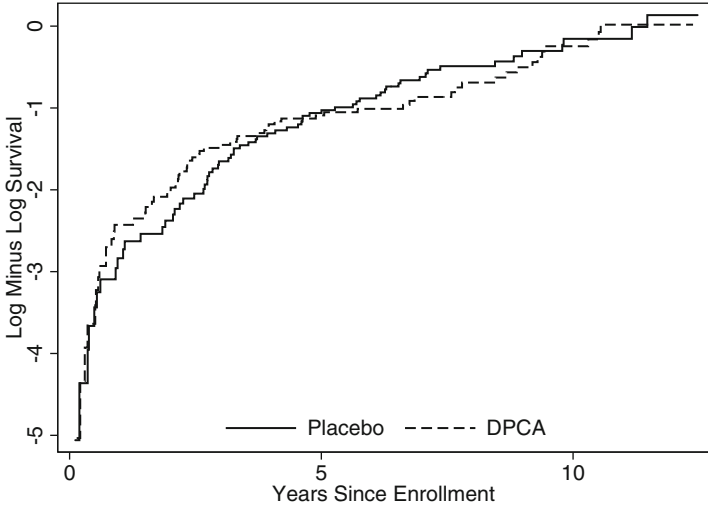


Fig. 6.9 Log-minus-log survival plot for DPCA treatment

This result enables us to use a simple graphical method for examining the proportional hazards assumption. Specifically, log-minus-log-transformed Kaplan–Meier estimates of the survival functions for the placebo and DPCA groups are plotted against follow-up time. In Stata, this plot is implemented in the `stphplot` command. The log-minus-log survival plot for DPCA is shown in Fig. 6.9.

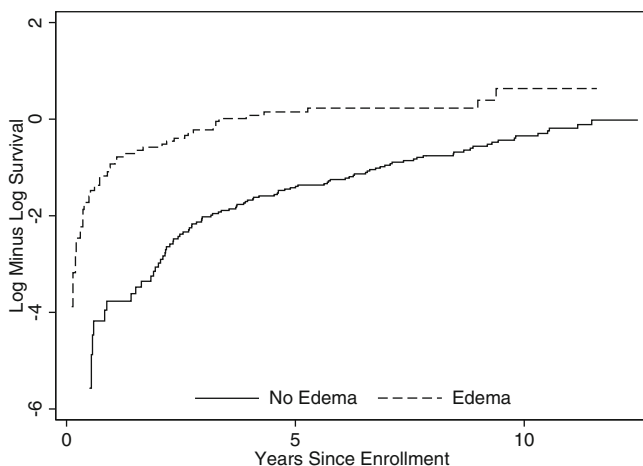


Fig. 6.10 Log-minus-log survival plot for edema

In assessing the *log-minus-log survival* plot for evidence of nonproportional hazards, the patterns to look for are convergence, divergence, or crossing of the curves. Converging curves suggest that the difference between the groups decreases with time; diverging curves suggest that differences increase with time. If the curves show pronounced crossing, then the nonproportionality may be more important; for example, this might indicate that treatment is harmful early on but protective later. In Fig. 6.9, however, the curves for DPCA and placebo remain close over the entire follow-up period and do not suggest nonproportionality.

In contrast, the log-minus-log survival plot for edema in Fig. 6.10 shows rather clear evidence of a violation of proportionality. While there is a pronounced difference between the groups at all time points, showing that patients with edema have poorer survival, the difference between the groups diminishes with follow-up. Specifically, the distances between the curves—that is, the implied log-hazard ratios—are 4.7, 1.8, 1.1, and 1.0 at years 1, 4, 7, and 10, respectively.

6.4.2.2 Smoothing the Hazard Ratio

Log-minus-log survival plots are good diagnostic tools for violations of the proportional hazards assumption. To address such a violation, however, we may need more information about how the log-hazard ratio changes with follow-up time. We can do this using a nonparametric, smoothed estimate of the hazard ratio against time, analogous to the LOWESS estimates of the regression function used in diagnosing problems in linear models in Sect. 4.7. If the smoothed estimate of the hazard ratio is nearly constant, then the assumption of proportional hazards is approximately satisfied. Conversely, when curvature is pronounced, the shape of the smooth line helps us determine how to model the hazard ratio as a function of time. In Stata,

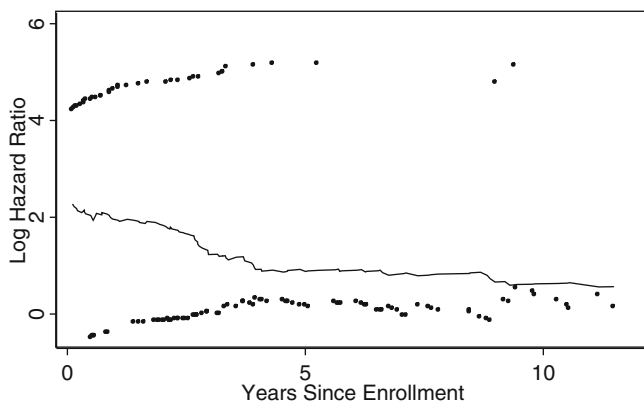


Fig. 6.11 Smoothed estimate of log-hazard ratio for edema

the plot can be generated using the `estat phtest` command with the `plot` option. Figure 6.11 shows the smoothed estimated plot of the hazard ratio over time for edema. A nonconstant trend is readily apparent: the log-hazard ratio decreases steadily over the first 4 years and then remains constant. This works by smoothing a specialized type of residual, *scaled Schoenfeld residuals*, for each predictor against time using LOWESS. The residuals appear as points in the plot. Smoothing them against time provides a nonparametric estimate of the log-hazard ratio for that predictor as it changes over time. An advantage of this approach is that it works for both categorical and continuous variables.

Relatively influential points are identifiable from the plots of the Schoenfeld residuals. DFBETA statistics, a measure of how much coefficients are changed by the deletion of individual observations (see Sect. 4.7.4 for an illustration of their applications in linear models), can be obtained for the Cox model in Stata by using the `predict` command.

6.4.2.3 Schoenfeld Test

Schoenfeld (1980) provides a test for violation of proportional hazards which is closely related to the diagnostic plot using LOWESS smooths of scaled Schoenfeld residuals just described. The test assesses the correlation between the scaled Schoenfeld residuals and time. This is equivalent to fitting a simple linear regression model with time as the predictor and the residuals as the outcome, and the parametric analog of smoothing the residuals against time using LOWESS. If the hazard ratio is constant, the correlation should be zero.

The Schoenfeld tests for `rx` and `edema` are shown in Table 6.19. Positive values of the correlation ρ suggest that the log-hazard ratio increases with time and vice versa. In accord with the graphical results, the Schoenfeld test finds strong evidence for a declining log-hazard ratio for edema ($\rho = -0.36$, $P = 0.0001$), but does not suggest problems with `rx` ($\rho = -0.07$, $P = 0.5$).

Table 6.19 Schoenfeld tests of proportional hazards assumption

```
. estat phtest, detail
```

Test of proportional-hazards assumption

Time: Time				
	rho	chi2	df	Prob>chi2
rx	-0.05862	0.43	1	0.5129
edema	-0.36107	14.63	1	0.0001
global test		14.71	2	0.0006

The Schoenfeld test is most sensitive in cases where the log-hazard ratio is linearly increasing or decreasing with time. However, because the test is based on a linear regression model, it is sensitive to a few large residual values. Such values should be evident on the scatterplot of the scaled Schoenfeld residuals against time. Useful examples and discussion of the application of the Schoenfeld test appear in Sect. 6.5 of [Therneau and Grambsch \(2000\)](#).

6.4.2.4 Graphical Diagnostics Versus Testing

We have described both graphical and hypothesis testing methods for examining the proportional hazards assumption. The Schoenfeld test is widely used and gives two easily interpretable numbers that quantify the violation of the proportional hazards assumption. However, as pointed out in Sect. 4.7, such tests may lack power to detect important violations in small samples, while in large samples they may find statistically significant evidence of model violations which do not meaningfully change the conclusions. While also lacking sensitivity in small samples, graphical methods give extra information about the magnitude and nature of model violation, and should be the first-line approach in examining the fit of the model.

6.4.2.5 Stratification

The stratified Cox model introduced in Sect. 6.3.2 is an attractive option for handling binary or categorical predictors which violate the proportional hazards assumption. We explained there that no assumption is made about the relationships between the stratified hazard functions specific to the different levels of the predictor. Because the resulting fit to the stratification variable is unrestricted, this is a particularly good way to rule out confounding of a predictor of interest by a covariate that violates the proportional hazards assumption. However, because no estimates, CIs, or *p*-values are obtained for the stratification variable, this approach is less useful for any predictor of direct interest.

Note that we can apply this approach to a continuous variable by first categorizing it. How many categories to use involves a trade-off (Problem 6.9). Using more strata more effectively controls confounding, but as we suggested in Sect. 6.3.2, precision and power can suffer if the confounder is stratified too finely, because strength is not borrowed across strata. Five or six strata generally suffice, but there should be at least 5–7 events per stratum.

6.4.2.6 Modeling Interactions with Time

In this section, we briefly outline a widely used approach to addressing violations of the proportional hazards assumption using interactions with time, and implemented using TDCs, as described above in Sect. 6.3.1. We return to the edema example and show how the declining hazard ratio can be modeled. To begin, let $h_1(t)$ and $h_0(t)$ denote the hazard functions for PBC patients with and without edema. Because proportional hazards does not hold, the hazard ratio

$$\text{HR}(t) = \frac{h_1(t)}{h_0(t)} \quad (6.17)$$

is a function of t . To address this, we define $\beta(t) = \log\{\text{HR}(t)\}$ as a coefficient for edema which changes with time. This is equivalent to a hazard function of the form

$$h(t|\text{edema}) = h_0(t) \exp\{\beta(t)\text{edema}\}, \quad (6.18)$$

where as before edema is a 0/1 indicator of the presence of edema. This can be modeled in one of two ways.

- We can model the log-hazard ratio for edema as a linear function of time. This is implemented using a main effect, edema , plus an interaction term, edemat , defined as a TDC, the product of edema and t . That is, we set

$$\begin{aligned} \beta(t)\text{edema} &= (\beta_0 + \beta_1 t)\text{edema} \\ &= \beta_0\text{edema} + \beta_1 t\text{edema} \\ &= \beta_0\text{edema} + \beta_1\text{edemat}. \end{aligned} \quad (6.19)$$

Alternatively, we could model the log-hazard ratio as linear in log time, defining the product term with $\log(t)$ in place of t ; this might be preferable in the edema example, since the decline in the log-hazard ratio shown in Fig. 6.11 grows less steep with follow-up (Sect. 4.7.1).

- We can split follow-up time into sequential periods and model the log-hazard ratio for edema as a step function with a different value in each period. For example, we could estimate one log-hazard ratio for edema in years 0–4, and

another in years 5–10, again motivated by Fig. 6.11. We could do this by defining two TDCs:

- `edema04`, equal to 1 during the first 4 years for patients with edema, and 0 otherwise.
- `edema5on`, equal to 1 during subsequent follow-up for patients with edema, and 0 otherwise.

Then we set

$$\beta(t)\text{edema} = \beta_1\text{edema04} + \beta_2\text{edema5on}. \quad (6.20)$$

This approach is analogous to categorizing a continuous predictor to model nonlinear effects (Sect. 4.7.1).

The first alternative is more realistic because it models the hazard ratio for edema as a smooth function of time. But it is harder to implement because the TDC `edemat` changes continuously for patients with edema from randomization forward; up to one record for every distinct time at which an outcome event occurs would be required for these patients in the “long” dataset used for the analysis in Stata, now easily obtained using the `stsplit` and `stjoin` commands. In contrast, the second alternative is less realistic but easier to implement, only requiring two records for patients with edema and more than 4 years of follow-up, and one record per patient otherwise. See Sect. 6.9 for discussion of another flexible approach.

6.5 Competing Risks Data

6.5.1 What Are Competing Risks Data?

The MrOS study (Orwoll et al. 2005) followed 5,993 men over the age 65 and examined predictors of bone fracture and low BMD (evidence of subclinical bone loss). At enrollment, all men underwent a dual X-ray absorptiometry (DEXA) scan to determine their BMD and were followed for an average of 5 years for risk of bone fracture. At the conclusion of follow-up, 531 participants had developed fracture, 4,805 remained alive without fracture and 657 had died prior to fracture. An important question is how well a baseline BMD measure predicts fracture risk over the follow-up period.

There are two possible sources of incomplete follow-up: (1) the end of the observation (due to loss to follow-up or short observation times due to staggered entry) and (2) death. To understand why it is important to distinguish between them, consider how our methods have handled incomplete follow-up. The approach that has been used so far in this chapter attempts to project forward the experience of a censored observation by representing their experience with those followed longer. Embedded in this approach is an assumption and an objective. The assumption

is the *independent censoring* assumption (see Sect. 6.6.4) that the future risk of those whose follow-up has ended can be represented by those who are followed longer (see Sect. 3.5.2 for a discussion of this in the context of the Kaplan–Meier survivor function). The implicit objective is to make an extrapolation to a setting in which the source of incomplete follow-up is eliminated. For incomplete follow-up due to patient dropout, the assumption may be suspect but the objective is highly relevant since we would like to estimate what would have happened if people had been followed completely. For death, both the assumption and the objective are in question. To extrapolate to a setting where death is not possible would be to project a new population or the ability to extend lives—altering the underlying conditions of the study. Instead we could acknowledge death as another possible outcome which can cut short the observation of fracture without attempting to project fracture experience beyond participant lifetimes. Thus, objectives and approaches to incomplete follow-up may differ depending on whether it is due to death or the inability of a study to retain or follow participants.

Definition: Competing risks data arise when multiple events can occur and follow-up can end due to occurrence of one or more of those types of events, precluding observation of at least one of the other event types.

The definition given is the most expansive possible definition. It could cover a situation in which observations are cut short due to patient dropout or due to end-of-study censoring. In the analysis of competing risks data, two major approaches can be taken—one which seeks to extrapolate to a scenario in which a type of event is not possible (typically, loss to follow-up or incomplete follow-up due to staggered entry). We call this approach *elimination*. Another family of methods is based on acknowledging and allowing for the competing risks in the analysis. This approach will be called *accommodation*.

In our example, we can observe fractures, death, or losses to follow-up. The objective of the analysis is to estimate the risk of fracture (in the presence of death) where there is no loss to follow-up.

6.5.2 Notation for Competing Risks Data

We denote competing risks outcome data using two variables: one which denotes the time of the first event and the other which denotes the type of event. Let

- Y : be the time of the first observed event of any type
- $\Delta = k$ if the k th event type occurs first

where each of the K possible types of failure are denoted by a numerical code. In the MrOS dataset, the failure types were coded as 0: loss to follow-up, 1: fracture, and 2: death. Using this, a participant who is followed for 18 months and dies (prior to fracture) will have $Y=18$ months and $\Delta=2$. Note, that this same notation is standard for ordinary survival analysis where there are two possible events: censoring and failure.

6.5.3 Summaries for Competing Risk Data

Two important summaries are typically available for competing risk data: these are analogs of the hazard and survival function.

6.5.3.1 Cause-Specific Hazard Functions

Definition: The *cause-specific hazard function* for event type k , $h_k(t)$, is the short-term rate at which subjects experience the onset of the k th event among those who have not yet experienced the event of interest (e.g., fracture) or a competing event (e.g., death) prior to t .

A hazard function can be thought of as a short-term rate of failure. Rate functions are ratios defined by which events get counted (in the numerator) and who is included in the “at risk” population (in the denominator). The k th cause-specific hazard only counts events of type k in the numerator (e.g., number of fractures). The denominator includes follow-up for all people who could have developed the event by time t . In the MrOS example, the cause-specific hazard function for fracture at a given time t calculates the rate of fracture among all people who are alive, without fracture, and uncensored prior to time t . Follow-up is counted in the denominator (the “at risk” population) until fracture, death, or loss to follow-up. Note that the cause-specific hazard reduces to the ordinary hazard function in the case that there is only one type of failure.

Estimating and modeling cause-specific hazard functions is straightforward. Simply set up the data as ordinary survival data with the k th failure type as the only type of failure and treat competing causes (even death) as “censored.” Counterintuitively, the cause-specific hazard function’s calculation does not make a distinction between events which are accommodated versus those which we attempt to eliminate. This will be in contrast to the cumulative incidence function for which this distinction will be very important.

If the data are analyzed this way, it is possible to examine the effect of predictor like bone-mineral density on the cause-specific hazard of fracture using a standard Cox model-type formulation. This will be discussed further in Sect. [6.5.3.3](#).

6.5.3.2 Cumulative Incidence Functions

The extension of the hazard function to competing risks data is given by the cause-specific hazard functions. The extension of the survivor function is given by what is called the *cumulative incidence function*. The cumulative incidence function at time t for the event type k is the proportion of the sample who have experienced the k th event by time t . For instance, at 5 years the cumulative incidence estimate of fracture is 0.080 and for death it is 0.093. This implies that after 5 years, about 8.0% of the study population developed a fracture prior to death, that 9.3% died without fracture and the remaining 82.7% are alive without fracture.

Table 6.20 Deaths and fractures in the MrOS cohort

Months in cohort	No. in follow-up	No. fractured	No. died	No. lost to FU	Pr event-free
1	5,993	7	2	1	0.9985
2	5,983	7	5	0	0.9965
3	5,971	10	1	0	0.9947
4	5,960	5	3	0	0.9933
5	5,952	4	6	0	0.9917
6	5,942	8	6	1	0.9893
7	5,927	10	1	1	0.9875
8	5,915	11	4	1	0.9850
9	5,899	6	5	1	0.9831
10	5,887	4	3	1	0.9818

Definition: The *cumulative incidence function* for cause type k at time t , $F_k(t)$, is the proportion who have developed the k th event prior to t .

The cumulative incidence function is a measure of prevalence of a particular event at each time t for a population which started with none.

This would be easy to estimate if there are no competing causes which we plan to eliminate. If we wanted to calculate the estimated cumulative incidence function at 1 year, we would simply count the number of people who at 1 year were alive without fracture (n_0), had experienced a fracture (n_1), or had died without experiencing a fracture (n_2). The cumulative incidence of fracture at 1 year would be $n_1/(n_0 + n_1 + n_2)$.

However, with incomplete follow-up, estimation requires more care. It requires that we consider increments of time as we did in the calculation of the Kaplan–Meier survivor estimate in Table 6.1.

Table 6.20 traces the first 10 months of the MrOS follow-up period and gives a summary by month. The second column is the number of participants who start the month alive, unfractured and under follow-up. The third through fifth column gives the number of fractures, losses to follow-up and deaths in that month, respectively. The final column is the estimated fraction of the proportion of the cohort who are alive and free of fracture at the end of each month. This quantity can be calculated by the Kaplan–Meier method by combining death and fracture into a single event.

Table 6.21 shows the cumulative incidence of fracture for the MrOS cohort during the first 10 months of follow-up. The cumulative incidence function for each month is the cumulative sum over the time-specific probabilities of a new fracture. This probability of a new fracture in a given month is the probability that someone is alive and fracture free at the beginning of the month (note that this is the probability from the end of the prior month) and develops a fracture during the month. For instance, the probability of an incident fracture in month seven is the chance of being alive and unfractured at the end of month 6 (0.9893) times the rate of failure in the seventh month. This rate is simply the number of fractures during the month divided by the number of participants under follow-up during the seventh month—

Table 6.21 Estimating the cumulative incidence of fracture in the MrOS cohort

Months in cohort	Event-free start of mo	Rate new fracture	Pr new fracture	Cum. incidence fracture
1	1.0000	7/5,993	0.0012	0.0012
2	0.9985	7/5,983	0.0012	0.0023
3	0.9965	10/5,971	0.0017	0.0040
4	0.9947	5/5,960	0.0008	0.0048
5	0.9933	4/5,952	0.0007	0.0055
6	0.9917	8/5,942	0.0013	0.0068
7	0.9893	10/5,927	0.0017	0.0085
8	0.9875	11/5,915	0.0018	0.0103
9	0.9850	6/5,899	0.0010	0.0113
10	0.9831	4/5,887	0.0005	0.0118

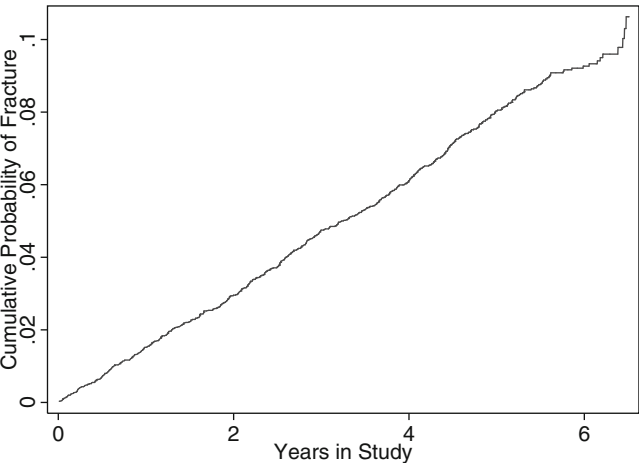


Fig. 6.12 Cumulative incidence of fracture in the MrOS cohort

10 of out 5,927 followed. The product of these two numbers is 0.0017 and means that about 0.17% of the cohort develops a new fracture in the seventh month. The cumulative incidence function is the sum over all the previous months. It estimates that the probability of developing a fracture by the end of the seventh month of the study is 0.085. Figure 6.12 graphs the cumulative incidence of fracture in MrOS cohort over a 6 year period.

The cumulative incidence function for the k th event at time t_j , $F_k(t_j)$, equals

$$F_k(t_j) = F_k(t_{j-1}) + \tilde{S}(t_{j-1})h_k(t_j), \tag{6.21}$$

where $\tilde{S}(t_{j-1})$ is the chance of being free of events at time t_{j-1} (just prior to time t_j). In the MrOS data, it is the chance of being both alive and unfractured. This can

be calculated by combining death and fracture into a composite and calculating the usual Kaplan–Meier estimator of being event-free and is given by values in the second column of Table 6.21. The hazard $h_k(t_j)$ denotes the cause-specific hazard for the k th event at time t_j which is given by the values in the third column of Table 6.21. The risk of a new event of type k at time t_j is the product of the k th cause-specific hazard at time t_j by the cumulative incidence of the k th event at time t_{j-1} and appears as a fourth column. The cumulative incidence is the total of new events over all time periods and in the final column.

The cause-specific *hazard* function can be obtained by censoring competing events. However, censoring competing causes and calculating a *survival* function lead to estimates which have an awkward interpretation. It can only be interpreted as probability of event type k if (1) all competing causes have been eliminated and (2) the competing events can be assumed to be independent of the k th event. Note, this is typically the assumption made with people who are lost to follow-up. However, it would be highly speculative to extrapolate the likelihood of fracture if death could be eliminated from the MrOS cohort and contrary to the objective of accommodating competing causes.

From the calculations in Table 6.21 and (6.21), it is evident that the cumulative incidence function for the k th event depends on two things—the k th cause-specific hazard function and the Kaplan–Meier curve for remaining event-free. The event-free curve combines those who have not had the k th event or any other event. Hence, the cumulative chance of developing a fracture can be decreased (holding the k th cause-specific hazard constant) by increasing the rate of other types of failures, putting far fewer participants at risk for a fracture.

For instance, if age has no effect on the risk of fracture but does increase the risk of death, then a comparison of the cumulative incidence function by age would show a lower cumulative incidence of fracture among older men. This follows from the fact that older men are less likely to develop a fracture over the follow-up period. However, the lower number of fractures is due to the fact that fewer older men live long enough to develop fractures. This effect, while real, happens through age's effect on death rather than on fracture.

In such a scenario, modeling the age effect on the cause-specific hazard of fracture will show no effect. A model for the effect of age on the cumulative incidence of fracture would show a lower incidence of fracture with increased age. Both descriptions are faithful to the situation but reflect different aspects. Hence, the analyses are complementary and we discuss regression models for both.

6.5.3.3 Cox Model for Cause-Specific Hazard Functions

One approach to allowing predictors to affect the onset of competing risks is to model the cause-specific hazard function using proportional hazard formulation. Covariate effects can be interpreted as ratios of cause-specific hazards. Let $h_k(t)$

Table 6.22 Cox model for effect of BMD on fracture risk adjusted for body weight

stset time, failure(status==1)

stcox i.bmd3 weight

No. of subjects =	5993	Number of obs =	5993
No. of failures =	531		
Time at risk =	30483.46339		
Log-likelihood =	-4442.2904	LR chi2(3) =	121.22
		Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
bmd3						
2	.4193745	.0447953	-8.14	0.000	.3401585	.5170384
3	.3290229	.0396476	-9.23	0.000	.2598098	.4166743
weight	1.004146	.00362	1.15	0.251	.997076	1.011266

be the k th cause-specific hazard and let $x_1, \dots, x_p$ be a set of predictors. The model has the form

$$h_k(t|\mathbf{x}) = h_{0k}(t) \exp(\beta_1 x_1 + \dots + \beta_p x_p),$$

(6.22)

where $h_{0k}(t)$ is a baseline hazard function. The model incorporates covariates into the model just as they appear in (6.5). The only difference is the hazard ratios apply to the k th cause-specific hazard while the interpretation of predictor effects are identical.

Consider a model for baseline BMD as a predictor of fracture in the MrOS cohort. The variable time in this dataset is the time to event variable Y and the type of failure is given by status which is coded as 0 if a person is free of death and fracture at the end of observation, 1 if the follow-up ends with fracture, and 2 if follow-up ends with a death prior to the onset of fracture. The predictors are BMD bmd3 categorized into levels <0.895 g/cm², 0.895 to 1.01 g/cm², and >1.01 g/cm² and baseline weight (weight) measured in kilograms.

In Table 6.22, we see that compared to the group with BMD < 0.895 g/cm², those with BMD 0.895 to 1.01 g/cm² and >1.01 g/cm² have relative hazards of fracture of 0.42 (a 58% reduction) and 0.33 (a 67% reduction) adjusting for body weight at enrollment.

6.5.3.4 Fine–Gray model for Cause-Specific Hazard Functions

The result of the cause-specific regression in Table 6.22 shows a higher rate of fractures with lower BMD, but what if BMD is correlated with a series of unmeasured factors which make death more likely? If the death rate was high enough, men with low BMD might develop fewer fractures after 2 years than high-BMD men simply because the low-BMD men are more likely to die before fracture. This would be apparent from a comparison of the cumulative incidence function

Table 6.23 Fine and Gray model for effect of BMD on cumulative incidence of fracture adjusted for body weight

```
stset time, failure(status==1)

stcrreg i.bmd3 weight, compete(status==2)
```

Competing-risks regression		No. of obs	=	5993
		No. of subjects	=	5993
Failure event : status == 1		No. failed	=	531
Competing event: status == 2		No. competing	=	657
		No. censored	=	4805
Log pseudolikelihood = -4472.9261		Wald chi2(3)	=	119.64
		Prob > chi2	=	0.0000

_t	SHR	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
bmd3						
2	.4219663	.0443983	-8.20	0.000	.3433337	.5186079
3	.3369128	.03982	-9.20	0.000	.2672472	.4247387
weight	1.004669	.0036759	1.27	0.203	.9974896	1.011899

but not from the cause-specific hazard function. This is the disconnect between the hazard and cumulative incidence scale—it is helpful to have a way to describe regression effects on the cumulative incidence scale.

The Fine and Gray model (Fine and Gray 1999) adapts the spirit of a proportional hazards model to the cumulative incidence formulation. The idea is to model a different kind of rate function. The k th cause-specific hazard function is the rate of event among those who have experienced no event. For the MrOS data, this means that those who die are no longer counted in the rate (or hazard). The type of hazard that Fine and Gray construct retains cohort members who succumb to the competing risk in the denominator of their rate. For the MrOS data, this can be thought of as the rate of developing fracture among those without a previous fracture who are not lost to follow-up and, in particular, including those who have died. Maintaining those who die in the risk set acknowledges that someone who succumbs to a competing risk will not develop the event of interest and does not require the kind of extrapolation used for someone who is lost to follow-up.

Denote this new rate function for the k th type of failure as $f_k(t)$

$$f_k(t|\mathbf{x}) = f_{0k}(t) \exp(\beta_1 x_1 + \dots + \beta_p x_p),$$

(6.23)

where $f_{0k}(t)$ takes the place of the usual baseline (or cause-specific) hazard function. The model incorporates covariates just as they appear in (6.22). Fitting the Fine and Gray model in Stata is done with the `stcrreg` command which is illustrated in Table 6.23.

The result of the Fine and Gray regression in Table 6.23 shows the hazard ratios for the model in (6.23) under the column marked “SHR” and the standard errors are labeled as “robust” because it is calculated in a way which is not model-based. The results are striking similar to the regression based on the cause-specific hazard function in Table 6.22. This is not surprising—the methods differ on whether people who die are retained in the risk set. The risk of death is not large and hence the two approaches give very similar results.

6.6 Some Details

In this section, we discuss some useful additional topics.

6.6.1 Bootstrap Confidence Intervals

The ACTG 019 dataset includes 880 observations but only 55 failures. Stata provides Wald-based CIs for the Cox model which require sample size which are “large.” The effective sample size is determined by the number of failures rather than the number of observation. Hence, it can be useful to check the validity of the Wald-based CIs for the Cox model for ZDV treatment (rx) and baseline CD4 cell count (cd4) using the bootstrap (Sect. 3.6). The results are reported on the coefficient scale in Table 6.24.

The standard and bias-corrected bootstrap CIs, based on 1,000 resampled datasets, yield very similar results, confirming that the semi-parametric model works well in this case, even though there are only moderate numbers of events.

Table 6.24 Cox model for ZDV and CD4 with bootstrap confidence intervals

stcox i.rx cd4, vce(bootstrap, bca reps(1000) nodots seed(881))

Cox regression -- Breslow method for ties

No. of subjects = 880

No. of failures = 55

Time at risk = 354872

Log likelihood = -314.17559

Number of obs = 880

Wald chi2(2) = 32.34

Prob > chi2 = 0.0000

	Observed	Bootstrap			Normal-based	
_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
1.rx	.4560671	.138132	-2.59	0.010	.2518951	.8257293
cd4	.9934464	.0013741	-4.75	0.000	.9907569	.9961432

6.6.2 Prediction

Evaluating prediction error using some form of cross-validation, as described in Sect. 10.1, is more complicated with time-to-event outcomes. Comparing observed to expected survival times is ruled out for censored observations in the test set; moreover, as we explained above in Sect. 6.2.13, expected—that is, mean—survival times are usually undefined under the Cox model. Comparing the *occurrence* of events in the test set with predictions based on the learning set, as with binary outcomes analyzed using a logistic model, is relatively tractable, but complicated by variations in follow-up time, in particular extrapolations for any follow-up times in the test set that exceed the longest times in the learning set.

Dickson et al. (1989) give one way in which predictions based on a Cox model can be cross-validated using a test dataset. The basic idea is to use coefficients estimated from a development dataset to classify observations in a test dataset. To see how this works, first note that the predictors are associated with the hazard ratio only through what is called the *linear predictor*, $\beta_1 x_1 + \dots \beta_p x_p$, as demonstrated in (6.5). The larger the value of the linear predictor, the larger the hazard and the shorter survival times tend to be. Obtaining the estimated coefficients from a development dataset, it is possible to calculate what is called a *risk score*, namely $\hat{\beta}_1 x_1 + \dots \hat{\beta}_p x_p$, for each observation in either the development or test datasets. The investigators grouped the patients in the development dataset into four predicted survival categories on the basis of the risk score, with the cutpoints determined to give approximately equal numbers of events in each category. They then demonstrated the models ability to discriminate by calculating the Kaplan–Meier survival curves for the test set using the groups defined by cutpoints from the development set. The pronounced separation of the survival curves in the test set is evidence of the ability of the model to stratify by risk.

6.6.3 Adjusting for Nonconfounding Covariates

If a covariate is strongly predictive of survival but uncorrelated with a predictor of interest, omitting it from a Cox model will nonetheless attenuate the estimated hazard ratio for the predictor of interest, as discussed in Sect. 10.2.6 (Gail et al. 1984; Schmoor and Schumacher 1997; Henderson and Oman 1999). Omitting important covariates from logistic models also induces such attenuation. Although the gain in precision is usually modest at best, it can be advantageous to include such a prognostic factor in order to avoid the attenuation.

A compelling example is provided by ACTG 019, the randomized clinical trial of ZDV for prevention of AIDS and death in HIV infection discussed in Sect. 6.1. As expected in a clinical trial, there was no between-group difference in mean baseline CD4 count, known to be an important prognostic variable. Thus by definition, baseline CD4 count could not have confounded the effect of ZDV. However, when

CD4 count is added to the model, the estimated reduction in risk of progression to AIDS or death afforded by ZDV goes from 49% to 54%, an increase of about 12%. More discussion of whether to adjust for covariates in a clinical trial is given in Sect. 10.2.6.

6.6.4 *Independent Censoring*

To deal with right-censoring, we have made the assumption of *independent censoring*. The essence of this assumption is that after adjustment for covariates, future event risk for a censored subject does not differ from the risk among other subjects who remain in follow-up and have the same covariate values. Under this assumption, subjects are censored independent of their future risk.

To see how this assumption may be violated, consider a study of mortality risk among patients followed from admission to the intensive care unit until hospital discharge. Suppose no survival information is available after discharge, so subjects have to be censored at that time. In general, subjects are likely to be discharged because they have recovered and are thus at lower risk than patients who remain hospitalized. Unless we can completely capture the differences in risk using baseline and TDCs, the assumption of independent censoring would be violated.

Dependent censoring can also arise from informative loss to follow-up. In prospective cohorts, it is not unlikely that prognosis for dropouts differs from that for participants remaining in follow-up in ways that can be difficult to capture with variables routinely ascertained.

It can also be difficult to diagnose dependent censoring definitively, because that would require precisely the information that is missing—for example, mortality data after discharge from the ICU. But that is a case where an experienced investigator might recognize on substantive grounds that censoring is likely to be dependent. Furthermore, the problem could be addressed in that study by ascertaining mortality for a reasonable period after discharge. Similarly, losses to follow-up are best addressed by methods to maximize study retention; but it also helps to collect as much information about censored subjects as possible. Inverse weighting methods can be used in situations where the dependence between failure and censoring is explained by a series of measured variables and where a model for censoring can be specified in terms of these measured (possibly time-dependent) covariates (see Sect. 9.5).

6.6.5 *Interval Censoring*

We also assume that the time of events occurring during the study is known more or less exactly. This is almost always the case for well-documented events like death, hospitalization, or diagnosis of AIDS. But the timing of many events is not

observed with this level of precision. For example, in prospective cohort studies of people at risk for HIV infection, it is common to test participants for infection at semi-annual visits (Buchbinder et al. 1996). Thus the actual time of an incident infection is only known up to an interval of possible values; in technical terms, it is *interval-censored* between the last visit at which the participant tested negative and the first at which the result was positive. Another example is development of abnormal cellular changes in the cervix, which must be assessed by clinical exam. These exams may be performed periodically, perhaps months or even years apart. As with HIV infection, newly observed changes may have occurred at any time since the last exam. In settings where intervals arise because of the study follow-up schedule and are regularly spaced, pooled logistic regression Sect. 5.5.2 can be used to handle the interval censoring. Interval censoring becomes more complex when the time between intervals is unequal and/or vary by individual requiring specialized methods beyond the scope of this book.

6.6.6 Left-Truncation

Survival times are measured from some initial time with more than one possible choice of origin. In the PBC study, we defined the survival time as the time from cohort enrollment until death. We could, instead, chose to measure survival time from the diagnosis of disease. Diagnosis is a more meaningful event biologically and easily aligning the time scales on this initial time will lead to more interpretable results.

The PBC study recruited patients from a referral center and months or years may have elapsed between diagnosis and entry into the cohort. A patient with a rapid disease course is less likely to be enrolled simply because they may die prior to referral to the center or prior to recruitment into the study. When this type of selection is active, there can be an undercounting of short survival times. The setting where some survival times are not observed because the sampling scheme tends to miss short survival times is known as *left-truncation*. To avoid bias, we need to consider the length of the period between the time origin and entry into the cohort. We denote this *truncation time* by V . Staggered entry into a cohort does not imply left-truncation; the key feature of left-truncation is the truncation time, V —there must be some time delay between the event which defines the origin event and entry into the cohort. For instance, if a PBC cohort was able to enroll participants at the time of their diagnosis, truncation times would be 0 and there would be no left-truncation. However, because patients have their diagnoses at different times, the cohort will still exhibit staggered entry.

The nature of the incomplete data from truncation is different from censored data. For censoring, it is incomplete because the event time falls outside of follow-up. Truncated values outside the follow-up period are not merely incomplete; rather, they are not observed at all. Because a left-truncated patient dies before enrollment, they leave no trace in the study—truncation is said to result in “ghosts.”

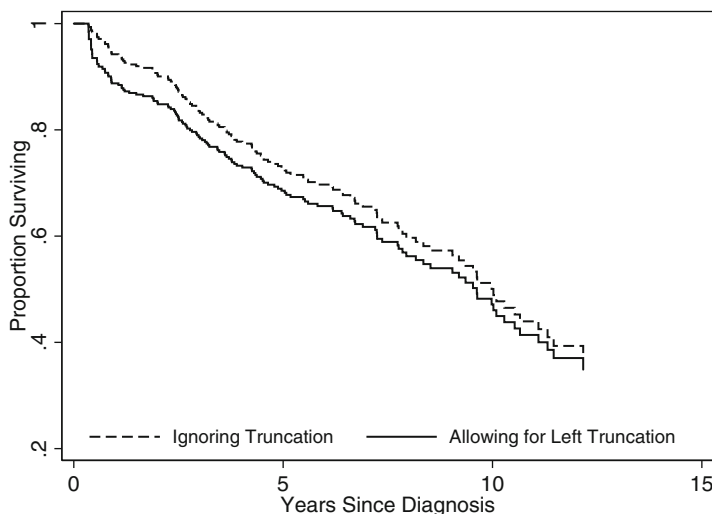


Fig. 6.13 Kaplan–Meier curves for the PBC data incorporating and ignoring left-truncation

Right-truncation can also arise if a study recruits based on an endpoint and people with large event times (or who never had the event) are not recruited. An example of right-truncation is a fecundability study that excludes couples who never conceive.

Survival analysis of risk factors can be conducted on the natural time-scale with origin at HIV infection under an assumption of *independent truncation*. This assumption is that the time of delayed entry and subsequent survival are independent and it is satisfied when the incidence of a disease and survival post-diagnosis are independent. Under independent truncation, the analysis uses the truncation time V along with the time and censoring indicators (X, Δ) , where X is the follow-up time relative to diagnosis. The PBC dataset does not include truncation times but we created the truncation time `disease_dur` for illustrative purposes.

In Stata, we introduce the censoring and truncation into the analysis using the `stset` command in Stata as follows:

```
stset years_since_diag, failure(status) entry
      (disease_dur)
```

Figure 6.13 graphs two survival curves based on time since diagnosis—one which uses the `stset` statement to account for left-truncation and naive calculation which ignores truncation. The estimator which ignores truncation estimates higher post-diagnosis survival probabilities. By ignoring the truncation, the analysis fails to account for undersampling of short survival times and, thus, overestimates survival. The effect of ignoring truncation on hazard ratios in a Cox model is less predictable but can often attenuate them.

Note that both survival estimators in Fig. 6.13 make drops at the same event times but the size of the drops for the truncation-based estimator are larger at earlier time points. This reflects the importance of short event times under left-truncation just as

long event times are important under right-censoring. A key assumption under left-truncation is that there is a positive probability that even the shortest failures could make it into the sample. If they are completely excluded, only strong parametric assumptions can account for their absence.

Fortunately, Stata can handle (independent) censoring and truncation simultaneously and once the `stset` command has been used it is possible to use the full set of survival analysis techniques without taking further account of the nature of the incomplete data.

6.7 Sample Size, Power, and Detectable Effects

Sections 4.8 and 5.7 provide formulas for calculating sample size, power, and minimum detectable effects for the linear and logistic models. Analogous results hold for the Cox model. To compute the sample size that will provide power of γ in two-sided tests with type-1 error of α to reject the null hypothesis $\beta_j = 0$ for the effect of a predictor X_j , accounting for the loss of precision due to adjustment for covariates, we can use

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2}{(\beta_j^a \sigma_{x_j})^2 \psi (1 - \rho_j^2)}, \quad (6.24)$$

where β_j^a is the hypothesized value of β_j under the alternative, $z_{1-\alpha/2}$ and z_γ are the quantiles of the standard normal distribution corresponding to the specified type-1 error and power, σ_{x_j} is the standard deviation of X_j and ρ_j is its multiple correlation with the other covariates, and ψ is the probability that an observation is uncensored, so that the expected number of events $d = n\psi$ (Hsieh and Lavori 2000; Schmoor et al. 2000; Bernardo et al. 2000). The variance inflation factor $1/(1 - \rho_j^2)$ in (6.24) accounts for the potential loss of precision due to the inclusion of other predictors in the model (Hsieh et al. 1998). For problems with fixed values of n and ψ , power is given by

$$\gamma = 1 - \Phi \left[z_{1-\alpha/2} - |\beta_j^a| \sigma_{x_j} \sqrt{n\psi(1 - \rho_j^2)} \right]. \quad (6.25)$$

Finally, the minimum detectable effect (on the log-hazard scale) is

$$\pm \beta_j^a = \frac{z_{1-\alpha/2} + z_\gamma}{\sigma_{x_j} \sqrt{n\psi(1 - \rho_j^2)}}. \quad (6.26)$$

Some additional points:

- Sample size (6.24) and minimum detectable effect (6.26) calculations simplify considerably when we specify $\alpha = 0.05$ and $\gamma = 0.8$, β_j^a is the effect of a one standard deviation increase in continuous x_j , and we do not need to penalize for covariate adjustment. In that case,

$$n = \frac{7.849}{(\beta_j^a)^2 \psi}. \quad (6.27)$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = \frac{2.802}{\sqrt{n\psi}}. \quad (6.28)$$

For 90% power, substitute 10.51 for 7.849 and 3.242 for 2.802.

- Similarly, for a two-arm clinical trial with equal allocation to arms, so that β_j^a is the log-hazard ratio for treatment and $s_{x_j}^2 = 0.25$, we can calculate

$$n = \frac{4 \times 7.849}{(\beta_j^a)^2 \psi}. \quad (6.29)$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = \frac{2 \times 2.802}{\sqrt{n\psi}}. \quad (6.30)$$

- Power calculations using (6.25) simplify analogously, but still require a statistical calculator or computer package to evaluate the normal cumulative distribution function $\Phi(\cdot)$.
- Power in the Cox model is driven by the expected number of events $d = n\psi$, with little or no independent influence of n once d is fixed. For the same reason, early censoring has relatively little influence. Some calculators may return d rather than n , or require d rather than n and ψ as inputs.
- Sample size, power, and minimum detectable effects can be calculated using the `stpower cox` command in Stata as well as many other statistical packages. Alternatively, (6.24)–(6.26) can easily be programmed in Stata, R, or Excel, or evaluated by hand if values of $z_{1-\alpha/2}$, z_γ , and $\Phi(\cdot)$ are available.
- When X_j is a binary predictor with prevalence f_j , $\sigma_{x_j} = \sqrt{f_j(1-f_j)}$ in (6.24)–(6.26).
- When X_j is a continuous predictor with standard deviation σ_{x_j} , it is important to recognize that sample size, power, and minimum detectable effects do not depend in any real way on the units in which X_j is measured. This is most clearly seen in (6.26). Suppose X_j is usually measured in grams. Changing the unit to milligrams increases σ_{x_j} by a factor of 1,000, and shrinks β_j^a by the same factor. But of course the effect on the outcome of a 1-mg increase in the predictor is 1,000 times smaller than the effect of a 1-g increase. One way to avoid confusion is to consider the minimum detectable effect size for a one standard deviation change in X_j , which is often a reasonable-sized change to consider. That effect size is obtained by setting $\sigma_{x_j} = 1$ in (6.26).

- As in calculations for the linear and logistic model, we need to use $|\beta_j^a|$ in (6.25) if $\beta_j^a < 0$. It follows that the negative of the value given by (6.26) is also a valid solution for the minimum detectable effect.
- The use of the factor $1 - \rho_j^2$ to account for covariate adjustment carries over from linear to Cox models. However, there is no analog to the reduction in residual variance that can result from including covariates in linear models, so that the adjustment to these calculations using $1 - \rho_j^2$ is less likely to be conservative.
- The `stpower cox` command does incorporate the factor $1 - \rho_j^2$ to account for covariate adjustment, via the `r2` option. In using sample size calculators that do not allow for this adjustment, unadjusted estimates of n or d should be inflated by $1/(1 - \rho_j^2)$; similarly the minimum detectable effect estimate should be inflated by $\sqrt{1/(1 - \rho_j^2)}$. To calculate power in such calculators, use $n\psi(1 - \rho_j^2)$ in place of $n\psi$ as an input.
- In Sect. 4.8, we showed how the standard error $SE(\hat{\beta}_j)$ plays a central role in sample size, power, and minimum detectable effect calculations for regression problems. $SE(\hat{\beta}_j)$ is a large-sample approximation in Cox models, and more exact small-sample computations using the t -distribution do not carry over from the linear model. Simulations of power may be a more reliable guide when the calculated or available sample size is small.
- The alternative calculations (4.15)–(4.17) presented in Sect. 4.8, which use an estimate $\tilde{SE}(\hat{\beta}_j)$ based on published results for an appropriately adjusted model using $\tilde{n}$ observations, carry over directly. There we showed that

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \tilde{n} [\tilde{SE}(\hat{\beta}_j)]^2}{(\beta_j^a)^2}. \quad (6.31)$$

Similarly, power in a new sample of size n is given by

$$\gamma = 1 - \Phi \left[z_{1-\alpha/2} - |\beta_j^a| / [\sqrt{\tilde{n}/n} \tilde{SE}(\hat{\beta}_j)] \right]. \quad (6.32)$$

Finally, the minimum detectable effect in a new sample of size n can be obtained as

$$\pm \beta_j^a = (z_{1-\alpha/2} + z_\gamma) \sqrt{\tilde{n}/n} \tilde{SE}(\hat{\beta}_j). \quad (6.33)$$

In implementing these calculations, care must be taken to obtain the SE of the regression coefficient β_j , not the SE of the hazard ratio e^{β_j} . This can be computed from a 95% CI for the hazard ratio as $\tilde{SE}(\hat{\beta}_j) = \log(UL/LL)/3.92$, where UL and LL are the upper and lower confidence bounds. We must also ensure that X_j is measured on the same scale as in the published results.

To illustrate these calculations, we first calculate the sample size providing 80% power in a two-sided test with α of 5% to detect an effect of bilirubin levels on survival, adjusting for the effects of hepatomegaly, edema, and spiders, as suggested by the analysis shown previously in Table 6.12.

Table 6.25 Sample size calculation for effect of bilirubin on mortality risk

```
. stpower cox, failprob(.15) hratio(1.15) sd(4.5) r2(0.2025)

Estimated sample size for Cox PH regression
Wald test, log-hazard metric
Ho: [b1, b2, ..., bp] = [0, b2, ..., bp]

Input parameters:
      alpha =      0.0500    (two sided)
      b1 =      0.1398
      sd =      4.5000
      power =      0.8000
      Pr(event) =      0.1500
      R2 =      0.2025

Estimated number of events and sample size:
      E =      25
      N =     166

. display (invnormal(.975)+invnormal(0.8))^2/((log(1.15)*4.5)^2*0.15*
(1-.2025)) 165.87573
```

The new study will have a shorter 2-year follow-up, as compared to the average 5.5 year follow-up in the DPCA Trial, with an estimated 15% cumulative mortality. Based on the DPCA results, we estimate that $\sigma_{x_j} \approx 4.5$ mg/dL and that $\rho_j \approx 0.45$ (so $\rho_j^2 = 0.2025$), indicating substantial variance inflation. We hypothesize that the hazard ratio per mg/dL increase in bilirubin level will be 1.15 (so $\beta_j^a = \log 1.15$). Table 6.25 shows results using the `stpower cox` command in Stata as well as a calculation using (6.24). The two estimates are essentially identical; a quick calculation using $\psi = 0.15$ shows that the expected number of events based on (6.24) is 25.

The `stpower cox` command can also be used to calculate minimum detectable effects. In the DPCA trial, suppose an ancillary study is being considered to evaluate the independent association of mortality with a novel risk marker, to be measured using stored baseline specimens. There were 125 deaths among 312 participants, so $\psi = 125/312 = 0.40$; equivalently, $d = n\psi = 125$. We hypothesize that the new marker will be highly correlated with available prognostic measures ($\rho_j \approx 0.5$), yet hope that it will provide additional predictive information. Initial testing suggests that the SD of the new marker is approximately 1.5 mg/dL. We hypothesize that higher levels of the marker will be associated with lower risk. What hazard ratio per mg/dL increase in the new marker will be detectable with 80% power in a two-sided test with α of 5%?

Table 6.26 shows that `stpower cox` and (6.26) give essentially the same result: for the DPCA sample to provide 80% power to reject $\beta_j = 0$, the mortality hazard must be independently reduced by approximately 18% for each mg/dL increase in the novel marker.

Table 6.26 Minimum detectable effect of a novel marker

```
. stpower cox, n(312) failprob(.40) sd(1.5) r2(0.25) power(.8) hr

Estimated hazard ratio for Cox PH regression
Wald test, hazard metric
Ho: [b1, b2, ..., bp] = [0, b2, ..., bp]

Input parameters:
      alpha =      0.0500    (two sided)
      sd =      1.5000
      N =      312
      power =      0.8000
Pr(event) =      0.4000
      R2 =      0.2500

Estimated number of events and hazard ratio:
      E =      125
hratio =      0.8244

. display exp(-(invnormal(.975)+invnormal(0.8))/(1.5*sqrt(125*(1-0.25))))
.82456636
```

6.8 Summary

Survival data exhibit novel features including right-censoring, interval censoring, truncation, and competing risks. The Cox proportional hazards model is suited to the special features of survival data and summarizes the effects of covariates through hazard ratios. The Cox model has much in common with other regression models; in particular, issues of confounding, mediation, and interaction are dealt with in similar ways. Specialized techniques are required to calculate predicted survival and to examine the assumption of proportional hazards. The Cox model can be ended to handle TDCs and stratification. Competing risks arise when other events may preclude observing the event of interest. Extensions to the proportional hazards model for competing risks data can be based on the cause-specific hazard function (which models the effect of covariates directly on the event of interest) or can be based on the Fine–Gray model (which allows for the effect of covariates which occur through competing events). The two approaches provide complementary perspectives on the effect of covariates in the presence of competing risks.

6.9 Further Notes and References

The Cox model has proven popular because it is computationally feasible and flexible. Alternatives include the *accelerated failure time model* (Wei 1992) or the *additive hazards model* (Aalen 1989). These models are less popular and statistical techniques for them are less well developed. By contrast, there are extensively developed techniques for parametric survival regression (implemented in Stata with the `streg` package). Parametric models require us to make assumptions about the form

of the baseline hazard function and have proved less popular because the parametric assumptions sacrifice robustness without substantial efficiency gains. Useful references include Chap. 5 of [Marubini and Valsecchi \(1995\)](#) and Chap. 12 of [Klein and Moeschberger \(1997\)](#).

Some more complex survival data settings are beyond the scope of chapter. For instance, there may be more than a single event per subject, yielding clustered or hierarchical survival data. See [Wei and Glidden \(1997\)](#) for an overview of possible approaches, including analogs of the *marginal* and *random effects* models described for repeated continuous and binary outcomes in Chap. 7. The are both available options in Stata `stcox` command—the marginal by using the `vce(cluster)` option and random effects by using the `shared` option.

Stata provides extensive capabilities for fitting and assessing Cox models. For instance, more flexible model for time-varying hazards than those discussed in Sect. 6.4.2.6 could be developed by treating time as continuous (using the `tvvc`) option in conjunction with splines. A complete suite of parametric survival analysis methods are also provided. The flexible `stset` command handles complex patterns of censoring and truncation.

Applied book-length treatments on survival analysis are available by [Miller et al. \(1981\)](#) and [Marubini and Valsecchi \(1995\)](#). These two texts strike a nice balance in their completeness and orientation toward biomedical applications. The texts by [Klein and Moeschberger \(1997\)](#) and [Therneau and Grambsch \(2000\)](#) are very complete in their coverage of tools for survival analysis in general and the Cox model in particular. Chap. 3 of [Klein and Moeschberger \(1997\)](#) provides a complete discussion on left-truncation, interval censoring, and general censoring patterns.

Sometimes time-to-event data can be more effectively handled using an alternative framework. In particular, consider cohort studies in which interval-censored outcomes are ascertained at each follow-up visit. One alternative is to use the continuation ratio model, referenced in Chap. 5, for time to the first such event. This can be seen as a discrete-time survival model, where the time scale is measured in visits (or intervals). Where appropriate, another, often more powerful, alternative is to use a logistic model for repeated binary measures, covered in Chap. 7. A closely related issue is the handling of Finally, some time-to-event data has no censored values. In that situation, techniques covered in Chap. 8 can provide a useful regression framework for dealing with the skewness and heteroscedasticity such data are likely to exhibit.

6.10 Problems

Problem 6.1. Divide the hazard ratio for `bilirubin` by its standard error in Table 6.4 and compare the result to the listed value of `z`. Also compute a CI for this hazard ratio by adding and subtracting 1.96 times its standard error from the hazard ratio estimate. Are the results very different from the CI listed in the output, which is based on computations on the coefficient scale?

Problem 6.2. In the ACTG 019 data, treatment `rx` is coded `ZDV = 1` and placebo = 0. Define a new variable `rxplus11` which is coded `ZDV = 12` and placebo = 11; this can be done using the Stata command `generate rxplus11=rx+11`. Fit a Cox model with `rxplus11` as the only predictor, then fit a second Cox model with `rx` as the only predictor. How do the two results compare?

Problem 6.3. Using the ACTG 019 data from Problem 6.2, recode treatment so it is coded `ZDV = 0` and placebo = 1. How do the hazard ratios, CIs, likelihood ratio (LR), and Wald tests compare to the original coding? If any are different, how are they different?

Problem 6.4. Using the PBC dataset, calculate the hazard ratio for values of albumin = 2.5, 3.5, and 4.0, using albumin = 3 as the reference level assuming the log-hazard is linear in albumin. The PBC dataset is available at <http://www.biostat.ucsf.edu/vgsm>.

Problem 6.5. For the PBC dataset, fit a model with cholesterol and bilirubin. Interpret the results, as you would in a paper, reporting the hazard ratios for a 100 mg/dL increase in cholesterol and a 10 mg/dL increase in bilirubin. Is the relationship between cholesterol and survival confounded by bilirubin?

Problem 6.6. Calculate a hazard ratio and CI for a 5-year increase in age by computing the fifth power of the estimated hazard ratio and its confidence limits, using the results for a 1-year increase in Table 6.9. Compare the result to a fit of the Cox model using a re-scaled version of the variable.

Problem 6.7. Using the model in Table 6.14 and taking Table 6.15 as your guide, calculate the effect of hepatomegaly among those on placebo. Then, derive and calculate the contrast required to identify the effect of hepatomegaly among those on DPCA. Given these, derive and fit the linear contrast to test for interaction. How does it compare with the test of interaction for comparing the effect of DPCA treatment across hepatomegaly that was given in Sect. 6.2.10?

Problem 6.8. For the ACTG 019 dataset, write out the Cox model allowing for an interaction between ZDV treatment `rx` and the baseline CD4 cell count `cd4`.

- Express the test of the null hypothesis of no interaction between CD4 and treatment in terms of the parameters of the model.
- Again using the parameters of the model, what is the hazard ratio for a ZDV-treated subject with x CD4 cells compared with a placebo-treated subject with x CD4 cells?
- Fit the model. Does there appear to be an interaction between treatment and CD4 stratum? If so, what is the interpretation?
- What are the hazard ratios for ZDV as compared to placebo for patients with 500, 109, and 50 CD4 cells, respectively?

Problem 6.9. We can also control for the effect of bilirubin in the PBC mortality data using stratification rather than adjustment. One way to categorize is to create approximately equal-size groups. In Stata, for example, you can categorize by

quintile of bilirubin using the command `xtile cat5=bilirubin, nq(5)`. Try fitting a Cox model for `cholesterol` stratified by `bilirubin`, stratified at 2, 3, 10, and 50 levels. What is the trade-off in increasing the number of levels? What number of levels works best? (Hint: Balance adjustment against the size of the standard error).

Problem 6.10. Using the PBC dataset, apply the methods of Sect. 6.4.2 for examining proportional hazards to the variable `hepatomegaly` and interpret the results.

6.11 Learning Objectives

- (1) Define right-censoring, hazard function, proportional hazards, left-truncation, competing risks data, and TDCs.
- (2) Be able to:
 - Convert a predictor to a new unit scale
 - Derive the hazard ratio between two groups defined by their predictor values
 - Interpret hazard ratio estimates, Wald test p -values, and CIs
 - Calculate and interpret the likelihood-ratio test comparing two nested Cox models
 - Detect and model interaction using the Cox model
 - Detect nonproportional hazards using log-minus-log and smoothed hazard ratio plots, and the Schoenfeld test
 - Use stratification to control for a covariate with nonproportional effects
- (3) Understand:
 - When to use survival techniques
 - Why the semi-parametric form of the Cox model is desirable
 - Why the Cox model is “multiplicative”
 - How the stratified Cox model relaxes the proportional hazard assumption
 - How to address confounding, mediation, and interaction using a Cox model
 - The difference between modeling cause-specific hazards and cumulative incidence functions for competing risk data
 - Recognize settings which are beyond the scope of this chapter, including interval and dependent censoring, and repeated-events data

Chapter 7

Repeated Measures and Longitudinal Data Analysis

Knee radiographs are taken yearly in order to understand the onset of osteoarthritis. Troponin (which is an indicator of heart damage) is measured from blood samples 1, 3, and 6 days following a brain hemorrhage. Groups of patients in a urinary incontinence trial are assembled from different treatment centers. Susceptibility to tuberculosis is measured in family members. All of these are examples of what is called repeated measures data or hierarchical or clustered data. Such data structures are quite common in medical research and a multitude of other fields.

Two features of this type of data are noteworthy and significantly impact the modes of statistical analysis. First, the outcomes are correlated across observations. Yearly radiographs on a person are more similar to one another than to radiographs on other people. Troponin measurements on the same person are more similar to one another than to those on other people. And groups of patients from a single center may yield similar responses because of treatment protocol variations from center-to-center, the persons or machines providing the measurements, or the similarity of individuals that choose to participate in a study at that center.

A second important feature of this type of data is that predictor variables can be associated with different levels of a hierarchy. Consider a study of the choice of type of surgery to treat a brain aneurysm either by clipping the base of the aneurysm or implanting a small coil. The study is conducted by measuring the type of surgery a patient receives from a number of surgeons at a number of different institutions. This is thus a hierarchical dataset with multiple patients clustered within a surgeon and multiple surgeons clustered within a hospital. Predictor variables can be specific to any level of this hierarchy. We might be interested in the volume of operations at the hospital, or whether it is a for-profit or not-for-profit hospital. We might be interested in the years of experience of the surgeon or where she was trained. Or we might be interested in how the choice of surgery type depends on the age and gender of the patient.

Accommodation of these two features of the data, predictors specific to different levels in the data structure, and correlated data, are the topics of the chapter.

We begin by illustrating the basic ideas in a simple example and then describe hierarchical models through a series of examples. In Sect. 7.4, we introduce the first of the methods of dealing with correlation structures, namely generalized estimating equations. Section 7.4.1 introduces an example that we use throughout the rest of the chapter to illustrate the use of the models. Section 7.5 considers an alternative to generalized estimating equations, called random effects modeling, and the following sections contrast these approaches. We close with a section on power and sample size for some repeated measures and clustered data scenarios (Sect. 7.10).

7.1 A Simple Repeated Measures Example: Fecal Fat

Lack of digestive enzymes in the intestine can cause bowel absorption problems. This will be indicated by excess fat in the feces. Pancreatic enzyme supplements can be given to ameliorate the problem. The data in Table 7.1 come from a study to determine if there are differences due to the form of the supplement: a placebo (none), a tablet, an uncoated capsule (capsule), and a coated capsule (coated).

We can think of this as either a repeated measures dataset, since there are four measurements on each patient or, alternatively, as a hierarchical dataset, where observations are clustered by patient. This simple example has as its only predictor pill type, which is specific to both the person and the period of time during which the measurement was taken. We do not have predictors at the patient level, though it is easy to envision predictors like age or a history of irritable bowel syndrome.

We identify a continuous outcome variable, fecal fat, and a single categorical predictor of interest, pill type. If we were to handle this analysis using the tools of Chap. 3, the appropriate technique would be a one-way ANOVA, with an overall *F*-test, or, perhaps better, a preplanned set of linear contrasts. Table 7.2 gives the one-way ANOVA for the fecal fat example.

Following the prescription in Chap. 3, the *F*-test indicates ($p = 0.1687$) that there are not statistically significant differences between the pill types. But this analysis is incorrect. The assumptions of the one-way ANOVA require that all observations be independent, whereas we have repeated measures on the same

Table 7.1 Fecal fat (g/day) for six subjects

Subject number	Pill type				Subject Average
	None	Tablet	Capsule	Coated	
1	44.5	7.3	3.4	12.4	16.9
2	33.0	21.0	23.1	25.4	25.6
3	19.1	5.0	11.8	22.0	14.5
4	9.4	4.6	4.6	5.8	6.1
5	71.3	23.3	25.6	68.2	47.1
6	51.2	38.0	36.0	52.6	44.5
Pill type average	38.1	16.5	17.4	31.1	25.8

Table 7.2 One-way ANOVA for the fecal fat example

anova fecfat pilltype

		Number of obs =	24	R-squared =	0.2183
		Root MSE =	18.9649	Adj R-squared =	0.1010
Source	Partial SS	df	MS	F	Prob > F
Model	2008.6017	3	669.533901	1.86	0.1687
pilltype	2008.6017	3	669.533901	1.86	0.1687
Residual	7193.36328	20	359.668164		
Total	9201.96498	23	400.085434		

Table 7.3 Two-way ANOVA for the fecal fat example

anova fecfat subject pilltype

		Number of obs =	24	R-squared =	0.8256
		Root MSE =	10.344	Adj R-squared =	0.7326
Source	Partial SS	df	MS	F	Prob > F
Model	7596.98166	8	949.622708	8.88	0.0002
subject	5588.37996	5	1117.67599	10.45	0.0002
pilltype	2008.6017	3	669.533901	6.26	0.0057
Residual	1604.98332	15	106.998888		
Total	9201.96498	23	400.085434		

six subjects, which are undoubtedly correlated. The one-way ANOVA would be appropriate if we had collected data on six *different* subjects for each pill type.

Should we have conducted the experiment with different subjects for each pill type? Almost certainly not. We gain precision by comparing the pill types within a subject rather than between subjects. We just need to accommodate this fact when we conduct the analysis. This is analogous to the gain in using a paired *t*-test.

In this situation, the remedy is simple: we conduct a two-way ANOVA, additionally removing the variability between subjects. Table 7.3 gives the two-way ANOVA.

The results are now dramatically different, with pill type being highly statistically significant. In comparing Tables 7.2 and 7.3, we can see that a large portion (about 5,588 out of 7,193 or almost 78%) of what was residual variation in Table 7.2 has been attributed to subject-to-subject variation in Table 7.3, thus sharpening the comparison of the pill types.

This is an illustration of a very common occurrence: failure to take into account the correlated nature of the data can have a huge impact on both the analysis strategy and the results.

7.1.1 Model Equations for the Fecal Fat Example

We next write down model equations appropriate for the fecal fat example to more precisely represent the differences between the two analyses from the previous section. The analysis in Table 7.2 follows the one-way ANOVA model from Chap. 3.

$$\begin{aligned}\text{FECFAT}_{ij} &= \text{fecal fat measurement for person } i \text{ with pill type } j \\ &= \mu + \text{PILLTYPE}_j + \epsilon_{ij},\end{aligned}\tag{7.1}$$

where, as usual, we would assume $\epsilon_{ij} \sim \text{i.i.d } \mathcal{N}(0, \sigma_\epsilon^2)$.

As noted above, there is no account taken of the effect of each subject. We would expect some subjects to generally have higher values and others to generally have lower values. To accommodate this we include a subject effect in the model, which simultaneously raises or lowers all the measurements on that subject:

$$\begin{aligned}\text{FECFAT}_{ij} &= \text{fecal fat measurement for person } i \text{ with pill type } j \\ &= \mu + \text{SUBJECT}_i + \text{PILLTYPE}_j + \epsilon_{ij},\end{aligned}\tag{7.2}$$

with

$$\epsilon_{ij} \sim \text{i.i.d } \mathcal{N}(0, \sigma_\epsilon^2).$$

To this we add one more piece. We assume that the subject effects are also selected from a distribution of possible subject effects: $\text{SUBJECT}_i \sim \text{i.i.d } \mathcal{N}(0, \sigma_{\text{subj}}^2)$, independently of ϵ_{ij} .

This additional piece serves two purposes. First, it captures the idea that the subjects in our experiment are assumed to be a random sample from a larger population of subjects to which we wish to draw inferences. Otherwise, the conclusions from our experiment would be scientifically uninteresting, as they would apply only to a select group of six subjects. Second, as we will examine in detail in the next section, the inclusion of a subject effect (along with an assigned distribution) models a correlation in the outcomes. Once we added this subject effect to our model, we modified our analysis to accommodate it using a two-way ANOVA.

7.1.2 Correlations Within Subjects

The main reason the results in Tables 7.2 and 7.3 differ so dramatically is the failure of the analysis in Table 7.2 to accommodate the repeated measures or correlated nature of the data. How highly correlated are measurements within the same person? The model given in (7.2) gives us a way to calculate this. The observations on the same subject are modeled as correlated through their shared random subject effect.

The larger the subject effects in relation to the error term, the larger the correlation (relatively large subject effect means the observations on one subject are quite different than those on another subject, but, conversely, that observations *within* a subject tend to be similar). More precisely, there is a covariance between two observations on the same subject:

$$\begin{aligned}\text{cov}(\text{FECFAT}_{ij}, \text{FECFAT}_{ik}) &= \text{cov}(\text{SUBJECT}_i, \text{SUBJECT}_i) \\ &= \text{var}(\text{SUBJECT}_i) \\ &= \sigma_{\text{subj}}^2.\end{aligned}\tag{7.3}$$

The first equality in (7.3) is because the μ and pilltype terms are assumed to be fixed constants and do not enter into the covariance calculation. The ϵ_{ij} terms drop out because they are assumed to be independent of the subject effects and of each other. The second equality is true because the covariance of any term with itself is a variance and the last equality is just the notation for the variance of the subject effects.

As we recall from Chap. 3, this is just one ingredient in the calculation of the correlation. We also need to know the standard deviations for the measurements. Model (7.2) also indicates how to calculate the variance and hence the standard deviation:

$$\begin{aligned}\text{var}(\text{FECFAT}_{ij}) &= \text{var}(\text{SUBJECT}_i) + \text{var}(\epsilon_{ij}) \\ &= \sigma_{\text{subj}}^2 + \sigma_{\epsilon}^2\end{aligned}\tag{7.4}$$

so that

$$\text{SD}(\text{FECFAT}_{ij}) = \sqrt{\sigma_{\text{subj}}^2 + \sigma_{\epsilon}^2},$$

which is assumed to be the same for all observations. The result, (7.4), is noteworthy by itself, since it indicates that the variability in the observations is being decomposed into two pieces, or components, the variability due to subjects and the residual, or error, variance.

We are now in a position to calculate the correlation as the covariance divided by the standard deviations.

$$\begin{aligned}\text{corr}(\text{FECFAT}_{ij}, \text{FECFAT}_{ik}) &= \frac{\text{cov}(\text{FECFAT}_{ij}, \text{FECFAT}_{ik})}{\text{SD}(\text{FECFAT}_{ij})\text{SD}(\text{FECFAT}_{ik})} \\ &= \frac{\sigma_{\text{subj}}^2}{\sqrt{\sigma_{\text{subj}}^2 + \sigma_{\epsilon}^2} \sqrt{\sigma_{\text{subj}}^2 + \sigma_{\epsilon}^2}} \\ &= \frac{\sigma_{\text{subj}}^2}{\sigma_{\text{subj}}^2 + \sigma_{\epsilon}^2}.\end{aligned}\tag{7.5}$$

While the methods of the calculations are not so important, the intuition and results are. Namely that subject-to-subject variability simultaneously raises or lowers all the observations on a subject, thus inducing a correlation, and that the variability of an individual measurement can be separated into that due to subjects and residual variance.

Looking at the ANOVA table in Table 7.3, we have an estimate of σ_ϵ^2 , which is 106.99888. But what about an estimate for σ_{subj}^2 ? It would be almost correct to calculate the variance of the subject averages in the last column of Table 7.1, but this would be a bit too large since each subject average also has a small amount of residual variation as well. Taking this into account (see Problem 7.1) gives an estimate of 252.67.

Using this in (7.5) gives a correlation of $0.70 = 252.67/(252.67 + 107.00)$, not a particularly high value. So even a moderate value of the correlation can have a fairly dramatic effect on the analysis, which is why it is so important to recognize repeated measures or clustered-data situations. In this instance, the analysis ignoring the correlation led to nonsignificant results and inflated p -values. Unfortunately, the effect of ignoring the correlation can also make the p -values appear incorrectly small, as will be demonstrated in Sect. 7.4.4. So ignoring the correlation does not always produce a “conservative” result.

In this example, we are mainly interested in comparing the effect of the different pill types and the correlation within subjects must be accommodated in order to perform a proper analysis. The correlation is more of a nuisance. In other studies, the correlation will be the primary focus of the analysis, such as repeatability or validation studies or in analysis of familial aggregation of a disease. In the knee osteoarthritis example, the same radiographs were sent to different reading centers to check consistency of results across the centers. One of the primary parameters of interest was the correlation of readings taken on the same image.

7.1.3 *Estimates of the Effects of Pill Type*

What about estimating the effects of the various pill types or differences between them? The simple averages across the bottom of Table 7.1 give the estimates of the mean fecal fat values for each pill type. There is nothing better we can do in this balanced-data experiment. The same is true for comparing different pill types. For example, the best estimate of the difference between a coated capsule and an uncoated capsule would be the simple difference in means: $31.07 - 17.42 = 13.65$. That is, we do nothing different than we would with a one-way ANOVA (in which all the observations are assumed independent). This is an important lesson that we extend in the next section: the usual estimates based on the assumption of independent data are often quite good. It is the estimation of the standard errors and the tests (like the F -test) that go awry when failing to accommodate correlated data.

7.2 Hierarchical Data

The data structures we describe in this chapter and the analysis strategies are designed for hierarchical data. This is a somewhat vague term, but we now attempt a more formal definition.

Definition: *Hierarchical data* is data (responses or predictors) collected from or specific to different levels within a study.

Other terminologies for the same or related ideas are repeated measures data, longitudinal data, clustered data, and multilevel data. We next illustrate this definition in the context of two examples.

7.2.1 Example: Treatment of Back Pain

A more complicated example of a hierarchical model was first introduced in Chap. 1. In [Korff et al. \(1994\)](#), 44 primary care physicians in a large HMO were classified according to their practice style in treating back pain management (low, moderate, or high frequency of prescription of pain medication and bed rest). An average of 24 patients per physician were followed for 2 years (1 month, 1 year, and 2 year follow-ups) after the index visit. Outcomes included functional measures (pain intensity, activity limitation days, etc.), patient satisfaction (e.g., “After your visit with the doctor, you fully understood how to take care of your back problem”), and cost. Two possible questions are (1) Do physicians with different practice styles differ in function, satisfaction, or cost? and (2) How much of the variability in the responses is due to physician? In this example, there are three levels to the data structure: physicians, patients, and visits. Predictors could be physician-level variables like practice style and years of experience, patient-level variables like age and reason for the back pain, and visit-level variables like time since index visit. The data set is hierarchical because it has variables that are specific to each of the different levels (physician, patient, or visit) of the data.

7.2.2 Example: Physician Profiling

Common methods for the assessment of individual physicians’ performance at diabetes care were evaluated in [Hofer et al. \(1999\)](#). They studied 232 physicians from three sites caring for a total of 3,642 patients, and evaluated them with regard to their ability to control HbA_{1c} levels (a measure of control of blood sugar levels) and with regard to resource utilization. Various methods for obtaining physician level predictions were compared including age- and sex-adjusted averages, the calculation of residuals after adjusting for the case-mix of the patients, and

hierarchical modeling. They found that the first two methods overstate the degree to which physicians differ. This could have adverse consequences in falsely suggesting that some physicians (especially those with small numbers of patients) are over using resources or ineffectively treating patients.

As we will see explicitly later in the chapter, hierarchical analysis is more effective in this situation because it “borrows strength” across physicians in order to improve the predicted values for each physician. Said another way, we can use knowledge of the variation between and within physicians in order to quantify the degree of unreliability of individual physician’s averages and, especially for those with small numbers of patients, make significant adjustments.

7.2.3 Analysis Strategies for Hierarchical Data

As has been our philosophy elsewhere in this book, the idea is use simpler statistical methods unless more complicated ones are necessary or much more advantageous. That raises the basic question: Do we need hierarchical models and the attendant more complicated analyses? An important idea is the following. Observations taken within the same subgroup in a hierarchy are often more similar to one another than to observations in different subgroups, other things being equal. Equivalently, data which are clustered together in the same level of the hierarchy (data on the same physician, or on the same patient or in the same hospital) are likely to be correlated. The usual statistical methods (multiple regression, basic ANOVA, logistic regression, and many others) assume observations are independent. And we have seen in Sect. 7.1 the potential pitfalls of completely ignoring the correlation.

Are there simple methods we can use that accommodate the correlated data? Simpler approaches that get around the issue of correlation include separate analyses for each subgroup, analyses at the highest level in the hierarchy, and analyses on “derived” variables. Let us consider examples of each of these approaches using the back pain example.

7.2.3.1 Analyses for Each Subgroup

Analysis for each subgroup would correspond to doing an analysis for each of the 44 doctors separately. If there were sufficient data for each doctor, this might be effective for some questions, for example, the frequency with which patients for that physician understood how to care for their back. For other questions it would be less satisfactory, for example, how much more it cost to treat older patients. To answer this question, we would need to know how to aggregate the data across doctors. For yet other questions it would be useless. For example, comparing practice styles is a between-physician comparison and any within-physician analysis is incapable of addressing it.

7.2.3.2 Analysis at the Highest Level in the Hierarchy

An analysis at the highest level of the hierarchy would proceed by first summarizing the data to that level. As an example, consider the effect of practice style on the cost of treatment. Cost data would be averaged across all times and patients within a physician, giving a single average value. A simple analysis could then be performed, comparing the average costs across the three types of physicians. And by entering into the analysis a single number for each physician, we avoid the complication of having correlated data points through time on the same patient or correlated data within a physician.

There are several obvious drawbacks to this method. First, there is no allowance for differences in patient mix between physicians. For example, if those in the aggressive treatment group also tended to have older, higher cost patients we would want to adjust for that difference. We could consider having additional variables such as average age of the patients for each physician to try to accommodate this. Or a case mix difference of another type might arise: some physicians might have more complete follow-up data and have different proportions of data at the various times after the index visit. Adjusting for differences of these sorts is one of the key reasons for considering multipredictor models.

A second drawback of analysis at the highest level of the hierarchy is that some physicians will have large numbers of patients and others will have small numbers. Both will count equally in the analysis. This last point bears some elaboration. Some data analysts are tempted to deal with this point by performing a weighted analysis where the physician receives a weight proportional to the number of observations that went into their average values or the number of patients that contributed to the average. But this ignores the correlated nature of the data. If the data are highly correlated within a physician then additional patients from each physician contribute little additional information and all physicians' averages should be weighted equally regardless of how many patients they have. At the other extreme, if each patient counts as an independent data point, then the averages *should* be weighted by the numbers of patients.

If the data are correlated but not perfectly correlated, the proper answer is somewhere in between these two extremes: a physician with twice as many patients as another should receive more weight, but not twice as much. To determine precisely how much more requires estimation of the degree of correlation within a physician, i.e., essentially performing a hierarchical analysis.

7.2.3.3 Analysis on “Derived Variables”

A slightly more sophisticated method than simple averaging is what is sometimes called the use of “derived variables.” The basic idea is to calculate a simple, focused variable for each cluster or subgroup that can be used in a more straightforward analysis. A simple and often effective example of this method is calculation of a change score. Instead of analyzing jointly the before and after treatment values on a subject (with a predictor variable that distinguishes them), we instead calculate the change score.

Here are two other examples of this methodology. In a pharmacokinetic study, we might sample a number of subjects over time after administration of a drug and be interested in the average value of the drug in the bloodstream and how it changes with different doses of the drug. One strategy would be to analyze the entire data set (all subjects and all times) but then we would need to accommodate the correlated nature of the data across time within a person. A common alternative is to calculate, for each person, the area under the curve (AUC) of the concentration of the drug in the bloodstream versus time. This AUC value would then be subjected to a simpler analysis comparing doses (e.g., a linear regression might be appropriate). In the fecal fat example, the derived variable approach is quite effective. Suppose we were interested in the effect of coating a capsule. We can calculate the six differences in fecal fat measurements between the uncoated and coated capsule (one for each person) and do a one-sample or paired t -test on the six differences. See Problem 7.5. For the back pain example, the derived variable approach is not as successful. The unbalanced nature of the data makes it difficult to calculate an effective derived variable.

In summary, the use of hierarchical analysis strategies is clearly indicated in any of three situations:

- (1) When the correlation structure is of primary interest,
- (2) When we wish to “borrow strength” across the levels of a hierarchy in order to improve estimates, and
- (3) When dealing with highly unbalanced correlated data.

7.3 Longitudinal Data

In *longitudinal* studies, we are interested in the change in the value of a variable within a “subject” and we collect data repeatedly through time. For example, a study of the effects of alcohol might record a measure of sleepiness before and after administration of either alcohol or placebo. Interest is in quantifying the effect of alcohol on the *change* in sleepiness. This is often a good design strategy since each subject acts as their own control, allowing the elimination of variability in sleepiness measurements from person-to-person or even occasion-to-occasion within a person. For this strategy to be effective, the before and after measurements need to be at least moderately strongly positively correlated (otherwise taking differences increases the variability rather than reducing it).

As another example, the Study of Osteoporotic Fractures (SOF) is a longitudinal, prospective study of osteoporosis, breast cancer, stroke, and mortality. In 1986, SOF enrolled 9,704 women and continues to track these women with clinical visits every two years. Data from the first seven visits are now available to the public. The data include measures of BMD, BMI, hormones, tests of strength and function, cognitive exams, use of medication, health habits, and much more.

Some of the questions SOF can be used to answer are:

- (1) Is change in BMD related to age at menopause? Considered more generally, this is an analysis relating a time-invariant predictor, age at menopause, with changes in the outcome, BMD.
- (2) Is change in BMD related to change in BMI? This is an analysis relating a time-varying predictor, BMI, with changes in the outcome, BMD. BMI varies quite a lot between women, but also varies within a woman over time.
- (3) Which participants are likely to maintain cognitive function into their 9th and 10th decades of life? This involves predicting the cognitive trajectory of each of the participants from covariates and previously measured values of cognitive function.

We next consider how longitudinal data can be used to answer questions like (1) and (2) above. We deal with questions of prediction in Sect. 7.7.3.

7.3.1 Analysis Strategies for Longitudinal Data

Including a time variable (such as time since enrollment or visit number if they are approximately equally spaced in time) as a predictor captures the idea of change over time in the outcome. This is because the regression coefficient for a time variable measures the change in the outcome per unit change in time, just as in any regression. For example, if the outcome in a linear regression model was BMD and the time variable was years since enrollment in SOF, the meaning of the regression coefficient for time would be the change in mean BMD per year. If the outcome in a logistic regression model was use of hypertensive medication then the regression coefficient for time would be the change in log odds of using hypertensive medication per year.

But suppose, as above, interest focuses not on the change in BMD *overall*, but instead on whether it is related to age at menopause. Then the regression coefficient for time will vary with age at menopause. In statistical parlance, there will be an *interaction* between time and age at menopause as described in Sect. 4.6. Therefore, to capture the association of change in outcome over time with a time-invariant predictor, we need to include in our model an interaction term with the time variable. For example, to assess whether age at menopause was associated with the change in BMD, the regression model would need to include an interaction between time and age at menopause.

To graphically investigate whether there was an interaction, we divided age at menopause as above or below age 52 and fit a restricted cubic spline in visit with three knots, allowing interactions between age at menopause and visit and derived the predicted values. The commands and results are given in Table 7.4. The fitted model was then plotted versus visit and is given in Fig. 7.1. The relationship between BMD and visit appears curvilinear and those women with age at menopause greater

Table 7.4 Fitting of a restricted cubic spline relating BMD to age at menopause and visit in SOF

```
. mkspline visit_spl=visit, cubic nknots(3)

. regress totbmd i.meno_ov visit_spl* i.meno_ov#c.visit_spl*
```

Source	SS	df	MS	Number of obs =	22372
Model	2.82075406	5	.564150812	F(5, 22366) =	32.64
Residual	386.56757	22366	.017283715	Prob > F =	0.0000
Total	389.388324	22371	.017405942	R-squared =	0.0072
				Adj R-squared =	0.0070
				Root MSE =	.13147

totbmd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
1.meno_ov_52	.0168294	.008409	2.00	0.045	.0003471 .0333116
visit_spl1	-.0070843	.0011079	-6.39	0.000	-.009256 -.0049127
visit_spl2	.0037694	.0015891	2.37	0.018	.0006546 .0068841
meno_ov_52#					
c.visit_spl1					
1	-.0001347	.0025039	-0.05	0.957	-.0050424 .0047731
meno_ov_52#					
c.visit_spl2					
1	-.0002443	.0035122	-0.07	0.945	-.0071284 .0066398
_cons	.7549819	.0036908	204.56	0.000	.7477477 .762216


```
. predict pred_spl
```

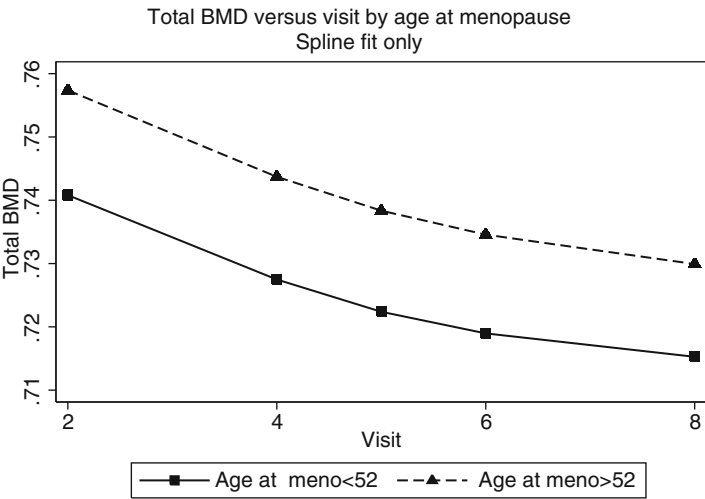


Fig. 7.1 Plot of spline fit to SOF BMD data by age at menopause category

than 52 may have slightly higher BMD values. However, the relationship of the change over time appears remarkably similar between the age at menopause groups, suggesting no time by age at menopause interaction. The analysis in Table 7.4 is

Table 7.5 Summary statistics for first- and last-born babies and the change score

summ initwght lastwght delwght

Variable	Obs	Mean	Std. Dev.	Min	Max
initwght	1000	3016.555	576.2185	815	4508
lastwght	1000	3208.195	578.3356	1210	5018
delwght	1000	191.64	642.3062	-1551	2700

adequate for visualizing the relationship between change in BMD over time and age at menopause, but improper for conducting a formal statistical analysis, since it does not accommodate the repeated measures nature of the data. We return to this example in Sect. 7.7 after we describe appropriate analysis strategies.

7.3.2 Analyzing Change Scores

In simple situations, there is a straightforward approach to analyzing longitudinal data—calculate the change scores (subtract the before measurement from the after measurement) as a derived variable and perform an analysis on the changes. In the alcohol example, we could simply perform a two-sample *t*-test using the change scores as data to compare the alcohol and placebo subjects.

We consider three approaches to analysis of before/after data that are commonly used: (1) analysis of change scores, (2) repeated measures analysis, and (3) analysis using the after measurement as the outcome and using the baseline measurement as a covariate (predictor). The justification for this last strategy is to “adjust for” the baseline value before looking for differences between the groups. How do these approaches compare?

7.3.2.1 Example: Birthweight and Birth Order

We consider an analysis of birthweights of first-born and last-born infants from mothers (each of whom had five children) from vital statistics in Georgia. We are interested in whether birthweights of last-born babies are different from first-born and whether this difference depends on the age of the woman when she had her first-born.

For the first question, we begin with the basic descriptive statistics given in Table 7.5, where lastwght in the variable containing the last-born birthweights, initwght indicates the first-born and delwght are the changes between last- and first-born within a woman. These show that last-born tend to be about 191 g heavier than first-born (the same answer is obtained whether you average the differences or take the difference between the averages). To accommodate the correlated data, we either perform a one-sample *t*-test on the differences or, equivalently, a paired *t*-test of the first and last births. A paired *t*-test gives a

Table 7.6 Regression of change in birthweight on centered initial age

regress delwght cinitage

Source	SS	df	MS	Number of obs = 200		
Model	163789.382	1	163789.382	F(1, 198) =	0.39	
Residual	82265156.7	198	415480.589	Prob > F =	0.5308	
				R-squared =	0.0020	
				Adj R-squared =	-0.0031	
Total	82428946.1	199	414215.809	Root MSE =	644.58	

delwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cinitage	8.891816	14.16195	0.63	0.531	-19.03579	36.81942
_cons	191.64	45.57854	4.20	0.000	101.7583	281.5217

t -statistic of 4.21, with 199 degrees of freedom (since there are 200 mothers) with a corresponding p -value that is approximately 0.

What about the relationship of the change in birthweight to the mother's initial age? For this, we conduct a simple linear regression of the change in birthweight regressed on initial age, where we have centered initial age (`cinitage`) by subtracting the mean initial age. The results are displayed in Table 7.6 with the interpretation that each increase of one year in initial age is associated with an additional 8.9 g difference between the first and last birthweights. This is not statistically significant ($p = 0.53$). When centered age is used, the intercept term (`_cons`) is also the average difference.

To conduct a repeated measures analysis, the data are first reordered to have a single column of data containing the birthweights and an additional column, birth order, to keep track of whether it is a first, second, third, fourth, or fifth birth. The output for the repeated measures analysis using only the first and last births is displayed in Table 7.7, for which we leave the details to the next section. However, many of the elements are similar to the regression analysis in Table 7.6. The term listed under `birthord#c.cinitage` is the interaction of birth order and centered initial age. It thus measures how the *difference* in birthweights between first- and last-born is related to centered initial age, that is, whether the change score is related to initial age, the same question as the regression analysis. As is evident, the estimated coefficient is identical and the standard error is virtually the same. They are not exactly the same because slightly different modeling techniques are being used (regression versus GEE, short for generalized estimating equations). The overall difference between first- and last-born is also displayed in the repeated measures analysis (again with the same coefficient and a very similar standard error and p -value) and is associated with the birth order term in the model. Finally, the average for first births is displayed as the intercept (see Problem 7.7). So, at a cost of more complication, the repeated measures analysis answers both questions of interest.

A different sort of analysis is to conduct a multiple regression with two predictor variables, initial age (centered) and first-born birthweight. The idea is to “adjust” the values of last-born weight by the first-born weight and then look for an effect due

Table 7.7 Repeated measures regression of birthweight on birth order and centered initial age

```
. xtgee bweight i.birthord cinitage i.birthord#c.cinitage
> if birthord==1|birthord==5, i(momid)
```

GEE population-averaged model

Group variable:	momid	Number of obs	=	400
Link:	identity	Number of groups	=	200
Family:	Gaussian	Obs per group: min	=	2
Correlation:	exchangeable	avg	=	2.0
		max	=	2
		Wald chi2(3)	=	26.47
Scale parameter:	323645.4	Prob > chi2	=	0.0000

bweight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
5.birthord	191.64	45.35007	4.23	0.000	102.7555 280.5245
cinitage	25.13981	12.4992	2.01	0.044	.6418238 49.6378
birthord# c.cinitage					
5	8.891816	14.09096	0.63	0.528	-18.72596 36.50959
_cons	3016.555	40.22719	74.99	0.000	2937.711 3095.399

Table 7.8 Regression of final birthweight on centered initial age, adjusting for first birthweight

```
regress lastwght cinitage initwght if birthord==5
```

Source	SS	df	MS	Number of obs	=	200
Model	10961363.1	2	5480681.54	F(2, 197)	=	19.33
Residual	55866154.3	197	283584.54	Prob > F	=	0.0000
				R-squared	=	0.1640
				Adj R-squared	=	0.1555
Total	66827517.4	199	335816.67	Root MSE	=	532.53

lastwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
cinitage	24.90948	11.81727	2.11	0.036	1.604886 48.21408
initwght	.3628564	.0660366	5.49	0.000	.232627 .4930858
_cons	2113.619	202.7309	10.43	0.000	1713.817 2513.42

to initial age. Table 7.8 gives the results of that analysis, which are quite different than the previous analyses. Now, initial age has a much larger coefficient and is statistically significant ($p = 0.036$).

The intuitive explanation for why this analysis is so different starts with the observation that the coefficient for birthweight of the first-born is approximately 0.363. So, using BW_k to denote the birthweight of the k th born child, we can think of the fitted model as

$$BW_5 = 2113.619 + .363BW_1 + 24.909 \text{ Centered initial age} \quad (7.6)$$

or, taking BW_1 to the left side of the equation,

$$BW_5 - .363BW_1 = 2113.619 + 24.909 \text{ Centered initial age.} \quad (7.7)$$

That is, this analysis is not purely looking at differences between last and first birthweight since we are only subtracting off a fraction of the initial birthweight. Since birthweights are more highly correlated with initial age than is the difference, this stronger relationship reflects the fact that the results are close to a regression of BW_5 on initial age.

In observational studies, such as this one, using baseline values of the outcome as a predictor is not a reliable way to check the dependence of the change in outcome on a predictor. In randomized studies, where there should be no dependence between treatment effects and the baseline values of the outcome, this may be a more reasonable strategy.

7.3.2.2 When to Use Repeated Measures Analyses

In the Georgia birthweight example, we see that analysis by change scores or by a repeated measures analysis gives virtually identical and reasonable results. The analysis using the baseline value as a predictor is more problematic to interpret.

If the analysis of change scores is so straightforward, why consider the more complicated repeated measures analysis? For two time points and no (or little) missing data, there is little reason to use the repeated measures analysis. However, in the birthweight example there are three intermediate births we have ignored that should be included in the analysis. In the alcohol example, it would be reasonable to measure the degree of sleepiness at numerous time points post-administration of alcohol (or placebo) to track the speed of onset of sleepiness and when it wears off. When there are more than two repeated measures, when the measurements are recorded at different times and/or when there is missing data, repeated measures analysis can more easily accommodate the data structure than attempting change score analyses. We now consider methods for multiple time points.

7.4 Generalized Estimating Equations

There are two main methods for accommodating correlated data. The first we will consider is a technique called *generalized estimating equations*, often abbreviated GEE. A key feature of this method is the option to estimate the correlation structure from the data without having to assume that it follows a prespecified structure.

Before embarking on an analysis, we will need to consider five aspects of the data:

- (1) What is the distributional family (for fixed values of the predictors) that is appropriate to use for the outcome variable? Examples are the normal, binary, and binomial families.
- (2) Which predictors are we going to include in the model?

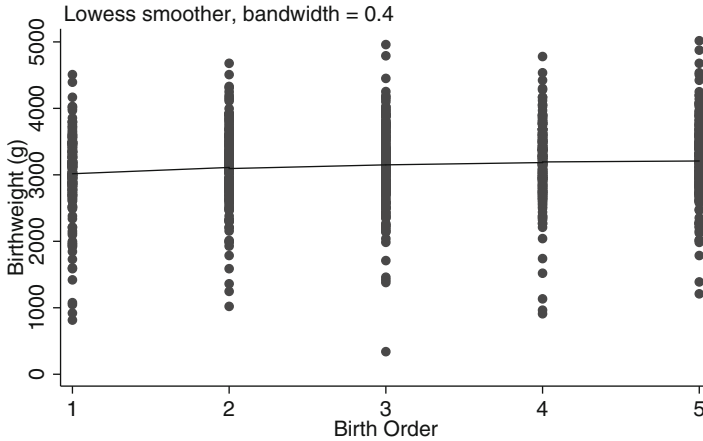


Fig. 7.2 Plot of birthweight (g) versus birth order

- (3) In what way are we going to link the predictors to the data? (Through the mean? Through the logit of the risk? Some other way?)
- (3) What correlation structure will be used or assumed temporarily in order to form the estimates?
- (4) Which variable indicates how the data are clustered?

The first three of these decisions we have been making for virtually every method described in this book. For example, the choice between a logistic and linear regression hinges on the distribution of the outcome variable, namely logistic for binary outcome and linear for continuous, approximately normal outcomes. Chapter 10 discusses the choice of predictors to include in the model (and is a focus of much of this book) and the third has been addressed in specific contexts, e.g., the advantage of modeling the log odds in binary data. The new questions are really the fourth and fifth and have to do with how we will accommodate the correlations in the data. We start by considering an example.

7.4.1 Example: Birthweight and Birth Order Revisited

We return to the Georgia birthweight example and now consider all five births. Recall that we are interested in whether birthweight increases with birth order and mothers' age. Figure 7.2 shows a plot of birthweight versus birth order with both the average birthweights for a given birth order and a LOWESS smooth superimposed. Inspection of the plot suggests we can model the increase as a linear function. A simple linear regression analysis of birthweight versus birth order gives a t -statistic for the slope coefficient of 3.61, which is highly statistically significant. But this analysis would be wrong (why?).

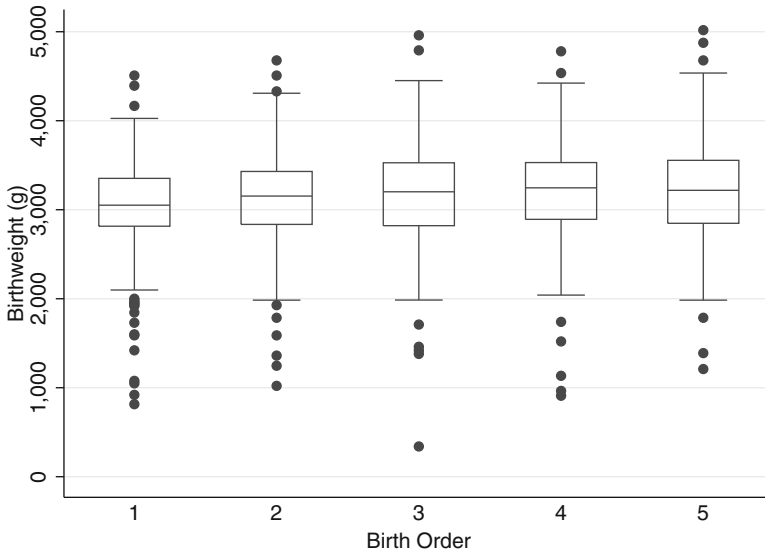


Fig. 7.3 Boxplots of birthweight (g) versus birth order

Recall that the paired t -test using just the first and last births gave a t -statistic of 4.21, even more highly statistically significant. This is perhaps a bit surprising since it discards the data from the three intermediate births.

The explanation for this apparent paradox is that the paired t -test, while using less of the data, does take advantage of the fact that birth order is a within mother comparison. It exploits the correlation of birthweights within a mom in order to make a more precise comparison. Of course, an even better analysis is to use all of the data and accommodate the correlated structure of the data, which we now proceed to do.

7.4.1.1 Analysis

To analyze the Georgia babies dataset, we need to make the decisions outlined above. The outcome variable is continuous, so a logical place to start is to assume it is approximately normally distributed. Figure 7.3 shows boxplots of birthweight by birth order, suggesting that the normality and equal variance assumptions are reasonable. Figure 7.2 has suggested entering birth order as a linear function, which leaves us with the accommodation of the correlation structure.

The data are correlated because five birthweights come from each mother and hence the clustering aspect is clear, leaving us with the decision as to how to model the correlation of measurements taken through time. Figure 7.4 gives a matrix plot of each birthweight against each of the others while Table 7.9 gives the values of the

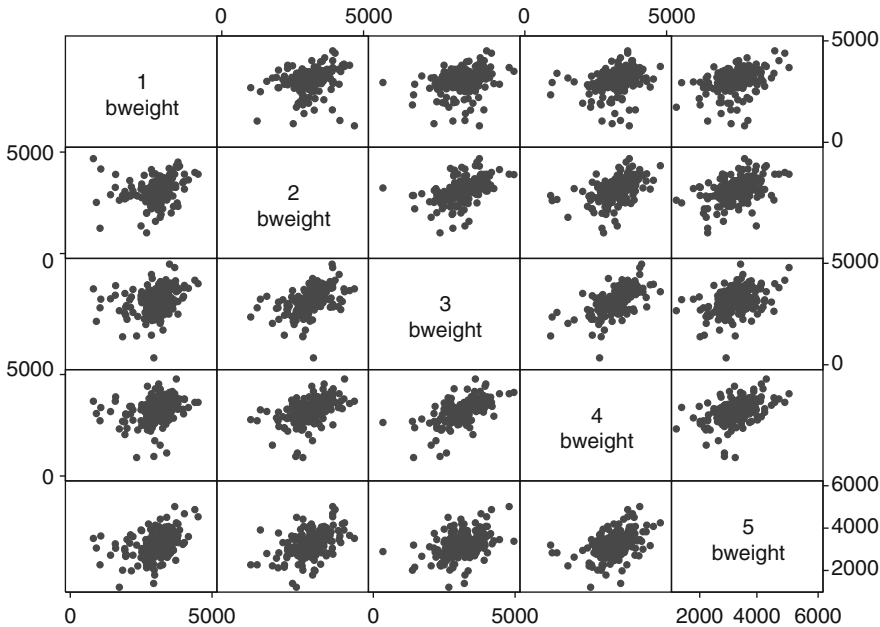


Fig. 7.4 Matrix plot of birthweights for different birth orders

Table 7.9 Correlation of birthweights for different birth orders

```
. corr bweight1 bweight2 bweight3 bweight4 bweight5 (obs=200)
```

	bweight1	bweight2	bweight3	bweight4	bweight5
bweight1	1.0000				
bweight2	0.2282	1.0000			
bweight3	0.2950	0.4833	1.0000		
bweight4	0.2578	0.4676	0.6185	1.0000	
bweight5	0.3810	0.4261	0.4233	0.4642	1.0000

correlation coefficients. Correlations with the first birthweight might be a bit lower, but the graphs suggest that a tentative assumption of all the correlations being equal would not be far off.

7.4.2 Correlation Structures

Dealing with correlated data typically means making some type of assumption about the form of the correlation among observations taken on the same subject, in the same hospital, on the same mouse, etc. For the Georgia babies data set in the previous section, we noted that assuming all the correlations to be equal might be a

reasonable assumption. This form of correlation is termed exchangeable and means that all correlations (except those variables with themselves) are a common value, which is typically estimated from the data. This type of structure is suitable when there is nothing to distinguish one member of a cluster from another (e.g., patients within a physician) and is the genesis for its name (patients within a doctor can be regarded as interchangeable or exchangeable). This sort of assumption is appropriate in the absence of other data structure, such as measurements taken through time or space.

If measurements are taken through time on the same person, it may be that observations taken more closely in time are more highly correlated. Another common correlation structure is the autoregressive structure, which exhibits this feature. In the simplest form of an *auto regressive* process (first order or AR(1)) the correlation between observations one time unit apart is a given value ρ , that between observations two time units apart ρ^2 , three time units apart ρ^3 , etc. Simple arithmetic calculation shows this drops off rapidly to zero (e.g., $0.6^5 = 0.08$) so this assumption would only be appropriate if the correlation between observations taken far apart in time was small and would not be appropriate in cases where stable over time characteristics generated the association. For example, SBP would be relatively stable over time for an individual. Even though observations taken more closely together in time would be slightly more highly correlated, an exchangeable correlation structure might come closer to the truth than an autoregressive one.

Other, less structured, assumptions can be made. In Stata, other options are *unstructured*, *non-stationary*, and *stationary*. All are related to the idea of observations within a cluster being ordered, such as by time. As its name suggests, the unstructured form estimates a separate correlation between observations taken on each pair of “times”. The non-stationary form is similar, but assumes all correlations for pairs separated far enough in time are zero. The stationary form assumes equal correlation for all observations a fixed time apart and, like non-stationary, assumes correlations far enough apart in time have correlation zero. For example, stationary of order 2 would assume that observations taken at time points 1 and 3 would have the same correlation as time points 2 and 4, but this might be different from the correlation between observations taken at times 2 and 3. Also, correlations for observations 3 or more time periods apart would be assumed to be zero.

If the correlation structure is not the focus of the analysis, it might seem that the unstructured form is best, since it makes no assumptions about the form of the correlation. However, there is a cost: even with a small number of time points, we are forced to estimate quite a large number of correlations. For instance, with measurements on five time points for each subject, there are ten separate correlations to estimate. This can cause a decrease in the precision of the estimated parameters of interest, or, worse yet, a failure in being able to even fit the model.

This is especially true in situations where the data are not collected at rigid times. For example, in the Nutritional Prevention of Cancer trials (Clark et al. 1996), long-term follow-up was attempted every six months. But the intervals varied widely in practice and quickly were out of synchronization. Estimation of the correlations

between all pairs of distinct times would require literally hundreds of estimated correlations. Use of the unstructured, and, to some extent, the stationary and non-stationary correlation assumptions should be restricted to situations where there are large numbers of clusters, e.g., subjects, and not very many distinct pairs of observation times.

Diagnosis and specification of the “correct” correlation structure is very difficult in practice. One method of addressing these problems is via a *working* correlation assumption and the use of “robust” standard errors, which is the next topic.

7.4.3 *Working Correlation and Robust Standard Errors*

Given the difficulty of specifying the “correct” correlation structure, a compromise is possible using what are called *robust standard errors*. The idea is to make a temporary or working assumption as to the correlation structure in order to form the estimates but to properly adjust the standard errors of those estimates for the correlation in the data. For example, we might temporarily assume the data are independent and conduct a standard logistic regression. The estimates from the logistic regression will be fairly good, even when used with correlated data, but the standard errors will be incorrect, perhaps grossly so. The solution is to use the estimates but empirically estimate their proper standard errors. Another possibility is to make a more realistic assumption, such as an exchangeable working correlation structure; in some circumstances a gain in efficiency may result.

Then, after the model coefficients have been estimated using the working correlation structure, within-subject residuals are used to compute robust standard errors for the coefficient estimates. Because these standard errors are based on the data (the residuals) and not the assumed working correlation structure, they give valid (robust) inferences for large sized samples as long as the other portions of the model (distribution, link and form of predictors) are correctly specified, even if our working correlation assumption is incorrect. Use of robust standard errors is not quite the same as using an unstructured correlation since it bypasses the estimation of the correlation matrix to directly obtain the standard errors. Avoiding estimation of a large number of correlations is sometimes an advantage, though in cases where both approaches can be used they often give similar results.

The key to the use of this methodology is to have sufficient numbers of subjects or clusters so that the empirical estimate of the correlation is adequate. The GEE approach, which goes hand in hand with estimation with robust standard errors, will thus work best with relatively few time points and relatively more subjects. It is hard to give specific guidelines, but this technique could be expected to work well with 100 subjects, each measured at 5 time points but much less well with 20 subjects, each measured at 12 time points, especially if the times were not the same for each subject.

Table 7.10 Generalized estimating equations analysis using robust standard errors

```
. xtgee bweight birthord initage, i(momid) corr(exch) robust
```

GEE population-averaged model

Group variable: momid

Link: identity

Family: Gaussian

Correlation: exchangeable

Scale parameter: 324458.3

Number of obs = 1000

Number of groups = 200

Obs per group: min = 5

avg = 5.0

max = 5

Wald chi2(2) = 27.95

Prob > chi2 = 0.000

standard errors adjusted for clustering on momid

bweight	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	46.608	10.02134	4.65	0.000	26.96653	66.24947
initage	26.73226	10.1111	2.64	0.008	6.914877	46.54965
_cons	2526.622	177.2781	14.25	0.000	2179.164	2874.081

7.4.4 Tests and Confidence Intervals

Hypothesis testing with GEE uses Wald tests, in which the estimates divided by their robust standard errors are treated as approximately normal to form *z*-statistics. Likewise, approximate 95% confidence intervals are based on normality by calculating the estimate plus or minus 1.96 standard errors. Table 7.10 shows the analysis with an exchangeable working correlation structure and robust standard errors. Some comments are in order about the form of the command. `xtgee` is a regression type command with numerous capabilities. In its basic form, exhibited in Table 7.10, it performs a linear regression (link of identity) of birthweight (`bweight`) on birth order (`birthord`) and mother’s age at first birth (`initage`) with an assumed exchangeable correlation structure (`corr(exch)`) within mother (`i(momid)`). The `robust` option requests the use of robust standard errors.

For comparison sake, Table 7.11 gives the analysis without robust standard errors. There is little difference, though this is to be expected since the preliminary look at the data suggested that the exchangeable assumption would be a reasonable one.

Looking at the analysis with the robust standard errors, the interpretation of the coefficient is the same as for a linear regression. With each increase of initial age of one year, there is an associated increase in average birthweight of about 26.7 g. This result is highly statistically significant, with a *p*-value of 0.008.

Lest the reader think that the analysis is impervious to the correlational assumptions, Table 7.12 shows what happens to the estimates and standard errors under three different correlation structures both with and without the use of robust standard errors. As expected, the estimates are all similar (the independence and exchangeable are equal because of the balanced nature of the data—five observations per mom with the same values of birth order), though there are

Table 7.11 Generalized estimating equations analysis without robust standard errors

```
. xtgee bweight birthord initage, i(momid) corr(exch)
```

GEE population-averaged model

Group variable: momid

Link: identity

Family: Gaussian

Correlation: exchangeable

Scale parameter: 324458.3

Number of obs = 1000

Number of groups = 200

Obs per group: min = 5

avg = 5.0

max = 5

Wald chi2(2) = 30.87

Prob > chi2 = 0.000

(standard errors adjusted for clustering on momid)

bweight	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	46.608	9.944792	4.69	0.000	27.11657	66.09943
initage	26.73226	8.957553	2.98	0.003	9.175783	44.28874
_cons	2526.622	162.544	15.54	0.000	2208.042	2845.203

Table 7.12 Comparison of the estimated coefficients for `initage` and its standard error for various forms of correlation, with and without robust standard errors

Working correlation	Robust SE?	Coefficient estimate	Standard error	Z-statistic	p-value
Independence	No	26.73	5.60	4.78	0.000
Exchangeable	No	26.73	8.96	2.98	0.003
Autoregressive(1)	No	27.41	7.82	3.51	0.000
Independence	Yes	26.73	10.11	2.64	0.008
Exchangeable	Yes	26.73	10.11	2.64	0.008
Autoregressive(1)	Yes	27.41	9.69	2.83	0.005

slight variations depending on the assumed working correlation. The estimates are unaffected by the use of robust standard errors.

However, the standard errors and hence Wald statistics and *p*-values are quite different. Those using the incorrect assumptions of independence or autoregressive structure (given in the rows without robust standard errors) are too small, yielding Wald statistics and *p*-values that are incorrect. Looking at the rows corresponding to the use of robust standard errors shows how the incorrect working assumptions of independence or autoregressive get adjusted and now have standard errors that are much more alike. As with any different methods of estimation slight differences do, however, remain.

For the `initage` coefficient the *p*-values assuming independence or, to a lesser extent, autoregressive, are falsely small, but standard errors and *p*-values can, in general, be incorrect in either direction. For example, the `birthorder` effect has a standard error of almost 13 assuming independence, but a standard error of about 10 under an exchangeable correlation (Table 7.11) or under a working exchangeable correlation structure using robust standard errors (Table 7.10).

Table 7.13 Generalized estimating equation logistic regression

```
. xtgee lowbrth birthord initage, i(momid) corr(exch) family(binomial) ///
> link(logit) robust ef
```

GEE population-averaged model		Number of obs	=	1000
Group variable:	momid	Number of groups	=	200
Link:	logit	Obs per group: min	=	5
Family:	binomial	avg	=	5.0
Correlation:	exchangeable	max	=	5
		Wald chi2(2)	=	10.64
Scale parameter:	1	Prob > chi2	=	0.0049

(standard errors adjusted for clustering on momid)

lowbrth	Odds Ratio	Semi-robust Std. Err.	z	P> z	[95% Conf. Interval]
birthord	.9204098	.03542	-2.16	0.031	.8535413 .9925168
initage	.9148199	.0312663	-2.60	0.009	.8555464 .9781999

7.4.5 Use of xtgee for Clustered Logistic Regression

As mentioned above, `xtgee` is a very flexible command. Another of its capabilities is to perform logistic regression for clustered data. We again analyze the Georgia birthweight data but instead use as our outcome the binary variable `lowbirthweight` (`lowbrth`) which is one if the birthweight is less than 3,000g and zero otherwise. Since the data are binary, we adapt `xtgee` for logistic regression by specifying `family(binomial)` and `link(logit)`. As before, we specify `i(momid)` to indicate the clustering, `corr(exch)` for an exchangeable working correlation, and `robust` to calculate robust standard errors; also we add the option `ef` to get odds ratios instead of log odds. Table 7.13 displays the analysis. The estimated odds ratio for birth order is about 0.92, with the interpretation that the odds of a low-birthweight baby decrease by 8% with each increase in birth order. We see that `initage` is still statistically significant, but less so than in the analysis of actual birthweight. This serves as a warning as to the loss of information possible by unnecessarily dichotomizing a variable.

7.5 Random Effects Models

The previous section discussed the use of generalized estimating equations for the accommodation of correlated data. This approach is limited in that

- (1) It is restricted to a single level of clustering,
- (2) It is not designed for inferences about the correlation structure,
- (3) It does not give predicted values for each cluster or level in the hierarchy.

A different approach to this same problem is the use of what are called *random effects* models.

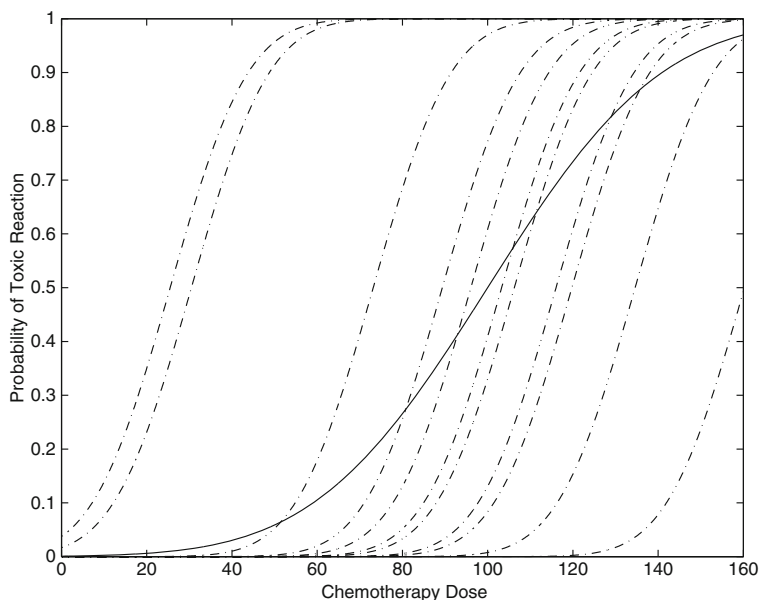


Fig. 7.5 Marginal versus conditional logistic models

First we need to consider two different modeling approaches that go by the names marginal and conditional. These are two common modeling strategies with which to incorporate correlation into a statistical model:

Marginal: Assume a model, e.g., logistic, that holds averaged over all the clusters (sometimes called population averaged). Coefficients have the interpretation as the average change in the response (over the entire population) for a unit change in the predictor. Alternatively, we can think of the coefficient as the difference in the mean values of randomly selected subjects that differ by one unit in the predictor of interest (with all the others being the same).

Conditional: Assume a model specific to each cluster (sometimes called subject-specific). Coefficients have the interpretation as the change in the response for each cluster in the population for a unit change in the predictor. Alternatively, we can think of the coefficient as representing the change within a subject when the predictor of interest is increased by one (holding all the others constant).

In the conditional modeling approach, marginal information can be obtained by averaging the relationship over all the clusters.

On the face of it, these would seem to be the same. But they are not. Here is a hypothetical example. Suppose we are modeling the chance that a patient will be able to withstand a course of chemotherapy without serious adverse reactions. Patients have very different tolerances for chemotherapy, so the curves for individual subjects are quite different. Those patients with high tolerances are shifted to the right of those with low tolerances (see Fig. 7.5). The individual curves are

subject-specific or conditional on each person. The population average or marginal curve is the average of all the individual curves and is given by the solid line in Fig. 7.5 and has quite a different slope than any of the individual curves. This emphasizes that it is important to keep straight which type of model is being used so as to be able to provide proper interpretations and comparisons.

The generalized estimating equations (GEEs) approach most always (always when using `xtgee`) fits a marginal model. Random effects models typically adopt the conditional approach.

Conditional models are usually specified by declaring one or more of the categorical predictors in the model to be *random factors*. (Otherwise they are called *fixed factors*.) Models with both fixed and random factors are called *mixed models*.

Definition: If a distribution is assumed for the levels of a factor, it is a *random factor*. If the values are fixed, unknown constants (to be estimated as model coefficients) it is a *fixed factor*.

The declaration of a factor to be random has several ramifications:

- Scope of inference: Inferences can be made on a statistical basis to the population from which the levels of the random factor have been selected.
- Incorporation of correlation in the model: Observations that share the same level of the random effect are being modeled as correlated.
- Accuracy of estimates: Using random factors involves making extra assumptions but gives more accurate estimates.
- Estimation method: Different estimation methods must be used.

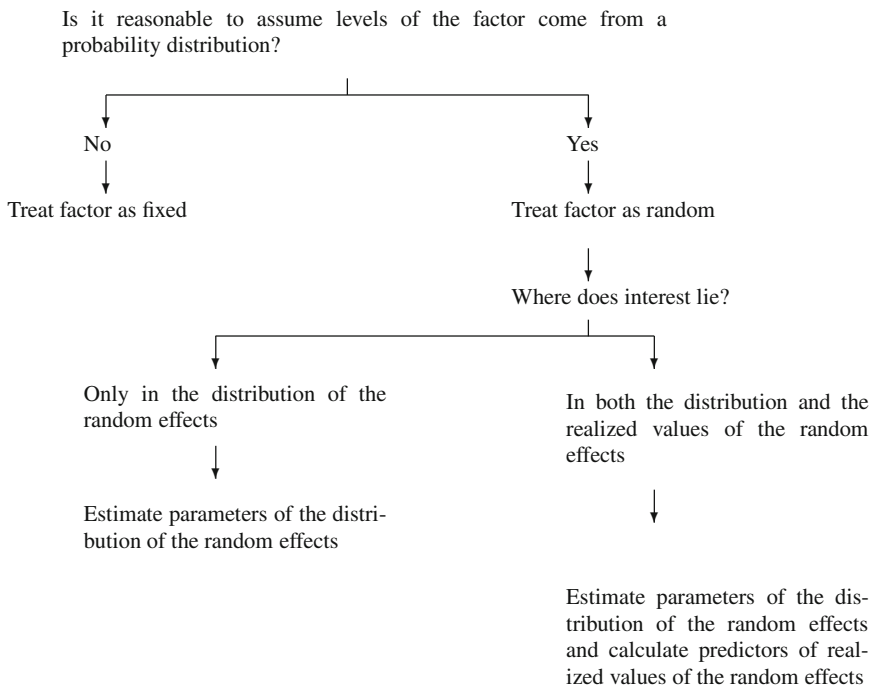
How do we decide in practice as to which factors should be declared random versus fixed? The decision tree in Table 7.14 may be useful in deciding whether the factor is to be considered as fixed or random.

7.6 Re-Analysis of the Georgia Babies Data Set

For the Georgia babies dataset, a random effects assumption for the moms is quite reasonable. We want to regard these particular moms as a sample from a larger sample of moms. Correspondingly the moms' effects on birthweights are easily envisioned as being selected from a distribution of all possible moms.

Stata has a number of commands for conducting random effects analyses; we will focus on two of them: `xtmixed` and `xtmelogit`. The first, `xtmixed`, fits linear mixed models to approximately normally distributed outcomes. The latter, `xtmelogit`, is for mixed models with binary outcomes.

The command syntax is somewhat different from that of `xtgee` because of the need to distinguish the fixed from the random factors. The fixed effect predictors follow the outcome variable in the commands, as is typical of regression commands.

Table 7.14 Decision tree for deciding between fixed and random

However, the random effects are listed after two vertical bars, as displayed in Table 7.15. The colon following the random effect indicates that the model should include random intercepts for each level of that random effect.

The random effects model we fit is similar to that of (7.2):

$$\begin{aligned} \text{BWEIGHT}_{ij} &= \text{birthweight of baby } j \text{ for mom } i \\ &= \beta_0 + \text{MOM}_i + \beta_1 \text{BIRTHORD}_{ij} + \beta_2 \text{INITAGE}_i + \epsilon_{ij}, \end{aligned}$$

with

$$\begin{aligned} \epsilon_{ij} &\sim \text{i.i.d } \mathcal{N}(0, \sigma_\epsilon^2) \\ \text{MOM}_i &\sim \text{i.i.d } \mathcal{N}(0, \sigma_M^2). \end{aligned} \tag{7.8}$$

Table 7.15 gives the analysis fitting this clustered-data linear regression model. For a linear regression model, the random effects assumption is equivalent to an exchangeable correlation structure as demonstrated in (7.5). Furthermore, for linear models with identity link functions, the marginal and conditional models are equivalent. Hence the random effects analysis reproduces the analysis with an assumed exchangeable correlation structure as given in Table 7.11.

Table 7.15 Linear mixed model analysis of the birthweight data

. xtmixed bweight birthord initage|| momid:

Mixed-effects REML regression
Group variable: momid

Number of obs = 1000
Number of groups = 200
Obs per group: min = 5
 avg = 5.0
 max = 5

Log restricted-likelihood = -7649.3763

Wald chi2(2) = 30.75
Prob > chi2 = 0.0000

bweight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	46.608	9.951013	4.68	0.000	27.10437	66.11163
initage	26.73226	9.002682	2.97	0.003	9.087332	44.3772
_cons	2526.622	163.3388	15.47	0.000	2206.484	2846.761

Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
momid: Identity					
	sd(_cons)	358.1761	23.71804	314.5799	407.8142
	sd(Residual)	445.0228	11.13253	423.7297	467.3859

LR test versus linear regression: chibar2(01) = 209.20
Prob >= chibar2 = 0.0000

We do, however, have extra output in the random effects analysis. First, the standard deviation of the mom effects, σ_M is equal to 358.1761. This is listed in the output as `sd(_cons)` because it is the standard deviation of the intercepts (or constant terms) associated with each mom. The interpretation of the standard deviation of the mom effects is that it is the standard deviation (across moms) of the true average birthweight per mom. Second is an estimate of the residual standard deviation of 445.0228, from which we can calculate the intramom correlation. Using (7.5), the within mom correlation of any two birthweights is estimated to be $358.1761/(358.1761 + 445.0228) = 0.45$. And third, a test of the null hypothesis of whether the mom-to-mom variation can be considered to be zero, which can be easily rejected using a $\bar{\chi}^2$ -test. This is given at the bottom of the Stata output and labeled `chibar2`, short for chi-bar-squared, which has a p -value of approximately 0.

7.7 Analysis of the SOF BMD Data

We return to the Study of Osteoporotic Fractures analysis of the relationship between change in BMD over time and age at menopause (categorized as over or under age 52) that we introduced in Sect. 7.3.1. A primary consideration is how to handle the time variable, visit, which takes on the discrete values, 2, 4, 5, 6, and

8. If we think of the outcome (in this case BMD) evolving smoothly over time, we are naturally led to modeling the trajectory of change using a functional form, for example, a linear trend over time. We would generally like to characterize the trajectory as simply as we can while still using an adequately fitting model. This leads to a natural “ladder” of handling a time predictor like visit, starting from a simple (to model and interpret) linear relationship with time. But this can be quite restrictive and we may need to move up to more flexible models to obtain an adequate fit, for example, also including quadratic functions of time (or even higher degree polynomials) or using a spline (flexible smooth) fit. Failing a simple description with polynomials or splines, and in cases where the times take on a small number of discrete values it may be most expedient to simply handle the time variable as categorical. Moving up the “ladder,” we can test statistical significance of the need to utilize the more complicated models.

Recall that the strategy is to include interactions of baseline variables (in this case age at menopause over age 52) with the time variable(s) to check whether there are interactions. Figure 7.1 makes it clear that we need to consider the possibility of a non-linear relationship with visit, so we accommodate visit by using restricted cubic splines. Table 7.16 gives the analysis using GEEs. Neither of the interaction terms with the spline variables is statistically significant so there is no evidence that age at menopause is related to *change* in BMD over time. The spline terms for visit are, themselves, highly statistically significant, indicating that there are changes in BMD over time (unrelated to age at menopause). A comparison with a linear relationship (not shown here) indicates that it is inadequate for describing the changes over time. Consistent with Fig. 7.1, there is a statistically significant difference of about 0.017 between the age at menopause groups across all the visits.

7.7.1 Time Varying Predictors

Age at menopause does not change over time and so is a time-invariant or baseline predictor and we checked for its relationship with changes in BMD by including interactions with the time variables. We next consider the relationship of BMD with BMI, which does change over time within a participant (as well as between participants) and so is a time-varying predictor. How should we include it in the model? Consider a simple model for the measurement on the i th woman at time t with the only predictor being BMI:

$$\text{BMD}_{it} = \beta_0 + \beta_1 \text{BMI}_{it} + \epsilon_{it}. \quad (7.9)$$

Using (7.9) at time $t + 1$ and subtracting (7.9) from it gives

$$\begin{aligned} \text{BMD}_{i,t+1} - \text{BMD}_{it} &= (\beta_0 + \beta_1 \text{BMI}_{i,t+1} + \epsilon_{i,t+1}) - (\beta_0 + \beta_1 \text{BMI}_{it} + \epsilon_{it}) \\ &= \beta_1 (\text{BMI}_{i,t+1} - \text{BMI}_{it}) + \epsilon_{i,t+1} - \epsilon_{it}. \end{aligned} \quad (7.10)$$

Table 7.16 Generalized estimating equations analysis of the SOF BMD data

```
. xtgee totbmd i.meno_ov_52 visit_spl* i.meno_ov_52#c.visit_spl*, ///
> i(id) robust
```

GEE population-averaged model

Group variable: id

Link: identity

Family: Gaussian

Correlation: exchangeable

Scale parameter: .0174184

Number of obs = 22372

Number of groups = 7004

Obs per group: min = 1

avg = 3.2

max = 5

Wald chi2(5) = 2529.68

Prob > chi2 = 0.0000

(Std. Err. adjusted for clustering on id)

	totbmd	Coef.	Semirobust Std. Err.	z	P> z	[95% Conf. Interval]	
1.meno_ov_52		.0174495	.0040542	4.30	0.000	.0095033	.0253956
visit_spl1		-.0088637	.0003067	-28.90	0.000	-.0094648	-.0082626
visit_spl2		-.000053	.0004967	-0.11	0.915	-.0010265	.0009206
meno_ov_52#							
c.visit_spl1							
1		.0000433	.0006456	0.07	0.947	-.0012221	.0013086
meno_ov_52#							
c.visit_spl2							
1		-.0001972	.0010286	-0.19	0.848	-.0022132	.0018188
_cons		.757436	.0017974	421.40	0.000	.7539131	.760959

In words, the change in BMD is related to the change in BMI. The import of (7.10) is that, if we fit a model relating the outcome to a time-varying predictor, the regression parameter for the time-varying predictor has the interpretation as the change in the outcome associated with a change in the predictor. That is, it is inherently able to address a longitudinal question.

Table 7.17 gives a mixed model analysis of the relationship between BMD and BMI. Several comments are in order. The model being fit allows for flexible trends over visit, by using a restricted cubic spline in visit. It also allows each participant to have their own intercept and linear trend over visits, through the `id:visit` option. The `cov(uns)` option allows those random intercepts and trends to have arbitrary (unstructured) standard deviations and correlations. This is generally appropriate: the intercepts and trends are measured on completely different scales and are unlikely to have the same standard deviation and, in practice, they are often correlated.

The analysis indicates that there is a highly statistically significant relationship between BMD and BMI. BMD (and perhaps BMI) are measured in units that are unfamiliar to many and so it is difficult to interpret the value of the coefficient for BMI in Table 7.17. It is sometimes easier to consider the changes measured in standard deviation units, namely, the change in BMD (measured in standard deviations of BMD) associated with a single standard deviation change in BMI. This can be simply derived by multiplying the regression coefficient by the standard

Table 7.17 Mixed model analysis of the SOF BMD and BMI data

```
. xtmixed totbmd bmi visit_spl* || id: visit, cov(uns)
```

Computing standard errors:

Mixed-effects REML regression

Group variable: id

Log restricted-likelihood = 41482.617

Number of obs = 26829

Number of groups = 8468

Obs per group: min = 1

 avg = 3.2

 max = 5

Wald chi2(3) = 7837.49

Prob > chi2 = 0.0000

totbmd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
bmi	.0080668	.0001296	62.26	0.000	.0078128	.0083207
visit_spl1	-.0091991	.0002191	-41.98	0.000	-.0096286	-.0087696
visit_spl2	-.0008798	.0002741	-3.21	0.001	-.001417	-.0003425
_cons	.5538826	.0036393	152.19	0.000	.5467496	.5610156

Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
id: Unstructured					
	sd(visit)	.0096978	.0001461	.0094156	.0099883
	sd(_cons)	.1146832	.0009979	.112744	.1166559
	corr(visit,_cons)	-.0491474	.0169775	-.0823559	-.0158298
	sd(Residual)	.023423	.0001517	.0231275	.0237224

LR test versus linear regression: chi2(3) = 45049.07 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

deviation of BMI (which is about 4.70 for this data set) and dividing it by the standard deviation of BMD (which is about 0.13), giving a result of 0.28. So a change in BMD of a single standard deviation is associated with a change of 0.28 standard deviations in BMD, a practically important effect.

7.7.2 Separating Between- and Within-Cluster Information

One could just as well fit a model like (7.9) to a time-invariant predictor, in which case it would not address a longitudinal question. A variable like BMI does vary within an individual over time, but varies even more between individuals. This raises the concern that the coefficient in Table 7.17 might mostly reflect differences between individuals rather than the association of the change in BMD with the associated change of BMI *within an individual*. Between individual associations are often more susceptible to confounding.

Table 7.18 Mixed model separating within and between person BMI

*Separate between and within person changes in BMI
bysort id: egen meanbmi=mean(bmi)
gen bmi_dev=bmi-meanbmi

xtmixed totbmd meanbmi bmi_dev visit_spl* || id: visit, cov(uns)

Mixed-effects REML regression
Group variable: id

Number of obs = 26829
Number of groups = 8468

Obs per group: min = 1
 avg = 3.2
 max = 5

Wald chi2(4) = 8282.99
Prob > chi2 = 0.0000

totbmd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
meanbmi	.0130329	.0002722	47.88	0.000	.0124994	.0135664
bmi_dev	.006695	.0001454	46.04	0.000	.00641	.00698
visit_spl1	-.0090266	.0002192	-41.17	0.000	-.0094562	-.0085969
visit_spl2	-.001178	.0002732	-4.31	0.000	-.0017134	-.0006425
_cons	.4226782	.0072925	57.96	0.000	.4083853	.4369712

Fortunately there are simple ways to isolate the within individual (or more generally within cluster) changes. The first step is to decompose the predictor into two pieces:

$$\begin{aligned} \text{BMI}_{it} &= (\text{BMI}_{it} - \overline{\text{BMI}}_i) + \overline{\text{BMI}}_i \\ &= \text{BMI\_dev}_{it} + \overline{\text{BMI}}_i, \end{aligned}$$

(7.11)

where $\overline{\text{BMI}}_i$ represents the average BMI for person i and BMI_dev_{it} is the deviation of the BMI measurement at time t from their mean BMI. In Stata, the mean and deviation forms of the predictor can easily be calculated using the `bysort` and `egen` commands. Next, both of them are entered in the model as predictors. The deviation form of the predictor represents the within-cluster association and the mean form of the predictor represents the between-cluster association. Another approach that works equally well is to describe the between-cluster portion of the predictor using its baseline value and the within-cluster portion using the difference between each value of the predictor and the baseline value. That is, use BMI_{i1} for between and $\text{BMI}_{it} - \text{BMI}_{i1}$ for within, again entering both predictors in the model.

Table 7.18 shows how to calculate the between and within forms of the predictor and displays a portion of the output from the analysis.

Both the within and between coefficients are highly statistically significant, though this is not surprising given the large sample size. But the between coefficient is about 0.013 and almost twice the size of the within person coefficient, which is about 0.007. This could easily be due to confounding at the person level. The previous analysis reported a weighted average of these two coefficients.

Even in situations in which confounding is not an issue, it may be of substantive interest to conduct such a decomposition of a predictor. For example, [Haas et al. \(2004\)](#) studied the influence of county level race and ethnic composition on access to health care over and above the influence of an individual's race or ethnicity. In this case, interest focused on separating the county level effect (cluster-level effect) of race and ethnicity from the individual level effects.

7.7.3 Prediction

One of the advantages of the random effects approach is the ability to generate predicted values for each of the random effects, which we do not get to observe directly. Returning to the Georgia babies data set, we consider obtaining predicted values for each of the mom effects, MOM_i .

First, let us consider how we might go about estimating the mom effect from first principles. The first mom in the data set had an initial age of 15 and hence, using the estimated coefficients from Table 7.15, has predicted values for the five births (in grams) of 2974.2, 3020.8, 3067.4, 3114.0, and 3160.6 (for example, the first of these is $2974.214 = 2526.622 + 46.608(1) + 26.732(15)$) and actual values of 3720, 3260, 3910, 3320, and 2480, respectively. Her residuals, defined as actual minus predicted, were 745.8, 239.2, 842.6, 206.0, and -680.6 with an average of 270.6. So we might guess that this mom has babies that are, on average, about 271 g heavier than the “average” mom.

Using software to get the predicted effect (deviation from average) for the first mom gives 206.7, only about 76% of the raw data value. Calculation for the other moms shows that all the predicted values are closer to zero than the raw data predicts. Why?

Predicted values from random effects models are so-called *shrinkage estimators* because they are typically less extreme than estimates based on raw data. The shrinkage factor depends on the degree of similarity between moms and, for simple situations, is given by

$$\text{shrinkage factor} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2/n_i}, \quad (7.12)$$

where n_i is the sample size for the i th cluster, σ_u^2 is the between cluster variance, and σ_ϵ^2 is the error variance. In our case, this factor is equal to (taking the estimates from Table 7.15)

$$\begin{aligned} \text{shrinkage factor} &= \frac{358.1761^2}{358.1761^2 + 445.0228^2/5} \\ &= \frac{128,290.1}{128,290.1 + 39,609.1} = 0.76. \end{aligned} \quad (7.13)$$

It is instructive to consider the form of (7.12). Since all the terms in the equation are positive, the shrinkage factor is greater than zero. Further, since the denominator is bigger than the numerator by the factor σ_ϵ^2/n_i , the shrinkage factor is less than 1. So it always operates to shrink the estimate from the raw data to some degree.

What is the magnitude of the shrinkage? If σ_u^2 is much larger than σ_ϵ^2/n_i then the shrinkage factor is close to 1, i.e., almost no shrinkage. This will occur when (a) subjects are quite different (i.e., σ_u^2 is large), and/or (b) results are very accurate and σ_ϵ^2 is small, and/or (c) when the sample size per subject, n_i , is large. So little shrinkage takes place when subjects are different or when answers are accurate or when there is much data.

On the other hand, in cases where subjects are similar (and hence σ_u^2 is small) there is little reason to believe that any individual person deviates from the overall. Or in cases of noisy data (σ_ϵ^2 large) or small sample sizes, random fluctuations can make up the majority of the raw data estimate of the effect and are naturally de-emphasized with this shrinkage approach.

The advantage of the shrinkage predictions are twofold. First, they can be shown theoretically to give more accurate predictions than those derived from the raw data. Second (which is related), they use the data to balance the subject-to-subject variability, the residual variance and the sample size to come up with the best combination of the subject-specific information and the overall data.

Examples of uses of this prediction technology include prediction for prostate cancer screening (Brant et al. 2003) and the use of shrinkage estimators in the rating of individual physicians (Hofer et al. 1999) in treatment of diabetes.

7.7.4 A Logistic Analysis

Turning to the binary outcome variable `lowbrth`, we use the Stata command `xtmelogit`. This model is similar to (7.8) with the needed changes for a logistic model for binary data. This model is:

$$\begin{aligned} \text{LOWBRTH}_{ij} &= 1 \text{ if baby } j \text{ for mom } i \text{ is } < 3,000 \text{ g and } 0 \text{ otherwise} \\ &\sim \text{Bernoulli}(p_{ij}) \end{aligned}$$

with

$$\text{logit}(p_{ij}) = \beta_0 + \text{MOM}_i + \beta_1 \text{BIRTHORD}_{ij} + \beta_2 \text{INITAGE}_i, \quad (7.14)$$

and

$$\text{MOM}_i \sim \text{i.i.d } \mathcal{N}(0, \sigma_u^2).$$

This analysis is given in Table 7.19 with a syntax similar to that of `xtmixed`. The fixed effects are listed after the outcome and the vertical bar notation separates the fixed effects from the random effects, again with the `momid:` indicating the

Table 7.19 Random effects logistic regression analysis for the birthweight data

. xtmelogit lowbirth birthord initage|| momid:, or

Mixed-effects logistic regression
Group variable: momid

Integration points = 7
Log likelihood = -588.07113

Number of obs = 1000
Number of groups = 200
Obs per group: min = 5
 avg = 5.0
 max = 5

Wald chi2(2) = 11.85
Prob > chi2 = 0.0027

lowbirth	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
birthord	.8872745	.0500702	-2.12	0.034	.7943711	.9910432
initage	.8808974	.0406081	-2.75	0.006	.8047967	.9641941

Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
momid: Identity					
	sd(_cons)	1.60859	.1676556	1.31138	1.973158

LR test versus logistic regression: chibar2(01) = 123.21 Prob>=chibar2 = 0.0000

inclusion of random intercepts for each mother. The option `or` requests odds ratios in the output table as opposed to log odds. This gives somewhat different results than the GEE analysis, as expected, since it is fitting a conditional model. More specifically (as predicted from Fig. 7.5), the coefficients in the conditional analysis are slightly farther from 1 than the marginal coefficients, for example the odds ratio for birth order is now 0.89 as compared to 0.92 in the marginal model. The tests are, however, virtually the same, which is not unusual.

The interpretation of the `birthord` coefficient in the conditional model is that the odds of a low-birthweight baby decreases by about 11% for each increase of birth order of one for each woman.

This is opposed to the interpretation of the odds-ratio estimate from the marginal fit given in Table 7.13 of 0.92. The interpretation in the marginal model is the decrease in the odds (averaged across all women) is about 8% with an increase in birth order of one.

7.8 Marginal Versus Conditional Models

The previous section has demonstrated that, for non-linear models like the logistic model, it is important to distinguish between marginal and conditional models since the model estimates are not expected to be equal. Conditional models have a more mechanistic interpretation, which can sometimes be useful (being careful, of course, to remember that many experiments do not strongly support

mechanistic interpretations, no matter what model is fit). Marginal models have what is sometimes called a “public health” interpretation since the conclusions only hold averaged over the entire population of subjects.

7.9 Example: Cardiac Injury Following Brain Hemorrhage

Heart damage in patients experiencing brain hemorrhage has historically been attributed to preexisting conditions. However, more recent evidence suggests that the hemorrhage itself can cause heart damage through the release of norepinephrine following the hemorrhage. To study this, [Tung et al. \(2004\)](#) measured cardiac troponin, an enzyme released following heart damage, at up to three occasions after patients were admitted to the hospital for a specific type of brain hemorrhage (subarachnoid hemorrhage or SAH).

The primary question was whether severity of injury from the hemorrhage was a predictor of troponin levels, as this would support the hypothesis that the SAH caused the cardiac injury. To make a more convincing argument in this observational study, we would like to show that severity of injury is an independent predictor, over and above other circulatory and clinical factors that would predispose the patient to higher troponin levels. Possible clinical predictors included age, gender, body surface area, history of coronary artery disease (CAD), and risk factors for CAD. Circulatory status was described using systolic blood pressure, history of hypertension (yes/no) and left ventricular ejection fraction (LVEF), a measure of heart function. The severity of neurological injury was graded using a subject’s Hunt–Hess score on admission. This score is an ordered categorical variable ranging from 1 (little or no symptoms) to 5 (severe symptoms such as deep coma).

The study involved 175 subjects with at least one troponin measurement and between 1 and 3 visits per subject. Figure 7.6 shows the histogram of troponin levels. They are *severely* skewed right with over 75% of the values equal to 0.3, the smallest detectable value and many outlying values. For these reasons, the variable was dichotomized as being above or below 1.0, as is labeled in the output as `CTOver1`. Table 7.20 lists the proportion of values above 1.0 for each of the Hunt–Hess categories and Table 7.21 gives a more formal analysis using GEE methods, but including only the predictor Hunt–Hess score and not using data from visits four or greater (there were too few observations to use those later visits).

The reference group for the Hunt–Hess variable in this analysis is a score of 1, corresponding to the least injury. So the odds of heart damage, as evidenced by troponin values over 1, is over two times higher for a Hunt–Hess score of 2 as compared to 1 and the odds go up monotonically with the estimated odds of heart damage for a Hunt–Hess score of 5 being over 70 times those of a score of 1. Even though the odds ratio of a score of 5 is poorly determined, the lower limit of the 95% CI is still over 16.

The primary goal is to assess the influence of a single predictor variable, Hunt–Hess score, which is measured only once per subject. Since it is only measured once,

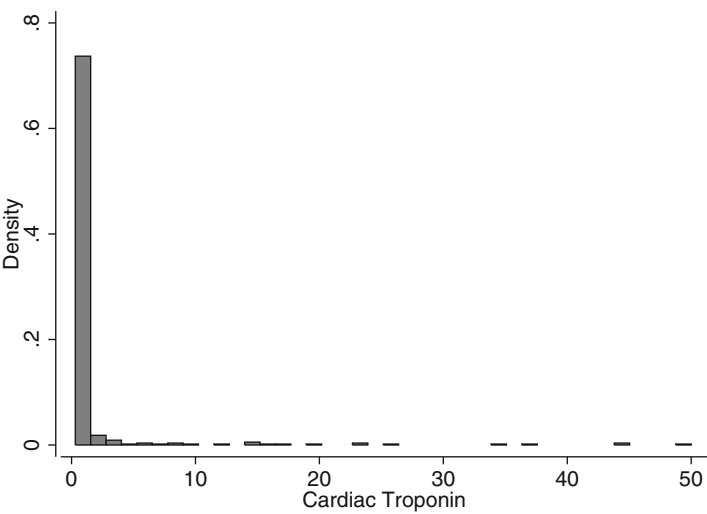


Fig. 7.6 Histogram of cardiac troponin levels

Table 7.20 Proportion of troponin levels over 1.0 and sample size versus Hunt–Hess score

```
. table hunt, c(mean CTover1 n CTover1)
```

Initial Hunt--Hess	mean (CTover1)	N (CTover1)
1	.0318471	157
2	.0615385	65
3	.1269841	126
4	.1692308	65
5	.6818182	22

rather than repeatedly, a marginal model and the use of GEE methods is attractive. Since we are interested in a single predictor, we will be more liberal in including predictors for adjustment. We certainly would like to adjust for the amount of time after the SAH occurred, as captured by the visit number, `stday`, since troponin levels drop over time. We also want to adjust for fundamental differences that might be due to age, sex, and body surface area (`bsa`), which may be related to troponin levels.

In addition, we choose to adjust for preexisting conditions that might influence the troponin levels, including left ventricular ejection fraction, standardized (`lvef_std`), SBP (`sbp`), heart rate (`hr`), and history of hypertension (`hxhtn`). Quadratic functions of left ventricular ejection fraction (`lvef_std2`) and SBP (`sbp2`) are included to model non-linear (on the logit scale) relationships.

Table 7.22 gives the output after dropping some nonstatistically significant predictors from the model and using the `xtgee` command. It also gives an overall test of whether troponin levels vary with Hunt–Hess score.

Table 7.21 Effect of Hunt–Hess score on elevated cardiac troponin levels

. xtgee CTover1 i.hunt if stday<4, i(stnum) family(binomial) ef

GEE population-averaged model

Group variable: stnum

Link: logit

Family: binomial

Correlation: exchangeable

Scale parameter: 1

Number of obs = 434

Number of groups = 168

Obs per group: min = 1

avg = 2.6

max = 3

Wald chi2(4) = 39.03

Prob > chi2 = 0.0000

CTover1	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
hunt						
2	2.036724	1.669731	0.87	0.386	.4084194	10.15682
3	4.493385	2.820396	2.39	0.017	1.313088	15.37636
4	6.542645	4.347658	2.83	0.005	1.778774	24.065
5	70.66887	52.16361	5.77	0.000	16.63111	300.286

Even after adjustment for a multitude of characteristics, the probability of an elevated troponin level is associated with Hunt–Hess score. However, the picture is a bit different as compared to the unadjusted analysis. Each of the categories above 1 has an estimated elevated risk of troponin release, but it is not a monotonic relationship. Also, only category 5, the most severely damaged group, is statistically significantly different from category 1.

What is the effect of adjusting for the large number of predictors in this model? We might be worried that CIs for some of the coefficients have gotten quite wide due to correlations among the predictors and the Hunt–Hess score. Table 7.23 gives the analysis after minimally adjusting for just `stday`.

While it is not clearly evident from the output on the odds-ratio scale, standard errors for the log odds values are not appreciably larger in the adjusted analysis (see Problem 7.10). The minimally adjusted and unadjusted analyses have similar pattern of estimated odds ratios. However, both of them may have overestimated the association with Hunt–Hess score slightly and so the adjusted analysis reported in Table 7.22 would be preferred.

7.9.1 Bootstrap Analysis

We might also be concerned about the stability of the results reported in Table 7.22 given the modest sized dataset with a binary outcome and the large number of predictors. This is exactly a situation in which bootstrapping can help understand the reliability of standard errors and CIs.

Correspondingly, we conducted a bootstrap analysis and we focus on the stability of the result for the comparison of Hunt–Hess score of 5 compared to a value of 1. Bootstrapping is conducted for the log odds (which can be transformed easily back to the odds scale) since that is the basis of the calculation of CIs.

Table 7.22 Adjusted effect of Hunt–Hess score on elevated troponin levels

```

. xtgee CTover1 i.hunt i.stday sex lvef_std lvef_std2 hxhtn sbp sbp2 if
  stday<4, i(stnum)
> family(binomial) ef

GEE population-averaged model
Group variable:          stnum      Number of obs      =      408
Link:                  logit      Number of groups   =      165
Family:                binomial   Obs per group: min =        1
Correlation:           exchangeable      avg      =      2.5
                                      max      =        3
                                      Wald chi2(12)   =      44.06
Scale parameter:        1         Prob > chi2      =      0.0000

```

	CTover1	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
hunt						
2		1.663476	1.334533	0.63	0.526	.3452513 8.014895
3		1.830886	1.211796	0.91	0.361	.5003595 6.69947
4		1.560879	1.241708	0.56	0.576	.3282637 7.421908
5		74.99009	69.48431	4.66	0.000	12.19825 461.0097
stday						
2		.5258933	.2163491	-1.56	0.118	.2348112 1.177813
3		.374303	.1753685	-2.10	0.036	.1494233 .9376232
sex		8.242847	6.418324	2.71	0.007	1.791785 37.92002
lvef_std		.5438802	.1290215	-2.57	0.010	.3416472 .8658223
lvef_std2		1.388986	.1863399	2.45	0.014	1.067836 1.806721
hxhtn		3.11661	1.572135	2.25	0.024	1.15959 8.376457
sbp		1.143139	.0771871	1.98	0.048	1.001438 1.30489
sbp2		.9995246	.0002293	-2.07	0.038	.9990753 .9999742

```

. testparm i.hunt

( 1)  2.hunt = 0
( 2)  3.hunt = 0
( 3)  4.hunt = 0
( 4)  5.hunt = 0
      chi2( 4) =    23.87
      Prob > chi2 =    0.0001

```

A complication with clustered data is what to resample. By default, bootstrapping will resample the individual observations. However, the basis of sampling in this example (which is common to clustered-data situations) is subjects. We thus need to resample *subjects* not observations. Fortunately, this can be controlled within Stata by using a `cluster` option on the `bootstrap` command. The analysis was run using a robust variance estimate and independence working correlation, which improved the stability of the estimates. Table 7.24 gives the portion of the output associated with the Hunt–Hess scores. The bias-corrected bootstrap (using the `ef` option to generate odds ratios) gives a CI for the odds ratio for a Hunt–Hess of 2 compared to 1 of 0.23–7.91. This compares with the interval from 0.35 to 8.01 from Table 7.22 in the original analysis. For comparing a Hunt–Hess score of 5 to that of 1, the bootstrap analysis gives a CI of 14.66–472.09 compared to 12.19–461.00. The results are quite similar and give qualitatively the same results, giving us confidence in our original analysis.

Table 7.23 Effect of Hunt–Hess score on elevated troponin levels adjusting only for stday

```
. xtgee CTover1 i.hunt i.stday if stday<4, i(stnum) family(binomial) ef
```

GEE population-averaged model

Group variable:	stnum	Number of obs	=	434
Link:	logit	Number of groups	=	168
Family:	binomial	Obs per group: min	=	1
Correlation:	exchangeable	avg	=	2.6
		max	=	3
		Wald chi2(6)	=	40.75
Scale parameter:	1	Prob > chi2	=	0.0000

CTover1	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
hunt						
2	2.136339	1.711752	0.95	0.343	.4442634	10.27306
3	4.312505	2.68268	2.35	0.019	1.274157	14.59609
4	6.41448	4.228072	2.82	0.005	1.762367	23.34676
5	60.09793	44.25148	5.56	0.000	14.19385	254.4595
stday						
2	.5564922	.1968294	-1.66	0.098	.2782224	1.113079
3	.5170812	.2016593	-1.69	0.091	.2407654	1.110512

Table 7.24 Bootstrap analysis of adjusted Hunt–Hess model

```
. bootstrap_b, reps(1000) cluster(stnum) seed(2718):xtgee CTo i.hunt i.stday sex
> lvef_std lvef_std hxhtn sbp sbp2 if stday <4, i(stnum) family(bin) robust corr(inde)
. estat boot, ef
```

Bootstrap results

Number of obs	=	408
Replications	=	921

(Replications based on 165 clusters in stnum)

(1) 1b.hunt = 0
(2) 1b.stday = 0

CTover1	Observed exp(b)	Bias	Bootstrap Std. Err.	[95% Conf. Interval]	
1b.hunt	1	0	0	.	. (BC)
2.hunt	1.5943809	.2109551	1.4253662	.233916	7.909329 (BC)
3.hunt	2.2787955	.1224353	1.8531043	.461805	11.75678 (BC)
4.hunt	2.3847466	.0586308	2.4034528	.219347	14.8611 (BC)
5.hunt	78.628816	66.77691	99.854523	14.66172	471.0894 (BC)
1b.stday	1	0	0	.	. (BC)
2.stday	.57482247	-.0672446	.26831764	.2518986	1.369652 (BC)
3.stday	.41069747	-.0548035	.23561377	.1419608	1.170843 (BC)

(BC) bias-corrected confidence interval

Note: one or more parameters could not be estimated in 79 bootstrap replicates; standard-error estimates include only complete replications.

7.10 Power and Sample Size for Repeated Measures Designs

Planning the sample size or calculating power for a repeated measures analysis can be challenging, due to the need to specify the correlation structure (which can be difficult) and because the calculations are different for different types of predictors. We present some results here for the simple situation in which there is a single level of clustering, observations within a cluster are equally correlated and all have the same variability, and the sample size per cluster is the same. This serves as the starting point for many calculations and illustrates some features of power and sample size for repeated measures designs.

An important distinction is whether the sample size calculation is for a *between- or within-cluster predictor*. A purely between-cluster predictor is one that may vary between clusters but is constant within a cluster. A within-cluster predictor is one that may vary within a cluster, but whose average is constant across clusters. For example, in a longitudinal study in which the clusters are participants, the participant's race, age at entry to the study, and genetic information are all between-cluster predictors. If every participant was measured at every visit, then visit would be a purely within-cluster predictor. In practice, most predictors that vary within a cluster are not purely within-cluster predictors; their average varies at least somewhat across clusters. Section 7.7.2 shows how to separate a predictor into its purely between and purely within components.

7.10.1 Between-Cluster Predictor

In the situation in which the cluster sample sizes are equal, the analysis of between-cluster predictors are, in essence, based on the cluster level means. This realization also serves to temper the number of between-cluster predictors that can be included in an analysis, because the effective sample size is the number of clusters.

When the data are equally correlated, the variance of a cluster-level mean is given by $\sigma^2[1 + (n - 1)\rho]/n$, where σ^2 is the variability of the outcome, ρ is the within-cluster (intraclass) correlation, and n is the sample size per cluster. In contrast, when the data are independent, the variance would be σ^2/n . That is, the cluster-level mean has a variance that is larger by a factor of $[1 + (n - 1)\rho]$. Since required sample sizes are proportional to the variability of the measurements, the consequence is that sample sizes must be larger by this factor, compared to an experiment using independent data. Because of the central role this factor plays, it has been named the *design effect* and is often abbreviated as DEFF, i.e., $\text{DEFF} = 1 + (n - 1)\rho$. This also gives a convenient way to do sample size calculations. Namely, a calculation is conducted assuming independent data, then it is multiplied by the DEFF to find the required sample size for the repeated measures design.

Here is an illustration of planning a new study, but patterned after [Whelan et al. \(2004\)](#), which was a randomized controlled trial of a decision-making aid (versus

Table 7.25 Sample size and power calculation examples

```
. sampsi 1.7 1.4, sd1(0.5) power(0.8)
```

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
and m2 is the mean in population 2

Assumptions:

alpha =	0.0500	(two-sided)
power =	0.8000	
m1 =	1.7	
m2 =	1.4	
sd1 =	.5	
sd2 =	.5	
n2/n1 =	1.00	

Estimated required sample sizes:

n1 =	44
n2 =	44


```
. sampsi 1.546 1.4, sd1(0.5) n1(130)
```

Estimated power for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
and m2 is the mean in population 2

Assumptions:

alpha =	0.0500	(two-sided)
m1 =	1.546	
m2 =	1.4	
sd1 =	.5	
sd2 =	.5	
sample size n1 =	130	
n2 =	130	
n2/n1 =	1.00	

Estimated power:

power =	0.6533

not) for physicians to help them counsel breast cancer patients on surgical options. The outcome is *decisional conflict* and will be assessed using a numerical scale and measures the degree to which patients are well-informed about their choices concerning treatment for breast cancer. This is a repeated measures design because the outcome will be measured at the patient level and there will be multiple patients per physician. The predictor (decision aid or not) is a between-cluster (physician-level) predictor. We use input values from [Whelan et al. \(2004\)](#): an average of about 7.5 patients per physician, standard deviation of the outcome of 0.5 (measured across patients and physicians), and an intraclass correlation, ρ , of 0.3.

We use the corresponding independent samples comparison (a two sample *t*-test), with a detectable effect size of 0.3 and a desired power of 0.8. Using the `sampsi` command as illustrated in Table 7.25 shows that 44 observations per group would be needed. With 7.5 patients per physician, the design effect is $DEFF = 1 + (n - 1)\rho = 1 + (7.5 - 1)0.3 = 2.95$ and about $2.95(44)$ or about 130 patients would be needed per treatment group, working out to about $130/7.5$ or 18 physicians for each of the two treatment groups.

While the calculation of the required sample size for a between-cluster predictor is not numerically difficult, in the absence of preliminary data, specifying the intraclass correlation coefficient can be an issue. It is sometimes slightly easier to consider the within-cluster variability in the outcome across observations and the variability in the true cluster-level means across clusters. In the decision aid example, this would mean considering the variation in the decisional conflict scale across patients within a physician and the variation in the true physician level means (i.e., the average value if an unlimited number of patients were measured for each physician). The intraclass correlation coefficient can then be calculated (see Sect. 7.1.2) as the ratio of the between-cluster variability and the sum of the between- and within-cluster variances. For the decision aid example, the within-cluster variance in the outcome is $\sigma_e^2 = 0.175$ and the between-cluster variation is $\sigma_u^2 = 0.075$, giving an intraclass correlation coefficient of $0.3 = 0.075/(0.075 + 0.175)$ and an overall variance of $0.25 = 0.075 + 0.175$, corresponding to the standard deviation of $0.5 = \sqrt{0.25}$.

The design effect can be used for power calculations, as opposed to sample size calculations, by reducing the detectable effect size by the square root of the design effect and using power calculations assuming independent data. Continuing the decision aid example, suppose the effect size was 0.25 instead of 0.3 and we have 130 patients per treatment group. How much would the power decrease? The reduced detectable effect size would be $0.25/\sqrt{2.95} = 0.146$. The power calculation for the *t*-test shown in Table 7.25 gives a power of 0.65.

The calculations above have been illustrated for a single predictor and for a numerical outcome, but the general principle extends to the other scenarios described in the book, including different outcome types and the use of multiple predictors. That is, for sample size, preliminary calculations are performed assuming independent data, the result of which is multiplied by the design effect to find the required sample size for the repeated measures design. For power, detectable effect sizes are reduced by dividing by the square root of the design effect and that is used in an independent sample size power calculation to find the power for the repeated measures design. In the binary outcomes outcome case, the design effect can be applied to (5.16) which also would accommodate multiple, correlated, between-cluster predictors through the factor $1/(1 - \rho_j^2)$.

7.10.2 Within-Cluster Predictor

In the typical case where the correlation within a cluster is positive and for the same sample size and detectable effect size, power for within-cluster predictors is higher than for between-cluster predictors; for the same power and detectable effect size, required sample sizes are smaller. In fact, this is a common rationale in using cluster designs such as longitudinal studies, which are often described as

“using each person as their own control” in order to increase precision. In contrast to between-cluster predictors, the effective sample size for purely within-cluster predictors is the total sample size, not the number of clusters.

As with between-cluster predictors, sample size or power calculations can be obtained by modifying the results from an independent sample size calculation. Again, with ρ being the intraclass correlation coefficient, the sample size can first be calculated assuming independent data and then reduced by the factor $1 - \rho$. Alternatively, if the within-cluster standard deviation is known, this can be used directly to perform a sample size calculation, ignoring the clustered design. Going back to the decision aid example, suppose we are interested in a within-physician predictor, such as the age of the patient, and we divide patients according to whether they are above or below the median value for that physician. A power of 0.8 is desired and the detectable effect size is 0.2. Using the `sampsi` command indicates that 99 observations are needed per group. But this can be reduced by $1 - \rho$ to arrive at a final sample size of 70 per group. That would mean that we would need an overall sample size of 140. Equivalently, we can use the within-cluster standard deviation of $0.418 = \sqrt{0.175}$ to directly perform a sample size calculation (see Exercise 7.12).

For calculating power for a within-cluster predictor, the detectable effect size is *increased* by multiplying it by $1/\sqrt{1 - \rho}$ and then an independent sample size calculation is conducted. Or, if the within-cluster standard deviation is available, this is used to directly perform the power calculation, ignoring the clustered nature of the design.

As with the between-cluster calculations, this approach extends to other scenarios covered in this book. That is, for sample size, preliminary calculations are performed assuming independent data, the result of which is reduced by $1 - \rho$ to find the required sample size for the repeated measures design. Again, in the binary outcomes outcome case, the multiplier can be applied to (5.16) which handles multiple, correlated, within-cluster predictors through the factor $1/(1 - \rho_j^2)$. Or the calculations are conducted using the within-cluster standard deviation, which is smaller than the overall standard deviation. For power, detectable effect sizes are increased by multiplying by $1/\sqrt{1 - \rho}$ and that is used in an independent sample size power calculation to find the power for the repeated measures design. Or, the within-cluster standard deviation is used directly to calculate the power assuming independent samples.

7.11 Summary

The main message of this chapter has been the importance of incorporating correlation structures into the analysis of clustered, hierarchical, longitudinal, and repeated measures data. Failure to do so can have serious consequences. Two main methods have been presented, GEEs and random effects models.

A primary advantage of the GEEs approach is the availability of the robust variance estimate, which provides valid standard errors without having to explicitly model the nature of the correlations within a cluster. GEEs approaches typically fit models for estimating effects averaged across a population, called marginal models.

In contrast, mixed models incorporate correlation by introducing random effects. This may require more careful modeling and assessment of assumptions, but yield extra capabilities in the form of partitioning the variability, enabling calculation of intraclass correlation coefficients, testing for the presence of clustering, and generating predicted values of random effects. Mixed model approaches typically fit models for estimating effects specific to a cluster (e.g., an individual) and are conditional models.

For simple clustered-data situations, power and sample size calculations can be based on straightforward modifications of the calculations for independent data. These modifications depend on whether the predictor of interest is a between- or within-cluster predictor and require knowledge of the within-cluster correlation (or equivalent quantities).

7.12 Further Notes and References

For those readers desiring more detailed information on longitudinal and repeated measures analyses, there are a number of book length treatments, especially for continuous, approximately normally distributed data. Notable entries include [Raudenbush and Bryk \(2001\)](#), [Goldstein \(2003\)](#), [Verbeke and Molenberghs \(2000\)](#), [Diggle et al. \(2002\)](#), [Fitzmaurice et al. \(2004\)](#), and [McCulloch et al. \(2008\)](#). Unfortunately, many are more technical than this book.

7.12.1 *Missing Data*

The techniques in this chapter handle unequal sample sizes and unequal spacing of observations in time with aplomb. However, sample sizes are often unequal and observation times unequal because of missing outcome data. And data are often missing for a reason related to the outcome under study. As examples, sicker patients may not show up for follow-up visits, leading to overly optimistic estimates based on the data present. Or those patients staying in the hospital longer may be the sicker ones (with the better-off patients having been discharged). This might lead us to the erroneous conclusion that longer stays in the hospital produce poorer outcomes, so why check-in in the first place?

To a limited extent, the methods in this chapter cover the situation in which the missing data are systematically different from the data available. If the fact that data are missing is related to a factor in the model (i.e., more missing data for males, which is also a factor in the model) then there is little to worry about. However, the

methods described here do *not* cover the situation where the missing data are related to predictors not in the model and can give especially misleading results if the fact that the data are missing is related to the value of the outcome that would have been measured.

See Chap. 11 for much more detail.

7.12.2 Computing

Stata has a wide array of clustered-data techniques. The commands `xtmixed`, `xtmelogit`, and `xtmepoisson` can fit mixed models for multilevel hierarchical data structures. The generalized estimating equations methods are limited to one level of clustering. So, for example, they can explicitly model repeated measures data on patients, but not repeated measures data on patients clustered within doctors. Of course, with sufficient numbers of doctors, even the clustering of patients within doctors could be accommodated with robust standard errors.

Other software packages can also conduct these analyses. For continuous, approximately normally distributed data, SAS Proc MIXED can handle a multitude of models (Littell et al. 1996) and SAS Proc GENMOD can fit models using GEEs and, for binary data, can fit two-level clustered binary data with a technique called alternating logistic regression (Carey et al. 1993). MLWin and HLM are two other clustered data packages with additional capabilities.

7.13 Problems

Problem 7.1. Using the fecal fat data in Table 7.1, calculate the sample variance of the subject averages. Subtract from this the residual variance estimate from Table 7.3 divided by four (why four?) to verify the estimate of σ_{subj}^2 given in the text.

Problem 7.2. Using the fecal fat data in Table 7.1, verify the F -tests displayed in Tables 7.2 and 7.3.

Problem 7.3. From your own area of interest, describe a hierarchical dataset including the outcome variable, predictors of interest, the hierarchical levels in the dataset and the level at which each of the predictors is measured. Choose a dataset for which not all of the predictors are measured at the same level of the hierarchy.

Problem 7.4. Could you successfully analyze the data from the fecal fat example using the idea of “analysis at the highest level of the hierarchy?” Briefly say why or why not.

Problem 7.5. For the fecal fat example of Table 7.1, analyze the difference between capsule and coated capsules in two ways. First, use the “derived variable” approach

to perform a paired t -test. Second, in the context of the two-way ANOVA of Table 7.3, test the contrast of coated capsule versus capsule. How do the two analyses compare? What differences do you note? Why do they come about? What are the advantages and disadvantages of each?

Problem 7.6. Consider an example (like the Georgia birthweight example) with before and after measurements on a subject. If the variability of the before and after measurements each have variance σ^2 and correlation ρ then it is a fact that the standard deviation of the difference is $\sigma\sqrt{2(1-\rho)}$.

- (1) The correlation of the first and last birthweights is about 0.381. Using Table 7.5, verify the above formula (approximately).
- (2) If we were to compare two groups, based on the difference scores or just the last birthweights (say, those with initial age greater than 17 versus those not), which analysis would have a larger variance and hence be less powerful? By how much?

Problem 7.7. The model corresponding to the analysis for Table 7.7 has an intercept, a dummy variable for the fifth birth, a continuous predictor of centered age (age minus the average age), and the product of the dummy variable and centered age.

- (1) Write down a model equation.
- (2) Verify that the intercept is the average for the first-born, and that the coefficient for the dummy variable is the difference between the two groups, both of these when age is equal to its average.
- (3) Verify that the coefficient for the product measures how the change in birthweight from first to last birth depends on age.

Problem 7.8. Reproduce the standard error calculations in Table 7.12, but for the coefficient of `birthorder`. How different are the standard errors when not using the robust option? When using the robust option? Are any of the analyses likely to give misleading results? If so, which ones?

Problem 7.9. Verify the calculation of the predicted values and residuals in Sect. 7.7.3.

Problem 7.10. Using the CIs for the odds ratios for the Hunt–Hess scores in Tables 7.22 and 7.21, calculate the confidence intervals for the log-odds ratios. Show that the width of the CIs in the adjusted analysis (Table 7.22) are not appreciably larger than those in the unadjusted analysis (Table 7.21).

Problem 7.11. Compare the bootstrap-based CI for the comparison of study day 1 and study day 2 from Table 7.24 to the CI from the original analysis reported in Table 7.22. Do they agree substantively? Do they lead to different conclusions?

Problem 7.12. Verify that a two independent sample t -test sample size calculation with a standard deviation of 0.5 when reduced by the factor $1 - \rho = 1 - 0.3 = 0.7$

gives virtually the same answer as a direct calculation using the standard deviation of 0.418.

7.14 Learning Objectives

- (1) Recognize a hierarchical data situation and explain the consequences of ignoring it.
- (2) Decide when hierarchical models are necessary versus when simpler analyses will suffice.
- (3) Define the terms hierarchical, repeated measures, clustered, longitudinal, robust variance estimator, working correlation structure, generalized estimating equations, fixed factor, and random factor.
- (4) Interpret Stata output for GEE and random effects analyses in hierarchical analyses for linear regression or logistic regression problems.
- (5) Explain the difference between marginal and conditional models.
- (6) Decide if factors should be treated as fixed or random.
- (7) Explain the use of shrinkage estimators and best prediction for random factors.
- (8) Perform power or sample size calculations for simple clustered-data situations.

Chapter 8

Generalized Linear Models

A new program for depression is instituted in the hopes of reducing the number of visits each patient makes to the emergency room in the year following treatment. Predictors include (among many others) treatment (yes/no), race, and drug and alcohol usage indices. A common and minimally invasive treatment for jaundice in newborns is exposure to light. Yet the cost of this is high, mainly because of longer hospital stays, which are expensive. Predictors of the cost include race, gestational age, and birthweight.

These analyses require special attention both because of the nature of the outcome variable (counts in the depression example and costs, which are positive and right-skewed, for the jaundice example) and because the models we would typically employ are not as straightforward as the linear models of Chap. 4.

On the other hand, many features of constructing an analysis are the same as we have seen previously. We have a mixture of categorical (treatment, race) and continuous predictors (drug usage, alcohol usage, gestational age, birthweight). There are the same issues of determining the goals of inference (prediction, risk estimation, and testing of specific parameters) and winnowing down of predictors to arrive at a final model as discussed in Chap. 10. And we can use tests and CIs in ways that are quite similar to those for previously described analyses.

We begin this chapter by discussing the two examples in a bit more detail and conclude with a look at how those examples, as well as a number of earlier ones, can be subsumed under the broader rubric of *generalized linear models*.

8.1 Example: Treatment for Depression

A new case-management program for depression is instituted in a local hospital that often has to care for the poor and homeless. A characteristic of this population is that they often access the health care system by arriving in the emergency room—an

expensive and overburdened avenue to receive treatment. Can the new treatment reduce the number of needed visits to the emergency room as compared to standard care? The recorded outcome variable is the number of emergency room visits for each patient in the year following treatment.

The primary goal of the analysis is to assess the treatment program, but emergency room usage varies greatly according to other factors. Secondary goals included association of emergency room usage with drug or alcohol abuse and to assess racial differences in use.

8.1.1 *Statistical Issues*

From a statistical perspective, we need to be concerned with the nature of the outcome variable: in the data set that motivated this example, about one-third of the observations are 0 (did not return to the emergency room within the year) and over half are either 0 or 1. This is highly nonnormal and cannot be transformed to be approximately normal—any transformation by an increasing function will merely move the one-third of the observations that are exactly 0 to another numerical value, but there will still be a “lump” of observations at that point consisting of one-third of the data. For example, a commonly recommended transformation for count data with zeros is $\log(y + 1)$. This transformation leaves the data equal to 0 unchanged since $\log(0 + 1) = 0$ and moves the observations at 1 to $\log(1 + 1) = \log(2)$, not appreciably reducing the nonnormality of the data. Over half the data take on the two values 0 and $\log(2)$.

Even if we can handle the nonnormal distribution, a typical linear model (as in Chap. 4) for the mean number of emergency room visits will be untenable. The mean number of visits must be a positive number and a linear model, especially with continuous predictors, may, for extreme values of the covariates, predict negative values. This is the same problem we encountered with models for the probability of an event in Sect. 5.1.

Another bothersome aspect of the analysis is that this is a hard-to-follow, transient population in generally poor health. It is not at all unusual to have subjects die or be unable to be contacted for obtaining follow-up information. So some subjects are only under observation (and hence eligible for showing up for emergency room visits) for part of the year.

Since not all the subjects are followed for the same periods of time, it is natural to think of a multiplicative model. In other words, if all else is equal, a subject that is followed for twice as long as another subject will have, on average, twice the emergency room utilization. This consideration, as well as the desire to keep the mean response positive, leads us to consider a model for the log of the mean response. Note that this is different from the mean of the log-transformed responses (See Problem 8.1, also Sects. 4.7.2 and 4.7.5).

8.1.2 Model for the Mean Response

To begin to write down the model more carefully, define Y_i as the number of emergency room visits for patient i and let $E[Y_i]$ represent the average number of visits for a year. For the moment we will ignore the fact that the observation periods are unequal. The model we are suggesting is

$$\log E[Y_i] = \beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i, \quad (8.1)$$

or equivalently (using an exponential, i.e., anti-log)

$$E[Y_i] = \exp\{\beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i\}, \quad (8.2)$$

where β_0 is an intercept, RACE_i is 1 for nonwhites and 0 for whites, TRT_i is 1 for those in the treatment group and 0 for usual care, ALCH_i is a numerical measure of alcohol usage and DRUG_i is a numerical measure of drug usage. We are primarily interested in β_2 , the treatment effect.

Since the mean value is not likely to be exactly zero (otherwise, there is nothing to model), using the log function is mathematically acceptable (as opposed to trying to log transform the original counts, many of which are zero). Also, we can now reasonably hypothesize models like (8.1) that are linear (for the log of the mean) in ALCH_i and DRUG_i since the exponential in (8.2) keeps the mean value positive.

This is a model for the number of emergency room visits per year. What if the subject is only followed for half a year? We would expect their counts to be, on average, only half as large. A simple way around this problem is to model the mean count per unit time instead of the mean count, irrespective of the observation time. Let t_i denote the observation time for the i th patient. Then, the mean count per unit time is $E[Y_i]/t_i$ and (8.1) can be modified to be

$$\log (E[Y_i]/t_i) = \beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i, \quad (8.3)$$

or equivalently (using the fact that $\log[Y/t] = \log Y - \log t$)

$$\log E[Y_i] = \beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i + \log t_i. \quad (8.4)$$

The term $\log t_i$ on the right-hand side of (8.4) looks like another covariate term, but with an important exception: there is no coefficient to estimate analogous to the β_3 or β_4 for the alcohol and drug covariates. Thinking computationally, if we used it as a predictor in a regression-type model, a statistical program like Stata would automatically estimate a coefficient for it. But, by construction, we know it must enter the equation for the mean with a coefficient of exactly 1. For this reason, it is called an *offset* instead of a covariate and when we use a package like Stata, it is designated as an offset and not a predictor.

8.1.3 Choice of Distribution

Lastly, we turn to the nonnormality of the distribution. Typically, we describe count data using the Poisson distribution. Directly modeling the data with a distribution appropriate for counts recognizes the problems with discreteness of the outcomes (e.g., the “lump” of zeros). While the Poisson distribution is hardly ever ultimately the correct distribution to use in practice, it gives us a place to start.

We are now ready to specify a model for the data, accommodating the three issues: nonnormality of the data, mean required to be positive, and unequal observation times. We start with the distribution of the data. Let λ_i denote the mean rate of emergency room visits per unit time, so that the mean number of visits for the i th patient is given by $\lambda_i t_i$. We then assume that Y_i has a Poisson distribution with log of the mean given by

$$\begin{aligned}\log E[Y_i] &= \log[\lambda_i t_i] \\ &= \log \lambda_i + \log t_i \\ &= \beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i + \log t_i. \quad (8.5)\end{aligned}$$

This shows us that the main part of the model (consisting of all the terms except for the offset $\log t_i$) is modeling the rate of emergency room visits per unit time:

$$\log[\lambda_i] = \beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i, \quad (8.6)$$

or, exponentiating both sides,

$$\lambda_i = \exp\{\beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{ALCH}_i + \beta_4 \text{DRUG}_i\}. \quad (8.7)$$

8.1.4 Interpreting the Parameters

The model in (8.7) is a multiplicative one, as we saw for the Cox model in Chap. 6, and has a similar style of interpretation. Recall that RACE_i is 1 for nonwhites and 0 for whites and suppose the race coefficient is estimated to be $\hat{\beta}_1 = -0.5$. The mean rate per unit time for a white person divided by that of a nonwhite (assuming treatment group, and alcohol and drug usage indices are all the same) would be

$$\begin{aligned}& \frac{\exp\{\beta_0 + 0 + \beta_2 \text{TRT} + \beta_3 \text{ALCH} + \beta_4 \text{DRUG}\}}{\exp\{\beta_0 - 0.5 + \beta_2 \text{TRT} + \beta_3 \text{ALCH} + \beta_4 \text{DRUG}\}} \\ &= \frac{e^{\beta_0} e^0 e^{\beta_2 \text{TRT}} e^{\beta_3 \text{ALCH}} e^{\beta_4 \text{DRUG}}}{e^{\beta_0} e^{-0.5} e^{\beta_2 \text{TRT}} e^{\beta_3 \text{ALCH}} e^{\beta_4 \text{DRUG}}}\end{aligned}$$

$$\begin{aligned}
&= \frac{e^0}{e^{-0.5}} \\
&= e^{0.5} \approx 1.65.
\end{aligned} \tag{8.8}$$

So the interpretation is that, after adjustment for treatment group and alcohol and drug usage, whites tend to use the emergency room at a rate 1.65 that of the nonwhites. Said another way, the average rate of usage for whites is 65% higher than that for non-whites. Similar, multiplicative, interpretations apply to the other coefficients.

In summary, to interpret the coefficients when modeling the log of the mean, we need to exponentiate them and interpret them in a multiplicative or ratio fashion. In fact, it is often good to think ahead to the desired type of interpretation. Proportional increases in the mean response due to covariate effects are sometimes the most natural interpretation and are easily incorporated by planning to use such a model.

8.1.5 Further Notes

Models like the one developed in this section are often called Poisson regression models, named after the distribution assumed for the counts. A feature of the Poisson distribution is that the mean and variance are required to be the same. So, if the mean number of emergency room visits per year is 1.5, for subjects with a particular pattern of covariates, then the variance would also be 1.5 and the standard deviation would be the square root of that or about 1.23 visits per year. Ironically, the Poisson distribution often fails to hold in practice since the variability in the data often exceeds that of the mean. A common solution (where appropriate) is to assume that the variance is proportional to the mean, not exactly equal to it, and estimate the proportionality factor, which is called the *scale parameter*, from the data. For example, a scale parameter of 2.5 would mean that the variance was 2.5 times larger than the mean and this fact would be used in calculating standard errors, hypothesis tests, and confidence intervals. When the scale parameter is greater than 1, meaning that the variance is larger than that assumed by the named distribution, the data are termed *overdispersed*. Another solution is to choose a different distribution. For example, the Stata package has a negative binomial (a different count data distribution) regression routine, in which the variance is modeled as a quadratic function of the mean.

The use of log time as an offset in model (8.5) may seem awkward. Why not just divide each count by the observation period and analyze Y_i / t_i ? The answer is that it makes it harder to think about and specify the proper distribution. Instead of having count data, for which there are a number of statistical distributions to choose from, we would have a strange, hybrid distribution, with “fractional” counts, e.g., with an

observation period of 0.8 of a year, we could obtain values of 0, 1.25 (which is 1 divided by 0.8), 2.5, 3.75, etc. With a different observation period, a different set of values would be possible.

8.2 Example: Costs of Phototherapy

About 60% of newborns become jaundiced, i.e., the skin and whites of the eyes turn yellow in the first few days after birth. Newborns become jaundiced because they have an increase in bilirubin production due to increased red blood cell turnover and because it takes a few days for their liver (which helps eliminate bilirubin) to mature. Newborns are treated for jaundice because of the possibility of bilirubin-induced neurological damage. What are the costs associated with this treatment and are costs also associated with race, the gestational age of the baby, and the birthweight of the baby?

Our outcome will be the total cost of health care for the baby during its first month of life. Cost is a positive variable and is almost invariably highly skewed to the right. A common remedy is to log transform the costs and then fit a multiple regression model. This is often highly successful as log costs are often well-behaved statistically, i.e., approximately normally distributed and homoscedastic. This is adequate if the main goal is to test whether one or more risk factors are related to cost.

However, if the goal is to understand the determinants of the actual *cost* of health care, then it is only the mean cost that is of interest (since mean cost times the number of newborns is the total cost to the health care system). One strategy is to perform the analysis on the log scale and then back transform (using an exponential) to get things back on the original cost scale.

However, since the log of the mean is not the same as the mean of the log, back-transforming an analysis on the log scale does not directly give results interpretable in terms of mean costs. Instead they are interpretable as models for median cost (Goldberger 1968). The reasoning behind this is as follows. If the log costs are approximately normally distributed, then the mean and median are the same. Since monotonic transformations preserve medians (the log of the median value *is* the median of the log values) back-transforming using exponentials gives a model for median cost. There are methods for getting estimates of the mean via adjustments to the back transformation (Bradu and Mundlak 1970) but there are also alternatives.

One alternative is to adopt the approach of the previous section: model the mean and assume a reasonable distribution for the data. What choices would we need to make for this situation?

A reasonable starting point is to observe that the mean cost must be positive. Additive and linear models for positive quantities can cause the problem of negative predicted values and hence multiplicative models incorporating proportional changes are commonly used. For cost, this is often a more natural characterization, i.e., “low birthweight babies cost 50% more than normal birthweight babies”

and is likely to be more stable than modeling absolute changes in cost (locations with very different costs of care are unlikely to have the same differences in costs, but may have the same ratio of costs). As in the previous section, that would lead to a model for the log of the mean cost (similar to but not the same as log-transforming cost).

8.2.1 *Model for the Mean Response*

More precisely, let us define Y_i as the cost of health care for infant i during its first month and let $E[Y_i]$ represent the average cost. Our model would then be

$$\log E[Y_i] = \beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{GA}_i + \beta_4 \text{BW}_i, \quad (8.9)$$

or equivalently (using an exponential)

$$E[Y_i] = \exp\{\beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{GA}_i + \beta_4 \text{BW}_i\}, \quad (8.10)$$

where β_0 is an intercept, RACE_i is 0 for whites and 1 for non-whites, TRT_i is 1 for those receiving phototherapy and 0 for those who do not, GA_i is the gestational age of the baby, and BW_i is its birthweight. We are primarily interested in β_2 , the phototherapy effect.

8.2.2 *Choice of Distribution*

The model for the mean for the jaundice example is virtually identical to that for the depression example in Sect. 8.1.2. But the distributions need to be different since cost is a continuous variable, while number of emergency room visits is discrete. There is no easy way to know what distribution might be a good approximation for such a situation, without having the data in hand. However, it is often the case that the standard deviation in the data increases proportionally with the mean. This situation can be diagnosed by looking at residual plots (as described in Chap. 4) or by plotting the standard deviations calculated within subgroups of the data versus the means for those subgroups. In such a case, a reasonable choice is the gamma distribution, which is a flexible distribution for positive, continuous variables that incorporates the assumption that the standard deviation is proportional to the mean.

When we are willing to use a gamma distribution as a good approximation to the distribution of the data, we can complete the specification of the model as follows. We assume that Y_i has a gamma distribution with mean, $E[Y_i]$, given by

$$\log E[Y_i] = \beta_0 + \beta_1 \text{RACE}_i + \beta_2 \text{TRT}_i + \beta_3 \text{GA}_i + \beta_4 \text{BW}_i. \quad (8.11)$$

8.2.3 Interpreting the Parameters

Since the model is a model for the log of the mean, the parameters have the same interpretation as in the previous section. For example, if $\hat{\beta}_2 = 0.5$ (positive since phototherapy increases costs) then the interpretation would be that, adjusting for race, gestational age, and birthweight, the cost associated with babies receiving phototherapy was $\exp(0.5) \approx 1.65$ as high as those not receiving it.

8.3 Generalized Linear Models

The examples in Sects. 8.1 and 8.2 have been constructed to emphasize the similarity of the models (compare Subjects. 8.1.4 and 8.2.3) for two very different situations. So even with very different distributions (Poisson versus gamma) and different statistical analyses, they have much in common.

A number of statistical packages, including Stata, have what are called *generalized linear model* commands that are capable of fitting linear, logistic, Poisson regression and other models. The basic idea is to let the data analyst tailor the analysis to the data rather than having to transform or otherwise manipulate the data to fit an analysis. This has significant advantages in situations like the phototherapy cost example where we want to model the outcome without transformation.

Fitting a GLM involves making a number of decisions:

- (1) What is the distribution of the data (for a fixed pattern of covariates)?
- (2) What function will be used to *link* the mean of the data to the predictors?
- (3) Which predictors should be included in the model?

In the examples in the preceding sections we used Poisson and gamma distributions, we used a log function of the mean to give us a linear model in the predictors and our choice of predictors was motivated by the subject matter. Note that choices on the predictor side of the equation are largely independent of the first two choices.

In previous chapters, we have covered linear and logistic regression. In linear regression, we modeled the mean directly and assumed a normal distribution. This is using an *identity link function*, i.e., we modeled the mean identically, without transforming it. In logistic regression, we modeled the log of the odds, i.e., $\log(p/[1 - p])$, and assumed a binomial or binary outcome. If the outcome is coded as zero for failure and one for success, then the average of the zeros and ones is p , the probability of success. In that case, we used a *logit link* to link the mean, p , to the predictors.

Generalized linear model commands give large degrees of flexibility in the choice of each of the features of the model. For example, current capabilities in Stata are to handle six distributions (normal, binomial, Poisson, gamma, negative binomial,

Table 8.1 Count regression example assuming a Poisson distribution

glm shared_syr i.homeless, family(poisson) link(log) eform

Generalized linear models		No. of obs	=	121
Optimization	: ML	Residual df	=	119
		Scale parameter	=	1
Deviance	= 1511.02467	(1/df) Deviance	=	12.69769
Pearson	= 3586.309617	(1/df) Pearson	=	30.13706
Variance function: V(u) = u		[Poisson]		
Link function : g(u) = ln(u)		[Log]		
Log likelihood = -805.0147598		AIC	=	13.33909
		BIC	=	940.3256

shared_syr	IRR	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
1.homeless	3.270615	.3985062	9.73	0.000	2.575819	4.152825

and inverse gaussian), and ten link functions (including identity, log, logit, probit, power functions).

8.3.1 Example: Risky Drug Use Behavior

Here is an example of modeling risky drug use behavior (sharing syringes) among drug users. The outcome is the number of times the drug user shared a syringe (`shared_syr`) in the past month (values ranged from 0 to 60!) and we will consider a single predictor, whether or not the drug user was homeless. Table 8.1 gives the results assuming a Poisson distribution. The Stata command, `glm`, specifies a Poisson distribution and a log link and we have specified the option `eform`, which automatically exponentiates the coefficients. The output contains a number of standard elements, including estimated coefficients, standard errors, Z-tests, P-values, and CIs. The homeless coefficient is highly statistically significant, with a value of about 3.27, meaning that being homeless is associated with over three times more use of shared syringes than nonhomeless.

However, these data are highly variable and the Poisson assumption of equal mean and variance is dubious. If we specify the `vce(robust)` a robust variance estimate will be used in the calculation of the standard errors. Just as described in Chap. 7, the robust variance estimate gives valid standard errors even when the assumed form of the variance is incorrect, in this case that the variance is equal to the mean.

Table 8.2 gives the result with the robust standard errors, which is not quite statistically significant. Standard errors have increased approximately fivefold using the `vce(robust)` option, so the assumption of a Poisson distribution is far from correct. In the terminology of generalized linear models, these data are highly

Table 8.2 Count regression example with scaled standard errors

```
. glm shared_syr i.homeless, family(poisson) link(log) eform vce(robust)
```

Generalized linear models		No. of obs	=	121
Optimization	: ML	Residual df	=	119
		Scale parameter	=	1
Deviance	= 1511.02467	(1/df) Deviance	=	12.69769
Pearson	= 3586.309617	(1/df) Pearson	=	30.13706
Variance function: V(u) = u		[Poisson]		
Link function : g(u) = ln(u)		[Log]		
Log pseudolikelihood = -805.0147598		AIC	=	13.33909
		BIC	=	940.3256

shared_syr	IRR	Robust Std. Err.	z	P> z	[95% Conf. Interval]
1.homeless	3.270615	2.005987	1.93	0.053	.9830072 10.88184

overdispersed, because the variance is much larger than that assumed for a Poisson distribution.

This example serves as a warning not to make strong assumptions, such as those embodied in using a Poisson distribution, blindly. It is wise at least to make a sensitivity check by using the robust variance estimator for count data as well as for binomial data with denominators other than 1 (with binary data, with a denominator of 1, no overdispersion is possible). Also, when there are just a few covariate patterns and subjects can be grouped according to their covariate values, it is wise to plot the variance within such groups versus the mean within the group to display the variance to mean relationship graphically. The mean values for the `shared_syr` variable are 4.7 and 1.4 for the homeless and nonhomeless groups, respectively, with corresponding standard deviations of 13.7 and 5.5. So the standard deviations are roughly three times the mean, as reflected by the robust standard errors being much larger in Table 8.2 compared to Table 8.1.

An alternative distribution for count data that allows more variability than the Poisson is the negative binomial distribution. Table 8.3 shows the negative binomial distribution fit. The estimated effect of being homeless is the same as the Poisson fit and the standard errors, *p*-value and CI are all similar to those in Table 8.2, which uses a robust standard error.

8.3.2 Modeling Data with Many Zeros

When analyzing count *or* numerical outcome data, it is not unusual to discover a large percentage of the data being zero. For example, in a study following the members of a health plan for use of emergency room visits, the vast majority would be zero, with the nonzero outcomes taking on integer values. If we change the

Table 8.3 Count regression using a negative binomial distribution

```
. glm shared_syr i.homeless, family(nbinomial ml) ef
```

Generalized linear models		No. of obs	=	121
Optimization	: ML	Residual df	=	119
		Scale parameter	=	1
Deviance	= 58.10151246	(1/df) Deviance	=	.488248
Pearson	= 94.44640018	(1/df) Pearson	=	.7936672
Variance function: V(u) = u+(14.1206)u^2		[Neg. Binomial]		
Link function : g(u) = ln(u)		[Log]		
Log likelihood = -154.8084869		AIC	=	2.591876
		BIC	=	-512.5976

shared_syr	IRR	Std. Err.	z	P> z	[95% Conf. Interval]
1.homeless	3.270616	2.270502	1.71	0.088	.8389088 12.751

Note: Negative binomial parameter estimated via ML and treated as fixed once estimated.

outcome to hospitalization costs, again the vast majority would be zero, but the nonzero values would likely be positive and skewed right. For the syringe sharing data, 78% of the outcomes are zero. How can these be handled in practice? As noted in Sect. 8.1.1, a transformation of the outcome will not help.

A simple strategy is to build separate models for the zeros and the nonzero values. These are sometimes called conditional, two-part or “hurdle” models, the latter name arising because, after the outcomes “hurdle” the value of zero, a different model is used. For the analysis of hospitalization costs, we could use a logistic regression for the probability of the cost being zero. And for the nonzero costs, we could fit a GLM assuming a gamma distribution or we could log transform the outcome to try to make it approximately normally distributed. The same predictors can be in both models or we can model each outcome with its own collection of predictors.

For the syringe sharing data, we could use a logistic regression to model the chance of sharing zero syringes with the predictor of being homeless. But what model could we use for the nonzero data? It does not fit any usual count data model, because there are no zeros allowed. Fortunately, Stata can accommodate either a Poisson or negative binomial distribution which has been *truncated* to only allow nonzero values through its *ztp* (zero-truncated Poisson) or *zt nb* (zero-truncated negative binomial) regression commands.

Table 8.4 shows the two fits, first modeling the probability of no syringe sharing with the predictor of being homeless and a logistic regression and then the number of times syringes are shared, using a zero-truncated negative binomial distribution. Because the fits provide two tests of the homeless effect based on the same data, we could use a Bonferroni correction (see Sect. 4.3.4) and test each at a significance level of 0.025 instead of 0.05; correspondingly, we have used the `level` option

Table 8.4 Fitting a two-part model to the syringe sharing data

```
. gen share0=(shared_syr==0)
. logistic share0 i.homeless, level(97.5)
```

Logistic regression

Log likelihood = -67.012808

Number of obs = 124

LR chi2(1) = 3.19

Prob > chi2 = 0.0740

Pseudo R2 = 0.0233

share0	Odds Ratio	Std. Err.	z	P> z	[97.5% Conf. Interval]	
1.homeless	.4676114	.2019848	-1.76	0.078	.1775875	1.231283

```
. ztnb shared_syr i.homeless if shared_syr>0, irr level(97.5)
```

Zero-truncated negative binomial regression

Dispersion = mean

Log likelihood = -91.29219

Number of obs = 27

LR chi2(1) = 1.02

Prob > chi2 = 0.3133

Pseudo R2 = 0.0055

shared_syr	IRR	Std. Err.	z	P> z	[97.5% Conf. Interval]	
1.homeless	2.122108	1.501492	1.06	0.288	.4345308	10.36369
/lnalpha	1.584369	1.081463			-.8396253	4.008363
alpha	4.876212	5.273443			.4318723	55.05666

Likelihood-ratio test of alpha=0: chibar2(01)= 413.32 Prob>=chibar2 = 0.000

to set the CIs to have 97.5% confidence. The logistic model estimates the odds ratio of *not sharing a needle*. Perhaps easier to interpret, the homeless have odds of sharing a needle which are a little over two times higher than the nonhomeless ($2.1385 = 1/.4676$). The zero-truncated regression estimate indicates that, among those that do share needles, the homeless share syringes at a rate a little over two times more often than the nonhomeless. Because of the Bonferroni correction, each of the tests would require a p -value of 0.025 to be declared statistically significant and neither is.

These two-part or zero-inflated (below) modeling approaches are especially attractive in situations where different predictors might influence the two parts of the model. For example, what determines whether or not someone is willing to share needles may be quite different from what determines how frequently they share needles when they do.

Another approach to modeling count data with many zeros is to use what are called *zero-inflated* models. Rather than breaking the data into two parts, a zero-inflated approach uses an underlying model in which two processes are operating: first a process that generates the zeros (like the logistic regression above) and then a count data model, such as the Poisson. This is slightly different from the two-part model since a zero in the data could have arisen either from the zero-generation process or from the count data process, which just happens to generate

Table 8.5 Fitting a zero-inflated negative binomial model to the syringe sharing data

```
. zinb shared_syr i.homeless, inflate(i.homeless) irr level(97.5)
```

Zero-inflated negative binomial regression	Number of obs	=	121
	Nonzero obs	=	27
	Zero obs	=	94
Inflation model = logit	LR chi2(1)	=	1.02
Log likelihood = -154.112	Prob > chi2	=	0.3133

shared_syr	IRR	Std. Err.	z	P> z	[97.5% Conf. Interval]
shared_syr					
1.homeless	2.122107	1.501493	1.06	0.288	.4345298 10.3637
inflate					
1.homeless	-.7708614	.7434234	-1.04	0.300	-2.437173 .8954498
_cons	.6342794	1.161001	0.55	0.585	-1.967991 3.236549
/lnalpha	1.584377	1.08147	1.47	0.143	-.8396342 4.008387
alpha	4.87625	5.27352			.4318685 55.05801

a zero. These models are more natural for some situations. For example, consider modeling the number of open nurse anesthetist positions per hospital, similar to the study of [Merwin et al. \(2009\)](#), with predictors being the log of the number of surgeries, log of the average daily number of patients, log of the number of operating rooms and the state in which it is located. The number of open positions could be zero because the hospital does not hire nurse anesthetists or because they do, but they have no open positions. Zero-inflated models can be used in situations in which we can justify the underlying two processes or in situations in which we merely need to accommodate the large percentage of zero outcome values.

Table 8.5 shows the results from fitting a zero-inflated negative binomial model, where the `inflate` option gives the predictors for the underlying process that generates the zeros and again we have set the level to 97.5% to accommodate the two tests. The estimates and interpretations are very similar to the two-part fit above, with the effect of being homeless on the count data model being identical and the effect of being homeless on the zero model being very similar. Table 8.5 reports the log odds of not sharing a syringe as -0.7708 , which corresponds to an odds ratio of $\exp\{-0.7708\} = 0.4626$, similar to Table 8.4.

8.3.3 Example: A Randomized Trial to Reduce Risk of Fracture

Osteoporosis (roughly porous bone, from the Greek) is a condition in which bones become weak and brittle and readily susceptible to fracture. It primarily affects postmenopausal women and can lead to chronic pain, skeletal deformities, and

Table 8.6 Fracture risk by fall risk and treatment group

glm numnosp ibn.trt_fall, family(poisson) offset(logyears) vce(robust) noconstant ef

Generalized linear models		No. of obs	=	6369
Optimization	: ML	Residual df	=	6365
		Scale parameter	=	1
Deviance	= 4116.885884	(1/df) Deviance	=	.6468006
Pearson	= 8002.406864	(1/df) Pearson	=	1.257252
Variance function: V(u) = u		[Poisson]		
Link function : g(u) = ln(u)		[Log]		
		AIC	=	.9180298
Log pseudolikelihood = -2919.465795		BIC	=	-51635.41

		Robust		z	P> z	[95% Conf. Interval]	
numnosp		IRR	Std. Err.				
trt_fall							
1		.041815	.0022419	-59.21	0.000	.0376439	.0464482
2		.0340865	.0020225	-56.95	0.000	.0303444	.0382902
3		.0509974	.0056819	-26.71	0.000	.0409931	.0634431
4		.0521462	.0055934	-27.54	0.000	.042259	.0643466
logyears		(offset)					

increased risk of death. The Fracture Intervention Trial ([Black et al. 1996a](#)) was a randomized controlled trial among postmenopausal women that showed that alendronate (a drug that increases bone density) was able to reduce the risk of fracture.

Falling is a major cause of fractures, but would alendronate prevent fractures from an event as traumatic as a fall? To answer this, women at high risk of falling were identified by poor performance on the “Timed Up and Go” test, which measures how long it takes to stand up from an armchair, walk 3 m, return and sit down, and has been shown to be a predictor of the risk of falling. The effect of alendronate on the number of nonspine fractures (numnosp) was then estimated separately for the high and low risk of falling groups.

Women were not followed for the same amount of time, so we use an offset of log of years in the trial (logyears). To get estimated rates for each of the groups, we created a four level, categorical variable (trt_fall) with values 1–4 representing, respectively, the low risk/placebo, low risk/alendronate, high risk/placebo, and high risk/alendronate groups. We use the ibn. prefix for the trt_fall variable (so there is no omitted baseline reference group) and noconstant options to force Stata to include all four groups and to not fit a constant term.

We fit the model using the glm command, specifying a Poisson distribution, using robust standard errors (in case of overdispersion), and reporting the results with the eform option to directly display the yearly fracture rates. The output is displayed in Table 8.6.

In the low risk of falling groups, the yearly rates of fracture are much less (0.042 for the placebo group and 0.034 for the alendronate groups). These correspond to about 4.2 and 3.4 fractures per 100 women over a year. In the high risk of falling groups, the rates are higher and about the same as one another (about 5.1 and 5.2 fractures per 100 woman years). We wish to compare the risk difference between the treated and untreated groups over the average followup time of 3.8 years. This is of interest because, unlike a relative risk or odds ratio, it is easily related to the number of fractures prevented by treatment.

To make the formal comparison, we fit a model with effects for treatment (`trt`), fall risk category (`fall_risk`), and their interaction. The `margins` command can be conveniently used with the `@` operator to compare the treatment groups within the fall risk categories. The results are given in Table 8.7.

The results show that, over a period of 3.8 years, the risk associated with being treated by alendronate in the low fall risk group is about 0.029 less than the untreated group. In other words, in the low fall risk group, treatment with alendronate prevents about 2.9 fractures per 100 women over a 3.8-year treatment period. However, in the high risk group, the difference between alendronate and the placebo group is not statistically significant and has a small estimated effect (the drug is estimated to increase the number of fractures per 100 women over 3.8 years by about 0.4 fractures). Because the high risk group is smaller, the confidence interval is wide, and so we cannot rule out clinically important differences.

The analysis above is appropriate if the focus is, a priori, on estimating the treatment effects separately in the low and high risk of falling groups. In this case, the results also suggest that there is a difference in the treatment effect in the two groups. However, if the goal is to directly compare the treatment effects in the low and high risk groups, a better approach is to test for the interaction between risk of falling and treatment (see the third point in Sect. 4.6). The test of interaction with this data in Table 8.7 does not provide strong evidence for a difference in treatment effects, with a p -value of 0.19. This is a caution not to interpret statistical significance of an effect in one group and lack of statistical significance in another group as evidence for a difference in the effects in the two groups. This is especially true with unequal sample sizes, such as in this example.

8.3.4 Relationship of Mean to Variance

The key to use of a GLM program is the specification of the relationship of the mean to the variance. This is the main information used by the program to fit a model to data when a distribution is specified. As noted above, this relationship can often be assessed by residual plots or plots of subgroup standard deviations versus means. Table 8.8 gives the assumed variance to mean relationship, distributional name, and situations in which the common choices available in Stata would be used.

Table 8.7 Fracture risk treatment comparisons within fall risk categories

glm numnosp i.trt##i.fall_risk, family(poisson) offset(logyears)
vce(robust) ef

Generalized linear models		No. of obs	=	6369
Optimization	: ML	Residual df	=	6365
Deviance	= 4116.885884	Scale parameter	=	1
Pearson	= 8002.406864	(1/df) Deviance	=	.6468006
		(1/df) Pearson	=	1.257252
Variance function: V(u) = u		[Poisson]		
Link function : g(u) = ln(u)		[Log]		
Log pseudolikelihood = -2919.465795		AIC	=	.9180298
		BIC	=	-51635.41

numnosp	IRR	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1.trt	.8151755	.0651883	-2.56	0.011	.6969183	.9534993
1.fall_risk	1.219596	.1507959	1.61	0.108	.9571279	1.55404
trt#fall_risk						
1 1	1.254364	.2183953	1.30	0.193	.8917067	1.764513
_cons	.041815	.0022419	-59.21	0.000	.0376439	.0464482
logyears	1	(offset)				

. margins r.trt@fall_risk

Contrasts of adjusted predictions
Model VCE : Robust

Expression : Predicted mean numnosp, predict()

	df	chi2	P>chi2
trt@fall_risk			
(1 vs 0) 0	1	6.55	0.0105
(1 vs 0) 1	1	0.02	0.8854
Joint	2	6.57	0.0374

	Contrast	Delta-method Std. Err.	[95% Conf. Interval]	
trt@fall_risk				
(1 vs 0) 0	-.0293294	.0114584	-.0517874	-.0068713
(1 vs 0) 1	.0043597	.0302577	-.0549444	.0636637

8.3.5 Non-Linear Models

Not every model fits under the GLM umbrella. Use of the method depends on finding a transformation of the mean for which the predictors enter as a linear model, which may not always be possible. For example, in drug pharmacokinetics, a common model for the mean concentration of a drug in blood, Y , as a function of time, t , is:

Table 8.8 Common distributional choices for generalized linear models in Stata

Distribution	Variance to mean ^a	Sample situation
Normal	Constant σ^2	Linear regression
Binomial	$\sigma^2 = n\mu(1 - \mu)$	Successes out of n trials
OD ^b Binomial	$\sigma^2 \propto n\mu(1 - \mu)$	Clustered success data
Poisson	$\sigma^2 = \mu$	Count data, variance equals mean
OD Poisson	$\sigma^2 \propto \mu$	Count data, variance proportional to mean
Negative binomial	$\sigma^2 = \mu + \mu^2/k$	Count data, variance quadratic in the mean
Gamma	$\sigma \propto \mu$	Continuous data, standard deviation proportional to mean

^aMean is denoted by μ and the variance by σ^2 .

^bOver-dispersed.

$$E[Y] = \mu_1 \exp\{-\lambda_1 t\} + \mu_2 \exp\{-\lambda_2 t\}. \quad (8.12)$$

In addition to time, we might have other predictors such as drug dosage or gender of the subject. However, there is no transformation that will form a linear predictor, even without the inclusion of dose and gender effects, and so a generalized linear model is not possible. As a consequence, more care must be taken in deciding how to incorporate the effects of predictor variables and building regression models is thus more complicated for nonlinear models. Software for fitting nonlinear models is relatively common for approximately normally distributed outcomes (such as `nl` in Stata) but less so for nonnormally distributed outcomes.

8.4 Sample Size for the Poisson Model

Section 5.7 provides formulas for calculating sample size, power, and minimum detectable effects for the logistic model. Similar results hold for the Poisson model. To compute the sample size that will provide power of γ in two-sided tests with type-1 error of α to reject the null hypothesis $\beta_j = 0$ for the effect of a predictor X_j , accounting for the loss of precision arising from multiple predictors, we can use

$$n = \frac{(z_{1-\alpha/2} + z_\gamma)^2 \theta}{(\beta_j^a \sigma_j)^2 \mu (1 - \rho_j^2)}, \quad (8.13)$$

where β_j^a is the hypothesized value of β_j under the alternative, $z_{1-\alpha/2}$ and z_γ are the quantiles of the standard normal distribution corresponding to the specified type-1 error and power, ρ_j is the multiple correlation of X_j with the other covariates, μ is the marginal mean of the count outcome, and θ is the scale parameter introduced in Sect. 8.3.1, defined as the ratio of variance of the outcome to μ . When X_j is binary with prevalence f_j , $\sigma_{x_j} = \sqrt{f_j(1 - f_j)}$. For problems with predetermined n , power is given by

$$\gamma = 1 - \Phi \left[z_{1-\alpha/2} - \beta_j^a s_{x_j} \sqrt{n\mu(1-\rho_j^2)/\theta} \right]. \quad (8.14)$$

Finally, the minimum detectable effect (on the log-mean scale) is

$$\pm \beta_j^a = \frac{z_{1-\alpha/2} + z_\gamma}{s_{x_j} \sqrt{n\mu(1-\rho_j^2)/\theta}}. \quad (8.15)$$

Some additional points:

- Sample size (8.13) and minimum detectable effect (8.15) calculations simplify considerably when we specify $\alpha = 0.05$ and $\gamma = 0.8$, β_j^a is the effect of a one standard deviation increase in continuous x_j , and we do not need to penalize for covariate adjustment. However, we do assume that over-dispersion may still need to be taken into account, via the parameter θ . In that case,

$$n = \frac{7.849}{(\beta_j^a)^2 \mu / \theta}. \quad (8.16)$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = \frac{2.802}{\sqrt{n\mu/\theta}}. \quad (8.17)$$

For 90% power, substitute 10.51 for 7.849 and 3.242 for 2.802.

- Similarly, for a two-arm clinical trial with equal allocation to arms, so that β_j^a is the log rate ratio for treatment, and $s_{x_j}^2 = 0.25$, we can calculate

$$n = \frac{4 \times 7.849}{(\beta_j^a)^2 \mu / \theta}. \quad (8.18)$$

For the minimum detectable effect, we have

$$\pm \beta_j^a = \frac{2 \times 2.802}{\sqrt{n\mu/\theta}}. \quad (8.19)$$

- Power calculations using (5.17) simplify analogously, but still require a statistical calculator or computer package to evaluate the normal cumulative distribution function $\Phi(\cdot)$.
- To our knowledge, these computations are not implemented in any statistical packages. However, (8.13)–(8.15) can easily be programmed in Stata, R, or Excel, or evaluated by hand if values of $z_{1-\alpha/2}$, z_γ , and $\Phi(\cdot)$ are available.
- As in calculations for other models, we need to use $|\beta_j^a|$ in (8.13) and (8.14) if $\beta_j^a < 0$.

Table 8.9 Sample size calculations for trial of behavioral intervention

```
. display (invnormal(.975) + invnormal(.9))^2 * 30 / ((log(0.5) * 0.5)^2 * 7.5)
349.91719
```

- The severe overdispersion evident in the example in Sect. 8.3.1 underlines the importance of obtaining a good estimate of θ , the scale parameter capturing overdispersion in (8.13)–(8.15). Note that $n \propto \theta$.
- The use of the variance inflation factor to account for covariate adjustment carries over to GLMs. However, there is no analog to the reduction in residual variance, so that the adjustment based on the variance inflation factor is less likely to be conservative for these models.
- $\text{SE}(\hat{\beta}_j)$ is a large-sample approximation, and more exact small-sample computations using the t -distribution do not carry over from the linear model. Simulations of power may be a more reliable guide in those circumstances.
- Equations (8.13)–(8.15) are based on the assumption that the conditional mean of the outcome does not vary strongly across observations; methods based on more complicated calculations or simulation avoid this simplification and perform slightly better in some circumstances (Vittinghoff et al. 2009). However, errors from these sources are usually small compared to errors arising from uncertainty about the required inputs.
- The alternative calculations (4.22)–(4.24) presented in Sect. 4.8, which use an estimate $\tilde{\text{SE}}(\hat{\beta}_j)$ based on published results for an appropriately adjusted model using $\tilde{n}$ observations, carry over directly. However, care must be taken to obtain the SE of the regression coefficient β_j , not the SE of the rate-ratio e^{β_j} . This can be computed from a 95% CI for the rate-ratio as $\tilde{\text{SE}}(\hat{\beta}_j) = \log(\text{UL}/\text{LL})/3.92$, where UL and LL are the upper and lower bounds. We must also ensure that β_j^a is based on the same predictor scale as in the published results.

To illustrate these calculations, suppose we are planning a randomized trial to assess the effectiveness of a behavioral intervention for reducing syringe sharing among drug users. Equal numbers will be allocated to the active intervention and a wait-list control, so that $f_j = 0.5$ and $s_{x_j} = \sqrt{0.5(1-0.5)} = 0.5$. Because the trial is randomized, we can assume that $\rho_j = 0$. Using the data shown in Tables 8.1 and 8.2, we estimate that among the wait-list controls, $\mu = 10$, and $\theta = 30$. We hypothesize that the intervention will reduce the frequency of sharing by 50%, so that overall, $\mu = 7.5$ and $\beta_j^a = \log 0.5$. In this case, we require power of 90% in a two-sided test with α of 5%.

Table 8.9 shows the sample size estimate of 350. This estimate has been inflated by a factor of $\theta = 30$ to account for overdispersion of the outcome. Clearly a naïve estimate assuming equality of the mean and variance would result in a badly underpowered trial.

8.5 Summary

The purpose of this chapter has been to outline the topic of GLMs, a class of models capable of handling a wide variety of analysis situations. Specification of the generalized linear model involves making three choices:

- (1) What is the distribution of the data (for a fixed pattern of covariates)? This must be specified at least up to the variance to mean relationship.
- (2) What function will be used to link the mean of the data to the predictors?
- (3) Which predictors should be included in the model?

Generalized linear models are similar to linear, logistic, and Cox models in that much of the work in specifying and assessing the predictor side of the equation is the same no matter what distribution or link function is chosen. This can be especially helpful when analyzing a study with a variety of different outcomes, but similar questions as to what determines those outcomes. For example, in the depression example we might also be interested in cost, with a virtually identical model and set of predictors.

8.6 Further Notes and References

There are a number of book-length treatments of generalized linear models, including [Dobson \(2001\)](#) and [McCullagh and Nelder \(1989\)](#). In Chap. 7, we extended the logistic model to accommodate correlated data by the use of generalized estimating equations and by including random effects. The GLMs described in this chapter can similarly be extended and fit using the `xtgee` command in Stata and GENMOD procedure in SAS, which can be used with a variety of distributions. Random effects models can be estimated for a number distributions using the cross-sectional time-series commands in Stata (these commands are prefixed by `xt`) and with the NLMIXED procedure in SAS.

There are a number of approaches to modeling data with many zeros; [Lachenbruch \(2002\)](#) provides an accessible survey. He also considers the issue of power compared to simpler analyses. For example, in the simple two-group comparison of Sect. 8.3.1, we could use a nonparametric test like the Wilcoxon rank sum test. He shows that using two part or zero-inflated models, which explicitly model zeros, will often have higher power than simpler approaches that merely accommodate an outcome distribution with many zeros.

8.7 Problems

Problem 8.1. We made the point in Sect. 8.1.1 that a log transformation would not alleviate nonnormality. Yet we model the log of the mean response. Let us consider the differences.

- (1) First consider the small data set consisting of 0, 1, 0, 3, 1. What is the mean? What is the log of the mean? What is the mean of the logs of each data point?
- (2) Even if there are no zeros, these two operations are quite different. Consider the small data set consisting of 2, 3, 32, 7, 11. What is the log of the mean? What is the mean of the logs of the data? Why are they different?
- (3) Repeat the above calculation, but using medians.

Problem 8.2. What would you need to add to model (8.5) to assess whether the effect of the treatment was different in whites as compared to non-whites?

Problem 8.3. Suppose the coefficient for $\hat{\beta}_2$ in (8.6) was -0.2 . Provide an interpretation of the treatment effect.

Problem 8.4. For each of the following scenarios, describe the distribution of the outcome variable (Is it discrete or approximately continuous? Is it symmetric or skewed? Is it count data?) and which distribution(s) might be a logical choice for a GLM.

- (1) A treatment program is tested for reducing drug use among the homeless. The outcome is injection drug use frequency in the past 90 days. The values range from 0 to 900 with an average of 120, a median of 90, and a standard deviation of 120. Predictors include treatment program, race (white/non-white), and sex.
- (2) In a study of detection of abnormal heart sounds the values of brain natriuretic peptide (BNP) in the plasma are measured. The outcome, BNP, is sometimes used as a means of identifying patients who are likely to have signs and symptoms of heart failure. The BNP values ranged from 5 to 4,000 with an average of 450, a median of 150, and a standard deviation of 900. Predictors include whether an abnormal heart sound is heard, race (white/non-white), and sex.
- (3) A clinical trial was conducted at four clinical centers to see if alendronate (a bone-strengthening medication) could prevent vertebral fractures in elderly women. The outcome is total number of vertebral fractures over the follow-up period (intended to be 5 years for each woman). Predictors include drug versus placebo, clinical center, and whether the woman had a previous fracture when enrolled in the study.

Problem 8.5. For each of the scenarios outlined in Problem 8.4, write down a preliminary model by specifying the assumed distribution, the link function, and how the predictors are assumed to be related to the mean.

8.8 Learning Objectives

- (1) State the advantage of using a GLMs approach.
- (2) Given an example, make reasonable choices for distributions, and link functions.
- (3) Given output from a GLMs routine, state whether predictors are statistically significant and provide an interpretation of their estimated coefficients.

Chapter 9

Strengthening Causal Inference

In Chaps. 4–8, we showed how multi-predictor regression can be used to control for confounding in observational data, with the purpose of estimating the independent association of an exposure with an outcome. The cautious language of associations notwithstanding, the underlying purpose is often to quantify causal relationships. In this chapter, we explain what is meant by the *average causal effect* of an exposure, and discuss the conditions under which regression might be able to estimate it. We also show the extra steps that are needed to estimate *marginal* effects, which sometimes differ from the *conditional* effects that regression models estimate by default.

We then present alternatives to regression that can be used when conditions for its successful use are not met. These include *propensity scores*, a robust alternative that is particularly useful when a binary or categorical exposure is common, but the binary or failure time outcome is not, and there are many confounders of exposure that must be accounted for. These scores are commonly estimated using ancillary logistic models for exposure, then incorporated in the analysis of the effect of the exposure on the outcome by means of stratification, regression adjustment, inverse weighting, or matching.

Regression adjustment can also fail when both the exposure and confounder are time-dependent, the confounder affects exposure and outcome, and exposure affects subsequent levels of the confounder. Cox and repeated measures models accommodate time-dependent exposures and confounders, but in this context cannot be used to estimate the overall effect of exposure. We focus on models using *inverse probability weights*, and briefly describe *nested new-user cohorts* and *G-estimation*.

In estimating causal effects from observational data, we usually need to assume that there are no unmeasured confounders—a condition that is difficult to meet and impossible to verify. One exception is analysis using *instrumental variables*. However, it does require other unverifiable assumptions. We also briefly discuss an extension of instrumental variables to clinical trials with poor adherence, and show its connection to another approach known as *principal stratification*. Finally, we point to newer developments in Sect. 9.10.

9.1 Potential Outcomes and Causal Effects

Consider the causal effect of exercise on glucose levels among post-menopausal women, first discussed in Chap. 4. Imagine that we could observe glucose levels for every member of this population under two conditions, with and without exercise. In reality, of course, one of the two outcomes would be an unobservable *potential outcome* or *counterfactual*. Nonetheless, an intuitively appealing definition of the causal effect of exercise on glucose levels is the difference between the actual and potential outcomes. Table 9.1 shows what this potential outcomes framework might look like.

In Table 9.1, $Y(1)$ and $Y(0)$ represent glucose levels with ($\mathcal{E} = 1$) and without ($\mathcal{E} = 0$) exercise, while the differences $Y(1) - Y(0)$ are interpretable as the causal effects of exercise on glucose levels for each woman.

9.1.1 Average Causal Effects

Potential outcomes are also central to the definition of the *average causal effect* (ACE) of the exposure. At the individual level, the causal effect of exposure is the difference between the potential outcomes with and without exposure. At the population level, the *average causal effect* is the mean of these differences. For the moment, think of the ten women in Table 9.1 as the entire population. The average causal effect of exercise, defined as the mean of the differences $Y(1) - Y(0)$, is to lower glucose levels by 2 mg/dL.

9.1.1.1 Average Causal Effect as a Difference in Marginal Means

We can also calculate the average causal effect as the difference between the so-called *marginal* means of the potential outcomes with and without exposure. In Table 9.1, we would calculate $E[Y(1)] - E[Y(0)] = 96 - 98 = -2$. This will

Table 9.1 Potential outcomes of exercise

Person	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	97	99	-2
2	98	99	-1
3	99	102	-3
4	100	105	-5
5	96	95	1
6	95	98	-3
7	93	95	-2
8	94	95	-1
9	96	93	3
10	92	99	-7
Mean	96	98	-2

be important in trying to estimate the average causal effect from observed data including actual but not potential outcomes, and also when we consider some other causal effect measures of interest, including the causal odds-ratio, which are *defined* in terms of the marginal means $E[Y(1)]$ and $E[Y(0)]$. In contrast to $E[Y(1)] - E[Y(0)] = E[Y(1) - Y(0)]$, some other causal measures cannot be defined as the mean of individual effects.

9.1.2 Marginal Structural Model

In our thought experiment, we can write a *marginal structural model* for the potential outcomes as

$$E[Y(\mathcal{E})] = \beta_0^* + \beta_1^* \mathcal{E}, \quad (9.1)$$

where $E[Y(\mathcal{E})]$ is the expected value of the potential outcome, $\beta_0^* = E[Y(0)]$ is the marginal mean when $\mathcal{E} = 0$, and $\beta_1^* = E[Y(1)] - E[Y(0)]$ is the average causal effect of $\mathcal{E}$. The marginal structural model resembles other linear models discussed in this book, beginning with (4.2). But in contrast to those models, it is a model for *potential*, not just observed outcomes. Accordingly, it can be unadjusted—exposure is unconfounded because we see both potential outcomes for each individual. The focus of this chapter is on obtaining valid estimates of the causal effect parameter β_1^* using observed data.

9.1.3 Fundamental Problem of Causal Inference

In the complete data shown in Table 9.1, including potential as well as actual outcomes, $E[Y(0)] = 98$ and $E[Y(1)] = 96$, so $\beta_1^* = -2$. But in reality, of course, each person contributes an actual but not a potential outcome. The missing potential outcomes are sometimes called the fundamental problem of causal inference (Holland 1986). Many causal effects of interest are defined in terms of the marginal means, but these means are difficult to estimate from observed data on actual outcomes only.

The problem arises because of what can be seen as selection bias. Suppose that a confounder $\mathcal{C}$ affects the outcome and also influences $\mathcal{E}$, which in turn determines which potential outcome is observed. In our example, the causal direct effect of $\mathcal{C}$, as defined in Sect. 4.5, is to lower glucose levels by 4 mg/dL; in addition, 60% of women with $\mathcal{C} = 1$ exercise, as compared to 40% of those with $\mathcal{C} = 0$.

$\mathcal{C}$ can be ignored in Table 9.1 and the marginal structural model (9.1) because each member of the population contributes an outcome when they do exercise ($\mathcal{E} = 1$) as well as when they do not ($\mathcal{E} = 0$). But this does not hold in Table 9.2, which shows the observed outcomes. The potential outcomes are missing, so we

Table 9.2 Observed outcomes

	Person	$\mathcal{E}$	$Y(1)$	$Y(0)$
$\mathcal{C} = 0$	1	0	–	99
	2	0	–	99
	3	0	–	102
	4	1	100	–
	5	1	96	–
	Mean		98	100
$\mathcal{C} = 1$	6	1	95	–
	7	1	93	–
	8	1	94	–
	9	0	–	93
	10	0	–	99
	Mean		94	96
Overall mean			95.6	98.4

cannot calculate the individual causal effects and average them. Nor can we compare the overall means of 95.6 and 98.4 in the exercise and no exercise groups, which differ substantially from the true marginal means of 96 and 98, as shown in Table 9.1. The difference in means is $95.6 - 98.4 = -2.8$ mg/dL, 40% larger than β_1^* , the average causal effect of exercise.

9.1.4 Randomization Assumption

We have just seen that bias arises because $\mathcal{C}$ affects $\mathcal{E}$ as well as Y , and thus which potential outcome we observe. This is a violation of the so-called *randomization assumption*. Technically, this assumption requires $\mathcal{E}$ to be independent of both potential outcomes, $Y(1)$ and $Y(0)$. In the glucose example, randomization would imply that exercising (or not) is independent of what glucose levels would be under either condition. The randomization assumption is generally met in randomized experiments, since in that setting, exposure is randomly assigned. The exposure we observe for each individual is not affected by confounders that influence the potential outcomes $Y(1)$ and $Y(0)$. When the randomization assumption holds, as in a successfully conducted randomized trial, the marginal means $E[Y(1)]$ and $E[Y(0)]$ can be *identified* or estimated using the sample means of observations with $\mathcal{E} = 1$ and $\mathcal{E} = 0$, respectively, thus providing an estimate of the causal effect β_1^* . Estimation of the marginal causal effect without having to make any modeling assumptions helps explain why experiments, including randomized clinical trials, are the gold standard for estimating marginal causal effects.

9.1.5 Conditional Independence

In contrast, the randomization assumption will rarely if ever hold in observational data. In our example, we know that $\mathcal{C}$ is a common cause of $\mathcal{E}$ and the potential

Table 9.3 Potential outcomes stratified by $\mathcal{C}$

	Person	$Y(1)$	$Y(0)$
$\mathcal{C} = 0$	1	97	99
	2	98	99
	3	99	102
	4	100	105
	5	96	95
	Mean	98	100
$\mathcal{C} = 1$	6	95	9
	7	93	95
	8	94	95
	9	96	93
	10	92	99
	Mean	94	96
Overall mean		96	98

outcomes $Y(1)$ and $Y(0)$, and $\mathcal{E}$, far from being randomized, is more common when $\mathcal{C} = 1$ than when $\mathcal{C} = 0$. However, observational data sometimes meet a weaker form of the randomization assumption, specifically that exposure is *conditionally independent* of the potential outcomes, given covariates. In our simple example, $\mathcal{C}$ is the only confounder, so that $\mathcal{E}$ is conditionally independent of $Y(1)$ and $Y(0)$, given $\mathcal{C}$. Or to put it another way: because there are no unmeasured confounders, $\mathcal{E}$ can be seen as randomly assigned within the strata defined by $\mathcal{C}$.

In our simple example, $\mathcal{E}$ is conditionally independent of the potential outcomes $Y(0)$ and $Y(1)$ given $\mathcal{C}$. The benefits of conditional independence can be seen by comparing Table 9.2 and Table 9.3, which shows the complete data stratified by $\mathcal{C}$. In particular, the conditional means of the potential outcomes in Table 9.3 *within the strata defined by $\mathcal{C}$* are equal to the observed conditional means in Table 9.2. Thus, when conditional independence holds, the conditional means can be estimated using the sample means for observations with $\mathcal{C} = c$ and $\mathcal{E} = e$.

9.1.6 Marginal and Conditional Means

The marginal means $E[Y(1)]$ and $E[Y(0)]$ in Table 9.3 can also be identified as appropriately weighted averages of the within-stratum means of $Y(1)$ and $Y(0)$ in Table 9.2, which we can estimate from the observed data under conditional independence. The weights are determined by the population prevalence of $\mathcal{C}$. To make this specific, the population prevalence of $\mathcal{C}$ in our simple example is 50%, so $E[Y(1)] = 0.5 \times 98 + 0.5 \times 94 = 96$; similarly, $E[Y(0)] = 0.5 \times 100 + 0.5 \times 96 = 98$. Thus, we can calculate the marginal means $E[Y(1)]$ and $E[Y(0)]$ from the observed data because conditional independence holds, and the prevalence of $\mathcal{C}$ is known.

Table 9.4 Regression model for estimating β_1^*

$\mathcal{E}$	$\mathcal{C}$	$E[Y \mathcal{E}, \mathcal{C}]$	Mean
0	0	β_0	100 mg/dL
1	0	$\beta_0 + \beta_1$	98 mg/dL
0	1	$\beta_0 + \beta_2$	96 mg/dL
1	1	$\beta_0 + \beta_1 + \beta_2$	94 mg/dL

9.1.7 Potential Outcomes Estimation

In our simple example with a single binary confounder $\mathcal{C}$, we were able to estimate the marginal means $E[Y(1)]$ and $E[Y(0)]$ by simple weighted averages of the conditional means within groups defined by $\mathcal{E}$ and $\mathcal{C}$. But in more complicated situations with many potential confounders, some of them continuous, there may be as many subgroups defined by the confounders, sometimes called *covariate patterns*, as there are observations.

In this situation, we could use a regression model to estimate the conditional means for each covariate pattern. Then, using the model parameter estimates, we would impute the missing potential outcome for each observation. Finally, $E[Y(1)]$ and $E[Y(0)]$ would be estimated by averages of the outcomes with and without exposure in the resulting “complete” data, including the imputed potential outcomes. These averages would implicitly be weighted by the overall sample distribution of the confounders included in the model.

Here is how potential outcomes estimation would work in our simple example. We can write a two-predictor linear model for the outcome as

$$E[Y|\mathcal{E}, \mathcal{C}] = \beta_0 + \beta_1\mathcal{E} + \beta_2\mathcal{C}. \quad (9.2)$$

This model determines mean glucose levels in each of the four groups defined by $\mathcal{E}$ and $\mathcal{C}$, as shown in Table 9.4. By modeling the effect of $\mathcal{C}$, regression achieves conditional independence for $\mathcal{E}$, so that estimates of the within-stratum means as specified by (9.2) would be unbiased for the within-group means in Table 9.3.

Then in the incomplete data shown in Table 9.2, potential outcomes estimation would work by imputing one of the four conditional means, as appropriate to the observed value of $\mathcal{E}$ and $\mathcal{C}$, for each of the ten missing potential outcomes. Specifically, the imputed values of $Y(1)$ would be 98 for persons 1–3 and 94 for persons 9 and 10. Then, $E[Y(1)]$ would be estimated by the simple average

$$\frac{(98 + 98 + 98 + 100 + 96) + (95 + 93 + 94 + 94 + 94)}{10} = 96. \quad (9.3)$$

Similarly, the imputed value of $Y(0)$ would be 100 for persons 4 and 5 and 96 for persons 6–8, and $E[Y(0)]$ would be estimated by

$$\frac{(99 + 99 + 102 + 100 + 100) + (96 + 96 + 96 + 93 + 99)}{10} = 98. \quad (9.4)$$

Finally, the causal parameter β_1^* , the average causal effect of exercise, is identified as the difference $96 - 98 = -2$. In effect, we have identified the parameters of the marginal structural model (9.1) by completing the potential outcomes data. Implementation of potential outcomes estimation based on direct regression adjustment as well as propensity scores is described in Sects. 9.3 and 9.4.2.

9.1.8 Inverse Probability Weighting

An alternative strategy for identifying the parameters of the marginal structural model (9.1) uses weighting to make the observed outcomes representative of the complete set of observed and potential outcomes. It can be shown that the weights should be inversely proportional to the probability of observed exposure, given confounders of the exposure–outcome relationship. Then, we can estimate $E[Y(1)]$ and $E[Y(0)]$ by weighted averages of the observed outcomes with and without exposure.

To illustrate how this works, note that in Table 9.2, the probability of exercise in the stratum with $\mathcal{C} = 0$ is $2/5$. Thus, the inverse probability (IP) weight for observations with $\mathcal{E} = 1$ and $\mathcal{C} = 0$ is $5/2$. Similarly, the probability of exercise in the stratum with $\mathcal{C} = 1$ is $3/5$, so the IP weight for observations with $\mathcal{E} = 1$ and $\mathcal{C} = 1$ is $5/3$. We would then estimate $E[Y(1)]$ by the weighted average

$$\frac{5/2 \times (100 + 96) + 5/3 \times (95 + 93 + 94)}{5/2 \times 2 + 5/3 \times 3} = 96. \quad (9.5)$$

For the observations with $\mathcal{E} = 0$, the probability of no exercise is $3/5$ in the stratum with $\mathcal{C} = 0$ and $2/5$ in the stratum with $\mathcal{C} = 1$. So, in this stratum the IP weights would be $5/3$ and $5/2$, respectively, and we would estimate $E[Y(0)]$ by the weighted average

$$\frac{5/3 \times (99 + 99 + 102) + 5/2 \times (93 + 99)}{5/3 \times 3 + 5/2 \times 2} = 98. \quad (9.6)$$

Calculating $\beta_1^* = 96 - 98 = -2$, we have again identified the parameters of the marginal structural model (9.1) by completing the potential outcomes data. Implementation of IP weighting in more complicated contexts with many confounders is described in Sects. 9.4.3 and 9.5.

9.2 Regression as a Basis for Causal Inference

Our examples in Sect. 9.1 greatly simplify the problem posed by confounding, in that all confounding effects are captured by a single binary factor $\mathcal{C}$, measured without error, with effects that are easily modeled. In practice, control of confounding is

difficult to achieve. In the following sections, we first consider the conditions under which regression modeling might succeed in achieving conditional independence for exposure and thus unbiased estimates of its effects. In subsequent sections, we describe alternatives that might work when those conditions are violated.

9.2.1 No Unmeasured Confounders

The assumption of no unmeasured confounders is common to most causal modeling methods, and is crucial to achieving conditional independence of exposure from potential outcomes. The main exception is instrumental variables, discussed in Sect. 9.7. The issue of unmeasured confounding is particularly critical in assessing small causal effects potentially accounted for by one or at most a few unmeasured confounders, themselves weak enough to have escaped notice. In addition, we need to assume that the confounders are measured more or less without error. Thus, carefully measuring all relevant confounders is a crucial and expensive part of observational studies.

9.2.2 Correct Model Specification

We also need to ensure that confounding effects are adequately modeled. In earlier chapters, we presented methods for capturing nonlinearities in the effects of continuous confounders and interactions, as well as for checking other model assumptions. However, those model checks can be insensitive, especially in small samples, potentially resulting in models that are at best only approximately right. Finally, we require that mediators of the effect of exposure, as well as certain so-called *colliders* defined in Sect. 10.2.5, are excluded from the model.

9.2.3 Overlap and the Positivity Assumption

In Sect. 9.1, causal effects were defined in terms of differences between actual and potential outcomes for the same individuals under different exposures or treatments. The crucial feature of that thought experiment was that each individual contributes an actual and a potential outcome, so that the distributions of individual-level covariates are identical for the exposed and unexposed outcomes.

At the opposite extreme, Rubin (1997) considers a hypothetical comparison of survival rates in 40-year-old smokers with 70-year-old nonsmokers. The lack of age overlap between smokers and nonsmokers implies that the data give essentially

no information about the effect of smoking in either age group; to do this, we would need smokers and nonsmokers in *both* age groups, because age is an important confounder of smoking, influencing both survival and smoking rates. Rubin's point is that we can only hope to estimate the causal effect of an exposure using observational data if we compare exposed and unexposed groups that are substantively comparable.

The need for overlap between the exposed and unexposed is known as the *positivity* or *experimental treatment assignment* assumption. This assumption implies that in every region of the data, there must be a positive probability of being exposed, and also a positive probability of *not* being exposed. If this assumption holds, then within all strata defined by covariates, there should be both treated and untreated observations, although this may not hold in small samples. This assumption also applies to approaches using propensity scores and inverse probability weights, discussed below.

9.2.3.1 Restriction to Address Positivity Violations

Restriction is a primary tool for causal inference. For example, suppose that the 40-year-old subsample included both smokers and nonsmokers, but there were almost no smokers in the 70-year-old subsample. In this case, we could proceed by focusing on the 40-year-olds, recognizing that the sample still provides no direct information about the effect of smoking in 70-year-olds. Moreover, if age were the *only* confounder of smoking, a simple comparison of survival rates by smoking status within the 40-year-old subsample might have a restricted causal interpretation, as the effect of smoking among 40-year-olds. This strategy also motivates estimating the *average treatment effect in the treated* (ATT) rather than ACE when the available data includes comparable controls for most treated observations, but also untreated observations in a region of poor overlap and unlike the treated group.

9.2.4 Lack of Overlap and Model Misspecification

The most common alternative to restriction is regression adjustment. If there is lack of overlap, the model essentially works by extrapolation to regions of poor overlap. The validity of those extrapolations depends on how well we deal with nonlinearity in the effects of continuous confounders, as well as interactions among confounders and with exposure. However, model misspecification is particularly hard to diagnose in regions of poor overlap, where the data are sparse.

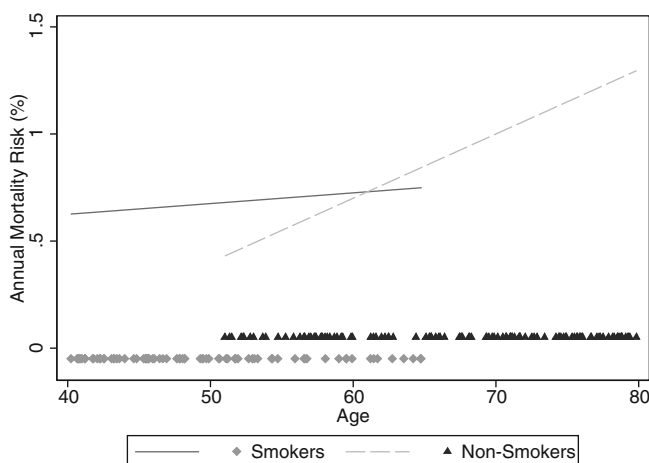


Fig. 9.1 Mortality risk by age in smokers and non-smokers

To illustrate this issue, we return to the example of the effects of smoking on mortality risk, potentially confounded by age. Suppose that the age range is 40–65 among smokers and 50–80 among non-smokers, as shown in Fig. 9.1. The diamonds and triangles along the x -axis show the age distribution of smokers and nonsmokers, respectively, while the solid and dashed lines show their mortality risk as a function of age.

Then, a logistic or Cox model adjusting for age as a continuous covariate would usually provide an age-adjusted estimate of the effect of smoking on survival. However, because age is a strong predictor of mortality risk, especially in this age range, the estimated effect of smoking would substantially depend on how we modeled the effect of age, and on whether or not we believed that smoking and age interact.

Under the assumed model, the effect of age is linear, but smoking and age interact, so that risk rises faster among nonsmokers than smokers. We could check for nonlinearity of the age effect and interaction between age and smoking, but would have little power to distinguish between them, except in large samples with high-mortality. The apparently safe course would be to allow for a nonlinear effect of the confounder age—as a result of which we would miss the effect of smoking. With less well-understood exposures, the potential for misleading conclusions can be substantial.

In multipredictor regression analyses, lack of overlap can be harder to detect. In this case, there may be substantial overlap on many or most prognostic covariates, so that the exposed and unexposed groups look fairly comparable by single measures. Nonetheless, for some individuals with anomalous combinations of covariates, there may be few if any truly comparable controls, so that for them the effects of exposure are estimated essentially by extrapolation. We show in Sect. 9.4.1.3 how propensity scores can help in detecting this kind of violation of the positivity assumption.

9.2.5 Adequate Sample Size and Number of Events

In estimating causal effects from observational data, we generally find ourselves between the extremes of the age and smoking example and the idealized case from Sect. 9.1 of a single binary covariate that captures all confounding effects and is well-represented in both exposure groups. In the usual context, more data make causal modeling easier. Although larger samples do nothing to address the problem of unmeasured confounders, adequate sample size is very important in deciding whether the observational sample can support regression modeling, and if so, how much confidence to place in the results.

In particular, larger samples, and relatively common binary or survival outcomes, make it easier to check the assumptions underlying regression adjustment, including linearity of the effects of powerful continuous confounders and the lack of interaction with exposure, as in the example of age and smoking. Furthermore, violations of normality and influential points are less likely to mislead us in larger samples.

A related question is whether the sample size or number of events is adequate to adjust for all relevant confounders. In Sect. 10.2, we argue for being inclusive when deciding which potential confounders to adjust for. Although the rule of thumb requiring ten events per variable (EPV) in logistic and Cox regression can sometimes be relaxed, regression adjustment for a large number of confounders is unquestionably more reliable and convincing with bigger samples and higher EPV. Having too few events to adjust for all relevant confounders is a principal motivation for the use of propensity scores, as we explain in Sect. 9.4.

9.2.6 Example: Phototherapy for Neonatal Jaundice

Newman et al. (2009) studied the efficacy of phototherapy (skin exposure to light) for the management of jaundice in a large cohort of newborn infants at twelve Northern California Kaiser Permanente hospitals between 1995 and 2004, and described in Table 9.5.

The infants in the original study sample, about 8% of all those born at these hospitals from 1995 to 2004, had qualifying total serum bilirubin (TSB) levels within 3 mg/dL of the American Academy of Pediatrics 2004 guideline threshold for phototherapy. Bilirubin is a product of the breakdown of heme from red blood cells, and causes jaundice at mild elevations and brain damage at very high levels. Phototherapy makes bilirubin more soluble in water and thus easier to excrete. The outcome of the study was a second TSB within 48 h that was over the higher academy threshold for so-called exchange transfusion, in which the infant's blood is replaced to reduce TSB. Among the infants studied, 5,251 (23%) received in-hospital phototherapy within 8 h of their qualifying TSB level, but only 187 (0.8%) crossed the threshold for exchange transfusion within 48 h.

Table 9.5 Characteristics of infants by receipt of phototherapy

Potential confounders of Phototherapy	Phototherapy			
	No		Yes	
	<i>N</i>	%	<i>N</i>	%
Gender				
Female	6,872	43	1,843	40
Male	9,275	57	2,741	60
Gestational Age (weeks)				
35	704	4	777	17
36	1,411	9	663	14
37	2,123	13	460	10
38	2,944	18	684	15
39	3,933	24	845	18
40	3,644	23	764	17
41	1,386	9	391	9
Qualifying TSB minus AAP threshold (mg/dL)				
−3 to less than −2	4,510	28	933	20
−2 to less than −1	4,127	26	889	19
−1 to less than 0	3,149	20	863	19
0 to less than 1	2,122	13	754	16
1 to less than 2	1,425	9	633	14
2 to less than 3	814	5	512	11
Age at qualifying TSB measurement (days)				
0	697	4	531	12
1	4,263	26	2,060	45
2	5,001	31	1,342	29
3	4,152	26	420	9
4	2,051	13	231	5

The investigators used multiple logistic regression to estimate the effect of phototherapy on this endpoint. They were convinced that they had measured most important potential confounders, although information on one potentially important co-intervention, feeding with formula, was unavailable. In addition, while the outcome rate was low, 187 outcomes were considered sufficient to model covariate effects accurately. Table 9.5 suggests good overlap between the treated and untreated samples, with at least several hundred infants in both groups in every row of the table. This was enhanced by restricting the sample to at-risk infants with starting TSB within 3 mg/dL of the guideline threshold for phototherapy.

We repeated their analysis, restricted to a subsample of 20,731 infants with negative direct anti-globulin test (DAT) results, the original analysis having shown that phototherapy was less effective in DAT-positive infants. There were 128 outcomes in the restricted sample. In unadjusted analysis, the odds of crossing the

Table 9.6 Multiple logistic regression analysis of phototherapy effect

```
. logistic over_thresh i.phototherapy male ib40.gest_age##c.birth_wt ///
>      ib4.qual_TSB ib2.age_days, cluster(hospital)
```

Logistic regression

Number of obs = 20731
Wald chi2(9) = .
Prob > chi2 = .
Pseudo R2 = 0.2849
Log pseudolikelihood = -556.91441
(Std. Err. adjusted for 11 clusters in hospital)

over_thresh	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1.photother	.1556457	.0572404	-5.06	0.000	.0757004	.320019
male	1.396058	.3245125	1.44	0.151	.8852021	2.201732
gest_age						
35	.0001092	.0004292	-2.32	0.020	4.95e-08	.2412867
36	.001609	.0057854	-1.79	0.074	1.40e-06	1.850252
37	.0031596	.0096163	-1.89	0.059	8.11e-06	1.230934
38	.0169247	.0696104	-0.99	0.321	5.34e-06	53.63804
39	.0023821	.0090549	-1.59	0.112	1.38e-06	4.097952
41	14.59515	35.12651	1.11	0.265	.130497	1632.362
birth_wt	.1056982	.111136	-2.14	0.033	.0134609	.8299667
gest_age#c.birth_wt						
35	33.33787	39.20356	2.98	0.003	3.326367	334.1224
36	14.4316	16.54287	2.33	0.020	1.526113	136.4716
37	10.33775	9.83796	2.45	0.014	1.600946	66.7537
38	4.99514	6.529896	1.23	0.219	.3853139	64.7561
39	7.404769	8.48014	1.75	0.080	.7846802	69.87637
41	.3629029	.3421455	-1.08	0.282	.0571842	2.303057
qual_TSB						
1	.049351	.0239477	-6.20	0.000	.0190654	.127745
2	.1378163	.068935	-3.96	0.000	.0517052	.3673392
3	.5232082	.173311	-1.96	0.051	.2733486	1.001457
5	3.988159	1.11587	4.94	0.000	2.304676	6.901366
6	8.252082	2.097117	8.30	0.000	5.014713	13.57941
age_days						
0	5.093211	3.017861	2.75	0.006	1.594529	16.26863
1	4.005234	1.203279	4.62	0.000	2.22282	7.216915
3	.4587072	.1313356	-2.72	0.006	.2617111	.8039869
4	.504136	.1664266	-2.07	0.038	.2639654	.9628272

threshold were 53% lower among infants receiving phototherapy (odds-ratio 0.47, 95% CI 0.24, 0.90, $P = 0.023$).

The fully adjusted model is shown in Table 9.6. In the Stata output, the categories of qualifying TSB correspond in order to the differences between qualifying TSB and the AAP threshold in Table 9.5; the reference category is 0 to less than 1. After adjusting for sex, gestational age, qualifying TSB, birth weight, and age in days at the qualifying TSB, the odds-ratio for phototherapy was 0.16 (95% CI 0.08–0.32). The fact that the adjusted estimate suggests even stronger protection shows that the unadjusted estimate is confounded by factors associated with higher risk of crossing the threshold for exchange transfusion.

In the following section, we show that the odds-ratio for phototherapy based directly on the logistic models is a *conditional effect* with an interesting but different interpretation from marginal causal effects defined in terms of the overall population means $E[Y(1)]$ and $E[Y(0)]$. We then explain the additional steps needed to estimate the marginal causal effects of phototherapy, including the marginal risk difference and odds-ratio. In addition, we briefly consider situations in which covariate-specific or conditional causal effects might be of equal or greater interest than marginal effects.

9.3 Marginal Effects and Potential Outcomes Estimation

We pointed out in Sect. 9.1 that in experiments where the randomization assumption is met, the marginal means $E[Y(1)]$ and $E[Y(0)]$ can be identified by within-group sample means. In this context, we can estimate the parameters of the marginal structural model (9.1) directly. In particular, the average causal effect β_1^* can be estimated by the difference between the within-group sample means. Similarly, when the outcome is binary, an unadjusted logistic model for the effect of treatment would estimate the marginal odds-ratio, as we explain below.

Thus, the familiar summary effect measures commonly used for experiments, which are regarded as the gold standard in clinical research, estimate marginal causal effects. Moreover, causal questions are often framed in terms of clinical trials that might answer them. In this view, the relevant causal parameter of interest is a marginal effect, averaged over a well-defined target population meeting the inclusion criteria for the implicit clinical trial.

The focus of this chapter is on estimating causal effects using observational data, in which the randomization assumption almost never holds. In that context, we may at best meet the weaker assumption of conditional independence. When we fit fully adjusted logistic models like those used by Newman et al. (2009) to estimate the effect of phototherapy, we obtain estimates of the conditional, not the marginal odds-ratio. In this section, we more carefully distinguish marginal from conditional effects, and present methods for using the conditional results to obtain the marginal causal effects that would be estimated by a clinical trial of phototherapy.

9.3.1 Marginal and Conditional Effects

In Sect. 9.1, we defined the average causal effect as a difference in the marginal means of potential outcomes, including the potential as well as actual outcomes. In the linear model (9.2) for continuous potential outcomes, the effect is directly captured by the regression coefficient β_1 . This effect is both *marginal*, because it is

the difference in the marginal means $E[Y(1)]$ and $E[Y(0)]$, and *conditional*, in also capturing the difference in conditional means within the subpopulations with $C = 0$ and $C = 1$.

With binary outcomes, the marginal means $E[Y(1)]$ and $E[Y(0)]$ are interpretable as outcome probabilities, and the average causal effect can still be defined as $E[Y(1)] - E[Y(0)] = E[Y(1) - Y(0)]$. However, the odds-ratio, not the difference in outcome probabilities, is the natural effect measure for the logistic model, which would most commonly be used to assess the effects of exposure on a binary outcome. For this case, we could define a logistic marginal structural model for the potential outcomes as

$$\log \left[\frac{E[Y(\mathcal{E})]}{1 - E[Y(\mathcal{E})]} \right] = \beta_0^* + \beta_1^* \mathcal{E}. \quad (9.7)$$

In this case, the marginal odds-ratio is directly defined in terms of the marginal means $E[Y(1)]$ and $E[Y(0)]$ —specifically, by

$$\frac{E[Y(1)]}{1 - E[Y(1)]} \times \frac{1 - E[Y(0)]}{E[Y(0)]}. \quad (9.8)$$

When the randomization assumption holds, as in a successfully conducted randomized trial, we could fit an unadjusted logistic model for the effect of exposure, and would obtain a direct estimate of the marginal odds-ratio (9.8) by exponentiating $\hat{\beta}_1^*$. Estimates of the marginal risk difference would also be easily obtained as the difference between the fitted outcome probabilities for the exposed and unexposed groups.

However, in observational data, as in our simple example, we could at best meet the assumption of conditional independence of $\mathcal{E}$, after adjustment for C . We would write the adjusted logistic model as

$$\log \left[\frac{E[Y|\mathcal{E}, C]}{1 - E[Y|\mathcal{E}, C]} \right] = \beta_0 + \beta_1 \mathcal{E} + \beta_2 C, \quad (9.9)$$

where $E[Y|\mathcal{E}, C]$ is the probability that $Y = 1$, given $\mathcal{E}$ and C . Under this model, $\exp(\beta_1)$, the odds-ratio for the effect of exposure $\mathcal{E}$ on Y , represents a *conditional effect*, assumed to be the same within both strata defined by C . This conditional odds-ratio would differ from the marginal odds-ratio (9.8) except when $\beta_1 = 0$ or $\beta_2 = 0$. In practice, these differences are often small, but the conceptual difference is important. Likewise, under (9.9), the conditional risk difference for any given observation depends on C , unless $\beta_1 = 0$ or $\beta_2 = 0$; this would hold even if C were unassociated with $\mathcal{E}$. Specifically, if $C = 1$, the conditional risk difference is

$$\frac{\exp(\beta_0 + \beta_1 + \beta_2)}{1 + \exp(\beta_0 + \beta_1 + \beta_2)} - \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}. \quad (9.10)$$

When $\mathcal{C} = 0$, the risk difference is

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} - \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}. \quad (9.11)$$

Thus, when we use an adjusted logistic model to meet the conditional independence assumption for $\mathcal{E}$, extra steps are needed to obtain estimates of the marginal risk difference $E[Y(1)] - E[Y(0)]$ and odds-ratio (9.8).

9.3.2 *Contrasting Conditional and Marginal Effects*

In the neonatal jaundice example, conditional effects would be more to the point when a clinician considers the potential effects of phototherapy *for a particular infant*. Newman et al. (2009) estimated that the absolute reduction in risk of crossing the threshold for exchange transfusion varied more than 200-fold among the infants treatment with phototherapy. In this context, good estimates of conditional risk reductions are especially useful for evidence-based clinical decision making. Note that if confounding is controlled, conditional independence implies that these conditional effects have a causal interpretation.

In contrast, marginal risk reductions, averaged across the target population of newborns with qualifying TSB near the current threshold, would be useful in assessing phototherapy treatment guidelines for exchange transfusion in the Kaiser system overall. In this context, some variability in individual effects may be taken as a given. More generally, marginal effect estimates are appropriate when we consider the effects of public health interventions or changes in policy.

Conditional estimates might still have a role in evaluating interventions or policy. In the phototherapy data, for example, Newman et al. (2009) interpreted the wide variability in the conditional risk differences as suggesting that the current guidelines allow for treatment of low-risk infants with too little expected benefit from phototherapy.

9.3.3 *When Marginal and Conditional Odds-Ratios Differ*

In the phototherapy example, the marginal and conditional odds-ratios will prove to be similar. However, this will not always hold. In particular, the difference will be larger when covariate effects are stronger. For an extreme example, consider hypothetical data in which $\mathcal{E}$ and $\mathcal{C}$ are uncorrelated, but the prevalence of the outcome Y is only 10% in the stratum with $\mathcal{C} = 0$, and 90% in the stratum with $\mathcal{C} = 1$. The conditional odds-ratio for $\mathcal{E}$ is more than 2.5 within both strata defined by $\mathcal{C}$, but the marginal odds-ratio is only 1.4.

Although the marginal and conditional odds-ratios are very similar in the phototherapy data, one of the principal findings of Newman et al. (2009) was that conditional risk differences varied widely among infants meeting guidelines for phototherapy. This commonly occurs in logistic models where covariates strongly affect the odds of the outcome, even when the odds-ratio for exposure is assumed constant—that is, not to interact with covariates.

9.3.4 Potential Outcomes Estimation

In Sect. 9.1.7, we showed how potential outcomes estimation could be used to estimate the marginal means of a continuous outcome in our simple example with a single binary confounder. Here, we extend this procedure to more complicated contexts with a binary outcome and many confounders, some of them continuous, with each observation potentially having a distinct covariate pattern.

To implement this procedure, we would fit a logistic model carefully adjusting for all measured confounders, then obtain two fitted probabilities for each observation: first with exposure, setting $\mathcal{E} = 1$, and then without exposure, setting $\mathcal{E} = 0$. Only one of these two values of $\mathcal{E}$ is observed; the other is potential. In both calculations, the covariate pattern for each observation would be held fixed, at the observed level. Then, assuming that the overall sample proportion with each covariate pattern is representative of the population, we can estimate $E[Y(1)]$ by the average of the estimated probabilities calculated after setting $\mathcal{E} = 1$. Crucially, this average would be taken over the entire sample, not just the observations with $\mathcal{E} = 1$. Likewise, we can estimate $E[Y(0)]$ by the average of the estimated probabilities calculated after setting $\mathcal{E} = 0$, again taken over the entire sample. In turn, we can use these two estimates to calculate the marginal risk difference or odds ratio.

Potential outcomes estimation can be implemented using a simple algorithm, which we applied to the phototherapy data in Table 9.7. In brief, we first used the Stata `expand` command to make a duplicate of each observation, then reversed the coding of `phototherapy` on the duplicate data records, so that the duplicates of the treated are coded as untreated and vice versa. In fitting the regression model, we restricted the estimation sample to the actual observations (i.e., if `potential==0`).

We then took advantage of the fact the `predict` postestimation command calculates predicted values for every observation with complete predictor data, regardless of whether they were used in estimation of the coefficients. Next, we obtained estimates $\hat{E}[Y(0)] = .00956$ and $\hat{E}[Y(1)] = .00164$ by averaging the predicted values for the treated and untreated observations, including the observations introduced by the duplication. That step ensured that the distribution of covariates was the same for both sets of predicted outcomes.

Then in a final step, we can calculate the marginal risk difference as $0.00956 - -0.00164 = 0.0079$. This amounts to fitting the marginal structural model (9.1) to

Table 9.7 Potential outcomes estimation

```

. * Duplicate each observation, identifying the second as potential
. expand 2, gen(potential)
(20731 observations created)

. * Assign the opposite exposure for the potential outcome
. replace phototherapy = 1-phototherapy if potential==1
(20731 real changes made)

. * Estimate the logistic model using only the actual outcomes
. quietly logistic over_thresh i.phototherapy male i.gest_age#c.birth_wt///
> i.qual_TSB i.age_days if potential==0, cluster(hospital)

. * Obtain expected values for both actual and potential outcomes
. predict Y, pr

. * calculate EY by treatment
. tab phototherapy, sum(Y)

```

Phototherap	Summary of Pr(over_thresh)		
y	Mean	Std. Dev.	Freq.
no	.00955488	.02960949	20731
yes	.00164365	.005798	20731
Total	.00559927	.02169805	41462

the complete data, with $\hat{\beta}_0^* = 0.00956$ and $\hat{\beta}_1^* = 0.0079$. We can also calculate the marginal odds-ratio as $0.00164/(1 - 0.00164)/(0.00956/(1 - 0.00956)) = 0.17$. As we would expect based on Sect. 7.5, the marginal odds-ratio of 0.17 for phototherapy is slightly closer to the null value of 1.00 than the conditional odds-ratio of 0.16 given directly in the model output shown in Table 9.7.

The Stata `margins` command implements potential outcomes estimation, and provides valid CIs for the parameters of the marginal structural model (9.1). Like the potential outcomes estimation procedure implemented by hand in Table 9.7, the `margins` command averages the expected values of the outcome under both the actual and potential value of phototherapy, holding all other covariates fixed at their observed values. (Note that for this Stata procedure to give the correct marginal result, phototherapy must have been treated as a so-called *factor* in the regression model, using the `i.phototherapy` syntax, not as a continuous variable.)

Table 9.8 shows the results of a re-analysis of the logistic model for the effect of phototherapy first shown in Table 9.6. The resulting estimates of $E[Y(1)]$ and $E[Y(0)]$, and accordingly of the marginal risk difference and odds-ratio, are identical to those in Table 9.7. This also provides valid CIs for the marginal means, although the tests of $E[Y(1)] = 0$ and $E[Y(0)] = 0$ are hard to interpret.

Table 9.9 shows direct calculation of the marginal risk difference, first using the `postestimation` command `margins, dydx(phototherapy)`, then using the `r.` contrast operator, which gives the same result. This procedure provides a valid CI and *P*-value for the marginal risk difference.

Table 9.8 Direct estimation of marginal means

```
. quietly logistic over_thresh i.phototherapy male ///
>   ib40.gest_age##c.birth_wt ib4.qual_TSB ib2.age_days, ///
>   cluster(hospital)

. margins phototherapy
Predictive margins                                Number of obs   =      20731
Model VCE      : Robust
Expression     : Pr(over_thresh), predict()
```

		Delta-method				
		Margin	Std. Err.	z	P> z	[95% Conf. Interval]
phototherapy	0	.0095549	.0009868	9.68	0.000	.0076208 .011489
	1	.0016437	.0006048	2.72	0.007	.0004582 .0028291

Table 9.9 Direct estimation of marginal risk difference

```
. margins, dydx(phototherapy)
Average marginal effects                                Number of obs   =      20731
Model VCE      : Robust
Expression     : Pr(over_thresh), predict()
dy/dx w.r.t.   : 1.phototherapy
```

		Delta-method				
		dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]
1.photother~y		-.0079112	.0010376	-7.62	0.000	-.0099448 -.0058777

```
. margins r.phototherapy
Contrasts of predictive margins
Model VCE      : Robust
Expression     : Pr(over_thresh), predict()
```

		df	chi2	P>chi2
phototherapy		1	58.14	0.0000

		Delta-method		
		Contrast	Std. Err.	[95% Conf. Interval]
phototherapy	(1 vs 0)	-.0079112	.0010376	-.0099448 -.0058777

Confidence intervals for the marginal odds-ratio can be obtained using the bootstrap, as shown in Table 9.10. This requires a short program to calculate the marginal odds-ratio from the margins results. Note that for this example, the bootstrap re-sampling was by hospital, to account for clustering, as in the other analyses. The bias-corrected percentile CI (0.09–0.36) is slightly wider than the CI for the conditional odds-ratio shown in Table 9.7, and shifted upward, reflecting the slight attenuation of the marginal odds-ratio.

Table 9.10 Bootstrap confidence interval for the marginal odds-ratio

```
. program define marginal_OR, rclass
1. logistic over_thresh i.phototherapy male i.gest_age##c.birth_wt ///
   i.qual_TSB i.age_days
2. margins phototherapy
3. matrix b = r(b)
4. scalar EY0 = b[1, 1]
5. scalar EY1 = b[1, 2]
6. * marginal odds-ratio
. return scalar marginal_OR = EY1/(1-EY1)*(1-EY0)/EY0
7. end

. bootstrap "marginal_OR" r(marginal_OR), reps(1000) cluster(hospital)
command:      marginal_OR
statistic:    _bs_1      = r(marginal_OR)
```

Bootstrap statistics

Number of obs = 20731

N of clusters = 11

Replications = 1000

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]	
_bs_1	1000	.1705817	.0108846	.0679137	.0373119	.3038515 (N)
					.0870933	.3547933 (P)
					.0889122	.3603035 (BC)

Note: N = normal
P = percentile
BC = bias-corrected

9.3.5 Marginal Effects in Longitudinal Data

So far we have focused on continuous and binary outcomes. Potential outcomes estimation of the marginal means $E[Y(1)]$ and $E[Y(0)]$ carries over directly to count outcomes that would be analyzed using Poisson or negative binomial models; in Stata, the margins command can be used to obtain both marginal means and rates. In contrast, extensions to repeated measures and survival outcomes are more complicated.

9.3.5.1 Repeated Measures Outcomes

For repeated measures in a longitudinal study with regular measurement times, we can posit analogous potential outcomes *at each measurement time*. Then, the average causal effect of exposure can be defined in terms of the marginal means specific to each time point. Marginal causal effects might vary over time point; averaging across occasions might be appropriate as long as the variation is not too great.

Potential outcomes estimation can sometimes be used to estimate marginal effects in this setting. However, this straightforward approach cannot be used when

both exposure and its confounders change over time, *and the confounders mediate part of the effect of exposure*. In that setting, with what we will call *time-dependent confounder–mediators*, IP weighting is one alternative for estimating marginal effects, as we explain in Sect. 9.5.

9.3.5.2 Survival Outcomes

For survival outcomes, we can define the potential outcomes $Y(1)$ and $Y(0)$ as failure times with and without exposure, and write marginal structural models for the potential outcomes analogous to (9.1) and (9.7). One strategy for estimating marginal effects in this setting uses so-called *structural nested failure time models*; we briefly describe one such method, *G-estimation*, in Sect. 9.10.

An alternative for estimating marginal effects with survival outcomes uses IP weighting, and is based on a proportional hazards marginal structural model similar in form to (6.5). A primary motivation for this approach, described in Sect. 9.5, is that it accommodates time-dependent confounder–mediators. But IP weighting has drawbacks and difficulties, as we also explain, and more reliable methods are the focus of ongoing statistical research.

9.3.5.3 Potential Outcomes Estimation for Cumulative Risks

With fixed exposures, and more generally *in the absence of time-dependent confounder–mediators*, potential outcomes estimation can be used to estimate marginal effects on the cumulative risk of the outcome at some fixed time point, estimated using survival data. In cancer studies, for example, treatment effects are often described in terms of differences in 5-year survival; in heart disease, 10-year risk of cardiovascular events is a common benchmark. These cumulative risks can be estimated using censored survival data.

Potential outcomes estimation can be implemented by fitting an adjusted Cox model for the effects of exposure or treatment, controlling for confounders, analogous to the adjusted logistic model used in the phototherapy example. Then predicted cumulative risks at the selected time point would be obtained for each observation under the alternative exposure or treatment histories of interest. This is analogous to predicting the cross-sectional risk of crossing the threshold for exchange transfusion for each infant with and without phototherapy.

One complication is these cumulative risk predictions depend on the base-line survival function. While estimates are available from most Cox model implementations, including `stcox` in Stata, implementation requires data duplication, as shown in Table 9.6, with additional programming to obtain the baseline survival function estimate at the selected time point. We sketch an implementation in Problem 9.5.

9.4 Propensity Scores

As illustrated by the phototherapy example presented in Sects. 9.2 and 9.3.4, regression methods can be used, in many cases, to estimate causal effects for binary exposures in observational studies. The outcome in this example was fairly rare, with only 128 cases in more than 20,000 observations, but common enough for regression adjustment. But if the sample size had been 5,000, with only 32 outcomes, this approach would have led to unstable or biased results.

Propensity score methods address this problem by splitting the analysis into two steps. First, the relationship of confounders with exposure is summarized using a regression model with exposure as the outcome; any of the binary regression models introduced in Chap. 5 can be used. The goal of this model is to estimate the influence of the confounding variables on the probability of exposure for each individual. The exposure probability predicted by this model *is* the propensity score.

It can be shown that individuals with similar propensity scores will have similar patterns of the confounding variables. This suggests that an estimate of the effect of the exposure on the outcome that accounts for values of the propensity score will also account for the influence of the confounders. This is the basis for the second step of propensity score analysis, in which we estimate the effect of exposure on the outcome. There are several ways to use the propensity scores in the second step, all of which resolve problems with controlling for multiple predictors. As long as exposure is common, this has clear advantages when outcomes are rare or the number of potential confounders is large.

Depending on how propensity scores are incorporated in the second step of the analysis, we obtain estimates of the conditional or marginal effect of exposure. In particular, when we stratify on or adjust for the scores, we obtain conditional effect estimates, and have to use potential outcomes estimation to obtain marginal effect estimates. In contrast, when the scores are used as inverse weights or for matching, we obtain direct estimates of marginal effects.

In the remainder of this section, we describe analysis using propensity scores more fully, and illustrate the approach using the phototherapy data set introduced in Sect. 9.2.6. Although the phototherapy outcome is binary, the methods illustrated apply directly to continuous, survival, and count outcomes.

9.4.1 *Estimation of Propensity Scores*

Model selection and specification, fitting the model, and then checking balance and overlap are all part of estimating propensity scores.

9.4.1.1 Model Specification

A crucial assumption of analysis using propensity scores is that the model for the scores is correctly specified. Accordingly, care should be taken to control for the confounders of the exposure–outcome relationship, to include interaction terms as needed, and to model nonlinearities adequately. In moderate to large samples, any potential confounder of the effect of exposure should be considered. For continuous and count outcomes, it may also be valuable to include covariates associated with the outcome but not exposure (Brookhart et al. 2006a); the rationale is to decrease residual variance.

However, in smaller samples or if exposure is uncommon, including too many predictors may actually exacerbate lack of overlap (Kang and Schafer 2007), and require selecting a smaller propensity score model. Furthermore, the model used to estimate the propensity scores should not include so-called *instrumental variables* associated with exposure but lacking any independent association with the outcome (Austin et al. 2007; Brookhart et al. 2006a). Finally, as in standard regression adjustment, mediators of the effect of exposure, as well as so-called *colliders* defined in Sect. 10.2.5, should be excluded from the propensity score model.

9.4.1.2 Propensity Score Model for Phototherapy

In the Kaiser sample of 20,731 newborns, only 128 infants crossed the threshold for exchange transfusion, limiting the complexity of the logistic model used to estimate the effect of phototherapy directly adjusting for confounders in Sect. 9.2.6. In contrast, 4,584 newborns were treated with phototherapy, allowing us to develop a relatively complicated propensity score model, as recommended by Schneeweiss et al. (2009). Our final propensity score model used the same covariates included in the model for crossing the exchange therapy threshold, but modeled the effect of birth weight using a 5-knot restricted cubic spline, and included almost all possible two-way interactions; both the nonlinearity of the birth weight effect and the interactions were highly statistically significant. However, we excluded hospital and year, which we will use as instrumental variables in Sect. 9.7. The Hosmer–Lemeshow test indicated satisfactory fit for the final model ($P = 0.33$).

9.4.1.3 Checking Covariate Balance

A key property of good propensity scores is that the distribution of measured confounding variables within strata defined by the scores is, on average, balanced between the two exposure groups (Rosenbaum and Rubin 1983).

Table 9.11 shows that average values of the major confounders of phototherapy differ much less between exposed and unexposed infants *within* quintiles of the propensity score than overall, illustrating the balancing property of the scores.

Table 9.11 Checking covariate balance

Predictor	Phototherapy	Overall mean	Propensity score quintile				
			1	2	3	4	5
Male sex	No	0.57	0.55	0.53	0.59	0.62	0.62
	Yes	0.60	0.51	0.52	0.58	0.64	0.60
Gestational	No	38.7	38.7	38.9	38.7	38.4	37.2
Age (weeks)	Yes	37.9	38.5	39.0	38.8	38.3	37.0
Birth	No	3.35	3.38	3.39	3.41	3.40	3.08
Weight (kg)	Yes	3.22	3.38	3.41	3.44	3.38	3.00
Qualifying TSB	No	2.64	2.02	2.34	2.62	3.31	3.48
(Category #)	Yes	3.17	2.25	2.33	2.52	3.35	3.55
Age (days) at	No	2.16	3.31	2.33	1.74	1.53	1.22
Qualifying TSB	Yes	1.51	3.36	2.31	1.68	1.57	1.12

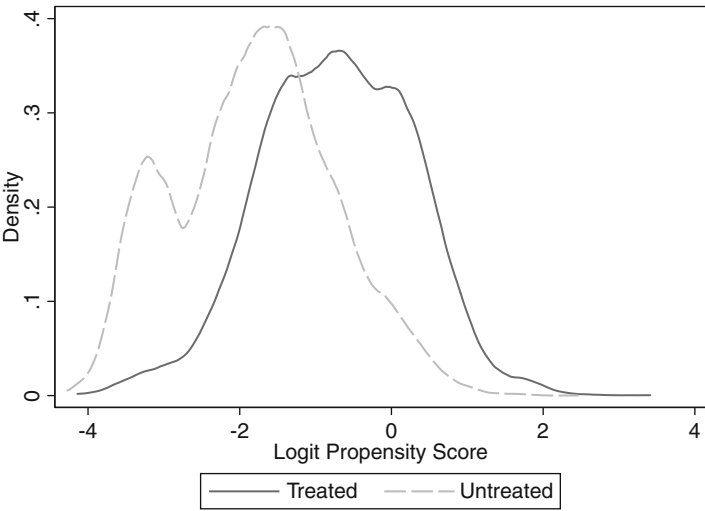


Fig. 9.2 Propensity scores in treated and untreated infants

9.4.1.4 Checking the Positivity Assumption

Like regression adjustment, propensity score analyses depend on the positivity assumption, introduced in Sect. 9.2.3. Fortunately, they also make it easier to diagnose positivity violations. Figure 9.2 shows the distribution of propensity scores (on the log odds or logit scale) for the treated and untreated samples. In contrast to the reassuring evidence for overlap in the rows of Table 9.5, the figure shows that untreated infants with logit scores < -3 had very few treated counterparts. Similarly, treated infants with logit propensity scores > 1 had almost no untreated counterparts. In the next section, we present methods for addressing this potential problem.

Table 9.12 Numbers of infants and events

Phototherapy	Overall	Propensity score quintile				
		1	2	3	4	5
No	113/16,147	2/3,999	8/3,715	19/3,386	28/3,010	56/2,037
Yes	15/4,584	0/150	1/432	1/757	2/1,137	11/2,108

9.4.2 Effect Estimation Using Propensity Scores

The next step in the analysis is to use propensity scores to estimate the causal effect of exposure. The scores may be incorporated using stratification, adjustment, inverse weighting, and matching, each with advantages and disadvantages. Stratification and adjustment require us to use potential outcomes estimation to obtain marginal effects. In contrast, inverse weighting and matching directly estimate marginal effects.

9.4.2.1 Quintile of Propensity Score

Analysis using quintile of the propensity score is often a good place to start. One advantage is that we can use contingency tables to look at the data. Table 9.12 gives little reason for concern, although there are no events among treated infants in the first quintile.

Next, we used quintile of the propensity scores as the only adjustment variable in a logistic model so that we could account for clustering by hospital. In a final step, we obtained marginal risk difference using potential outcomes estimation. In Table 9.13, the conditional odds-ratio (0.20, 95% CI 0.10, 0.42) and marginal risk difference (0.71%, 95% CI 0.50–0.92%) suggest slightly less protection than the standard logistic regression model. In this case, the conditional and marginal odds-ratios barely differ. The marginal risk difference could also be obtained using the command `margins r.phototherapy` used in Table 9.9.

9.4.2.2 Restricted Cubic Splines

Modeling the propensity score as a categorical variable may result in residual confounding. To address this possible shortcoming, we repeated this analysis adjusting for a 5-knot restricted cubic spline in the logit propensity score. In this analysis, we rescaled the logit scores before calculating the splines so that the corresponding parameter estimates would appear reasonable, but this makes no difference to the conditional or marginal estimates we obtain for the effect of phototherapy. Again, after estimating the conditional odds-ratio, we use potential outcomes estimation to obtain the marginal risk difference.

Table 9.13 Analysis using propensity score quintiles

```
. logistic over_thresh i.phototherapy i.ps_quintile, cluster(hospital)
Logistic regression
Log pseudolikelihood = -706.10698
(Std. Err. adjusted for 11 clusters in hospital)
```

over_thresh	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1.photothe~y	.2015758	.0751063	-4.30	0.000	.0971146	.4184008
ps_quintile						
2	4.777587	3.700018	2.02	0.043	1.047111	21.79839
3	11.44334	8.659595	3.22	0.001	2.596672	50.42992
4	18.81359	14.75728	3.74	0.000	4.043839	87.52853
5	56.11242	44.62975	5.06	0.000	11.80442	266.7309

```
. * Marginal risk difference
. margins, dydx(phototherapy)
Average marginal effects
Model VCE      : Robust
Expression     : Pr(over_thresh), predict()
dy/dx w.r.t.   : 1.phototherapy
```

	Delta-method					
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
1.photothe~y	-.0071373	.0010718	-6.66	0.000	-.009238	-.0050365

Results shown in Table 9.14 are consistent with the analysis using quintiles, including conditional and marginal odds-ratios of 0.20 and risk difference of 0.73% (95% CI 0.52–0.94%). There was also clear evidence for a nonlinear effect of the propensity score, showing the need for using a spline. The marginal risk difference could also be obtained using the command `margins r.phototherapy`.

9.4.3 Inverse Probability Weights

In the analyses using propensity scores as quintiles and splines, we obtain estimates of the conditional effect of phototherapy, and then use potential outcomes estimation to obtain marginal risk differences and odds-ratios. Another way to obtain marginal estimates, introduced in Sect. 9.1.8, uses the propensity scores to define so-called *inverse probability (IP) weights*—literally, the inverse of the estimated probabilities of observed exposure, conditional on confounders. Using $\Pr(\mathcal{E}|\mathcal{C})$ to denote the propensity score, IP weights are defined as $1/\Pr(\mathcal{E}|\mathcal{C})$ for the exposed, and as $1/(1 - \Pr(\mathcal{E}|\mathcal{C}))$ for the unexposed.

Using IP weights creates comparable weighted samples of exposed and unexposed observations, sometimes called *pseudo populations*, both with the same

Table 9.14 Analysis using restricted cubic splines

```

. gen lps100 = logit_ps*100
. mkspline lps_rcs = lps100, cubic
. logistic over_thresh 1.phototherapy lps_rcs*, cluster(hospital)
Logistic regression

```

	Number of obs	=	20731
	Wald chi2(5)	=	63.07
	Prob > chi2	=	0.0000
Log pseudolikelihood = -707.00778	Pseudo R2	=	0.0922
(Std. Err. adjusted for 11 clusters in hospital)			

over_thresh	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1.photother~y	.1934407	.0706752	-4.50	0.000	.0945267	.3958598
lps_rcs1	1.017154	.010613	1.63	0.103	.9965642	1.038169
lps_rcs2	.9824837	.0465576	-0.37	0.709	.8953419	1.078107
lps_rcs3	1.102977	.2458561	0.44	0.660	.7125771	1.707266
lps_rcs4	.8289924	.2502172	-0.62	0.534	.4588069	1.49786

```

. * check non-linearity of response to propensity score
. testparm lps_rcs2-lps_rcs4
    Prob > chi2 =    0.0008

. * Marginal risk difference
. margins, dydx(phototherapy)
Average marginal effects
Model VCE      : Robust
Expression    : Pr(over_thresh), predict()
dy/dx w.r.t.  : 1.phototherapy

```

	Delta-method dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
1.photother~y	-.0073261	.0010548	-6.95	0.000	-.0093936	-.0052587

distribution of estimated propensity for exposure as the overall sample. Ideally, exposure is unconfounded in the overall weighted sample—assuming no unmeasured confounders, correct specification of the model used to estimate the propensity scores, and positivity.

Table 9.15 shows the analysis using IP weights. As we explained in Sect. 9.1.8, using IP weights means that we directly obtain estimates of marginal causal effects from the weighted model for the outcome, including the marginal odds-ratio when the model for the outcome is logistic. Like procedures based on potential outcomes estimation in Sect. 9.3.4, fitting the weighted model can be seen as fitting the marginal structural model (9.1) to the complete potential outcomes data. The data are “completed” by inverse weighting in this case, rather than by imputation of the missing potential outcomes.

An advantage of IP weighting is that it easily accommodates survival outcomes. If time to crossing the threshold for exchange transfusion were the outcome in the phototherapy data, we could use an IP-weighted Cox model to obtain a direct estimate of the marginal hazard ratio for the effect of phototherapy. In contrast, calculation of the marginal effects on cumulative risk using the potential outcomes approach would be complicated.

Table 9.15 Analysis using propensity scores as IP weights

```
. gen iptw = phototherapy/prop_score + (1-phototherapy)/(1-prop_score)
. logistic over_thresh i.phototherapy [pweight=iptw], cluster(hospital)
Logistic regression
Log pseudolikelihood = -1403.0863
(Std. Err. adjusted for 11 clusters in hospital)
```

over_thresh	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1.photother~y	.2220519	.084231	-3.97	0.000	.1055767	.4670259

```
. * Marginal risk difference
. margins, dydx(phototherapy)
Conditional marginal effects
Model VCE      : Robust
Expression     : Pr(over_thresh), predict()
dy/dx w.r.t.   : 1.phototherapy
```

	dy/dx	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
1.photother~y	-.0072356	.0013033	-5.55	0.000	-.00979	-.0046811

A drawback of IP weighting is that extreme weights are fairly common, possibly reflecting violations of the positivity assumption, and lead to highly unstable estimates. An initial check on the weights showed that there were 91 observations with weights of more than 20, all of them in the phototherapy group, reflecting less than 5% estimated probability of treatment received; the largest weight was 64. In part as a result of the large weights, this analysis gave a somewhat different and less precise estimate of the marginal odds-ratio (0.22, 95% CI 0.11–0.47), although the marginal risk difference (0.72%, 95% CI 0.47–0.98%) was similar to earlier results based on the propensity score.

9.4.4 Checking for Propensity Score/Exposure Interaction

An advantage of propensity scores is that it is easy to check for interaction between exposure and the propensity for exposure, which may be easier to detect than interactions between exposure and covariates, and thus uncover meaningful variability in the effects of exposure.

Table 9.16 presents an assessment of the interaction, including estimates of the odds-ratio for phototherapy within each propensity score quintile. This analysis gave reassuring results ($P = 0.54$ for interaction); although the point estimate in the second quintile did not suggest benefit, the CI was very wide. The Mantel–Haenszel (M–H) weights make explicit the influence of the fifth quintile in the overall estimate.

Table 9.16 Checking for propensity score/exposure interaction

. cc over_thresh phototherapy, by(ps_quintile)

Propensity score	OR	[95% Conf. Interval]		M-H Weight	
1	0	0	51.4532	.0723066	(exact)
2	1.075116	.0241739	8.051342	.8314444	(exact)
3	.2344055	.0056341	1.47967	3.467053	(exact)
4	.1876652	.0216347	.7464934	7.663371	(exact)
5	.1855627	.0874299	.3593499	28.331	(exact)
Crude	.4658365	.2521401	.8023481		(exact)
M-H combined	.2081478	.1197724	.3617317		
Test of homogeneity (Tarone) chi2(4) = 3.13 Pr>chi2 = 0.5356					

In contrast to our relatively reassuring results, [Kurth et al. \(2006\)](#) found important interaction between propensity for treatment with tissue-plasminogen activator (t-PA), which dissolves blood clots, and mortality among 6,269 patients with ischemic strokes caused by blood clots. In contrast to randomized trials showing no benefit, they found evidence for substantial adverse effects, with harm concentrated among patients with propensity scores of less than 5%. As in our analysis, they estimated the effect of t-PA using logistic models incorporating the propensity score both as continuous and categorical (using deciles rather than quintiles). But it was only analyses using the methods we present next—restriction, matching, or using so-called *standardized mortality ratio* (SMR) weights—that results were consistent with trial findings. These alternative methods estimate the effects of exposure in restricted target populations of possibly greater interest.

9.4.5 Addressing Positivity Violations Using Restriction

Our check on overlap of the propensity scores in Sect. 9.4.1 gave some evidence for positivity violations. One strategy for addressing such violations is to restrict the analysis to observations with predicted probabilities of exposure between, say, 5% and 95% ([Mortimer et al. 2005](#)). This will exclude individuals who are almost always or almost never exposed; in studies of treatments, this sensibly focuses the analysis on patients for whom consensus about the value of treatment is lacking. We re-analyzed the phototherapy data, including only infants with logit propensity scores between -3 and 1 , corresponding to propensity scores between 4.7% and 73%, as motivated by the regions of poor overlap in Fig. 9.2. This gave reasonably similar estimates of the conditional odds-ratio (0.21 95% CI 0.10–0.44), marginal odds-ratio (0.21) and marginal risk difference (0.79%, 95% CI 0.54–1.04%), suggesting that positivity violations do not substantially affect our estimates of the effect of phototherapy.

In contrast, restriction to patients with propensity scores of at least 5% in the analysis of the effects of t-PA among ischemic stroke patients gave results very different from the analysis of the complete data, but consistent with randomized trials ([Kurth et al. 2006](#)).

9.4.6 Average Treatment Effect in the Treated (ATT)

In some cases, it may make more sense to estimate the causal effect of treatment *in the treated*, or ATT, defined as the average causal effect in a population with the same distribution of propensities for exposure as the exposed individuals in the sample. One example is the effect of smoking cessation, which only makes sense for smokers. Of course, estimating the ATT for cessation would require comparable nonsmoking controls, but would exclude nonsmokers who never would have smoked and differ from smokers on many dimensions. In the ischemic stroke example, this focuses the analysis on the relatively small group of low-risk patients who are more commonly treated with t-PA, excluding the much larger group of high-risk patients in whom t-PA is rarely used.

In contrast to estimating ACE, in which we average the exposure effects across the distribution of covariates in the entire population, in estimating ATT we average the exposure effect across the distribution of covariates *among the exposed*. A secondary effect of focusing on the exposed is that it will address positivity violations stemming from unexposed individuals with few, if any, counterparts in the exposed sample. Propensity scores make it possible to estimate ATT in three ways, using potential outcomes estimation restricted to the exposed subpopulation, matching, and standardized mortality ratio weights.

9.4.6.1 Potential Outcomes Estimation

To estimate ATT using potential outcomes estimation, we used the model adjusting for the propensity score as a restricted cubic spline. Then we used the `margins` command with option `subpop(phototherapy)` to estimate ATT. This could also be done using the command `margins r.phototherapy, subpop(phototherapy)`.

Results are shown in Table 9.17. The ATT risk difference is 1.3%, almost twice as large as the ACE estimate of 0.73% given by the propensity score analysis using restricted cubic splines. This suggests that pediatricians are more likely to use phototherapy among higher risk infants with greater expected benefit.

Table 9.17 ATT using potential outcomes estimation

```
. qui logistic over_thresh i.phototherapy lps_rcs*, cluster(hospital)

. * Marginal risk difference
. margins, dydx(phototherapy) subpop(phototherapy)
Average marginal effects                                Number of obs =      20731
                                                         Subpop. no. obs =      4584

Model VCE      : Robust
Expression     : Pr(over_thresh), predict()
dy/dx w.r.t.   : 1.phototherapy
```

		Delta-method				
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
1.photother~y	-.0133132	.0020652	-6.45	0.000	-.0173609	-.0092655

Table 9.18 Matching to estimate ATT

```
. psmatch2 phototherapy, out(over_thresh) pscore(prop_score) noreplace
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
over_thresh	Unmatched	.003272251	.006998204	-.003725953	.001310775	-2.84
	ATT	.003272251	.016143106	-.012870855	.002043817	-6.30

9.4.6.2 Matching

A second way to estimate ATT is to match unexposed to exposed observations on values of the propensity score. Only exposed observations that can be matched and unexposed observations matched to exposed observations contribute to the analysis. As compared to matching on two or more confounders of exposure, matching on propensity score is relatively easy, since we need only match on a single continuous variable.

We implemented propensity score matching in the phototherapy data using the downloadable Stata `psmatch2` package. Table 9.18 shows the results. Again, the ATT estimate of 1.3% is about twice as large as the ACE estimate, and is close to the estimate obtained using potential outcomes estimation.

9.4.6.3 Standardized Mortality Ratio Weights

Again using $\Pr(\mathcal{E}|\mathcal{C})$ to denote the propensity score, SMR weights are defined as 1 for the exposed and $\Pr(\mathcal{E}|\mathcal{C})/(1 - \Pr(\mathcal{E}|\mathcal{C}))$ for the unexposed. SMR weights create a weighted sample of the unexposed with the same distribution of propensities for being exposed as the exposed sample. Thus, an analysis using SMR weights, like the matched analysis, estimates ATT. Furthermore, a logistic model using SMR weights directly estimates the marginal odds-ratio.

Table 9.19 Estimation of ATT using SMR weights

```
. gen smrw = phototherapy + (1-phototherapy)*prop_score/(1-prop_score)
. logistic over_thresh i.phototherapy [pweight=smrw], cluster(hospital)
Logistic regression
Log pseudolikelihood = -506.4758
(Std. Err. adjusted for 11 clusters in hospital)
```

over_thresh	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1.photother~y	.1848805	.0652173	-4.79	0.000	.0926033	.36911

```
. * Marginal risk difference
. margins, dydx(phototherapy)
Conditional marginal effects
Model VCE      : Robust
Expression     : Pr(over_thresh), predict()
dy/dx w.r.t.   : 1.phototherapy
```

	dy/dx	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
1.photother~y	-.0141753	.0022128	-6.41	0.000	-.0185124	-.0098382

Table 9.19 shows an analysis of the phototherapy data using SMR weights. In contrast to the IP weights, which exceeded 20 for many exposed infants, the largest SMR weight was less than 12. Like the potential outcomes and matched analyses, the risk difference of 1.4% (95% CI 0.98–1.9%) was larger than in the overall analysis, but the marginal odds-ratio of 0.18 (95% CI 0.09–0.37) was similar.

In summary, our propensity score analysis suggested slightly less—though unquestionably great—protection from phototherapy than the analysis using regression adjustment. In the light of restrictions imposed by the limited number of outcomes on direct regression adjustment, the propensity score results using stratification, splines, matching, and SMR weights have somewhat greater credibility. We have less confidence in the analysis using IP weights because of the presence of some large weights.

9.4.7 Recommendations for Using Propensity Scores

In most cases, propensity score quintiles are a good place to start. This makes it easy to check numbers of events, covariate balance, and interaction between exposure and the propensity score. Using more than five categories should reduce residual confounding, but categories with no events may be a bigger problem, and checks for balance and interaction may be hard to interpret in all but the

largest data sets. More generally, because categorization by quantile models the effect of the propensity score as a step function, as shown in Fig. 4.7, this may allow for residual confounding. As a result, we recommend an additional analysis incorporating the propensity score as a restricted cubic spline (Kang and Schafer 2007). If the estimated exposure effects are similar to those using categories, the simpler analysis has the advantage of being easier to understand and present. If the results are inconsistent, the spline analysis is worth the extra trouble.

In general, we are reluctant to recommend using propensity scores as inverse probability weights. Potential problems include loss of precision, large, influential weights that need to be dealt with using ad hoc approaches, and difficulty obtaining correct standard errors in some packages other than Stata. Although approaches have been developed to address these issues, they are generally complex to implement and still the subject of ongoing research.

Matching on propensity scores may be particularly effective in control of confounding (Austin 2007, 2009), but can also lead to a loss of observations in cases where matching criteria are stringent. Estimation of ATT using SMR weights avoids that difficulty, but can entail the same difficulties as IP weights, although the SMR weights were well-behaved in the phototherapy example. More generally, the resulting ATT effect estimates have a special interpretation that may not always be appropriate.

9.4.7.1 Advantages and Limitations of Propensity Scores

Propensity scores are particularly useful in analyses of uncommon binary or failure time outcomes where there are more confounders than can realistically be adjusted for using conventional regression adjustment. In addition, balance and covariate overlap can be checked and improved without looking at outcomes, helping to avoid overfitting and inflation of the type-I error rate (Rubin 2001). Sometimes these checks may lead to restriction, estimation of ATT using matching or SMR weights, or even to the recognition that the exposed and unexposed in the available sample are too unlike to be usefully compared.

Despite their applicability and relative simplicity, propensity scores do have limitations. First, there is some subjectivity in deciding whether to incorporate the scores in the second step of the analysis by stratification, regression adjustment, or inverse weighting. This decision can sometimes have major effects on resulting estimates. Second, the propensity score approach involves two statistical models, one for the relationship between the probability of exposure and predictors, and a second for the relationship between exposure and the outcome, accounting for the propensity scores. If either (or both) of these models is incorrect, biased estimates of the causal effect may result.

9.5 Time-Dependent Treatments

Estimation of average causal effects is more complicated when we consider assessing the effects of long-term treatments on long-term outcomes. For example, high blood pressure, or hypertension, is a risk factor for declines in kidney function, as measured by the estimated glomerular filtration rate (eGFR). So we might be interested in the effect of antihypertensive drugs on decline in eGFR over time. Alternatively, patients are classified as having chronic kidney disease (CKD) when eGFR falls to less than $60 \text{ mL/min/1.73 m}^2$. So we might also be interested in evaluating the efficacy of antihypertensive drugs for preventing progression to CKD.

To estimate the average causal effect of antihypertensive treatment on eGFR and CKD, we could use longitudinal data from an observational study in which blood pressure, antihypertensive use, and eGFR are measured regularly, and incidence of CKD is observed. Antihypertensives will typically be started at varying times, on the basis of clinical indications and patient preferences. We might handle this by treating antihypertensive use as a time-dependent covariate (TDC) in one of the longitudinal models introduced in Chap. 7 for the repeated eGFR measurements, or in a Cox model for time to onset of CKD.

Clearly, blood pressure is a potential confounder of antihypertensive use in our observational cohort, driving initiation of treatment as well as risk of CKD. But because blood pressure is variable over time, we would be faced with a time-dependent confounder. To achieve conditional independence of current treatment, we would likely need to condition on current and possibly past blood pressure values. Supposing that both blood pressure and antihypertensive use are measured at frequent intervals over follow-up of the cohort, an apparent solution is to treat them *both* as TDCs.

Why Time-dependent Covariates May Not Work

As a means of controlling for confounding, use of TDCs in a repeated measures or Cox model appears reasonable, but there are difficulties with this approach. In our example, the problem is that the prognostic variable we would use to control for confounding is also affected by treatment. Specifically, updated blood pressure measurements made *after* treatment is begun would reflect earlier treatment. As a result, in a Cox model with TDCs capturing current blood pressure and antihypertensive treatment, we would adjust away some part of the treatment effect, so the hazard ratio for treatment would not estimate the overall effect of treatment with antihypertensive medication.

In the best-known approach for dealing with time-dependent treatments, confounding by time-dependent confounder–mediators is controlled using time-dependent IP weights rather than TDCs. We also briefly describe two alternatives to models using IP weights, *nested cohorts of new users* and *G-estimation*.

9.5.1 Models Using Time-dependent IP Weights

The IP weights used in this context are time-dependent extensions of the IP weights introduced in Sect. 9.4.3. Ideally, use of IP weights creates comparable weighted samples of treated and untreated patients so that treatment is unassociated with the confounders in the overall weighted sample. This also means that the causal effect estimates provided by these models are intrinsically marginal, not conditional, without explicit potential outcomes estimation.

The rationale for using time-dependent IP weights is that because the confounders are not included *as covariates* in the model, updated after the initiation of treatment, we do not remove the indirect effect of treatment mediated by its downstream effects on those confounders—but we do remove the confounding. This is in contrast to standard approaches to mediation, in which we would add the mediator to the model in order to estimate the direct effect of the primary predictor via other pathways.

9.5.1.1 Inverse Probability of Censoring Weights

In addition to IP weights, *inverse-probability-of-censoring (IPC) weights* may be used to reduce bias potentially stemming from so-called *dependent censoring*, discussed in Sect. 6.6.4. The effect of the IPC weights is to maintain the comparability of the IPC-weighted treated and untreated samples. If the model for the IPC weights is correct, this avoids selection bias due to dependent censoring.

Note that TDCs affected by treatment—that is, potential mediators of the treatment effect—may have to be included in the model used to estimate the IPC weights, in order to reduce bias from dependent censoring; baseline covariates may not suffice for this purpose. However, using IPC weights rather than including these mediators as TDCs has the same benefit as using IP weights rather than TDCs to control confounding by time-dependent confounder–mediators: specifically, they allow us to estimate the overall effect of treatment without adjusting away the indirect effects mediated by the TDCs.

9.5.1.2 Stabilized and Final Weights

In many applications, so-called *stabilized* weights are used. The purpose of the stabilization is to reduce the variability of the weights, thus increasing the precision of the treatment effect estimate.

Stabilization of IP weights requires estimation of two models for the probability of current treatment status, a denominator model including baseline and time-dependent confounders, possibly including treatment history, and a numerator model including the baseline confounders and treatment history from the denominator model, but excluding other time-dependent confounders. Then the stabilized

IP weight is calculated as the ratio of the estimated probabilities of current treatment status from the numerator and denominator models. Analogous numerator and denominator models are used to estimate stabilized IPC weights. In a final step, the combined stabilized weight for each observation is calculated as the product of the stabilized IP and IPC weights.

The rationale for this method is that the numerator of the stabilized weights is correlated with the denominator, because the two models share predictors. Accordingly, the combined weight should be less variable than the denominators alone. In our experience, stabilization does not always substantially reduce variability, but in cases where very large weights are a problem, this approach may be useful.

9.5.1.3 Checking for Positivity Violations

Models using IP weights require the positivity assumption, introduced in Sect. 9.2: in this case, that at every time point, each participant must have a positive probability of being treated, and also a positive probability of *not* being treated. Violations of the positivity assumption can lead to large weights, loss of efficiency, and bias. Since the probability of treatment is estimated in calculating the IP weights, this assumption can be checked.

Positivity violations may sometimes be avoided by more careful development of the models used to calculate the weights, or by restricting the analysis to observations with predicted probabilities of current treatment status between 5% and 95%, as in our analysis using propensity scores in Sect. 9.4.5. Again, this will exclude participants who are almost always or almost never treated, focusing inferences on a target population in which the risk and benefits of treatment are unclear. Note that a stabilized weight of 20 no longer corresponds to a 5% probability of treatment received, so care must be taken in implementing this procedure. Petersen et al. (2010) provide in-depth guidance on responding to violations of this crucial assumption in models using IP weights to deal with time-dependent confounder–mediators.

9.5.1.4 Checking the Proportional Hazards Assumption

A common focus in fitting models using IP weights for time-dependent treatments is the marginal hazard ratio for the comparison of continuous treatment for the entire study period, compared to no treatment. In several published reports (Hernán et al. 2000; Cole et al. 2003; Fewell et al. 2004); this is modeled using a single parameter for current treatment, under the assumption that treatment has a constant effect—essentially the proportional hazards assumption introduced in Sect. 6.1.4.

It is important to check whether the treatment effect is in fact time-dependent, violating the proportional hazards assumption. In our example concerning treatments for hypertension and CKD, this might hold if the reduction in CKD risk increased with duration of antihypertensive treatment. A simple model assuming

a constant treatment effect would, under these circumstances, provide biased estimates of the effect of continuous treatment for the entire period. The assumption of a constant treatment effect can be checked by assessing the (possibly nonlinear) effects of treatment duration. If the effect of treatment changes with treatment duration, then it may make more sense to target the cumulative treatment effect.

9.5.2 *Implementation*

Models using IP and IPC weights to deal with time-dependent confounder–mediators require a repeated-measures extension of the methods used to implement a cross-sectional propensity score analysis in which the scores are incorporated as IP weights, as shown in Sect. 9.4.2.

9.5.2.1 Repeated Measures Outcomes

For repeated measures outcomes ascertained at each study visit, the extension to the longitudinal setting is immediate. For each participant contributing an outcome at each visit, we would define one or more TDCs for treatment, as well as a time-dependent combined stabilized weight dependent on the history of treatment, the confounder–mediator, and other baseline and time-dependent confounders up to that visit. Then, the data would be pooled across visits and analyzed using robust standard errors to account for clustering within individuals. Covariates in the model would include the TDCs for treatment and optionally baseline covariates; information from other time-dependent confounders and confounder–mediators is incorporated via the combined weight.

9.5.2.2 Survival Outcomes

For survival outcomes, the analysis would typically use pooled logistic regression (PLR), introduced in Sect. 5.5.2, rather than the Cox model. The rationale for using PLR is that suitable software typically accommodates time-dependent weights, in contrast to the Cox model implementations in most statistical packages.

To implement PLR, we would first need to split the time axis into relatively short intervals, so that information on the timing of events is not lost. For example, in a cohort study of six years duration with a survival endpoint, the time scale might be divided into 72 one-month intervals. Then for each participant still at risk of the outcome in each monthly interval, we would define one or more TDCs for treatment, a time-dependent combined stabilized weight as in the repeated measures case, and an indicator of whether the outcome occurs in the interval. As in the Cox model, individuals would not contribute to intervals after failure or censoring.

Again, the data would be pooled across intervals for analysis. In contrast to the Cox model, the baseline event rate cannot be left unspecified in PLR. Instead, some parsimonious modeling is required; one often-workable solution is to include interval number as a restricted cubic spline. The model would include the TDCs for treatment and optionally baseline covariates, with information from other time-dependent confounders and confounder–mediators incorporated via the combined weight. Robust standard errors must be used.

9.5.2.3 Worked Example

The programming required to set up these analyses is moderately complicated and particular to the package used. Thus, we have only outlined the implementation here, but provide a fully annotated Stata example with a survival outcome on the website for this book. Do-files as well as annotated code are included.

9.5.3 Drawbacks and Difficulties

Implementing a model using inverse weighting to deal with time-dependent confounder–mediators can be complicated. In particular, there may be more than one confounder–mediator to deal with, and many predictors of treatment status will generally need to be taken into account. Furthermore, the appropriate form for all five models will be unknown, although the specification must be approximately correct for the model to provide consistent estimates. Chapters 4 and 5 provide guidance on developing good models, but power to detect model misspecification may be low. Missing values pose additional challenges, although not qualitatively different from more conventional survival analyses using time-dependent covariates. Finally, very large weights reflecting positivity violations may strongly influence the results and need to be dealt with, either by improvement of the weights or by restriction to a subsample where the positivity assumption is more clearly met.

The problem of estimating the effects of time-dependent treatments in the presence of time-dependent confounder–mediators is a topic of current statistical research, and in our view there is currently no established, straightforward solution broadly applicable to survival as well as repeated continuous, binary, and count outcomes. As noted in Sect. 9.4.2, more recent statistical research (Lunceford and Davidian 2004; Kang and Schafer 2007; Schafer and Kang 2008; Freedman and Berk 2008) has pointed out drawbacks in the use of IP weights for estimation of causal effects. These include loss of precision when the weights are highly variable, the potential need for ad hoc trimming of large weights, and vulnerability to bias when the models underlying the weights are misspecified.

These considerations lead us to recommend that analysis using IP weights be considered only for estimation of the effects of time-dependent treatments

or exposures with time-dependent confounder–mediators—the case where special methods are needed to obtain an estimate of the overall effect of treatment. In the absence of time-dependent confounder–mediators, other approaches, including other methods for using propensity scores, avoid the inefficiency and difficulties of inverse weighting, yet often provide comparable control of confounding. In addition, marginal rather than conditional effect estimates are often easily calculated using potential outcomes estimation, as shown in Sect. 9.3.4.

9.5.4 *Focusing on New Users*

Our discussion of time-dependent treatments has implicitly assumed that we would observe cohort participants before treatment is begun. In cases where the time-dependent confounder is subsequently affected by treatment, we need to measure the confounder *before* treatment is initiated to remove confounding. For example, in estimating the effects of antihypertensive use on risk of developing CKD, on-treatment blood pressure levels would be a misleading measure of baseline risk. Likewise, our discussion of choosing an appropriate causal target assumed that the focus would be on the effect of a treatment from initiation forward, although the effect may vary over time. Parenthetically, we recognize that other analyses might focus on the effect of discontinuing treatment among prevalent users, entailing a different study design.

These considerations emphasize the importance of excluding prevalent users in most analyses of the effect of time-dependent treatments. If this is done, estimates of the effect of treatment are based entirely on comparisons between new users observed to initiate treatment and appropriate controls. By focusing on new users, we can reduce several types of bias (Ray 2003):

- *Bias from time-dependent treatment effects.* HT, as an example, has early adverse effects, possibly followed by late benefit. If we assume that the treatment effect is constant, inclusion of prevalent users places too much weight on the late effects.
- *Bias from selection of survivors.* This issue is clearest for surgical treatments with perioperative mortality risk. A sample including patients recruited after surgery will include an unrepresentative proportion of survivors, and thus put too much weight on operative successes. Similarly, women dying from heart attacks in the first year of hormone therapy use will almost surely be under-represented in a cohort including prevalent users.
- *Adherence bias.* Placebo-controlled trials have shown that adherence *to placebo* is independently associated with better outcomes in many contexts. Including prevalent users puts too much weight on outcomes among the long-term users, by definition better adherers to treatment.

The primary disadvantage of excluding prevalent users is loss of precision.

9.5.5 Nested New-User Cohorts

[Hernán et al. \(2008\)](#) generalizes Ray’s new-user approach to time-dependent treatments, providing an alternative to models using IP weights to deal with time-dependent confounder–mediators of time-dependent treatments. Typically using data from a cohort study with visits at regular intervals, a nested cohort is selected at each sequential visit, consisting of new users who started treatment in the interval since the last visit, and controls who remain untreated up through that visit. Follow-up for the new users begins at the time of treatment initiation, and for controls at the *average* time of initiation among the new users in the nested cohort.

In the analysis, the resulting nested cohorts are pooled. Because observations as well as outcome events may figure in multiple cohorts, robust standard errors must be used. Survival or repeated measures models, depending on the outcome, are then used to control for confounders *as fixed covariates*, ascertained at the newly defined beginning of follow for each nested cohort participant. This is in contrast to the conventional Cox model with TDCs. As a result, we do not adjust away the indirect effect of treatment mediated by its subsequent effects on the confounder–mediator.

Of course, some patients included as new users in each nested cohort cease use, and some controls start. [Hernán et al. \(2008\)](#) resolve this problem by censoring follow-up at the time of cross-over, thus focusing comparisons on new users who continue use and controls who remain nonusers.

However, the censoring will often depend on time-dependent covariate values at the time of censoring—that is, on potential mediators of the treatment effect. Controlling for these confounder–mediators as TDCs might make the censoring conditionally independent, but would also adjust away the fraction of the treatment effect that they mediate. Thus, to estimate the overall treatment effect, we would need to use IPC weights rather than TDCs to account for the dependent censoring.

In summary, at the cost of some programming to set up the nested cohorts, we avoid having to model the IP weights. However, the models for the IPC weights must be correct, and large IPC weights may impose some of the same loss of efficiency and vulnerability to bias seen with IP weights in some applications. On the website for this book, we provide an example of a nested new-user cohort analysis with IPC weights, implemented in Stata and using simulated data. Do-files with annotated code are included.

9.6 Mediation

In Sects. 4.5, 5.2.3, and 6.2.9, we presented methods for assessing the mediating influence of predictors in regression models. Assigning a causal interpretation to related quantities such as direct and indirect effects involves extension of potential outcomes to include the mediating variable, and generalization of assumptions required for valid estimates to include the relationships between the mediator, outcome, and confounders.

Recall the example from the FIT study presented in Table 5.12 on estimating the effect of a treatment on new fracture risk in the presence of possible mediation through observed changes in BMD level. Although the original assignment to treatment was randomized, changes in BMD occur postrandomization. Thus, controlling for observed change in BMD raises the possibility of confounding by variables causally related to both change in BMD and fracture risk.

In addition to the assessment of the presence of mediating effects of changes in BMD summarized in Table 5.12, we may also want estimates of the impact of treatment not mediated through the BMD pathway. As introduced in Sect. 4.5, this is an example of a *direct effect* of treatment. Although a logistic regression model including treatment and change in BMD may be used to provide an estimate of this direct effect, in the presence of additional confounding variables (e.g., the model in Table 5.12), this will have a conditional interpretation discussed in Sect. 9.3.1. Marginal estimates that are interpretable as a causal direct effect can be obtained using a generalization of the potential outcomes approach described in Sect. 9.1.

The causal *controlled direct effect* of treatment is defined as a comparison of the potential fracture outcomes in treated and untreated women with change in BMD fixed at a specified level. This corresponds to the effect that would be observed if we could randomize treatment in women known to be homogeneous in their BMD response, and provides useful information about the effectiveness of treatment in this context. Note that potential outcomes of women in this situation need to account for both treatment alternatives and the specified level of change in BMD. The potential outcome for a woman assigned treatment $\mathcal{E}$ and mediating variable $\mathcal{Z}$ is defined as $Y(\mathcal{E}, \mathcal{Z})$. The controlled direct effect for a fixed value z of $\mathcal{Z}$, expressed as a causal risk difference, is then defined as

$$E[Y(1, z)] - E[Y(0, z)]. \quad (9.12)$$

Because the potential outcomes now depend on two variables, the definitions in Sect. 9.1 need to be extended accordingly. For example, the marginal structural model (9.1) for the mean potential outcomes must be specified as a function of both $\mathcal{E}$ and $\mathcal{Z}$. The additional conditional independence assumption required for valid estimation of related causal effects also must include observed confounding variables $\mathcal{C}$ of the relationship between $\mathcal{Z}$ and Y . These may be distinct from observed variables that confound the relationship between $\mathcal{E}$ and Y . This assumption specifies that potential outcomes $Y(\mathcal{E}, \mathcal{Z})$ are independent of $\mathcal{Z}$ conditional on $\mathcal{E}$ and $\mathcal{C}$.

When the assumptions outlined above hold, estimation of controlled direct effects can generally be accomplished using a modified version of the potential outcomes approach described in Sect. 9.1.7. Table 9.20 illustrates the potential outcomes approach for the example from Table 5.12. After fitting the model linking outcomes to both the mediator and potential confounders (and suppressing the output using the Stata prefix `quietly`), the `margins` command estimates the treatment group-specific marginal outcome probabilities with change in BMD fixed at zero for all women, using the `margins` option

Table 9.20 Estimating the controlled direct effect of treatment in the FIT study

```
. quietly logistic frac_new i.treat bmd_diff bmd_base i.frac_base ///
> i.smoking age_spl*
```

```
. margins treat, at(bmd_diff==0)
Predictive margins
Model VCE      : OIM
Expression     : Pr(frac_new), predict()
at             : bmd_diff = 0
```

	Margin	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]
treat					
0	.0681827	.0047664	14.30	0.000	.0588408 .0775247
1	.0430936	.004234	10.18	0.000	.0347952 .051392

```
. margins, dydx(treat) at(bmd_diff==0)
Average marginal effects
Model VCE      : OIM
Expression     : Pr(frac_new), predict()
dy/dx w.r.t.   : 1.treat
at             : bmd_diff = 0
```

	dy/dx	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]
1.treat	-.0250891	.0065736	-3.82	0.000	-.0379731 -.0122051

```
. margins r.treat, at(bmd_diff==0)
Contrasts of predictive margins
Model VCE      : OIM
Expression     : Pr(frac_new), predict()
at             : bmd_diff = 0
```

	df	chi2	P>chi2
treat	1	14.57	0.0001

```
. margins, dydx(treat) at(bmd_diff==0)
Contrasts of predictive margins
Model VCE      : OIM
Expression     : Pr(frac_new), predict()
at             : bmd_diff = 0
```

	Contrast	Delta-method Std. Err.	[95% Conf. Interval]
treat			
(1 vs 0)	-.0250891	.0065736	-.0379731 -.0122051

at (bmd_diff==0). Then the controlled direct effect on the risk difference scale is obtained two ways, first using dydx option, then using the r. contrast operator.

In the situation where there are observed variables that are mediators of the relationship between the exposure and the primary mediator of treatment effects $\mathcal{Z}$, estimation of controlled direct effects may require inverse weighting methods as described in Sect. 9.1.8. In the context of the FIT example, consider an intermediate biological factor that results from treatment that in turn affects both changes in BMD and fracture risk. Controlling for this variable as a confounder would effectively remove some of the effect of treatment on changes in BMD. Omitting it would result in residual confounding of the relationship between changes in BMD and

fracture risk. The need for inverse weighting methods in such situations thus echoes the motivations for their use in the context of marginal structural models for event time outcomes introduced in Sect. 9.5. In the mediation case, inverse weights are required for both the probability of treatment and the mediator (VanderWeele 2009).

The controlled direct effect is of limited interest in situations where the mediating variable cannot be interpreted as amenable to control via an intervention. The *natural direct effect* is an alternative measure that represents the effect of blocking the effect of exposure on the mediator, but allowing the value of the mediator to vary among individuals at levels that would have been observed in the absence of exposure. Causal interpretation requires potential versions of the mediator corresponding to possible exposure scenarios. The natural direct effect can then be defined as the average causal effect among individuals with the potential mediating variable fixed at the level indicating no exposure. Estimation of natural direct effects requires additional assumptions beyond those required for controlled direct effects, and valid estimates from standard regression approaches are possible only in fairly restricted situations. Some of these methods are implemented in the downloadable Stata package *mediation*. These issues also apply to decomposition of overall effects into direct and indirect components, illustrated for linear models for continuous outcomes in Sect. 4.5. Because methods for estimation are an area of active research, we refer readers to recent references provided in Sect. 9.10.

9.7 Instrumental Variables

A primary assumption of most methods for estimating the causal effects of an exposure or treatment using observational data is that there are no unmeasured confounders. This assumption underlies regression adjustment, the primary topic of this book, as well as propensity scores and the methods proposed for dealing with time-dependent treatments. The assumption of no unmeasured confounders cannot be directly verified, and arguments on substantive grounds that nothing important has been omitted will sometimes be unconvincing.

In contrast, the method of *instrumental variables* (IVs) may allow us to obtain valid estimates of causal effects when this assumption is not met. Instrumental variables have a long history in the social sciences, are an everyday tool of econometricians, political scientists, and sociologists, and may play an important role in comparative effectiveness research using administrative databases with limited confounder measurements.

For example, Hearst et al. (1986) used the draft lottery in the United States as an IV to estimate the effect of having served in the military on mortality risk *after* the Vietnam war. Not nearly enough information was available for veterans, not to mention appropriate controls, to attempt to answer this question using regression adjustment or propensity scores. However, draft lottery numbers had several properties that made them useful for the analysis: having a lottery number below the eligibility threshold was a strong determinant of military service, it was

randomly assigned, and it did not obviously influence subsequent life course except through its influence on service. Essentially, these are the defining characteristics of an IV:

- (1) It must be a strong predictor of exposure.
- (2) Its associations with both exposure and outcome must be unconfounded, at least conditionally on measured covariates.
- (3) All of its association with the outcome must be mediated by exposure.

Clearly, we have replaced the assumption that exposure is unconfounded with the assumption that the IV is unconfounded. But in some cases, this assumption is easier to accept for an IV than an exposure. Examples include certain natural experiments and treatment assignment as an IV for treatment received in clinical trials.

IVs from Natural Experiments

Well-justified IVs can come from *natural experiments*. The Vietnam-era draft lottery is one example. Another is the intertwining of the pipelines of the Lambeth Waterworks with those of the Vauxhall and Southwark; [Snow \(1855\)](#) recognized that waterworks was effectively allocated at random to households. Waterworks could have served as an IV because it strongly influenced exposure to the cholera bacterium (assumption 1), was not associated with other cholera risk factors (assumption 2), and could have had no effect on cholera except through its influence on this exposure (assumption 3).

Similarly, [Smith and Ebrahim \(2004\)](#) show how *Mendelian randomization* can also be viewed as a natural experiment in which genetic variants that influence causal factors of interest are allocated at random. For example, [Katan \(1986\)](#) used one such variable allele linked to higher cholesterol levels as an IV to assess the possible causal effects of cholesterol on cancer risk. The allele can serve as an IV because it influences cholesterol levels (assumption 1), is not associated with other cancer risk factors (assumption 2) under Mendelian randomization, and presumably has no effect on cancer risk except through its influence on cholesterol levels (assumption 3).

Treatment Assignment as an IV

In clinical trials with excellent adherence, a simple comparison of average outcomes in the treatment and control groups often has a straightforward interpretation as the causal effect of treatment. However, in trials with incomplete adherence, the treatment that participants actually receive is often affected by patient characteristics that influence both adherence and outcomes. In this case, random treatment assignment can be a good IV for estimating the causal effect of treatment *received* rather than the effect of treatment *assignment*, which is generally attenuated by nonadherence. In most trials, assumption 1 holds because treatment assignment

is a strong determinant of treatment received. Assumption 2 holds provided the randomization was successful. And assumption 3 holds if the trial is successfully blinded, blocking plausible indirect causal pathways from treatment assignment to the outcome.

For example, [Permutt and Hebel \(1989\)](#) used random assignment of expectant mothers who smoked to a program encouraging them to stop smoking as an IV for the effect of smoking on the birth weight of their newborns. This analysis suggested that actual reductions in smoking resulted in substantially higher birth weights. Similarly, [Sommer and Zeger \(1991\)](#) and later [Greenland \(2000\)](#) used treatment assignment in a cluster-randomized trial as an IV to show that vitamin A supplementation reduced mortality among children in rural Indonesia.

IVs in Comparative Effectiveness Research

One context in which IV analysis might prove useful is comparative effectiveness research on the safety and efficacy of approved treatments. The crucial problem for such research is confounding of treatment effects by clinical indications that physicians use in deciding on a course of treatment. More effective treatments may be preferentially given to sicker patients, especially if they entail costs, risks, or side effects that are only acceptable in graver cases. However, many of the signs and symptoms identifying these patients are not adequately captured in observational and especially administrative databases. As a result, standard regression adjustment is commonly unable to adjust completely for differences in prognosis between patients given alternative treatments. The resulting treatment effect estimates are confounded.

In contrast, IVs hold out some hope, because in principle they do not require that all confounders be measured. Differences in practice patterns across regions, hospitals, or physicians are one possible IV for a treatment of interest. Assumption 1 holds because the varying practice patterns can be assumed to influence or at least reflect what treatments are used. Assumption 2 holds if practice patterns are conditionally independent of unmeasured risk factors for the disease outcome under consideration, given available covariates. And assumption 3 holds if practice patterns only affect outcomes via receipt of the treatment of interest.

As an example of using variation in practice patterns, [Brookhart et al. \(2006b\)](#) used physician preferences for prescribing Cox-2 inhibitors, a class of nonsteroidal anti-inflammatory drugs (NSAIDs), as an IV in estimating the effect of these pain relievers on gastrointestinal complications, relative to other NSAIDs.

9.7.1 Vulnerabilities

In many contexts it can be difficult to find an IV that unquestionably meets assumptions 2 and 3. For example, in using the draft lottery as an IV for military

service during the Vietnam war, assumption 3 could have been violated if men with low-lottery numbers stayed in school to retain draft deferments, which could have improved their life chances by means other than avoiding military service (Angrist and Krueger 1992; Angrist et al. 1996).

Similarly, in the Mendelian randomization example, assumption 2 could be violated in samples including people of different race or ethnicity, which might be associated with both allele frequency and exposure to other cancer risk factors—the well-known problem of *population stratification*. And assumption 3 could be violated if the allele of interest affects pathways other than cholesterol levels that are important for cancer risk, or is in so-called *linkage disequilibrium* and thus correlated with other alleles that do. We could control for race/ethnicity, but direct effects would be harder to rule out.

In the Cox-2 example, assumption 2 could be violated if physicians who are more likely to prescribe Cox-2 inhibitors also see higher risk patients on average, so that the association between practice style and gastrointestinal complications is confounded by differences in patient risk. In addition, assumption 3 could be violated if the physicians who more frequently prescribe Cox-2 inhibitors also tend to prescribe additional protective medications, such as H₂-blockers or proton pump inhibitors. In this case, a practice style favoring Cox-2 inhibitors would have direct effects on the outcome that are not mediated by the Cox-2 inhibitors themselves (Hernán and Robins 2006). This issue threatens the validity of our IV analysis of the phototherapy data, reported in Sect. 9.7.6.

Several other potential problems with the use of IV for estimation of causal effects are worth mentioning:

- IV methods are generally less efficient than direct regression adjustment, so make most sense when unmeasured confounding of exposure is a well-justified concern.
- The IV should be strongly associated with exposure. Weak correlation between them makes IV effect estimators less precise. This problem is generally worse when the measured IV is a noisy surrogate (Hernán and Robins 2006), as in the Brookhart et al. (2006b) example.
- IV regression coefficient estimates are not unbiased in small samples. At best, under assumptions 1–3, they are *consistent*—that is, the bias is negligible in large samples.
- Weak correlation between the IV and exposure inflates any bias.
- In cases where the exposure–outcome relationship is strongly confounded, IVs strongly associated with exposure may not exist. If a strong IV is found in this context, assumption 3 is likely violated (Martens et al. 2006).
- With continuous exposures and outcomes, the linearity and constant variance assumptions are important, with violations potentially inducing bias and invalidating CIs and *P*-values.

9.7.2 Structural Equations and Instrumental Variables

Instrumental variables were originally proposed in the context of linear *structural equation models*. In this section, which can be skipped without loss of continuity, we briefly sketch the underpinnings of IV analysis.

Suppose we would like to estimate the causal effect of an exposure $\mathcal{E}$ on an outcome Y , using observational data. We know that the effect of $\mathcal{E}$ on Y is confounded by a measured confounder $\mathcal{C}$, but also by an unmeasured confounder $\mathcal{U}$. Recall that a proposed instrumental variable $\mathcal{I}$ must be strongly associated with $\mathcal{E}$, its associations with both $\mathcal{E}$ and Y must be unconfounded, given $\mathcal{C}$, and its association with Y must completely mediated by $\mathcal{E}$.

We have two linked structural equations, the first for the effect of $\mathcal{E}$ on Y :

$$Y = \beta_0 + \beta_1\mathcal{E} + \beta_2\mathcal{C} + \epsilon. \quad (9.13)$$

Because $\mathcal{U}$ is omitted from this model, regressing Y on $\mathcal{E}$ and $\mathcal{C}$ would give a biased estimate of β_1 . So simple regression adjustment will not provide unbiased estimates of the causal effect of $\mathcal{E}$ on Y . The second structural equation is for the effect of $\mathcal{I}$ on $\mathcal{E}$:

$$\mathcal{E} = \gamma_0 + \gamma_1\mathcal{I} + \gamma_2\mathcal{C} + \eta. \quad (9.14)$$

Under our assumption that the association of $\mathcal{I}$ with $\mathcal{E}$ is unconfounded, given $\mathcal{C}$, a regression of $\mathcal{E}$ on $\mathcal{I}$ and $\mathcal{C}$ will provide an unbiased estimate of γ_1 . Next, substituting (9.14) in (9.13), we do some algebra to obtain an equation for the effect of $\mathcal{I}$ on Y .

$$\begin{aligned} Y &= \beta_0 + \beta_1(\gamma_0 + \gamma_1\mathcal{I} + \gamma_2\mathcal{C} + \eta) + \beta_2\mathcal{C} + \epsilon \\ &= \beta_0 + \beta_1\gamma_0 + \beta_1\gamma_1\mathcal{I} + (\beta_1\gamma_2 + \beta_2)\mathcal{C} + \beta_1\eta + \epsilon \\ &= \lambda_0 + \lambda_1\mathcal{I} + \lambda_2\mathcal{C} + \psi. \end{aligned} \quad (9.15)$$

Under our assumption that the association of $\mathcal{I}$ with Y is unconfounded, given $\mathcal{C}$, a regression of Y on $\mathcal{I}$ and $\mathcal{C}$ will provide an unbiased estimate of λ_1 . By definition, $\lambda_1 = \beta_1\gamma_1$, so we can estimate β_1 by $\hat{\lambda}_1/\hat{\gamma}_1$. This IV causal effect estimator is implemented in the `ivregress` command in Stata.

9.7.3 Checking IV Assumptions

To begin, it is straightforward to assess the strength of the relationship between the IV and exposure. For the case with continuous exposure and outcome, the `ivregress` post-estimation command `estat firststage` provides R^2 and an F -test to help make this assessment. For other cases, this can be done using

a linear or logistic regression, as appropriate, of the exposure on the IV as well as confounders of this association. Here, interest would focus on the increment in R^2 or pseudo- R^2 for the addition of the IV to the model.

Since IV analysis is less efficient than conventional regression adjustment, it makes sense to look at whether unmeasured confounding justifies its use. Although we can never rule out confounding by unmeasured factors, we can assess evidence for its existence. In particular, tests for residual confounding of exposure are available for both continuous and binary exposures and outcomes. When both are continuous, Stata's `ivregress` post-estimation command `estat endogenous` provides appropriate tests. When either or both are binary, residual confounding of exposure can be assessed by using certain likelihood-ratio or Wald tests. We implement these tests in Tables 9.21 and 9.22 below.

Finally, for continuous exposures and outcomes, methods exist for assessing the validity of the IV. Called tests for *overidentifying restrictions* and implemented in Stata's `ivregress` post-estimation command `estat overid`, these tests would only be applicable to the examples we have considered, with a single exposure variable of interest, if we had used more than one IV. More generally, they are only applicable in analyses where the number of IVs is larger than the number of exposure variables.

9.7.4 Example: Effect of Hormone Therapy on Change in LDL

To illustrate a basic IV analysis, we analyzed changes in LDL cholesterol during the first year of the HERS trial. A simple intention-to-treat (ITT) comparison by treatment assignment showed that average reductions in LDL were 15.6 mg/dL larger in the HT group. We conducted an observational analysis regressing change in LDL on `HT_use`, the proportion of days HT was taken, simulated to depend on unmeasured confounders associated with reductions in LDL. This analysis showed that taking HT daily would reduce LDL by almost 22 mg/dL.

To deal with the unmeasured confounding, we used treatment assignment as an IV to estimate the causal effect of HT use on change in LDL. Results are shown in Table 9.21. This analysis suggests that daily HT use would reduce LDL by an average of 17 mg/dL, more than the ITT estimate, but considerably less than the confounded estimate.

In checking IV assumptions, the `estat endogenous` post-estimation command gives very strong evidence ($P < 0.00005$) that HT use was confounded. In addition, `estat firststage` shows that the IV, treatment assignment, was very strongly associated with the exposure, HT use. However, there was some unblinding in HERS, because of the side effects of HT. This might violate the assumption that the entire association of the IV with the outcome is mediated by exposure, and would potentially bias an actual IV estimate of the effect of HT use.

Table 9.21 IV analysis of hormone use effect on change in LDL

```
. ivregress 2sls ldlch (HT_use = HT)
```

Instrumental variables (2SLS) regression					Number of obs = 2597	
					Wald chi2(1) = 143.00	
					Prob > chi2 = 0.0000	
					R-squared = 0.0846	
					Root MSE = 33.215	

ldlch	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
HT_use	-16.99995	1.421609	-11.96	0.000	-19.78626	-14.21365
_cons	-4.66981	.9199404	-5.08	0.000	-6.47286	-2.86676


```
Instrumented: HT_use
Instruments: HT
```



```
. estat endogenous
Tests of endogeneity
Ho: variables are exogenous
Durbin (score) chi2(1) = 305.91 (p = 0.0000)
Wu-Hausman F(1,2594) = 346.355 (p = 0.0000)
```



```
. estat firststage
```

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	F(1,2595)	Prob > F
HT_use	0.9569	0.9569	0.9569	57650.4	0.0000

9.7.5 Extension to Binary Exposures and Outcomes

So far we have assumed that both the exposure $\mathcal{E}$ and the outcome Y are continuous, as in the structural equations (9.13) and (9.14). In contrast, we have placed no restrictions on the distribution of the IV. The primary tool for accommodating binary exposures and outcomes in IV analysis is the probit model.

With a single outcome, the probit model is comparable to logistic regression, commonly gives similar results, and is implemented in the Stata `probit` command. Probit models can be thought of as arising from a *latent*, or unobserved, normally distributed outcome, Y^* , which follows the linear regression model:

$$Y^* = \beta_0 + \beta_1 \mathcal{E} + \beta_2 \mathcal{C} + \epsilon, \quad (9.16)$$

where $\mathcal{E}$ and $\mathcal{C}$ are defined as before, and ϵ has a standard normal distribution. However, we only observe the binary outcome Y , which takes on the value 1 if $Y^* > 0$ and 0 otherwise. For a binary exposure, the analogous probit model is

$$\mathcal{E}^* = \gamma_0 + \gamma_1 \mathcal{I} + \gamma_2 \mathcal{C} + \eta. \quad (9.17)$$

In some circumstances, the latent variable has a real interpretation. For example, many individual alleles may contribute to an observable phenotype ($Y = 1$). In this case, Y^* , the sum of the allelic contributions, might be approximately normal by the central limit theorem.

When exposure is continuous but the outcome is binary, we substitute (9.16) for (9.13). We then can use the Stata `ivprobit` command to obtain an IV estimate of the causal effect of continuous $\mathcal{E}$ on binary Y , based on (9.14) and (9.16). With binary exposure and continuous outcome, we substitute (9.17) for (9.14), then use the downloadable `cmp` (*conditional mixed process*) command. Finally, for binary exposure *and* outcome, we make both substitutions, then use the `biprobit` command. In Sect. 9.7.6, we use this method to re-estimate the effect of phototherapy on neonatal jaundice.

9.7.6 Example: Phototherapy for Neonatal Jaundice

In addition to the re-analysis using propensity scores in Sect. 9.2.6, we also estimated the causal effect of phototherapy on neonatal jaundice using IVs. In this analysis, we took advantage of variation in practice patterns, using hospital and year of birth jointly as an IV for phototherapy.

The estimates obtained from the IV analysis using the bivariate probit model, shown in Table 9.22, differ substantially from the adjusted logistic and propensity score results. The long model output is difficult to interpret directly and thus omitted. The likelihood ratio test of $\rho=0$ gives evidence ($P = 0.0162$) for the residual confounding of phototherapy and thus the need for IV analysis. After fitting the model, we used the `margins` command to implement potential outcomes estimation, then calculated the marginal odds-ratio and risk difference. As in the conventionally adjusted and propensity score analyses, the marginal risk difference can be obtained two ways.

The estimated marginal odds-ratio of 0.050 is an order of magnitude smaller than the marginal odds-ratio of 0.18 obtained using the results in Table 9.6. Similarly, the estimated risk difference is larger (1.8%, 95% CI 0.53–3.1%), and much less precisely estimated than results based on standard regression adjustment (0.79%, 95% CI 0.59–1.00%; Table 9.9) or using propensity scores as a restricted cubic spline (0.81%, 95% CI 0.61–1.0%; Table 9.14).

In a sensitivity analysis omitting the control variables in both `biprobit` equations, the estimated marginal odds-ratio for phototherapy was 0.049, very close to the adjusted IV estimate, lending support to the claim that IV analysis can control for unmeasured confounders.

9.7.6.1 Evaluating Assumptions

In this example, phototherapy use varied substantially across hospitals and years, so there was support for assumption 1. In addition, the IV was plausibly unconfounded, conditional on the other strongly predictive risk factors included in the analysis (assumption 2).

However, assumption 3, that TSB levels were unlikely to be influenced by hospital and year except through receipt of phototherapy, was called into question

Table 9.22 Instrumental variable analysis of phototherapy effect

```

. biprobit ///
> (over_thresh male i.gest_age##c.birth_wt i.qual_TSB i.age_days i.phototherapy) ///
> (phototherapy2 = i.hosp_year male i.gest_age##c.birth_wt i.qual_TSB i.age_days)

Seemingly unrelated bivariate probit      Number of obs   =      20731
                                           Wald chi2(150)    =      3441.26
Log likelihood = -9605.9172                Prob > chi2       =      0.0000
-----+-----
              |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      ....
1.phototherapy |   -1.359804   .2550643    -5.33  0.000    -1.859721    -.8598873
      ....
-----+-----
      /athrho |    .4720604   .1965266     2.40  0.016     .0868753     .8572454
-----+-----
      rho |    .4398626   .1585028                .0866574     .6948357
-----+-----
Likelihood-ratio test of rho=0:      chi2(1) =  5.77649      Prob > chi2 = 0.0162

. * Marginal risk difference
. margins, dydx(phototherapy) predict(pmarg1)
Average marginal effects      Number of obs   =      20731
Model VCE      : OIM
Expression     : Pr(over_thresh=1), predict(pmarg1)
dy/dx w.r.t.   : 1.phototherapy
-----+-----
              |      Delta-method
              |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
1.photothey |   -.0181195   .0065423    -2.77  0.006    -.0309422    -.0052968
-----+-----

. * Marginal risk difference using contrast operator
. margins r.phototherapy, predict(pmarg1)
Contrasts of predictive margins
Model VCE      : OIM
Expression     : Pr(over_thresh=1), predict(pmarg1)
-----+-----
              |      df      chi2    P>chi2
-----+-----
phototherapy |      1      7.67    0.0056
-----+-----

              |      Delta-method
              |      Contrast   Std. Err.     [95% Conf. Interval]
-----+-----
phototherapy |
(1 vs 0) |   -.0181195   .0065423    -.0309422    -.0052968
-----+-----

```

by an unmeasured co-intervention, switching from breast feeding to formula. In a matched case-control sample nested within the larger study (Kuzniewicz et al. 2008), use of this co-intervention was strongly correlated ($r = 0.56$, $P < 0.001$) with use of phototherapy across Kaiser facilities. However, adjusted estimates of the effect of phototherapy were similar with (odds-ratio 0.15, 95% CI 0.06, 0.40, $P < 0.001$) and without (odds-ratio 0.14, 95% CI 0.06, 0.35, $P < 0.001$) additional adjustment for the co-intervention, formula use, suggesting that the analysis using regression adjustment may not be badly biased. However, because the co-intervention is more common at hospitals where phototherapy is more often used, it would make phototherapy appear even more protective than it is in the IV analysis.

9.7.7 Interpretation of IV Estimates

In the original IV formulation using structural equation modeling, it was assumed that the causal effect of exposure on the outcome is constant across the population. Under this view, the IV analysis estimates the population-wide average causal effect of the exposure. This interpretation requires us to posit a mechanism under which the entire population is treated, or not.

In contrast, in the potential outcomes framework, IV effect estimates are commonly interpreted more narrowly. For example, [Greenland \(2000\)](#) interpreted the causal effect of Vitamin A supplementation assessed in the Indonesian trial as applying only to the children of families that would comply with the supplementation program, but not necessarily to children in other families. This is sometimes called the *local average treatment effect* (LATE).

9.8 Trials with Incomplete Adherence to Treatment

Randomization is well known to prevent confounding of treatment in an experiment, at least on average and in large enough samples. It follows that when adherence to assigned treatment (as well as follow-up) is complete, then unadjusted comparisons of outcomes in the treated and control groups provide unbiased estimates of the causal effect of treatment.

9.8.1 Intention-to-Treat

We know, of course, that adherence to assigned treatment in clinical trials is commonly incomplete, especially for treatments that have adverse side effects or are freely available to controls. Setting aside the complications posed by incomplete follow-up until Chap. 11, on missing data, incomplete adherence implies that an unadjusted comparison of mean values of the outcome in the treated and control groups, an *intention-to-treat* (ITT) analysis, only provides a consistent estimate of the causal effect of treatment *assignment*, which is sometimes interpretable as the effectiveness of a treatment program. (We note that there may be some attenuation of the effectiveness estimate in logistic and Cox models, arising from the omission of covariates uncorrelated with treatment assignment but strongly associated with treatment, as noted earlier in Sects. 3.4.5, 4.4, 5.2.3, and 6.6.3). However, it does not provide an unbiased estimate of the causal effect of treatment received.

To illustrate the difference between the causal effects of treatment assignment and treatment received, we return to our example of exercise and glucose levels. Now, we consider a potential outcomes experiment for the effect of treatment assignment, with the complication that there is incomplete adherence to assigned

Table 9.23 Potential outcomes with incomplete adherence

		Potential outcomes by treatment assignment					Observed	
		$T^r(1)$	$T^r(0)$	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$	T^a	$T^r(a)$
							Y	
$C = 0$	1	0	100	105	-5	0	0	105
	1	0	98	96	2	0	0	96
	1	0	96	99	-3	0	0	99
	0	0	102	102	0	0	0	102
	0	0	98	98	0	0	0	98
$C = 1$	1	0	96	94	2	0	0	94
	1	0	94	96	-2	0	0	96
	1	0	92	98	-6	0	0	98
	1	1	95	95	0	0	1	95
	1	1	93	93	0	0	1	93
Means	0.8	0.2	96.4	97.6	-1.2		0.2	97.6
$C = 0$	1	0	95	97	-2	1	1	95
	1	0	97	100	-3	1	1	97
	1	0	102	103	-1	1	1	102
	0	0	99	99	0	1	0	99
	0	0	101	101	0	1	0	101
$C = 1$	1	0	91	97	-6	1	1	91
	1	0	98	95	3	1	1	98
	1	0	93	96	-3	1	1	93
	1	1	97	97	0	1	1	97
	1	1	91	91	0	1	1	91
Means	0.8	0.2	96.4	97.6	-1.2		0.8	96.4

treatment. In Table 9.23, we represent this potential outcomes experiment, with each member of the population contributing an outcome under assignment to exercise as well as control.

9.8.1.1 Example: Exercise and Glucose Levels

As before, we use $Y(1)$ to denote outcomes under assignment to treatment, and $Y(0)$ for outcomes under assignment to control. Here, we also need to distinguish T^a , the indicator for assignment to treatment, from $T^r(1)$, the treatment received under assignment to treatment ($T^a = 1$), and $T^r(0)$, the treatment received under assignment to control ($T^a = 0$); we observe $T^r(a)$, the treatment received under the actual assignment $T^a = a$.

Now suppose that only 80% of women exercise when assigned to it, but 20% of women exercise even when assigned to control. We have also assumed that when women are assigned to exercise, nonadherence is concentrated in the group with $C = 0$, but when they are assigned to control, nonadherence is only seen in the

subgroup with $C = 1$. As a result, $T^r(1)$ and $T^r(0)$ are correlated with C . We again suppose that the causal effect of exercise is to lower glucose levels an average of 2 mg/dL, that the causal direct effect of C is to lower glucose 4 mg/dL, and that half of women are in the subgroup with $C = 1$.

The supposed data are shown in Table 9.23. Keeping in mind that the potential outcomes $Y(1)$ and $Y(0)$ are now defined in terms of treatment *assignment*, not treatment received, note that there is no difference in potential outcomes for the 8 women who are nonadherent, because the treatment they receive is unaffected by treatment assignment. Within each randomized group as well as overall, the average difference in potential outcomes is 1.2 mg/d, 40% less than the causal effect of exercise. This is the intention-to-treat effect of assignment to exercise.

9.8.2 As-Treated Comparisons by Treatment Received

Consider a comparison of outcomes in the trial shown in Table 9.23 according to $T^r(a)$, or treatment received, sometimes called an *as-treated* analysis. Here we assume that each row of the table represents two participants, one assigned to treatment, the other to control. In this context, an as-treated analysis would amount to comparing women who exercise with those who do not, without regard to treatment assignment.

Unless adherence to assigned treatment is perfect, this comparison would likely be biased for the causal effect of treatment. In making this comparison, we would lack any assurance that confounding variables would be balanced in those who exercise as compared to those who do not. In Table 9.23, 70% of women who exercised ($T^r(a) = 1$) were from the group with $C = 1$, as compared to only 30% of the women who did not exercise ($T^r(a) = 0$). As a result, the means defined by treatment received, equal to 98.8 mg/dL for $T^r(a) = 1$ and 95.2 mg/dL for $T^r(a) = 0$, differ by 2.6 mg/dL—failing to capture either the causal effect of exercise or assignment to exercise. The explanation is of course that $T^r(a)$ is confounded by C .

Thus, as in Sect. 9.1.4, we could only hope to obtain an unbiased estimate of the causal effect of treatment in an analysis according to treatment received by successfully modeling the effects of C . Table 9.24 shows the data from Table 9.23 rearranged. Within strata defined by C , the differences in mean glucose levels by $T^r(a)$, or treatment received, accurately estimate the causal effect of exercise. Of course, this depends on the fact that all confounding of adherence to treatment assignment is captured by the measured covariate C . In practice, this would be a substantial and unverifiable assumption.

Table 9.24 Analysis by treatment received, controlling for $\mathcal{C}$

$\mathcal{C} = 0$			$\mathcal{C} = 1$	
$T^r(a)$	Y		$T^r(a)$	Y
0	105		0	94
0	96		0	96
0	99		0	98
0	102		1	95
0	98		1	93
1		95	1	91
1		97	1	98
1		102	1	93
0	99		1	97
0	101		1	91
Means	100	98		96 94

9.8.3 Instrumental Variables

We saw in Sect. 9.7 that randomized treatment assignment can be used as an instrument for treatment received, meeting all three IV assumptions in a well-conducted trial. Following Sect. 9.7.2, the IV estimate of the causal effect of an exposure could be calculated as an estimate of the effect of the instrument on the outcome, divided by an estimate of the effect of the instrument on the exposure; if blinding is preserved, $\mathcal{C}$ could be omitted from (9.13) to (9.15). Thus, the IV estimate of the causal effect of treatment received is

$$\hat{\beta}_1^{\text{IV}} = \frac{\hat{\beta}_1^{\text{ITT}}}{\hat{\mathbb{E}}[T^r(1) - T^r(0)]}.$$

(9.18)

The numerator of (9.18) can be estimated using an unadjusted comparison by treatment assignment, and the denominator by the difference in the proportions receiving treatment among those assigned to treatment and control. In Sect. 9.8.1.1, we showed that the ITT estimate of the effect of exercise on glucose levels is -1.2 mg/dL, and that the proportions exercising in the groups assigned to treatment and control were 0.8 and 0.2, respectively. Thus, the IV estimate is $-1.2 / (0.8 - 0.2) = -2.0$ mg/dL, the causal effect of exercise on glucose levels in our example.

9.8.4 Principal Stratification

Another way to motivate (9.18) is through so-called *principal stratification* (Frangakis and Rubin 2002). Under this view, there are four unobservable principal strata in the population, defined by adherence to assigned treatment:

- (1) compliers, who comply with treatment or control as assigned
- (2) always-takers, who take treatment whether assigned to treatment or control
- (3) never-takers, who would not comply if assigned to treatment
- (4) defiers, who would take treatment if and only if assigned to control.

In many applications, defiers are assumed not to exist, under so-called *monotonicity* assumptions. The need for this assumption is made clear below. In Table 9.23, again viewed as a potential outcomes experiment, there are 12 compliers with $T^r(1) = 1$ and $T^r(0) = 0$, four never-takers, with $T^r(1) = T^r(0) = 0$, and four always-takers, with $T^r(1) = T^r(0) = 1$. Stratum membership is unobservable because in most trials we only get to see each study participant under one assignment.

Using our earlier notation, and making the standard assumption that there are no defiers, it is straightforward to check that $T^r(1) = 1$ for compliers as well as always-takers and 0 for never-takers, while $T^r(0) = 1$ for always-takers and 0 for compliers and never-takers. In addition, $E[T^r(1)]$, the proportion receiving treatment when assigned to it, includes compliers plus always-takers, while $E[T^r(0)]$, the proportion receiving treatment when assigned to control, only includes always-takers—provided there are no defiers. In that case, $E[T^r(1) - T^r(0)]$ is the proportion of compliers in the population. We note that more complicated estimation methods would make it possible to relax this requirement.

Finally, the causal effect of treatment assignment, $E[Y(1) - Y(0)]$ equals β_1 for compliers, but is 0 for always- and never-takers—because treatment received does not vary for these groups (assuming that there are no indirect effects of treatment assignment). Thus, under this stratification of the population, the ITT effect of treatment assignment can be viewed as the weighted average of β_1 , now defined as the causal effect of treatment *among compliers*—sometimes referred to as the *complier-averaged causal effect*, or CACE (Little and Rubin 2000)—and the null effects among always-takers and never-takers, where the weights are given by the proportions of the population in each subgroup. Letting $Pr(S = s)$ denote the proportion of the population in stratum 1 (compliers), 2 (always-takers), or 3 (never-takers), we can write

$$\begin{aligned}\beta_1^{\text{ITT}} &= \beta_1 \Pr(S = 1) + 0 \times \Pr(S = 2) + 0 \times \Pr(S = 3) \\ &= \beta_1 E[T^r(1) - T^r(0)].\end{aligned}\tag{9.19}$$

Thus, we can use a linear model to estimate β_1^{ITT} , the difference in the proportions actually receiving treatment by arm to estimate $\Pr(S = 1)$, and the ratio of these two estimates to estimate β_1 (Problem 9.12).

In summary, for this simple case, the IV and principal stratification estimators of the causal effect of treatment are the same. Finally, we note that principal stratification is a more general approach, applicable in many other settings.

9.9 Summary

In this chapter, we take one contemporary approach to understanding causation, based on *potential outcomes*, only one of which is the observed outcome at the actual level of exposure, while the others are outcomes that would be observed at other possible levels of exposure. This led naturally to the definition of causal effects as differences in potential outcomes, averaged across an appropriate population. We focused on estimating average causal effects in observational studies with a single binary exposure or treatment variable. The potential outcomes framework was also useful for clarifying confounding and mediation, both common themes throughout the book.

When all potential confounding variables are measured, standard regression techniques covered in other chapters can often be used to estimate average causal effects. For linear models this can be straightforward, but for non-linear models, in particular the logistic model for binary outcomes, additional steps are required. We focused on *potential outcomes estimation*, which can be seen as imputing the missing potential outcome of interest, and also discussed inverse probability weighting (IPW).

When the number of potential confounders is large but a binary or failure time outcome is uncommon, propensity scores are a robust method for strengthening causal inference. We showed why care must be taken in specifying the model used to estimate the scores, in checking balance and overlap, and in deciding how to use the scores in the estimating the causal effect of exposure—for example, as a 5-level category or restricted cubic splines. We also showed how propensity scores can be used to estimate average treatment effects in the treated, using potential outcomes estimation, matching, or *standardized mortality weights*.

Specialized methods are frequently required to strengthen causal inference when exposures and confounders are time-dependent. We focused on IPW as well as *nested new user designs*, and will sketch an alternative, *G-estimation*, in Sect. 9.10.

Finally, we described instrumental variables, which, in contrast to the other methods we discuss, can strengthen causal inference in contexts where all potential confounding variables have not been measured. While instrumental variables do require other substantial, unverifiable assumptions, they can be useful in randomized trials with incomplete adherence for estimating the causal effect of treatment among compliers, and in helping to clarify why a trial provides little or no information about the effect of treatment in noncompliers.

9.10 Further Notes and References

Causal inference is a rapidly expanding field, and many alternate approaches to estimation and inference are in active development. See [Pearl \(2009a\)](#) for an introduction to modern causal inference, and a useful discussion distinguishing

causal analyses from those that focus primarily on detecting associations. [Pearl \(2009b\)](#) provides a book-length treatment of these issues, and also illustrates the link between directed acyclic graph representation of causal relationships (covered in Sect. 10.2.5) and methods for estimation of causal effects. [Hernán and Robins \(2011\)](#) provide more complete coverage of many of the methods discussed here, and give more detail on the important topic of time-dependent confounding introduced in Sect. 9.5. [Gelman and Hill \(2007\)](#) also give more detail, and provide examples using R.

Potential Outcomes Estimation

This procedure has a long history, and has been variously called *standardization* ([Lane and Nelder 1982](#); [Hernán and Robins 2011](#)), *G-computation* ([Robins et al. 1999](#)), and most recently *regression estimation* ([Schafer and Kang 2008](#)).

Exposures and Treatments

In defining causal effects, we deliberately use the term *exposure* in most contexts, reserving *treatment* for specific cases, including the example used repeatedly in this chapter of phototherapy for treatment of neonatal jaundice. This terminology reflects our sense that we can reasonably consider the causal effects of exposures even when they are difficult or impossible to manipulate. For example, the BRCA1 and BRCA2 genetic mutations have solidly established causal effects on risk of breast and ovarian cancer. Our thought experiment makes it possible to think about potential outcomes with and without the mutations, even though they are unmodifiable.

Implicit Randomized Trials

In framing a causal question that we would like to answer using observational data, it is often helpful to think of an implicit randomized trial that might provide the answer. For example, quite different trials would be used to estimate the effect of new use of a treatment and the effect of continuing use among current users. If our interest is in the effect of new use, the implicit trial strongly suggests we should focus on new and never users in the observational cohort, and exclude prevalent users, as discussed in Sect. 9.5.4. Furthermore, as [Hernán and Robins \(2011\)](#) point out, this can help avoid posing ill-defined questions about the effects of conditions like obesity, which may reflect different sources including genetics as well as lifestyle. In our own simple example, exercise would benefit from sharper definition.

Propensity Scores

Improvements of propensity score methods are a topic of active research, and a number of alternatives addressing current problems have appeared in the scientific literature. One potential advance is use of data adaptive methods developed for prediction problems, as discussed in Sect. 10.1.4, to select the model for the propensity scores. This approach may minimize confounding without overfitting. Another promising avenue involves the use of so-called *doubly robust* methods, which provide consistent results even if one of the models is misspecified. For example, *targeted maximum likelihood* generalizes standard regression adjustment for the propensity score via an iterative procedure based on considerations from the theory of semiparametric models (Rosenblum and van der Laan 2010). The resulting estimates can be shown to improve on conventional propensity score adjustment in terms of bias and variance, especially in situations where one of the component models is wrong. Of course, even doubly robust approaches have limitations when important variables are omitted and/or when both models are misspecified (Kang and Schafer 2007).

Time-Dependent Treatments

Seminal work on models using IP weights to deal with time-dependent confounder–mediators of time-dependent treatments includes Robins et al. (1999); Robins et al. (2000); and Hernán et al. (2001); Fewell et al. (2004) give more detail about implementation of models using time-dependent IP weights in Stata. For a clear in-depth discussion of this approach, as well as an example of implementing these models in SAS, see Hernán et al. (2000); Ko et al. (2003) treat the repeated measures case with an HIV example, and show how to conduct sensitivity analyses assessing the possible influence of unmeasured confounding.

G-Estimation

An alternative for estimating the effects of time-dependent treatment with time-dependent confounder–mediators with survival outcomes is the *structural nested failure time model* (SNFTM). In contrast to proportional hazards models, including the Cox model, in which treatment is assumed to act multiplicatively on the baseline hazard for the untreated, this procedure is based on the *accelerated failure time* (AFT) model, under which treatment is assumed to act by expanding or contracting a baseline failure time that would be observed in the absence of treatment.

SNFTMs make use of an ancillary model for receiving treatment, assumed to depend on measured confounders, previous treatment history, and, in this case, one additional covariate. Specifically, using a procedure called *G-estimation* (not to be confused with G-computation), potential failure times that would be observed in the absence of treatment can be calculated under the assumed AFT model

from the observed failure times and treatment patterns, using a candidate value of the treatment effect parameter. These calculated potential no-treatment failure times are then included as the additional covariate in the ancillary model for receiving treatment. In practice, a transformation of the failure times must be used to accommodate censoring.

The rationale for G-estimation is that under the assumption of no unmeasured confounders, receiving treatment should not depend on the potential failure time that would be observed in the absence of treatment, after accounting for measured confounders and previous treatment history. Accordingly, the G-estimate of the causal effect of treatment is the candidate AFT treatment parameter value under which the calculated no-treatment potential failure times have no independent association with receiving treatment in the ancillary model. Thus, the G-estimate of the treatment effect is the value most consistent with no uncontrolled confounding of treatment. A special algorithm is required to obtain this estimate and a CI.

[Hernán et al. \(2005\)](#) provide a clear explication of SNFTMs and G-estimation, including methods for handling censored data and calculating confidence intervals. A downloadable Stata command `stgest`, detailed in [Sterne and Tilling \(2002\)](#), implements the procedure. Applications of SNFTMs include [Robins et al. \(1992\)](#); [Mark and Robins \(1993\)](#), [Robins and Greenland \(1994\)](#), [Witteman et al. \(1998\)](#), [Keiding et al. \(1999\)](#) and [Tilling et al. \(2002\)](#). As in models using IP weights, the models for treatment as well as outcome must be correctly specified.

Mediation

Causal approaches to assessment of mediation are under active development, and a range of solutions has been proposed. Estimation and inference for causal controlled and natural direct effects, including conditions for valid estimation using standard regression, are summarized in [Petersen et al. \(2006\)](#) and [VanderWeele \(2009\)](#).

Instrumental Variables

See [Martens et al. \(2006\)](#) for a clear explication of the roots of IV analysis in structural equation models. [Angrist et al. \(1996\)](#); [Heckman \(1997\)](#); [Martens et al. \(2006\)](#) and [Hernán and Robins \(2006\)](#) provide careful examinations of assumptions in several IV analyses, pointing out reasons to question them specific to the cases they examine, and showing the likely effects of potential violations. [Hernán and Robins \(2006\)](#) discuss the conditions under which the causal effect estimated using IVs might have wider interpretations. [Greene \(1998\)](#) and [Chib and Hamilton \(2002\)](#) motivate the extension to binary exposures and outcomes using probit models. [Angrist and Pischke \(2009\)](#) provide broad but non-technical coverage of IVs. [Baum et al. \(2003\)](#) explain methods of model assessment and their implementation in Stata.

Trials With Incomplete Adherence

In introducing methods that can be used to estimate the causal effects of treatment in clinical trials with incomplete adherence to assigned treatment, we have focused on the relatively simple case of all-or-nothing adherence, and on two of the more straightforward approaches that can be used to address it. [Bellamy et al. \(2007\)](#) explain in detail the assumptions underlying these approaches, and also describe an alternative approach using so-called *structural mean models*, of which the SNFTM assumed in G-estimation is one example.

More complicated approaches are required to estimate the causal effects of treatment in trials where adherence to assigned treatment can range from complete to nil; examples include trials of treatments that must be taken regularly over the course of the study, including medications, and, for that matter, exercise, as in our example. [Efron and Feldman \(1991\)](#) proposed an early solution to this problem by assuming a deterministic relationship between adherence under assignment to placebo and active treatment. [Jin and Rubin \(2008\)](#) show how principal stratification can be extended to cover this case, emphasizing how their approach clarifies the assumptions that underlie the analysis.

Other New Developments

A number of important topics were omitted from this chapter or covered only briefly, including applications to treatment variables that have more than two categories or are continuous, methods for investigating the causal effects of dynamic treatments ([Van Der Laan and Petersen 2007](#)), and causal estimation of direct and indirect effects ([Petersen et al. 2006](#)).

9.11 Problems

Problem 9.1. In the example in Sect. 9.1.3, the overall effect of C is in part mediated by its effect on $\mathcal{E}$. We defined the *direct* effect of C on Y as -4 mg/dL. Use the results in Table 9.2 to determine the *overall* causal effect of C on Y .

Problem 9.2. Show that in our simple example in Sect. 9.1, potential outcomes estimation and inverse weighting are doing essentially the same thing.

Problem 9.3. Using the WGCS data, posted on the book website, estimate the conditional odds-ratio for the effect of Type A temperament (`dibpat`) on CHD (`chd69`) using a logistic model to adjust for age, BMI, SBP, cholesterol levels, and smoking. Now use the `margins` command or data duplication to obtain estimates of the marginal odds-ratio and absolute risk difference. Do the conditional and marginal odds-ratios differ by much? Why or why not? Would you be willing to interpret the resulting estimates as causal? Why or why not?

Problem 9.4. Using the HT and statin use example in Sect. 4.6.1, show that if we first centered the `statins` indicator, β_1 in (4.10) would be interpretable as the average causal effect of HT. Contrast this with the interpretation of β_1 if `statins` is used in its original form as a 0–1 indicator for statin use. *Hint:* Derive the expression for the conditional effect of HT on LDL, then take the average of this expression across the entire sample.

Problem 9.5. Using the UNOS data on the book website, estimate the marginal effect of donor type (cadaveric vs living) on 5-year mortality risk, adjusting for recipient age and sex, donor age (`age_don`), HLA match (`hlamat`), graft status (`graf_stat`), and previous treatment (`prev_ki`). *Hint:* Use data duplication to estimate predicted 5-year risk for each participant with both the actual and potential donor type. The `basesurv` option for `stcox` returns an estimate of the baseline survival function at the observed follow-up time for each observation, whether it is an event or censored. Isolate the observation with the largest follow-up time less than 5 years, and use that value to calculate 5-year risk for each observation (both actual and potential) as

$$F(5) = 1 - S_0(5)^{\exp(\eta_{ij})}, \quad (9.20)$$

where $S_0(5)$ is the baseline survival estimate for 5 years, and $\eta_{ij} = \mathbf{X}_{ij}\beta$ is the linear predictor estimated using the postestimation `predict` command for each participant i with living ($j = 1$) and cadaveric ($j = 0$) donor. A do-file implementing a solution is also posted as Problem 9.5 do.

Problem 9.6. Suppose that in the phototherapy example, the co-intervention of switching to formula had been ascertained, but the overall sample is considerably smaller, with only 32 outcome events, rather than 128. What approach would you use for estimating the effect of phototherapy, and why?

Problem 9.7. In the propensity score analysis of the effect of phototherapy, we found some evidence for lack of overlap between treated and untreated infants. How would you address this problem?

Problem 9.8. Use propensity scores in combination with Cox models for time (`fu`) to death, to re-evaluate the effect of donor type (`txtype`) on survival following pediatric kidney transplant from Problem 9.5. Using your propensity scores, check balance, overlap of the living and cadaveric donor groups, and evidence for positivity violations. Implement models using quintile, decile, and a 5-knot restricted cubic spline in the propensity scores. Are the results consistent with standard adjustment? What would you do to address evidence for lack of overlap?

Problem 9.9. Consider an analysis using an IP weighted model. How would you check for violations of the assumption of constant treatment effects? If you found such a violation, how could the model be modified to accommodate it? And in that case, how would you estimate that hazard ratio for the comparison of always-on versus always-off treatment patterns?

Problem 9.10. Researchers at Kaiser in Northern California wanted to evaluate the effect of use of their mail-order pharmacy service on adherence to medications. Some confounder information was available from administrative databases, including age, sex, race/ethnicity, smoking, depression, and other co-morbidities, and whether the medication was covered by insurance, but there was concern about unmeasured confounders. Accordingly, they considered distance from the nearest brick-and-mortar Kaiser pharmacy to each member's residence as an instrument. Consider this potential instrument in terms of its association with mail-order use, unconfoundedness, and possible indirect effects on the outcome not mediated by mail order use. What, if anything, could we do statistically to assess these assumptions?

Problem 9.11. Suppose we tried to check the assumption that the entire effect of a proposed instrument on the outcome is mediated by the exposure of interest by regressing the outcome on exposure, the instrument, and measured confounders, on the hypothesis that if there is no direct effect of the instrument on the outcome, it should appear unimportant in this regression. Using a directed acyclic graph, as described in Sect. 10.2.5, show that in the presence of unmeasured confounding of the exposure–outcome relationship (the motivation for use of an instrumental variable), exposure is a collider on a backdoor path between the instrument and the outcome and thus controlling for it will induce an association between them.

Problem 9.12. Suppose we use the simple linear model

$$E[Y|T^a] = \beta_0 + \beta_1^{\text{ITT}} T^a, \quad (9.21)$$

to estimate the ITT effect of treatment assignment based on data from a randomized trial. Show that fitting (9.21) would result in a biased estimate of the causal effect of treatment. Specifically, show that

$$E\left[\hat{\beta}_1^{\text{ITT}}\right] = \beta_1 (E[T^r(1) - T^r(0)]). \quad (9.22)$$

where β_1 is the causal effect of treatment received, and $E[T^r(1) - T^r(0)]$ is the expected difference in the proportions of trial participants who receive treatment in the treatment and control groups respectively.

Problem 9.13. Consider a clinical trial in which women are randomized in equal proportions to a paced respiration intervention for the control of perimenopausal hot flashes, or a wait-list control. The ITT estimate of the treatment effect was a net reduction of four hot flashes per day, after controlling for baseline frequency. However, only 70% of women assigned to the paced respiration arm adhered to the intervention, and about 10% of women assigned to control crossed over. Obtain the IV estimate of the causal effect of paced respiration on hot flash frequency. Is this estimate valid for all women, or compliers only?

Problem 9.14. Consider a placebo-controlled trial of a nitroglycerin patch to increase bone mineral density (BMD) in women with osteoporosis. The outcome is change in BMD from randomization to 12 months. Numbers of patches used is available for the duration of the trial, in both groups, providing estimates of percent compliance to treatment. Clearly, percent compliance is a postrandomization variable potentially confounded by other behaviors that may be associated with changes in BMD, including smoking, exercise, and calcium supplement use. Consider how percent compliance could be used to estimate the causal effect of treatment received. How can percent compliance in the placebo group be used to remove confounding? What could invalidate this analysis?

Problem 9.15. Describe the sense in which the potential outcomes view of causal effects can be seen a missing data problem, as described in Chap. 11, and how potential outcomes estimation and inverse weighting can both be seen as solutions to this problem.

9.12 Learning Objectives

- (1) Define an average causal effect in terms of potential outcomes.
- (2) Describe the conditions under which standard regression methods are likely to give biased estimates of causal effects.
- (3) State the conditions under which propensity scores are most useful, and understand the advantages and disadvantages of various methods of incorporating the scores in estimating the effect of exposure or treatment.
- (4) Distinguish natural and controlled direct effects, and state the conditions under which standard adjustment for a mediator does not suffice to estimate direct effects.
- (5) Describe the context in which IP weight models are particularly useful, the assumptions on which they are based, and some problems that can arise in implementing them.
- (6) State the main assumptions of an instrumental variables analysis. Describe the sense in which this approach replaces the unverifiable assumption that treatment is unconfounded with the equally unverifiable assumption that the instrument is unconfounded.

Chapter 10

Predictor Selection

Walter et al. (2001) developed a model to identify older adults at high risk of death in the first year after hospitalization, using data collected for 2,922 patients discharged from two hospitals in Ohio. Potential predictors included demographics, activities of daily living (ADLs), the APACHE-II illness-severity score, and information about the index hospitalization. A “backward” selection procedure with a restrictive inclusion criterion was used to choose a multipredictor model, using data from one of the two hospitals. The model was then validated using data from the other hospital. The goal was to select a model that best predicted future events, with a view toward identifying patients in need of more intensive monitoring and intervention.

Grodstein et al. (2001) evaluated the efficacy of hormone therapy (HT) for secondary prevention of CHD, using observational data for 2,489 women with a history of heart attack or documented coronary artery disease in the Nurse’s Health Study (NHS), a prospective cohort followed from 1976 forward. In addition to measures of the use of HT, a set of known CHD risk factors were controlled for, including age, BMI, smoking, hypertension, LDL cholesterol levels, parental heart disease history, diet, and physical activity. The goal of predictor selection was to obtain a minimally confounded estimate of the effect of HT on risk of CHD events.

The Heart and Estrogen/Progestin Replacement Study (HERS), a randomized clinical trial addressing the same research question, was conducted among 2,763 postmenopausal women with clinically evident heart disease (Hulley et al. 1998). As in the NHS, a wide range of predictors were measured at study entry. Yet in the pre-specified analysis of the main HERS outcome, the only predictor was treatment assignment. The goal was to obtain a valid test of the null hypothesis as well as an unbiased estimate of the effectiveness of assignment to HT.

Orwoll et al. (1996) examined independent predictors of axial bone mass using data from the Study of Osteoporotic Fractures (SOF). SOF was a large ($n = 9,704$) observational cohort study designed to address multiple research questions about osteoporosis and fractures among ambulatory women aged 65 and up. Predictors considered by Orwoll had been identified in previous studies, and included weight, use of medications such as HT and diuretics, smoking history, alcohol and caffeine

use, calcium intake, physical activity, and various measures of physical function and strength. All variables that were statistically significant at $P < 0.05$ in models adjusting for age were included in the final multipredictor linear regression model. The goal was to identify all important predictors of bone mass.

In each of these examples, many more potential predictor variables had been measured than could reasonably be included in a multivariable regression model. The difficult problem of how to select predictors was resolved differently, to serve three distinct inferential goals:

- (1) *Prediction.* Here, the primary issue is minimizing prediction error rather than causal interpretation of the predictors in the model. The prediction error of the model selected by [Walter et al. \(2001\)](#) was evaluated using an independent data set from a second hospital.
- (2) *Evaluating a predictor of primary interest.* In pursuing this inferential goal, a central problem in observational data is confounding, which relatively inclusive models are more likely to minimize. Predictors necessary for *face validity* as well as those that behave like confounders should be included in the model. Randomized experiments like HERS represent a special case where the predictor of primary interest is the intervention; confounding is not usually an issue, but covariates are sometimes included in the model for other reasons.
- (3) *Identifying the important independent predictors of an outcome.* This is the most difficult of the three inferential goals, and one in which both causal interpretation and statistical inference are most problematic. Pitfalls include false-positive associations, the potential complexity of causal pathways, and the difficulty of identifying a single best model. We also endorse inclusive models in this context, and recommend a selection procedure that affords increased protection against false-positive results. Cautious interpretation of weak associations is key to this approach.

In summary, *predictor selection* is the process of choosing appropriate predictors for inclusion in a multipredictor regression model. A good model should be substantively motivated, appropriate to the inferential goal and sample size, interpretable, and persuasive.

10.1 Prediction

In selecting a good prediction model, candidate predictors should be considered in terms of their contribution to reducing prediction error.

Definition: *Prediction error* (PE) measures how well the model is able to predict the outcome for new observations not used in developing the prediction model.

10.1.1 *Bias–Variance Trade-off and Overfitting*

Inclusive models that minimize confounding may not work as well for prediction as models with smaller numbers of predictors. This can be understood in terms of the *bias–variance trade-off*. Bias in predictions is often reduced when more variables are included in the model, provided they are measured and modeled adequately. Moreover, the coefficients are often nearly unbiased under the assumptions commonly made in these analyses. But as less important covariates are added to the model, precision may start to erode, without commensurate decreases in bias. The larger models may be *overfitted* to the idiosyncrasies of the data, and, thus, more poorly predict new, independent observations. We can minimize PE by optimizing the bias–variance trade-off.

10.1.2 *Measures of Prediction Error*

For continuous outcomes, R^2 is a potential measure of PE. A function of the residual sum of squares (RSS), R^2 depends on the averaged squared distance between the predictions, or fitted values, and the observed outcomes, and so is a natural metric for PE.

For binary outcomes, the analogous Brier score, also given by the average of the squared distances between the predicted and observed outcomes, is not commonly used. A much more widely used PE measure is the area under the ROC curve, or equivalently the *C*-statistic, introduced in Sect. 5.2.6. The analogous PE measure for Cox models is the *C*-index. The *C*-statistic and *C*-index are both measures of *discrimination*—that is, how effectively the model can distinguish between events and nonevents, or correctly order the timing of two events.

Both the *C*-statistic and *C*-index are rank-based measures, and can be insensitive to improvements in prediction as a result (Pencina et al. 2008). To see this, note that in calculating the *C*-statistic, two correctly ranked event/nonevent pairs for which the predictions differ by five and 95 percentage points would be treated alike, although the model much more clearly distinguishes the second pair. Likewise, in calculating the *C*-index, we ignore differences between failure times as well as between fitted risks.

In addition to discrimination, measures of *calibration* for logistic and Cox models assess the agreement between fitted and observed risks. The Hosmer–Lemeshow statistic presented in Chap. 5 measures calibration of the logistic model, comparing fitted and observed events within deciles (or other groupings) of the fitted risks. Analogs have been proposed for the Cox model (Parzen and Lipsitz 1999; van Houwelingen 2000). One often-used measure of calibration for the Cox model is to compare average fitted probabilities of an event within a fixed time

period to observed probabilities nonparametrically estimated using Kaplan–Meier curves. For example, [Cook et al. \(2006\)](#) compared fitted and observed ten-year risks for cardiovascular events within two-point intervals of the model-based risk score.

10.1.3 Optimism-Corrected Estimates of Prediction Error

To select a model that minimizes prediction error, we need an accurate estimate of the target PE measure that does not overstate the ability of the model to predict the outcome for new, independent observations—in brief, one that is not *optimistic*.

10.1.3.1 Optimism of Naïve Estimates of PE

To see why optimism is an issue, consider R^2 , the proportion of variance explained by a linear regression model, and a potential measure of PE. It increases with each additional covariate, even if the added predictor provides minimal information about the outcome. At the extreme, $R^2 = 1$ in a model with one predictor for each observation. This happens because the same observations are used to estimate the model and assess its predictiveness. Selecting predictors simply to maximize R^2 would almost surely result in overfitted models.

10.1.3.2 Simple Alternatives to R^2

An alternative less subject to optimism is adjusted R^2 , which is calculated by penalizing R^2 for the number of predictors in the model. Thus, when a variable is added, adjusted R^2 increases only if the increment in R^2 is larger than the increment in the penalty. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are analogs which impose stiffer penalties for each additional variable—specifically, penalties against minus twice the log-likelihood, another potential measure of PE. With AIC, the penalty is $2p$, where p is the number of predictors in the model; with BIC, it is $p \log N$, where N is the sample size.

The AIC criterion is relatively liberal, allowing for the inclusion of simple continuous or binary predictors with P -values < 0.16 . In contrast, the P -value cutoff imposed by BIC for such predictors grows progressively stricter with sample size, requiring $P < 0.05$ in samples of about 50, $P < 0.01$ in samples of 500, and $P < 0.009$ in samples of 1,000, and, thus, leads to increasingly parsimonious models, relative to AIC. Both measures depend on the number of additional coefficients, and so set the bar higher for inclusion of restricted cubic splines or multicategory predictors.

In Stata the `regress` command prints adjusted R^2 by default, and AIC and BIC can be obtained for linear, logistic, Cox, and other models using the postestimation command `estat ic`. The best prediction model is taken to be the one that maximizes adjusted R^2 , or minimizes AIC or BIC.

10.1.3.3 Generalized Cross-Validation

In contrast to indirect, theoretically-based measures such as adjusted R^2 , AIC, and BIC, more direct methods for obtaining nonoptimistic estimates of PE are based on *cross-validation*, which uses distinct, independent sets of observations to estimate the model and to evaluate PE.

10.1.3.4 Development and Validations Sets

The most straightforward example of cross-validation is the split-sample approach, in which the parameter estimates are obtained from a so-called development set, but then PE is evaluated in an independent validation set by comparing observed outcomes to expected values calculated using development set parameter estimates in combination with validation set covariate values.

In some implementations, the development and validation sets are obtained by splitting a single data set, often with two-thirds of the observations randomly assigned to the development set. Other implementations, as in Walter's analysis of posthospitalization mortality among high-risk older adults, use an independent sample as the validation set. Precisely because the validation set is not sampled under exactly the same circumstances, this procedure may do a better job of forecasting the utility of the prediction model in practical use. [Altman and Royston \(2000\)](#) discuss the merits of internal and external validation sets.

Splitting one data set into development and validation sets is less efficient than the alternative discussed next, but also easier to implement, and commonly more credible to nonstatisticians, in particular when the validation set is truly external.

10.1.3.5 h -Fold Cross-Validation

A more efficient alternative to splitting the data into development and validation sets is *h-fold cross-validation*. With this method, the entire data set is used both for development and validation of the model. The procedure works in five basic steps.

- (1) The data are randomly divided into h mutually exclusive subsets of equal size.
- (2) Each of the h subsets is set aside in turn, and the model is estimated using the remaining observations.
- (3) Using the parameter estimates from each of those h models, the statistics necessary to calculate the target measure of PE are estimated for the corresponding set-aside observations.
- (4) A summary estimate of PE is then calculated using the statistics from all h subsets.
- (5) The h -fold procedure is repeated k times, using a new division of the data each time, and then the k summary estimates of PE are averaged.

Values of $h = 5-10$ and $k = 10-20$ are reasonable.

Table 10.1 Ten-fold cross-validation of the area under the ROC curve

```
. quietly logistic chd69 age chol sbp bmi smoke
. predict fitted, pr

. * Naive estimate of area under the ROC curve
. roctab chd69 fitted
```

Obs	ROC Area	Std. Err.	-Asymptotic Normal-- [95% Conf. Interval]	
3142	0.7333	0.0156	0.70270	0.76395

```
. Step 1: divide data into 10 mutually exclusive subsets
. xtile group = uniform(), nq(10)
. gen cv_fitted = .
. forvalues i = 1/10 {
2.
.     * Step 2: estimate model omitting each subset
.     qui logistic ytemp age chol sbp bmi smoke if group~='i'
3.     qui predict cv_fittedi, pr
4.
.     * Step 3: save cross-validated statistic for each omitted subset
.     qui replace cv_fitted = cv_fittedi if group==`i'
5.     qui drop cv_fittedi
6. }

.
. * Step 4: calculate cross-validated area under ROC curve
. roctab chd69 cv_fitted
```

Obs	ROC Area	Std. Err.	-Asymptotic Normal-- [95% Conf. Interval]	
3142	0.7277	0.0158	0.69386	0.75566

Cross-validation is easy to implement in Stata. In Table 10.1, we first re-run the logistic model for CHD risk shown in Table 5.6, save the fitted probabilities, and calculate the naïve estimate of the area under the ROC curve (ROC Area), equivalent to the *C*-statistic. Then, the WCGS data are randomly divided into ten mutually exclusive subsets, and the model is refitted ten times, omitting in turn each of the ten subsets from the data used in estimation of the model. However, predicted values are calculated for the entire data set; we also exploited this feature of Stata for potential outcomes estimation in Table 9.6. The cross-validation fitted values for the omitted subsets are collected in the new variable `cv_fitted`, and in a final step, the cross-validation estimate of the area under the ROC curve is calculated using these fitted values and the observed outcomes. For clarity, we have omitted the fifth step of repeating the procedure 10–20 times, but the additional programming is simple enough.

As expected, the optimistic naïve estimate of the area under the ROC curve shown in Table 10.1 is larger than the cross-validated estimate. However, the difference is small, suggesting that the simple logistic model for CHD events is not badly overfitted.

10.1.4 Minimizing Prediction Error Without Overfitting

A model that fits well, including all important predictors and accurately capturing nonlinear effects as well as interactions, should provide better prediction than a poorly specified model that excludes some important predictors, inaccurately models the effects of others, and includes unimportant predictors.

Earlier chapters have shown how to ensure that nonlinear effects of continuous predictors are adequately modeled, essentially by examining the relationship between predictor and outcome, using diagnostic plots or models including restricted cubic splines or interactions. And later in this chapter, in discussing predictor selection for the second inferential goal of evaluating the causal effect of a primary predictor of interest, we recommend methods to ensure that all measured confounders are included and adequately modeled, again by examining alternative models for the outcome.

However, in this context, the danger is that examining relationships with the outcome can easily lead to overfitting, resulting in a model that does not perform well in external validation data. Overfitting can be minimized using four strategies:

- (1) Pre-specify well-motivated predictors and how to model them
- (2) Eliminate predictors without using the outcome
- (3) Use the outcome, but cross-validate the target measure of PE
- (4) Use the outcome, and shrink the coefficient estimates.

10.1.4.1 Pre-specifying Well-Motivated Predictors

One primary strategy for avoiding overfitting is to depend so far as possible on a priori specification of well-motivated candidate predictors. In areas of clinical research where prognostic factors have been thoroughly studied, expert opinion, grounded in the literature, may provide considerable guidance, and meta-analyses can be especially reliable measures of variable importance. This strategy would also rely on the literature to determine how the effects of continuous covariates should be modeled—that is, to select functional form—rather than using the data to guide these decisions.

In some well-studied areas, this step may be sufficient to choose a good prediction model, without the need for subsequent elimination of predictors driven by the development data. Furthermore, while the bias–variance tradeoff may suggest the need for parsimony, a wisely-chosen set of pre-specified predictors may often work better in external validation data than a subset of those predictors chosen by looking at their relationships with the outcome in the data used for model development ([Harrell 2005](#); [Steyerberg 2009](#)).

10.1.4.2 Predictor Elimination Without Using the Outcome

A second-line strategy for avoiding overfitting is to eliminate candidates without looking at predictor–outcome relationships, but taking account of the effective sample size m , defined as the number of observations in linear regression, the number of events in Cox regression, and the smaller of the numbers of observations with or without the outcome in logistic models (Harrell 2005).

For example, summary variables can be chosen for predictor domains: LDL and HDL cholesterol levels might be chosen on substantive grounds from among the larger set of lipid measures including total cholesterol, triglycerides, and the HDL/LDL ratio. Practical considerations may also be important. In particular, expensive, invasive, risky, and relatively unreliable tests can be ruled out if more practical alternatives are available. Predictors with fewer missing values in the development data are also preferable, in particular, if missing values reflect the likely difficulty of obtaining the measurement in practice.

Linearity would of course be a concern in modeling the effect of continuous covariates such as LDL cholesterol. To address this issue, a related means of outcome-free predictor elimination is to allocate spline knots based on prior estimates of variable importance and m . Thus, if a predictor has been of primary importance and had strongly nonlinear effects in earlier research, and m allows it, a four- or five-knot spline may be pre-specified. In contrast, a less important predictor or one known to have approximately linear effects can be treated more simply. Smaller samples and fewer outcomes may also limit how flexibly we can model continuous effects.

Principal components is a more complicated alternative for reducing the number of parameters to be estimated without using the outcome, and has been shown to work well in some studies (Harrell et al. 1984, 1996). This method summarizes a large set of correlated continuous predictors by a much smaller set of uncorrelated summary variables, or principal components, chosen to explain most of the variance in the predictors. This simplification is achieved without reference to the outcome.

This approach does have some drawbacks. One is that the principal components may not be substantively interpretable, which is desirable for face validity, although not really needed for prediction. In addition, principal components capturing the greatest variability in the *predictors* are not guaranteed to capture the most variability in the *outcome*, although with well-chosen predictors this is likely. Finally, this procedure does not reduce the number of underlying variables that need to be measured, and so makes it more difficult to focus on easily-obtained predictors with fewer missing values.

A widely used guideline suggests that at most $m/10$ or even $m/20$ candidate predictors should be *considered* for inclusion in the prediction model. Note that each component of a complicated predictor counts as an additional candidate, so that if we pre-specify a restricted cubic spline with five knots to represent a continuous predictor, the number of candidates increases by four, the required number of spline basis variables. Motivated by simulation studies of the precision of predictions based

on Cox models, this guideline is approximate, but does suggest that large samples are necessary for developing valid prediction models, in particular, when variable selection is required.

10.1.4.3 Model Selection Using the Outcome and Cross-Validation

In the common case where the combination of prespecification and outcome-free predictor elimination does not adequately reduce the number of candidate predictors, an effective strategy is to use exhaustive screening of all subsets of the remaining candidate predictors. Crucially, to avoid overfitting, this final screening step must use cross-validation of a selected target measure of PE to help identify the most predictive of these models. This is the approach used in most modern algorithms for prediction model development, including the Deletion/Substitution/Addition (DSA) algorithm (Molinari and van der Laan 2004). In this procedure, implemented in R, the candidate predictors, including polynomial terms and interactions, are efficiently screened using h -fold cross-validation of a selected measure of PE.

Efficient screening is an important issue in this context. For example, even if the number of candidate predictors has been reduced to a seemingly tractable eight, the number of subsets of all sizes is $2^8 = 256$. And even if an indirect optimism-corrected measure of prediction error—adjusted R^2 , AIC, or BIC—is used in place of cross-validation, this represents an onerous computing task without programs like DSA that automate the screening.

However, screening can be made more practicable if some of the remaining candidates are always to be included on a priori grounds. For example, if five of eight candidate variables were to be included by default, then only $2^3 = 8$ models must be screened. But if many models have to be screened, programming of the procedure, including any intermediate steps, will almost surely be required. We illustrate this approach in Sect. 10.1.6 below.

While this screening procedure should help us find a good predictive model without overfitting, it is important to note that the cross-validated estimate of PE for the selected model will be at least slightly optimistic, not because we use the same data to estimate model parameters and evaluate PE—the source of optimism in naïve PE estimators—but because of the selection.

10.1.4.4 Shrinking the Coefficient Estimates

Dropping variables, on a priori or practical grounds or on the basis of a cross-validated PE measure, is equivalent to setting their coefficients equal to zero. An alternative approach is to shrink them only part way to zero. So-called *shrinkage* procedures can be motivated on the grounds that even the weaker candidate predictors specified a priori have some predictive value, and so should not be excluded outright from the model. However, because their coefficients may be less precisely estimated, better prediction may be achieved by reducing their influence. This approach is closely related to the shrinkage estimators introduced in Sect. 7.7.3.

In general shrinkage procedures impose penalties against the log-likelihood in model fitting, with the degree of penalization generally optimized using cross-validation. Le Cessie and Van Houwelingen (1992) and Verweij and Van Houwelingen (1994) discuss applications to logistic and Cox regression. These methods derive from *ridge regression* (Hoerl and Kennard 1970), which provides slightly biased but less variable estimates in linear models when the predictors are highly correlated. In ridge regression, the penalty is proportional to the sum of the squared values of the regression coefficients, with the proportionality factor commonly optimized using cross-validation. Coefficients are shrunken roughly in inverse proportion to the variance of the corresponding predictor, but no variables are omitted outright.

In contrast, the penalty imposed by the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1997) is proportional to the sum of the *absolute* values of the regression coefficients. Surprisingly, the result is that the LASSO can set the coefficients for the least important predictors to zero, effectively omitting those variables from the model, while differentially shrinking others. Thus, it is a selection as well as a shrinkage procedure. The LASSO has been implemented only for linear models in the Stata `lars` package, as so-called *least angle regression*. However, the `penalized` package in R extends both ridge regression and the LASSO to GLMs and Cox models, and incorporates cross-validation for selecting the penalty factor.

10.1.5 Point Scores

Unless a continuous predictor has strong threshold effects, we can generally achieve better prediction by keeping it continuous, modeling any nonlinearity in its effects, and avoiding dichotomization. However, one drawback, especially if splines are used to capture nonlinear effects, is that the predictions almost always need to be calculated using some electronic interface, or at least a nomogram. If the prediction model is intended for everyday clinical use, easily calculated scores assigning points to a small set of risk factors are more likely to be adopted.

For example, the Thrombosis in Myocardial Infarction (TIMI) risk score for predicting event-free survival in heart disease patients is simply calculated by counting up seven risk indications, including age ≥ 65 , having ≥ 3 CAD risk factors, coronary stenosis, ST-segment deviation, elevated serum cardiac markers, ≥ 2 recent angina episodes, and aspirin use in the last week (Antman et al. 2000). Each of the underlying predictors was dichotomized and assigned one point.

At some cost in complexity, more information can be retained by splitting continuous variables into more than two categories, with nonreference levels assigned different numbers of points. For example, D'Agostino et al. (2000) tabulate points assigned to each level of several multicategory predictors, and provide an additional table for translating the summed point scores into predicted risks.

Point systems allowing differing weights are commonly derived by rounding the regression coefficients for each binary indicator variable, after suitable rescaling so that each factor is assigned at least one point. In some cases, risk scores of this type may perform nearly as well as summary scores based on the underlying coefficients. However, considerable increases in prediction error may sometimes result ([Gordon et al. 2010](#)).

10.1.6 Example: Risk Stratification of Patients with Heart Disease

The Heart and Soul Study follows a prospective cohort of 1,024 adults with established CHD, recruited from several clinical centers in the San Francisco Bay Area in 2000–2002 ([Whooley et al. 2008](#)). Over 5,745 person-years of follow-up by the time of analysis, 272 outcome events, a composite defined by heart attack, heart failure, stroke, or death from cardiovascular causes, had been observed among 916 of these participants with complete baseline test data.

Starting from a wide range of baseline predictors, we developed two Cox models for risk stratification of this moderate-to-high risk patient population. One, requiring computer implementation, includes three continuous predictors, two of them represented by 3-knot restricted cubic splines. The second is a point score model. We selected Harrell's *C*-index as our target PE measure, and drove final model selection mainly by minimizing cross-validated estimates of this target.

Based on the knowledge of the investigators, an initial set of 36 candidate predictors was identified. On practical grounds and by choosing—without using the outcomes—the best predictor in several domains, the number was reduced to 18, under the $m/10$ upper bound of 27, but still exceeding the more conservative bound of $m/20$. While cut-points for dichotomizing continuous predictors, as required for the point score, were available from the literature, less information was available on functional form. On practical grounds, the investigators specified that the point score should include at most 7 predictors, and preferably 5 or 6, and were reluctant to consider larger continuous models.

Because the number of possible models was very large even before considering the functional form of continuous predictors, we used exploratory analysis to reduce the scope of the cross-validation screening. Specifically, using backward selection procedures, we decided that the four clearly most powerful predictors (age, left ventricular ejection fraction (LVEF), B-natriuretic peptide levels (BNP), and urinary creatinine/albumin ratios (UACR), would be included in any selected model, and that we could safely omit the four weakest (hypertension, history of heart attack, LDL, and HDL cholesterol). The remaining candidates for inclusion in the model included gender, BMI, current smoking, diabetes, C-reactive protein (CRP), chronic kidney disease (CKD), detectable troponin, congestive heart failure (CHF), physical inactivity, and poor adherence to medication.

Table 10.2 Top-scoring prediction models

Number of Predictors	Continuous			Point score		
	C-Index (%)		GOF	C-Index (%)		GOF
	CV ^a	Naïve	<i>P</i> -value ^b	CV	Naïve	<i>P</i> -value
5	76.2	76.6	0.90	73.1	74.0	0.002
6	76.2	76.9	0.50	73.9	74.5	0.07
7	76.2	76.8	0.72	73.0	74.8	0.03

^a Cross-validation.
^b Goodness of fit test due to [Parzen and Lipsitz \(1999\)](#).

In addition, we selected the functional form for continuous predictors by comparing AIC values for alternatives, in models adjusting for other powerful covariates. On this basis, we elected to treat age as linear, dichotomized LVEF at the established cutpoint of 50%, and used 3-knot restricted cubic splines for UACR and BNP as well as BMI and CRP. In additional exploratory analyses using four or five-knot splines, the cross-validated *C*-index decreased substantially, reflecting overfitting.

We then programmed algorithms in Stata to perform 10-fold cross-validation of the *C*-index for each of several hundred candidate continuous and point score models. For the point-score models, we used a simple automatic algorithm for calculating the scores based on each of the ten cross-validation development sets.

Table 10.2 shows results for 5, 6, and 7-predictor continuous and point score models with values of the cross-validated *C*-index at the observed maximum. Several comments are in order:

- The continuous models consistently do better than the point score models. The 2%–3% point improvements in the *C*-index are substantial and not easily achieved. Note that the number of *parameters* estimated for the continuous models is two greater than the number of predictors, because BNP and UACR were modeled using 3-knot splines.
- The larger models have at most slightly higher cross-validated values of the *C*-index. Moreover, continuous models with more than 7 predictors did not do substantially better than the 7-predictor model.
- The naïve and cross-validated *C*-index values are also very close, possibly reflecting optimism of the cross-validated estimate due to selection.
- Holding the number of predictors or parameters fixed, the *C*-indices for the top 5–10 models barely differed (data not shown). This illustrates that in prediction, models containing different sets of predictors may be quite competitive.

When different models are close in terms of the cross-validated target measure of PE, additional criteria may be used to decide between them, including calibration. Despite the evidence for poor fit of the point score models, model-based and Kaplan–Meier estimates of risk were in reasonably good agreement for the 6-predictor models, as shown in Fig. 10.1, as well as for the 7-predictor and larger models. Results are stratified by decile of predicted risk for the continuous model, and by point scores for the point score model.

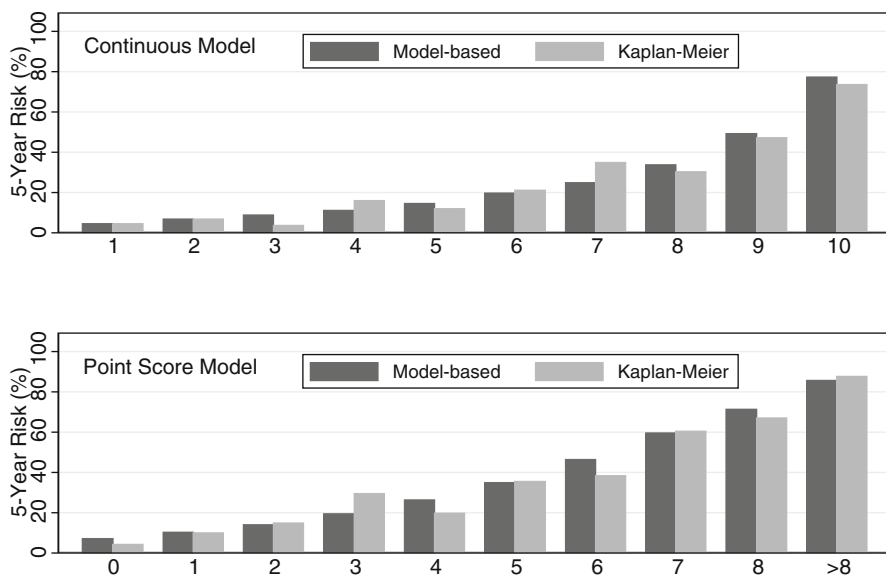


Fig. 10.1 Calibration of prediction models

Face validity and clinical convenience were also top priorities for the investigators. Current smoking and diabetes are accepted and reasonably strong cardiovascular risk factors. The best 7-predictor continuous models included either troponin or CRP, and so would have required an extra test, without improving discrimination or calibration. Accordingly, the investigators selected the 6-predictor model, including age, LVEF, BNP, UACR, current smoking, and diabetes.

10.2 Evaluating a Predictor of Primary Interest

In observational data, the main problem in evaluating a predictor of primary interest is to rule out confounding of the association between this predictor and the outcome as persuasively as possible. Potential confounders to be considered include factors identified in previous studies or hypothesized to matter on substantive grounds, as well as variables that behave like confounders by the statistical measures described in Sect. 4.4. Three classes of covariates would not be considered for inclusion in the model: covariates which are essentially alternative measures of either the outcome or the predictor of interest, and those hypothesized to mediate its effect. A diagram of the proposed causal model can be useful for clarifying hypotheses about these relationships, which can be complex, and for selecting variables for consideration.

In contrast, mediation of one confounder by another would not affect the estimate for the primary predictor nor its interpretation. Similarly, high correlation between pairs of adjustment of confounding variables would not necessarily be a compelling

reason for removing one of them, if both are seen as necessary on substantive or statistical grounds; the reason is that collinearity between confounding variables will not affect the estimate for the primary predictor or its precision. Covariates which are in some sense alternative measures of the outcome are not always easy to recognize, but should usually be excluded. For example, it would be problematic to include diabetes in a model for glucose, because diabetes is largely defined by elevated glucose. Another example is history of a potentially recurrent outcome like falling in a model for subsequent incidence of the outcome. In both examples, addition of the alternative outcome measure as a predictor to the model tends to attenuate the estimates for other, more interpretable predictors.

10.2.1 Including Predictors for Face Validity

Some variables in the hypothesized causal model may be such well-established causal antecedents of the outcome that it makes sense to include them, essentially to establish the face validity of the model and without regard to the strength or statistical significance of their associations with the primary predictor and outcome in the current data set. The risk factors controlled for in the Nurse's Health Study analysis of the effects of HT on CHD risk are well understood and meet this criterion.

10.2.2 Selecting Predictors on Statistical Grounds

In many areas of research, the potential confounders of a predictor of interest may be less well established, so that in the common case where there are many such potential confounders, a priori selection of a reasonable subset to adjust for is not a realistic option. However, the inclusion of too many predictors may unacceptably inflate the standard errors of the regression coefficients, especially in smaller samples; in logistic and Cox models bias can also be induced when too many parameters are estimated. We discuss collinearity and the numbers of predictors that can safely be included in Sects. 10.4.1 and 10.4.2. Because of these potential problems, we would like to eliminate variables that are effectively not confounders, because they demonstrate little or no independent association with the outcome after adjustment. Similarly, hypothesized interactions that turn out not to be important on statistical grounds would be eliminated, almost always before either of the interacting main effects are removed.

An easily implemented method for eliminating redundant predictors on statistical grounds is so-called backward selection. In brief, backward selection begins with full model including all pre-specified candidate predictors, then sequentially

eliminates the weaker candidates, at each step removing the predictor with the largest P -value. The advantages of backward over forward and stepwise procedures are explained in Sect. 10.4.3.

If P -value driven selection is used, we recommend a liberal criterion, to rule out confounding more effectively: in particular, only removing variables with P -values ≥ 0.2 (Maldonado and Greenland 1993). A comparably effective alternative is to retain variables if removing them changes the coefficient for the predictor of interest by more than 10% or 15% (Greenland 1989; Mickey and Greenland 1989). These liberal criteria are particularly important in small data sets, where even important confounders may not meet the usual $P < 0.05$ criterion for statistical significance.

10.2.3 Interactions With the Predictor of Primary Interest

A potentially important check on the validity of the selected model is to assess interactions between the primary predictor and important covariates, in particular, those that are biologically plausible. Especially for a novel or controversial main finding, it can add credibility to show that the association is similar across subgroups. There is no reason for concern if the association is statistically significant in one subgroup but not in the complementary group, provided the subgroup-specific estimates are similar. However, if a substantial and credible interaction is found, particularly such that the association with the predictor of interest differs qualitatively across subgroups, then the analysis would need to take account of this complexity. For example, Kanaya et al. (2004) found an interaction between change in obesity and HT in predicting CHD and mortality risk which substantively changed the interpretation of the finding. However, since such exploratory analyses are susceptible to false-positive findings, this unexpected and hard-to-explain interaction was cautiously interpreted.

10.2.4 Example: Incontinence as a Risk Factor for Falling

Brown et al. (2000) examined urinary incontinence as a risk factor for falling among 6,049 ambulatory, community-dwelling women in the SOF cohort also studied by Orwoll. The hypothesis was that incontinence might cause falling because of hasty trips to the bathroom, especially at night. But it was important to rule out confounding by physical decline, which is strongly associated with both aging and incontinence. The final model included all predictors which were associated with the outcome at $P < 0.2$ in univariable analysis and remained statistically significant at that level after multivariable adjustment. Alternative and more inclusive models with different sets of predictors were also assessed. After adjustment for 12 covariates

(age; history of nonspine fracture and falling; living alone; physical activity; use of a cane, walker, or crutch; history of stroke or diabetes; use of two classes of drugs; a physical performance variable; and BMD) weekly or more frequent urge incontinence was independently associated with a 34% increase in risk of falling (95% CI 6%–69%, $P = 0.01$).

In this example, falling was defined as a binary outcome, discussed in Chap. 5. In addition, because the outcome was observed over multiple time intervals for each SOF participant, methods presented in Chap. 7 for longitudinal repeated measures were used. A subsequent example in Sect. 10.4.2 uses a Cox proportional hazards model, covered in Chap. 6. In using these varied examples, we underscore the fact that predictor selection issues are essentially the same for all the regression models covered in this book.

10.2.5 Directed Acyclic Graphs

So-called *directed acyclic graphs* (DAGs) (Pearl 1995), a type of causal diagram, are potentially useful in determining which covariates need to be included in—and excluded from—regression models used for the second inferential goal of evaluating the effects of a predictor of primary interest. In the following example we briefly review the terminology and some key ideas, show how a DAG could be used to guide predictor selection for this inferential goal, and discuss some complications that can arise.

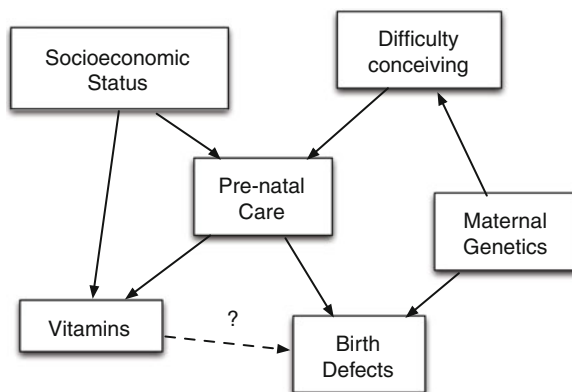
10.2.5.1 Example: Vitamin Use and Birth Defects

Suppose we would like to assess the causal effect of vitamin use on prevention of birth defects. The DAG in Fig. 10.2 identifies four common causes of vitamin use and birth defects, all of them potential confounders: pre-natal care, socioeconomic status (SES), difficulty conceiving, and maternal genetics. Vitamin use, birth defects, and the potential confounders are represented as *nodes* of the DAG, while the causal relationships between them are represented as arrows, or *directed edges*. The DAG is *acyclic* in the sense that no ordered sequence of arrows or directed edges leads back to the node from which the sequence began.

The DAG in Fig. 10.2 encodes several causal assumptions:

- Pre-natal care affects both vitamin use and risk of birth defects.
- A history of difficulty conceiving affects the likelihood that expectant mothers seek pre-natal care.
- Maternal genetics is a common cause of difficulty conceiving and birth defects.
- SES affects access to pre-natal care as well as vitamin use.

Fig. 10.2 Initial DAG for examining effects of vitamin use on birth defects



The preceding discussion of predictor selection for the second inferential goal suggests that we might want to control for all four hypothesized confounders of vitamin use. But do we really need to control for all of them? Not having to ascertain maternal genetics would save money and increase study participation, and a smaller model would likely be more efficient statistically.

10.2.5.2 Backdoor Paths

The DAG in Fig. 10.2 can be used to identify a minimum set of covariates we need to control for. To do this, we need to examine *backdoor paths* between vitamin use and birth defects. *Paths* are sequences of edges connecting two nodes, without regard to their direction. There are a total of five distinct paths connecting vitamin use and birth defects. Only one of these begins with a directed edge *from* vitamin use, specifically the path leading directly to birth defects, representing the hypothesized causal effect of interest; this is not a backdoor path. The other four paths connecting vitamin use and birth defects are backdoor paths, because they all include a directed edge leading *to* vitamin use:

- (1) Vitamin use $\leftarrow$ pre-natal care $\rightarrow$ birth defects
- (2) Vitamin use $\leftarrow$ SES $\rightarrow$ pre-natal care $\rightarrow$ birth defects
- (3) Vitamin use $\leftarrow$ pre-natal care $\leftarrow$ difficulty conceiving $\leftarrow$ maternal genetics $\rightarrow$ birth defects
- (4) Vitamin use $\leftarrow$ SES $\rightarrow$ pre-natal care $\leftarrow$ difficulty conceiving $\leftarrow$ maternal genetics $\rightarrow$ birth defects

Note that this DAG includes no paths beginning with a directed edge from vitamin use and passing through one or more nodes on the way to birth defects. Such indirect paths via mediators would not be considered backdoor paths.

10.2.5.3 Colliders

Pre-natal care is a so-called *collider* on the fourth backdoor path between vitamin use and birth defects, because it is the *common effect* of SES and difficulty conceiving. Note that pre-natal care is *not* a collider on any of the other three backdoor paths; likewise none of the other covariates are colliders on any of the four backdoor paths. Rules for determining what we need to control for treat colliders differently from other covariates along backdoor paths between exposure and outcome.

10.2.5.4 Blocking Backdoor Paths

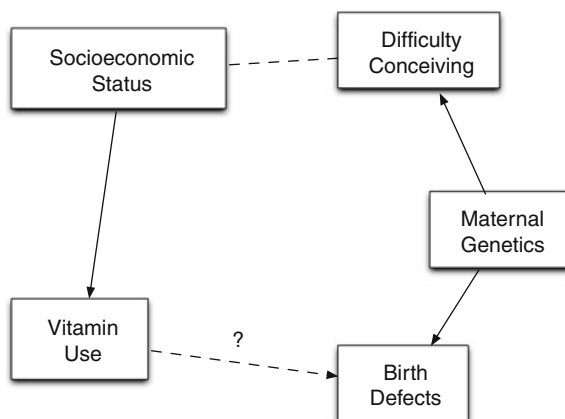
Backdoor paths between exposure and outcome may be *blocked* or remain *open*. If any of the four backdoor paths between vitamin use and birth defects remains open, we would expect to find a statistical association between them, even if there was no causal relationship; essentially, this is uncontrolled confounding. But if all the backdoor paths are blocked, then we would only expect a statistical association between vitamin use and birth defects if a causal relationship links them. Whether any of the four backdoor paths remain open depends on whether it includes a collider, and what we control for in the statistical model we use to estimate the effect of vitamin use on birth defects. Specifically,

- (1) A backdoor path is blocked, provided we control for at least one noncollider on the path. Thus, we can efficiently block the first three backdoor paths between vitamin use and birth defects by controlling for pre-natal care, because it is a noncollider on all those paths.
- (2) A backdoor path including a collider is blocked, provided we do *not* control for the collider in the statistical model. Controlling for a collider induces a negative correlation between its common causes, opening an additional backdoor path, as shown in Fig. 10.3. To block this path, the model must control for a noncollider on the newly opened path.

Thus, the DAGs in Figs. 10.2 and 10.3 imply that we could obtain an unbiased estimate of the causal effect of vitamin use on birth defects using a statistical model in which we parsimoniously controlled for pre-natal care as well as one of the other three hypothesized confounders: SES, difficulty conceiving, or maternal genetics.

This pattern of confounding relationships, examined in a slightly simpler form by Greenland et al. (1999), illustrates that controlling for one apparently sufficient confounder may not be enough, if it is also a collider. Nonetheless, the solution is simple: controlling for just one additional factor will block the new backdoor path opened by controlling for the collider. Thus, the insight gained from the DAG might still make it possible to increase the efficiency of our study, relative to the more inclusive model selection strategy discussed earlier in this section.

Fig. 10.3 Additional backdoor path opened by controlling for pre-natal care



10.2.5.5 Vulnerability to Assumptions

This result may be vulnerable to several assumptions implicit in the DAG in Fig. 10.2. Specifically, we may question whether

- *SES affects birth defects only through its effects on pre-natal care and vitamin use.* An additional pathway may result from environmental exposures, which are concentrated among the poor and minorities.
- *Difficulty conceiving affects vitamin use only through uptake of pre-natal care.* An additional pathway could be opened by the huge market for over-the-counter dietary supplements.
- *There is no direct link between maternal genetics and SES.* So-called population stratification suggests that the prevalence of genetic factors causing birth defects may differ by race/ethnicity. This opens a complicated causal pathway from maternal genetics to SES, mediated by racial and class discrimination.

If these concerns are valid, then there are three additional backdoor paths we might need to block, as shown in Fig. 10.4:

- (1) Vitamin use $\leftarrow$ SES $\rightarrow$ birth defects
- (2) Vitamin use $\leftarrow$ difficulty conceiving $\leftarrow$ maternal genetics $\rightarrow$ birth defects
- (3) Vitamin use $\leftarrow$ SES $\leftarrow$ maternal genetics $\rightarrow$ birth defects

Thus, we would need to control for pre-natal care and SES, as well as either difficulty conceiving or maternal genetics.

10.2.5.6 Colliders We Should Not Adjust For

DAGs can also help us avoid adjusting in cases where this will *induce* bias. For example, suppose we hypothesized the causal relationships in Fig. 10.5. In this

Fig. 10.4 Plausible additional causal pathways affecting birth defects

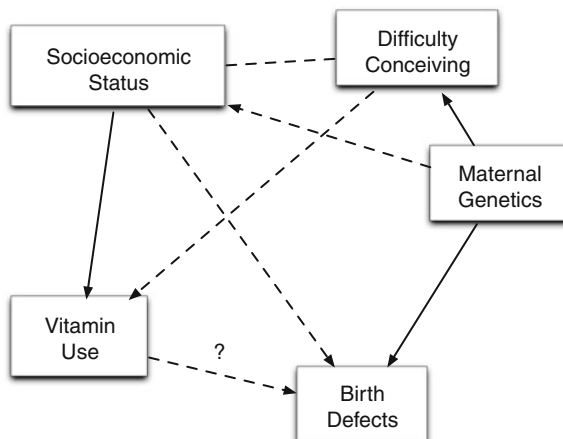
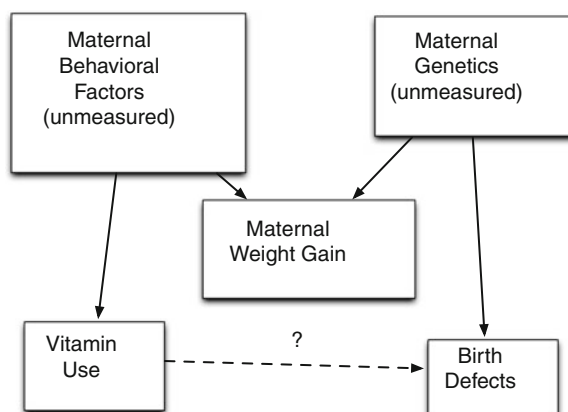


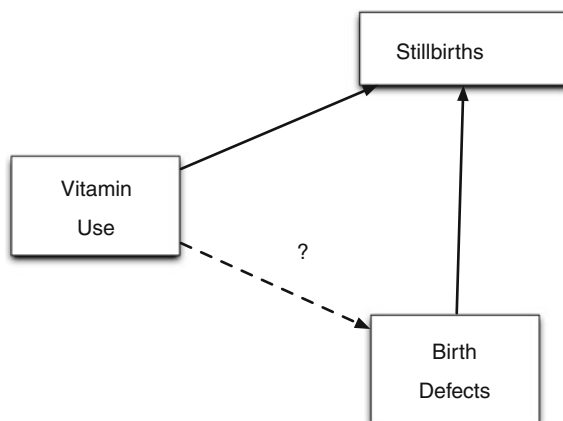
Fig. 10.5 Collider on pathway with unmeasured confounders



DAG, maternal weight gain is not a confounder of vitamin use, and so does not need to be adjusted for. As in Fig. 10.2, this depends on the *absence* of directed edges, in this case from maternal weight gain to vitamin use and birth defects. Their absence is based on substantive arguments: specifically, that most birth defects are caused by genetics and/or toxic exposures, with no prior evidence for an independent effect of weight gain; and that the perceptions of vitamin efficacy and inadequate diet, not maternal weight gain, are the primary motivations for vitamin use.

However, maternal weight gain *is* a collider on a backdoor path involving maternal behavioral and genetic factors, both unmeasured. Adjusting for maternal weight gain would *induce* bias in this case, by opening the backdoor path; moreover, we would be unable to block this path by adjusting for either of the two noncolliders, because they are unmeasured. Assuming the DAG in Fig. 10.5 is correct, it could prevent us from making this error, on the mistaken principle of adjusting for *any* possible confounder, without more carefully considering causal relationships.

Fig. 10.6 Common cause of exposure and outcome is a collider



Similarly, Fig. 10.6 shows stillbirths, potentially reduced by vitamin use and increased by birth defects, as a common *effect* of exposure and outcome; in contrast, a confounder is a common cause. As a collider on the backdoor path between vitamin use and birth defects, stillbirths should not be adjusted for.

10.2.5.7 More About DAGs

The case of vitamin use and birth defects shows that using DAGs to identify a minimum set of covariates that must be controlled for may rest on strong assumptions that certain directed edges are *absent* from the DAG, assumptions that may be easy to second-guess, especially in newer fields of research. In that case, the safe course is to include the additional directed edges in the DAG and make sure that the resulting backdoor paths are blocked, either by a collider that is not controlled for in the statistical model, or by a noncollider that is.

At the same time, backdoor paths of the kind shown in Fig. 10.5 should be interpreted with caution if evidence for the unmeasured factors is unconvincing, or their effects are thought to be weak. In this case, leaving the backdoor path unblocked may not induce substantial bias. Greenland (2003) shows that bias from controlling for the common effects of exposure and outcome, as in Fig. 10.6, may often be comparable in magnitude with bias from not controlling for a common cause of exposure and disease. In contrast, biases from controlling for a collider as shown in Fig. 10.5 may be smaller.

DAGs are also useful for determining whether to adjust for the baseline outcome in analyses of pre-post change scores, as discussed in Sect. 7.3.1. Glymour et al. (2005) use DAGs to show that if exposure affects outcome *levels* at baseline (regardless of whether it affects subsequent changes), and the baseline outcome is measured with error, then bias results from adjusting for baseline. Similarly, so-called *horse-racing bias* arises if changes have already begun at baseline, and

unmeasured causes of change affect both the baseline and follow-up outcomes. In both cases, the baseline outcome is a collider on a backdoor path from the primary predictor to the observed change. Since the common cause of the baseline outcome and change is by definition unmeasured, the resulting bias cannot be removed by adjustment.

In contrast, it is legitimate to adjust for the baseline outcome in estimating the effect of treatment on pre-post changes in a randomized trial, even though both outcomes are measured with error (Cragger 1987). In this case, the directed edge from treatment to the baseline outcome is absent, so there is no backdoor path from treatment to change, with the corollary that the baseline outcome is not a collider.

In addition, Hernán et al. (2004) show how DAGs can be used to analyze the potential for selection bias. In particular, they show that restricting study entry according to participant characteristics is equivalent to adjusting for a collider, if common causes link the qualifying characteristics to both exposure and outcome. This approach also explains why informative censoring or dropout in longitudinal studies can induce bias. In contrast to the biases analyzed by Glymour et al. (2005), these biases can potentially be avoided by measuring and adjusting for the common causes linking exposure, outcome, and selection.

In summary, DAGs are a useful tool for thinking through what we need to adjust for in analyses focusing on the effect of a primary predictor, as well as what needs to be omitted, at least at the initial stages of an analysis. At the same time, overcomplicated DAGs should not stop progress—small biases from residual confounding or collider bias may not result in qualitatively mistaken inferences.

10.2.6 Randomized Experiments

In clinical trials and other randomized experiments, the intervention is the predictor of primary interest. Other predictors are, in expectation, uncorrelated with the intervention, by virtue of randomization. Thus, in the regression model used to analyze an experiment, covariates do not usually need to be included to rule out confounding of assignment to the intervention. However, there are several other reasons for including covariates in the models used to analyze experiments.

- *Making valid inferences in stratified designs.* Design variables in stratified designs need to be included to obtain correct standard errors, CIs, and *P*-values. At issue is the potential for clustering of outcomes within strata, potentially violating the assumption of independence (Chap. 7). Thus, analyses of multicenter clinical trials now commonly take account of clinical center, even though random and equal allocation to treatment within center ensures that treatment is in expectation uncorrelated with this factor. Clustering within center can arise from differences in the populations studied and in the implementation of the intervention.
- *Increasing precision and power in experiments with continuous outcomes.* Adjusting for important baseline predictors of a continuous outcome can increase

the precision of the treatment effect estimate by reducing the residual error; because the covariates are in expectation uncorrelated with treatment, the variance inflation factor described in Sect. 4.2.2 is usually negligible. However, [Beach and Meier \(1989\)](#) use simulations to suggest that adjustment may on average increase squared error of the treatment effect estimate in smaller studies or when the selected covariates are not strongly predictive of the outcome. They also explore the difficulties in selecting a reasonable subset of the many baseline covariates typically measured, and conclude that adjusting for covariates which are both imbalanced and strongly predictive of the outcome has the largest expected effect on the statistical significance of the treatment effect estimate. We support adjustment for important prognostic covariates in trials with continuous endpoints, but also endorse the stipulation of [Hauck et al. \(1998\)](#) that the adjusted model should be pre-specified in the study protocol, to prevent *post hoc* “shopping” for the set of covariates which gives the smallest treatment effect *P*-value.

- *“De-attenuating” the treatment effect estimate and increasing power in experiments with binary or failure time outcomes.* In contrast to linear models for continuous outcomes, omission of important but balanced predictors, including the stratification variables mentioned previously, from a logistic ([Neuhaus and Jewell 1993](#); [Neuhaus 1998](#)) or Cox model ([Gail et al. 1984](#); [Schmoor and Schumacher 1997](#); [Henderson and Oman 1999](#)) used to analyze binary or failure time outcomes attenuates the treatment effect estimate. Hypothesis tests remain valid when the null hypothesis holds ([Gail et al. 1988](#)), but power is lost in proportion to the importance of the omitted covariates ([Lagakos and Schoenfeld 1984](#); [Begg and Lagakos 1993](#)). Note, however, that adjustment for imbalanced covariates can potentially move the treatment effect estimate *away* from as well as toward the null value, and can decrease both precision and power. In their review, [Hauck et al. \(1998\)](#) recommend adjustment for influential covariates in trials analyzed using logistic and Cox models. Their rationale is not only increased efficiency, but also that the adjusted or de-attenuated treatment effect estimates are more nearly interpretable as *subject specific*—in contrast to *population averaged*, a distinction that we explain in Sect. 7.5. We cautiously endorse adjustment for important covariates in trials with binary and failure time endpoints, but only if the adjusted model can be pre-specified and adjustment is likely to make the results more, not less convincing to the intended audience.
- *Adjusting for baseline imbalances.* Adjusted analyses are often conducted when there are apparent imbalances between groups, which can arise by chance, especially in small studies, or because of problems in implementing the randomization. The treatment effect estimate can be badly biased when strongly predictive covariates are imbalanced, even if the imbalance is not statistically significant. It is of course not possible to pre-specify such covariates, but adjustment is commonly undertaken in secondary analyses to demonstrate that the inferences about the treatment effect are not qualitatively affected by any apparent baseline imbalance. Note that the precision and statistical significance of the treatment effect estimate can be eroded by adjustment in this case, whether the endpoint

is continuous, binary, or a failure time. However, a difficult problem can arise when the selection of covariates to adjust for makes a substantive difference in interpretation, as [Beach and Meier \(1989\)](#) show in a re-analysis of time-to-event data from the Chicago Breast Cancer Surgery Study ([Meier et al. 1985](#)). In this small trial ($n = 112$), where the unadjusted treatment effect estimate just misses statistical significance ($P = 0.1$), different sets of covariates give qualitatively different results, with some adjusted models showing a statistically significant treatment effect and others weakening and even reversing the direction of the estimate.

10.3 Identifying Multiple Important Predictors

When the focus is on evaluating a predictor of primary interest, covariates are included in order to obtain a minimally confounded estimate of the association of the main predictor with the outcome. A good model rules out confounding of that association as persuasively as possible. However, broadening the focus to multiple important predictors of an outcome can make selecting a single best model considerably more difficult.

For example, inferences about most or all of the predictors retained in the model are now of primary interest, so overfitting and false-positive results are more problematic, particularly for novel associations not strongly motivated a priori. Effect modification or interaction will usually be of interest, but systematically assessing the large number of possible interactions can easily lead to false-positive findings, some at least not easily rejected as implausible. It may also be difficult to choose between alternative models that each include one variable from a collinear pair or set. Mediation is also more difficult to handle, to the extent that the overall effect of any predictor as well as its direct and indirect effects may be of interest. In this case, multiple, nested models may be required, as outlined in Sect. 4.4. Especially in the earlier stages of research, modeling these complex relationships is difficult, prone to error, and likely to be an iterative process. In some cases, a series of models, possibly including interactions, might be necessary to give a full and interpretable picture.

10.3.1 *Ruling Out Confounding Is Still Central*

In exploratory analyses to identify the important predictors of an outcome, confounding remains a primary concern—in this case, for any of the independent predictors of interest. Thus, some of the same strategies useful when a single predictor is of primary interest are likely to be useful here. In particular, relatively large models, including variables thought necessary for face validity, are preferable. Ideally, the model can be specified a priori. However, as in the previous section,

small sample size and high correlation between predictors may limit the number of variables that can be included. In this case, we recommend using backward selection with a liberal retention criterion. We discuss these issues in more detail in Sects. 10.4.1 and 10.4.2.

Simplifying the problem by treating each of the candidate predictors in turn as a predictor of primary interest, using the procedures from the previous section, is not a particularly satisfactory solution in our view. This can result in as many different models as there are predictors of interest, especially if covariates are retained because removing them changes the coefficient of the predictor of interest. Such a description of the data is uneconomical and hard to reconcile with an internally consistent causal model. Furthermore, missing values can result in the different models being fit to different subsets of the data.

10.3.2 *Cautious Interpretation Is Also Key*

What principally differs in this context is that *any* of the associations in the final model may require substantive interpretation, not just the association with a primary predictor. This may justify a more conservative approach to some minor aspects of the model; for example, poorly motivated and implausible interactions might more readily be excluded. In addition, well-motivated choices among any set of highly correlated predictors would need to be made.

However, we do not recommend “parsimonious” models that only include predictors that are statistically significant at $P < 0.05$ or even stricter criteria, especially with small samples, because the potential for residual confounding in such models is substantial. At the same time, we do not recommend explicit correction for multiple comparisons, since in an exploratory analysis it is far from clear how many comparisons to correct for, and by how much. This is in contrast to analyses evaluating multiple outcomes of a single treatment, as discussed in Sect. 13.4.1, where adjustment is almost certainly needed.

A better approach is to interpret the results of a larger model cautiously, especially novel, implausible, weak, and borderline statistically significant associations, to report model selection procedures, including the complete list of covariates considered, and to be aware of the potential inflation of type-I error, listing this as a limitation in published descriptions.

A more radical alternative, briefly discussed in Sect. 10.6, is to use methods for developing prediction models, based on minimizing prediction error, often via cross-validation. For example, the LASSO, discussed in Sect. 10.1.4, drops the least important variables and shrinks the less precisely estimated coefficients for others that are retained. Some of these methods provide direct measures of so-called *variable importance*, the implicit focus of this inferential goal. Drawbacks often include the lack of P -values and CIs, and the difficulty of accounting for mediating relationships and retaining variables for face validity.

10.3.3 *Example: Risk Factors for Coronary Heart Disease*

Vittinghoff et al. (2003) used multipredictor Cox models to assess the associations between risk factors and CHD events among 2,763 postmenopausal women with established CHD. Because of the large number ($n = 361$) of outcome events, it was possible to include all previously identified risk factors that were statistically significant at $P < 0.2$ in unadjusted models and not judged redundant on substantive grounds in the final multipredictor model. Among the 11 risk factors judged to be important on both substantive and statistical grounds were six noted by history (nonwhite ethnicity, lack of exercise, treated diabetes, angina, congestive heart failure, ≥ 2 previous heart attacks) and five that were measured (high blood pressure, lipids including LDL, HDL, and Lp(a), and creatinine clearance).

For face validity and to rule out confounding, the final model also controlled for other known or suspected CHD risk factors, including age, smoking, alcohol use, and obesity, although these were not statistically significant in the adjusted analysis. Mediation of obesity and diabetes, both shown to be associated with risk in single-predictor models, was covered in the discussion section of the paper. The model also controlled for a wide range of CHD-related medications, but because these effects were not of direct interest and hard to interpret, estimates were not presented. However, interactions between risk factors and relevant treatments were examined, on the hypothesis that treatments might modify the association between observed risk factor levels and future CHD risk; the final model included interactions that were statistically significant at $P < 0.2$.

10.3.4 *Allen–Cady Modified Backward Selection*

Flexible predictor selection procedures, including conventional backward selection, are known to increase the probability of making at least one type-I error. A backward selection procedure (Allen and Cady 1982) based on a ranking of the candidate variables by importance can be used to help avoid false-positive results, while still reducing the number of covariates in the model. In this procedure, a set of variables may be forced into the model, including predictors of primary interest, as well as confounding variables thought important for face validity. The remaining candidate variables would then be ranked in order of importance. Starting with an initial model including all covariates in these two sets, variables in the second set would be deleted in order of ascending importance until the first variable meeting a criterion for retention is encountered. Then the selection procedure stops.

This procedure is special in that only the remaining variable hypothesized to be least important is eligible for removal at each step, whereas in conventional backward selection, any of the predictors not being forced into the model is eligible. False-positive results are less likely because there is only one pre-specified sequence of models, and selection stops when the first variable not meeting the criterion for removal is encountered. In contrast, conventional stepwise procedures and especially best subsets search over broader classes of models.

10.4 Some Details

10.4.1 Collinearity

In Sect. 4.2, we saw that the variance of the regression coefficient estimate for predictor x_j , increases with r_j , the multiple correlation between x_j and the other predictors in the model. When r_j is large, the estimate of β_j can become quite imprecise. Consider the case where two predictors are fairly highly correlated ($r \geq 0.80$). When both are included in the model, the precision of the estimated coefficient for each can be severely degraded, even when both variables are statistically significant predictors in simpler models that include one but not both. In the model including both, an F -test for the joint effect of both variables may be highly statistically significant, while the variable-specific t -tests are not. This pattern indicates that the two variables jointly provide important information for predicting the outcome, but that neither is necessary over and above the other. With modern computers, problems in estimating the independent effects of highly correlated predictors no longer arise from numeric inaccuracy in the computations. Rather, the information is coming from both variables jointly, which makes them both seem unimportant in t -tests evaluating their individual contributions.

Definition: Collinearity denotes correlation between predictors high enough to degrade the precision of the regression coefficient estimates substantially for some or all of the correlated predictors.

How we deal with collinear predictors depends in part on our inferential goals. For a prediction model, inference on individual predictors is not of direct interest. Rather, if inclusion of collinear variables decreases prediction error, then it is legitimate to include them both. In this case, cross-validation of the target measure of PE can be used to decide which of a collinear set of predictors to include.

Alternatively, suppose that one of two collinear variables is a predictor of primary interest, and the other is a confounder that must be adjusted for on substantive grounds. If the predictor of interest remains statistically significant after adjustment, then the evidence for an independent effect is usually convincing. In small data sets especially, it would be necessary to demonstrate that the finding is not the result of a few influential points, and where the data do not precisely meet model assumptions, to show that the inferences are robust, possibly using the bootstrap methods introduced in Sect. 3.6. Alternatively, if the effects of the predictor of interest are clearly confounded by the adjustment variable, we would also have a clearcut result. However, in cases where neither is statistically significant after adjustment, we may need to admit that the data are inadequate to disentangle their effects.

In contrast, where the collinearity is between adjustment variables and does not involve the predictor of primary interest, then inclusion of the collinear variables can sometimes be justified. In this case, information about the underlying factor being adjusted for may be increased, but the precision of the estimate for the predictor

of interest is unaffected. To see this, consider evaluating the effect of diabetes on HDL, adjusting for BMI. In Sect. 4.7, we found that a quadratic term in BMI added significantly to the model. However, BMI and its square are clearly collinear ($r = 0.99$). If instead we first “center” BMI (i.e., subtract off its sample mean before computing its square), the collinearity disappears ($r = 0.46$). However, the estimate for diabetes and its standard error are unchanged whether or not we center BMI before computing the quadratic term. In short, collinearity between adjustment variables is unlikely to matter.

Finally, when we are attempting to identify multiple independent predictors, an attractive solution is to choose on substantive grounds, such as plausibility as a causal factor. Otherwise, it may make sense to choose the predictor that is measured more accurately or has fewer missing values. As in the case of a predictor of primary interest, the multivariable model may sometimes provide a clear indication of relative importance, in that one of the collinear variables remains statistically significant after adjustment, while the others appear to be unimportant. In this case, the usual course would be to include the statistically significant variable and drop the others.

10.4.2 *Number of Predictors*

The rationale for inclusive predictor selection rules, whether we are assessing a predictor of primary interest or multiple important independent predictors, is to obtain minimally confounded estimates. However, this can make regression coefficient estimates less precise, especially for highly correlated predictors. At the extreme, model performance can be severely degraded by the inclusion of too many predictors.

Rules of thumb have been suggested for number of predictors that can be safely included as a function of sample size or number of events. A commonly used guideline prescribes at least ten observations for each predictor; with binary or survival outcomes the analogous guideline specifies ten events per predictor (Peduzzi et al. 1995, 1996; Concato et al. 1995). The rationale is to obtain adequately precise estimates, and in the case of the logistic and Cox models (Chaps. 5 and 6), to ensure that the models behave properly.

Such guidelines are useful as flags for potential problems, but need not be inflexibly applied. Their primary limitation is that the precision of coefficient estimates depends on other factors as well as the number of observations or events per predictor (Vittinghoff and McCulloch 2007). In particular, recall from Sect. 4.2 that the variance of an estimated regression coefficient in a linear model depends on the residual variance of the outcome, which is generally reduced by the inclusion of important covariates. Precision also depends on the multiple correlation between a predictor of interest and other variables in the model. Thus, addition of covariates that are at most weakly correlated with the primary predictor but explain substantial outcome variance can actually improve the precision of the estimate for the predictor

Table 10.3 Cox models for DVT-PE

Predictor variable	RH (95% Confidence interval)				<i>P</i> -values	
	11-Predictor model		5-Predictor models		Wald	LR
HT vs. placebo	2.7	(1.4–5.2)	2.7	(1.4–5.1)	0.002	0.001
≥ 53 at LMP	3.6	(2.0–6.4)	3.3	(1.8–5.8)	< 0.001	< 0.001
Inpatient surgery	4.3	(2.1–8.7)	4.7	(2.3–9.5)	< 0.001	< 0.001
Hospitalization	5.6	(2.9–11)	6.7	(3.6–13)	< 0.001	< 0.001
Hip fracture	5.9	(0.8–46)	6.6	(0.9–51)	0.09	0.18
Leg fracture	17.3	(5.1–58)	14.1	(4.2–47)	< 0.001	< 0.001
Cancer	4.1	(1.7–9.7)	3.5	(1.5–8.4)	0.002	0.006
Nonfatal MI	6.0	(2.3–16)	4.4	(1.7–11)	< 0.001	0.002
Stroke/TIA	0.9	(0.1–6.5)	0.9	(0.1–6.4)	0.88	0.88
Aspirin use	0.4	(0.2–0.7)	0.4	(0.2–0.6)	0.003	0.004
Statin use	0.4	(0.2–0.9)	0.4	(0.2–0.7)	0.02	0.02

of interest. In contrast, addition of just one collinear predictor can degrade its precision unacceptably. In addition, the allowable number of predictors depends on effect size, with larger effects being more robust to multiple adjustment than smaller ones.

Rather than applying such rules categorically, we recommend that problems potentially stemming from the number of predictors be assessed by checking for high levels of correlation between a predictor of interest and other covariates, and for large increases in the standard error of its estimated regression coefficient when additional variables are included. For logistic and Cox models, consistency between Wald and LR test results is another useful measure of whether there are enough events to support the number of predictors in the model. Additional validation of a relatively inclusive final model is provided if a more parsimonious model with fewer predictors gives consistent results, in particular for the predictor of interest. If problems do become apparent, a first step would be to make the criterion for retention in backward selection more conservative, possibly $P < 0.15$ or $P < 0.10$. It would also make sense to consider omitting variables included for face validity which do not appear to confound a predictor of primary interest.

An analysis of risk factors for deep-vein thrombosis and pulmonary embolism (DVT-PE) among postmenopausal women in the HERS cohort ([Grady et al. 2000](#)) is an example of stable results despite violation of the rule of thumb that the number of events per predictor should be at least 10. In this survival analysis of 47 DVT-PE events, 11 predictors were retained in the final model, so that there were only 4.3 events per predictor. However, the largest pairwise correlation between the selected risk factors was only 0.16 and most were below 0.02. As shown in [Table 10.3](#), estimates from the 11-predictor model were consistent with those given by 5-predictor models, in accord with the rule of thumb, which omitted the less important predictors. Although CIs were wide for the strongest and least common risk factors, this was also true for the 5-predictor models. Finally, P -values for the Wald and LR tests based on the larger model were highly consistent.

10.4.3 Alternatives to Backward Selection

Some alternatives to backward selection include best subsets; sequential (so-called *greedy*) procedures, including forward and stepwise selection; and bivariate screening.

- *Best subsets* screens models including all possible subsets of the candidate predictors in a user-specified range of model sizes, using a summary measure such as adjusted R^2 to compare models. This computer-intensive procedure is implemented in SAS for some models, but not in Stata. It was also the underlying approach of the cross-validation screening described in Sect. 10.1.6, but did require prior simplification to reduce the computational burden.
- *Forward selection* begins with the null model with only the intercept, then adds variables sequentially, at each step adding the variable that promises to make the biggest additional contribution to the current model.
- *Stepwise* methods augment the forward procedure by allowing variables to be removed if they no longer meet an inclusion criterion after other variables have been added. Stata similarly augments backward selection by allowing variables to re-enter after removal. As compared to best subsets, these three sequential procedures are more vulnerable to missing good alternative models that happen not to lie on the sequential path. This implies that plausible alternatives to models selected by stepwise procedures should be examined.
- In *bivariate screening*, candidate predictors are evaluated one at a time in single-predictor models. In some cases, all predictors that meet the screening criterion are included in the final model; in other cases, screening is used as a first step to reduce the number of predictors then considered in a backward, forward, stepwise, or best subsets selection procedure. Orwoll et al. (1996) used a variant of this procedure, including all variables statistically significant at $P < 0.05$ in two-predictor models adjusting for age.

Note that only observations with complete data on all variables under consideration are used in automated selection procedures. The resulting subset can be substantially smaller than the data set used in the final model, and unrepresentative. When implemented by hand, different subsets are commonly used at different steps, for the same reason, and this can also affect results. Findings which depend on the inclusion or exclusion of subsets of observations should be carefully checked.

10.4.3.1 Why We Prefer Backward Selection

The principal advantage of backward selection is that negatively confounded sets of variables are less likely to be omitted from the model (Sun et al. 1999), since the complete set is included in the initial model. Best subsets shares this advantage. In contrast, forward and stepwise selection procedures will only include such sets if at least one member meets the inclusion criterion in the absence of the others.

Univariate screening will only include the complete set if all of them individually meet the screening criterion; moreover, this difficulty is made worse if a relatively conservative criterion is used to reduce the number of false-positive findings in an exploratory analysis.

A disadvantage of backward selection is that initial deletions may be badly determined if the list of candidate predictors is too large for the number of observations or events. In this case, bivariate screening with a liberal criterion can be used to eliminate the weakest predictors; in addition, the Stata stepwise procedure allowing variables to re-enter affords some protection against this problem. More generally, sensitivity analyses using forward and/or stepwise in addition to backward selection are useful for showing whether results are robust to the model selection procedure used

10.4.4 Model Selection and Checking

Section 4.7 focused on methods for checking the linear model which make use of the residuals from a multipredictor model rather than examining bivariate relationships. There, we took as a given that the predictors had already been selected. However, transformation of the outcome or of continuous predictors can affect the apparent importance of predictors. For example, in Sect. 4.6.4 we saw that the need for an interaction between treatment with HT and the baseline value of the outcome LDL was eliminated by analyzing treatment effects on percent rather than absolute change from baseline. Alternatively, detection of important nonlinearities in the model checking step can uncover associations that were masked by an initial linear specification. As a consequence, predictor selection should be revisited after changes of this kind are made. And then, of course, the fit of the modified model would need to be rechecked.

10.4.5 Model Selection Complicates Inference

Underlying the CIs and P -values which play a central role in interpreting regression results is the assumption that the predictors to be included in the model were determined a priori without reference to the data at hand. In *confirmatory* analyses in well-developed areas of research, including phase-III clinical trials, prior determination of the model is feasible and important. In contrast, at earlier stages of research, data-driven predictor selection and checking are reasonable, often necessary, and certainly widely used. However, some of the issues raised for inference include the following.

- The chance of at least one type-I error can greatly exceed the nominal level used to test each term, leading to false-positive results with too-small P -values and too-narrow CIs.

- In small data sets, precision and power are often poor, so important predictors may well be omitted from the model, especially if a restrictive inclusion criterion is used. Conversely, in large data sets unimportant predictors are commonly included, reinforcing the need for cautious interpretation of novel, implausible, weak, and borderline statistically significant findings.
- Parameter estimates can be biased away from the null, owing to selection of estimates that are large by chance, sometimes called *testimation bias* (Steyerberg 2009). This bias is greater for relatively weak predictors.
- Choices between predictors can be poorly motivated, especially between collinear variables. Univariate screening provides no guidance for this problem. Moreover, predictor selection is potentially sensitive to addition or deletion of a few observations, especially when the predictors are highly correlated. Altman and Andersen (1989) propose bootstrap methods for assessing this sensitivity.

Predictor selection driven by P -values is subject to these pitfalls whether it is automated or implemented by hand. How seriously do these problems affect inference for our three inferential goals?

- *Prediction.* In many modern prediction methods, potentially large sets of candidate predictors are aggressively screened, but P -values are not used as the criterion. We implemented one such procedure in Sect. 10.1.6, and Breiman (2001) briefly reviews other modern methods which even more aggressively search over candidate models. However, use of GCV measures of prediction error as a criterion for predictor selection effectively protects against both overfitting and invalid inferences. In short, predictor selection does not adversely affect modern procedures for this inferential goal.
- *Evaluating a predictor of primary interest.* Iterative model checking and selection should likewise have relatively small effects on inference about a predictor of primary interest, since it is included by default in all candidate models. In fact, iterative checking and predictor selection should result in better control of confounding, a primary aim for this inferential goal. However, when the primary predictor is of borderline statistical significance, the issue of P -value shopping raised in Sect. 10.2.6 needs to be conscientiously handled, and sensitivity of results to predictor selection reported.
- *Identifying multiple important predictors.* Model selection most clearly complicates inference for this inferential goal, since CIs and P -values for any of the predictors are potentially of direct interest. Note that inclusion of variables for face validity, use of a loose inclusion criterion ($P < 0.2$), and the Allen–Cady procedure all reduce the potential impact of predictor selection on inference. Nonetheless, selection procedures should *only* be used with prior consideration of hypothesized relationships, careful examination of alternative models with other sets of predictors, checks on model fit and robustness, skeptical review of the findings for plausibility, and cautious interpretation of the results, especially novel, borderline statistically significant, and weak associations.

10.5 Summary

We have identified three inferential goals, and recommend predictor selection procedures appropriate to each of them.

For prediction, we recommend identifying candidate predictors and appropriate transformations well-supported by prior research. But in the common case where expert opinion and the literature do not provide sufficient guidance, we recommend exhaustive screening of candidate models to find the few models that minimize a generalized cross-validation measure of prediction error.

For evaluating a predictor of primary interest, we recommend using DAGs to specify hypothesized relationships between the primary predictor, potential confounders and mediators, and the outcome; caution should be used in eliminating variables based on any DAG that omits plausible but unestablished causal pathways. The selected model should include all generally accepted confounders required to ensure its face validity. Other potential confounders that turn out not to be important on statistical grounds can optionally be removed from the model using a backward selection procedure, but with a liberal inclusion criterion to minimize the potential for confounding. Especially in smaller data sets, care must be taken with the inclusion of covariates highly correlated with the predictor of interest, since these can unduly inflate the standard errors of the estimate of its effect. Negative findings for the primary predictor should be carefully interpreted in terms of the point estimate and CI, as described in Sect. 3.7.

For identifying multiple important predictors of an outcome, we recommend a procedure similar to that used for a single predictor of primary interest. A DAG mapping out hypothesized relationships between variables can be particularly useful. Strongly motivated covariates may be included by default to ensure the face validity of the model. The Allen–Cady modification of the backward selection procedure is useful for selecting from among the remaining candidate variables while limiting false-positive results. Negative, weak, and/or borderline statistically significant associations retained in the final model as much to control confounding of other associations as for their intrinsic plausibility and importance should be interpreted with particular caution.

10.6 Further Notes and References

Predictor selection is among the most controversial subjects covered in this book. Book-length treatments include [Miller \(1990\)](#) and [Linhart and Zucchini \(1986\)](#), while regression texts including [Weisberg \(1985\)](#) and [Hosmer and Lemeshow \(2000\)](#) address predictor selection issues at least briefly. The central place we ascribe to ruling out confounding in the second and third inferential goals owes much to [Rothman and Greenland \(1998\)](#), a standard reference in epidemiology that describes how substantive considerations can be brought to bear on predictor selection.

One promising method for ensuring adequate control of confounding is more or less exhaustive screening of candidate models with different covariate sets, some including interactions between covariates and/or restricted cubic splines for continuous confounders. As described in Sect. 10.1.4, these procedures use cross-validated prediction error as a model selection criterion to avoid overfitting, and avoid some pitfalls of P -value driven selection procedures, as discussed in Sect. 10.4.5. However, these methods can be difficult to implement, and are a focus of ongoing statistical research.

Both the theory and application of causal diagrams and models have been advanced substantially in recent years (Pearl 1995; Greenland et al. 1999) and give additional insights into situations where confounding can be ruled out a priori. However, these more advanced methods appear to be most useful in problems where causal pathways are more clearly understood than is our usual experience. Jewell (2004) and Greenland and Brumback (2002) explore the connections between causal diagrams, potential outcomes, and some model selection issues.

Chatfield (1995) reviews work on the influence of predictor selection on inference, while Buckland et al. (1997) propose using weighted averages of the results from alternative models as a way of incorporating the extra variability introduced by predictor selection in computing CIs. These would be particularly applicable to the second inferential goal of evaluating a predictor of central interest.

For a sobering view of the difficulty of validly modeling causal pathways using the procedures covered in this book and particularly this chapter, see Breiman (2001). From this point of view, computer-intensive methods validated strictly in terms of prediction error not only give better predictions but may also be more reliable guides to “variable importance”—another term for our third inferential goal of identifying important predictors, and with obvious implications for assessing a predictor of central interest.

10.7 Problems

Problem 10.1. Characterize the following contexts for predictor selection as prediction, evaluation of a primary predictor of interest, or identifying the important predictors of an outcome:

- examining the effect of treatment on a secondary endpoint in an RCT
- determining which newborns should be admitted to the neonatal intensive care unit (NICU)
- comparing a measure of treatment success between two surgical procedures for stress incontinence using data from a large longitudinal cohort study
- identifying risk factors for incident hantavirus infection.

Problem 10.2. Consulting Stata documentation, describe how the `sw:` command prefix with the `lockterm1`, `hier`, and `pr()` options can be used to implement the Allen–Cady procedure.

Problem 10.3. Think of an outcome under preliminary investigation in the area of your expertise. Following Allen and Cady's prescriptions, try to rank predictors of this outcome in order of importance. Are there any variables that you would include by default? Why?

Problem 10.4. Do any of the variables you have selected in the previous problem potentially mediate the effects of others in your list? If so, how would this affect your decision about what to include in the initial model? What series of models could you use to examine mediation? (See Sect. 4.5.)

Problem 10.5. Suppose you included an indicator for diabetes in a multivariable model estimating the independent effect of exercise on glucose. How would you interpret the estimate for exercise? Would you want to consider interactions between exercise and diabetes in this model? How would you deal with use of insulin and oral hypoglycemics?

Problem 10.6. Why are univariate screening and forward selection more likely to miss negatively confounded variables than backward deletion and best subsets?

Problem 10.7. Give an example of a "biologically plausible" relationship that has turned out to be false. Give an example of a biologically *implausible* relationship that has turned out to be true.

Problem 10.8. Suppose you were using a logistic model to examine the association between a predictor and outcome of interest, and to rule out confounding you needed to include one or two more predictors than would be allowed by the rule of 10 events per variable. In comparing models with and without the two extra predictors, what might signal that you were asking the bigger model to do too much? How would the correlation between the extra variables and the predictor of interest influence your thinking?

10.8 Learning Objectives

- (1) Describe and implement strategies for predictor selection for
 - prediction
 - evaluation of a primary predictor
 - identifying multiple important predictors.
- (2) Use a DAG to define hypothetical relationships among confounders, mediators, and the outcome.
- (3) Be familiar with the drawbacks of predictor selection procedures.

Chapter 11

Missing Data

Missing data are a fact of life in medical research. Subjects refuse to answer sensitive questions (e.g., questions about income or drug use), are unable to complete an MRI exam because of metallic implants, or drop out of studies and do not contribute further data. In each of these cases, data are “missing” or not complete. How should this be accommodated in a data analysis? Statistical computing packages will typically drop from the analysis all observations that are missing any of the variables (outcomes or predictors). So, for example, a linear regression predicting a patient’s number of emergency room visits from their age, gender, race, income, and current drug use will drop any observation missing even one of those variables. Analysis of data using this strategy is called *complete case analysis* because it requires that the data be complete for *all* variables before that observation can be used in the analysis.

Complete case analysis is simple and the default for statistical analysis packages. But it can be inefficient and lead to biased estimates. Imagine a situation in which the first 20% of the sample is missing age information, the second 20% is missing gender information and so on, with the last 20% missing drug use information. Even though, in a sense, 80% of the predictor data is present, there will be no observations left for a complete case analysis.

Further, data are often missing for a reason related to the outcome under study. As examples, sicker patients may not show up for follow-up visits, leading to overly optimistic estimates based on the data present. Or those patients staying in the hospital longer may be the sicker ones (with the better-off patients having been discharged). This might lead us to the erroneous conclusion that longer stays in the hospital produce poorer outcomes, so why check-in in the first place? A basic message is that we need to think carefully about why the data are missing. This may influence how we will handle it and guide us to ways we can avoid biased estimates.

How can these drawbacks be overcome? If we could intelligently fill in the missing data to obtain a complete dataset then we could use standard methods without concern. Of course, we would need to account for the fact that the missing data are estimated and not actual measured quantities in our sample. This is the

basic idea behind *multiple imputation*, which we discuss in Sect. 11.5. Or perhaps, we could use members in the sample with complete data to represent those with missing data. For example, suppose heavy drug users tended to drop out of a study at twice the rate of other participants. Then we could “double-count” the heavy drug users who did not drop out of the study by weighting their contributions to the analysis more heavily. This is the basic idea behind *inverse probability weighting* (IPW) which we cover in Sect. 11.9.3. In either case, the key is to use the data on hand, along with anything we might know about why the data are missing in order to infer the missing data. Not surprisingly, this strategy will only work if the values of the missing data are, to some extent, predictable from the observed data.

We begin this chapter with some simple illustrations of what can go wrong when there is missing data. This naturally leads to consideration of why the data are missing and some more formal classifications of the missing data process in Sect. 11.2. We discuss some simple strategies that have been used in the past to accommodate missing data. We then consider common missing data scenarios: missing predictor values (with at least some of the associated outcomes being measured) and complete (or nearly complete) predictor values, but missing outcomes. For this latter situation, we consider three different ways in which the data came to be missing. The two strategies mentioned above—multiple imputation and inverse probability weighting—are then considered in more detail as principled approaches to missing data. In Sect. 11.9.1, we also describe situations with missing outcome data in longitudinal studies that can be addressed by using maximum-likelihood methods like mixed models. These “automatically” infer the missing data with the advantage of not requiring explicit modeling. Our focus throughout this chapter is on the effect that missing data has on estimation of regression coefficients, but missing data can also cause predictions to be biased.

11.1 Why Missing Data Can Be a Problem

To more clearly demonstrate why missing data can be a problem, we consider two examples using the HERS study (see Sect. 3.1). In the first, we consider linear regression of SBP on glucose level, BMI, and whether the person was Caucasian or not using only the data from the fourth visit. For that visit 443 of the 1,871 observations had missing data for glucose. The second considers a longitudinal data setting in which SBP is measured over two visits with the second one potentially missing, as would happen with participants dropping out of a study.

11.1.1 Missing Predictor in Linear Regression

Standard regression of SBP on blood glucose level (`glucose`), whether a person is Caucasian or not (`white`), and their BMI (`bmi`) using the 1,871 participants with

Table 11.1 Regression of SBP using a complete case analysis

```
. regress sbp glucose white bmi
```

Source	SS	df	MS	Number of obs =	1385
Model	2855.36663	3	951.788878	F(3, 1381) =	2.69
Residual	488496.255	1381	353.72647	Prob > F =	0.0450
				R-squared =	0.0058
				Adj R-squared =	0.0037
Total	491351.622	1384	355.022848	Root MSE =	18.808

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
glucose	.0294818	.0126344	2.33	0.020	.0046972 .0542665
white	-1.537906	1.689423	-0.91	0.363	-4.852019 1.776207
bmi	.0644021	.0934208	0.69	0.491	-.11886 .2476641
_cons	132.716	3.29506	40.28	0.000	126.2521 139.1799

Table 11.2 Regression of systolic blood pressure using imputed glucose values

```
. regress sbp imp_glucose white bmi
```

Source	SS	df	MS	Number of obs =	1750
Model	5766.65623	3	1922.21874	F(3, 1746) =	5.34
Residual	628318.844	1746	359.861881	Prob > F =	0.0012
				R-squared =	0.0091
				Adj R-squared =	0.0074
Total	634085.5	1749	362.541738	Root MSE =	18.97

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
imp_glucose	.0338782	.0122595	2.76	0.006	.0098333 .057923
white	-2.209204	1.49944	-1.47	0.141	-5.150092 .7316834
bmi	.1364681	.083551	1.63	0.103	-.0274025 .3003388
_cons	130.385	2.937854	44.38	0.000	124.6229 136.1471

data for visit 4 in HERS gives the output in Table 11.1. We can see that only 1,385 subjects are used in the complete case analysis. This is because, in addition to the 443 participants missing data on glucose, there are 85 missing values for SBP, 110 missing values for BMI, and 3 missing values for white (and some overlap in the missing data). We will concentrate on the missing glucose values to introduce the main ideas.

Glucose values are fairly strongly related to the other predictors, so there is some hope in filling in the missing values relatively accurately; a regression of glucose on SBP, BMI, white, current smoking status, and whether or not a woman develops diabetes has an R^2 of 0.44. We could use this regression to generate predicted values for 372 of the 443 of the missing glucose values—we cannot fill them all in because there is missing data for BMI, white, and diabetes. Using the predicted values in place of the missing glucose values, we can now use more of the data. Table 11.2 gives the regression results, where `imp_glucose` is equal to the actual value of glucose when it is available and the predicted (imputed) value of glucose when it is missing. Some of the regression coefficients are noticeably different,

Table 11.3 Regression of systolic blood pressure using multiply imputed glucose values

. mi estimate: regress sbp glucose white bmi

Multiple-imputation estimates
Linear regression

DF adjustment: Small sample

Model F test: Equal FMI

Within VCE type: OLS

Imputations = 5
Number of obs = 1750
Average RVI = 0.0106
Complete DF = 1746
DF: min = 1046.77
 avg = 1557.86
 max = 1743.23
F(3, 1644.5) = 4.57
Prob > F = 0.0034

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
glucose	.0269531	.0116979	2.30	0.021	.0039991	.049907
white	-2.199165	1.500637	-1.47	0.143	-5.142402	.7440727
bmi	.1467563	.0836029	1.76	0.079	-.0172179	.3107305
_cons	130.8553	2.930445	44.65	0.000	125.1077	136.6029

for example, the BMI coefficient has approximately doubled in size, has a smaller standard error, and has a smaller *p*-value. All the standard errors are smaller. This is an illustration of what is called *single imputation*, because we have filled in or imputed the missing data a single time.

But this is not quite legitimate. In this analysis, the software does not distinguish between the imputed glucose values and the actual measured values. So the information content of the dataset is overestimated and standard errors may be falsely small. A solution to this is to impute the glucose values but properly account for the actual amount of information available. One way to do this is to use *multiple imputation* which we describe in more detail in Sect. 11.5. Table 11.3 gives the results of such an analysis.

The results are very similar to the singly imputed analysis. Because we have not imputed a large portion of the data, the standard errors are only slightly increased in the multiply imputed approach compared to the singly imputed. Notably, the standard errors remain smaller than those from the complete case analysis.

Using imputation to handle the missing data for this example has had two benefits: it may have slightly reduced a bias in the original coefficients and we have been able to successfully utilize more of the data, thereby reducing the standard errors. Multiple imputation is a flexible methodology and can be used to impute not only the predictor, but also the outcomes.

11.1.2 Missing Outcome in Longitudinal Data

To illustrate the potential problems with drop out in longitudinal data, we used the HERS study, for which there is actually very little drop out. We consider the outcome of SBP using data only from baseline and year 1. In the complete dataset,

Table 11.4 Analysis of HERS data using complete data and generalized estimating equations

```
. xtgee sbp visit bmi baseline_dm, i(pptid) corr(exch) robust
```

GEE population-averaged model

Group variable: pptid

Link: identity

Family: Gaussian

Correlation: exchangeable

Scale parameter: 357.8178

Number of obs = 5368

Number of groups = 2761

Obs per group: min = 1

avg = 1.9

max = 2

Wald chi2(3) = 67.85

Prob > chi2 = 0.0000

(Std. Err. adjusted for clustering on pptid)

sbp	Coef.	Semirobust Std. Err.	z	P> z	[95% Conf. Interval]
visit	.2836451	.3388382	0.84	0.403	-.3804657 .9477558
bmi	.1385708	.0598246	2.32	0.021	.0213167 .255825
baseline_dm	5.511153	.7814551	7.05	0.000	3.97953 7.042777
_cons	129.66	1.723401	75.23	0.000	126.2822 133.0379

the average SBP at baseline was 135.1 and at year 1 was 135.2, so very little change from baseline to year 1.

To quantify the change from baseline to visit 1, we used regression analyses of SBP on visit (baseline, coded as 0, or the year 1 visit, coded as 1), BMI and whether the participant had diabetes at baseline (yes/no). Since we have repeated measures, we could use either GEEs (via `xtgee`) or mixed models (via `xtmixed`) to analyze the data. Tables 11.4 and 11.5 give the results using the complete data.

The two analyses give virtually the same results. Focussing on the visit term, there is a small and nonstatistically significant increase from baseline to year 1 (estimated to be about 0.28), consistent with the raw data.

We next simulated drop out at year 1 on the basis of either the baseline SBP or the year 1 SBP, but keeping all the data for the baseline visit. In either case, those with higher SBP were dropped at higher rates than those with lower SBP. In the situation where drop out depended on baseline SBP, we “dropped” 1,461 participants at year 1 and “retained” 1,302. Those retained had average SBP at year 1 of 127.5 (range 85–196) and those dropped had average SBP 143.9 (range 93–220). So there is a distinct difference between those dropped and retained, but there is also considerable overlap. Importantly, in the incomplete data, the average SBP drops from 135.1 at baseline to 127.5 at year 1, quite different from the complete data.

We, therefore, anticipate trouble with the analysis using the incomplete data since the average SBP drops between baseline and the year 1 visit. Ideally, a technique that handles missing data well will give results similar to the analysis of the complete data (e.g., Table 11.4). Table 11.6 gives the regression coefficient tables for the situation where drop out depends on SBP at baseline.

Now we see a completely different story. The generalized estimating equations (GEEs) approach incorrectly estimates a highly statistically significant drop in SBP

Table 11.5 Analysis of HERS data using complete data and maximum likelihood

. xtmixed sbp visit bmi baseline_dm || pptid:

Mixed-effects REML regression

Group variable: pptid

Number of obs = 5368

Number of groups = 2761

Obs per group: min = 1

avg = 1.9

max = 2

Log restricted-likelihood = -22872.471

Wald chi2(3) = 73.13

Prob > chi2 = 0.0000

sbp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
visit	.2843892	.338578	0.84	0.401	-.3792114	.9479898
bmi	.1392584	.0587622	2.37	0.018	.0240865	.2544302
baseline_dm	5.507891	.7583126	7.26	0.000	4.021625	6.994156
_cons	129.6413	1.677004	77.31	0.000	126.3544	132.9282

Random-effects Parameters

Estimate

Std. Err.

[95% Conf. Interval]

pptid: Identity					
sd(_cons)	14.40895	.2784187	13.87346	14.9651	
sd(Residual)	12.28843	.1702939	11.95916	12.62678	

LR test vs. linear regression: chibar2(01)= 1055.76 Prob >= chibar2 = 0.0000

of 1.32 from baseline to year 1. Interestingly, the mixed model approach (which uses maximum likelihood, or ML, to fit the model) gives estimates similar to the complete data analysis with a small estimated increase which is not statistically significant. For the other coefficients, the two analyses give similar results, both to one another and to the complete data analyses.

Finally, we also simulated a dataset where drop out at year 1 depended on year 1 SBP in a fashion similar to that described above. This differs from the previous case in that whether or not a participant was included in the dataset depended on *unobserved* quantities. Table 11.7 gives the results with drop out that depends on SBP at year 1. Now both the analyses give very severely biased estimates of the visit effect, though other coefficients are little affected.

There are several important messages from this example. When drop out is dependent on previous, observed values, some analysis methods such as GEEs can give badly biased estimates whereas others such as mixed model methods, based on maximum likelihood, are less affected. The situation when drop out depends on unobserved values is much more serious and leads to severe bias using either method.

Table 11.6 Analysis of HERS data with drop out depending on baseline outcome using GEEs and ML

```
. xtgee sbp visit bmi baseline_dm if miss_mar==0, i(pptid) corr(exch) robust
```

sbp	Coef.	Semirobust Std. Err.	z	P> z	[95% Conf. Interval]	
visit	-1.320828	.431427	-3.06	0.002	-2.166409	-.4752463
bmi	.1041894	.0622733	1.67	0.094	-.0178641	.2262428
baseline_dm	5.787856	.813257	7.12	0.000	4.193901	7.38181
_cons	130.5635	1.790897	72.90	0.000	127.0534	134.0736

```
. xtmixed sbp nvisit bmi baseline_dm if miss_mar==0 || pptid:
```

sbp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
visit	.5912612	.4179003	1.41	0.157	-.2278083	1.410331
bmi	.1084238	.0625694	1.73	0.083	-.0142101	.2310576
baseline_dm	5.894762	.801877	7.35	0.000	4.323111	7.466412
_cons	130.4142	1.779439	73.29	0.000	126.9266	133.9019

Table 11.7 Analysis of HERS data with drop out depending on unobserved outcome using GEEs and ML

```
. xtgee sbp visit bmi baseline_dm if miss_nmar==0, i(pptid)corr(exch) robust
```

sbp	Coef.	Semirobust Std. Err.	z	P> z	[95% Conf. Interval]	
visit	-9.889191	.3840043	-25.75	0.000	-10.64183	-9.136557
bmi	.0962627	.0574965	1.67	0.094	-.0164284	.2089539
baseline_dm	4.985786	.7507309	6.64	0.000	3.514381	6.457192
_cons	131.0006	1.656733	79.07	0.000	127.7534	134.2477

```
. xtmixed sbp visit bmi baseline_dm if miss_nmar==0 || pptid:
```

sbp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
visit	-8.35655	.4240134	-19.71	0.000	-9.187601	-7.525499
bmi	.1043524	.0573602	1.82	0.069	-.0080715	.2167762
baseline_dm	5.027966	.73204	6.87	0.000	3.593194	6.462738
_cons	130.7572	1.630634	80.19	0.000	127.5613	133.9532

11.2 Classifications of Missing Data

The previous example has shown that the mechanism that causes the data to be missing can be very important. It is, therefore, useful to develop categorizations of missing data mechanisms that either are or are not likely to cause misleading results.

To motivate some of the considerations, we use the Steroids for Corneal Ulcer Trial (SCUT). SCUT was a randomized clinical trial to gauge the effectiveness of a steroid treatment (steroid eye drop versus a placebo) on visual acuity (VA) in people with bacterial corneal ulcers. The primary outcome of VA for SCUT was measured on a scale called logmar, which is short for logarithm (base 10) of the minimum angle of resolution. A logmar of 0 corresponds to 20/20 vision, a logmar of 1 to 20/200 vision, and, in general, a logmar of x corresponds to a vision of $20/(20 \times 10^x)$ on an eye chart. Follow-up measures were taken at 3 weeks, 3 months, and 12 months. The predictors, all measured at the enrollment visit, are baseline VA, ulcer location (whether it covered the center of the eye), ulcer size (in square mm), and the type of infecting organism (gram positive versus gram negative). It is easy to envision what the full or “complete” data would consist of for this example: all participants have all their predictors measured at baseline and outcome information at baseline and each of the three follow-up times.

For regression analyses, a key distinction with regard to missing information is whether or not we have a considerable percentage of observations for which the predictors are missing but we have a measured outcome. This is important because, in regression analyses we typically model the distribution of the outcome variable and treat the predictor variables as fixed (see Sect. 3.3.3). If the predictor variables are missing for some observations (e.g., glucose values in the HERS example) then we need a method for inferring those missing values and assumptions will have to be made with respect to their distribution.

In the HERS example above, we used multiple imputation to build a model, temporarily treating glucose as an *outcome* variable in a linear regression model. That model assumes that glucose follows a normal distribution (for fixed values of the predictors of that model). That is, we have to make a distributional assumption about a variable that was a predictor in the original model (a regression of blood pressure on glucose), something we did not have to do before.

In more complicated examples, with multiple missing predictors, we would have to account for not only the distribution of each missing predictor by itself but also the joint distribution, including aspects such as correlations between predictors. In the not uncommon situation where the predictors with missing values consist of nominal, ordinal, and skewed variable types, specifying a distribution for how they are all jointly associated is a daunting task.

A simpler situation to handle is when there is little or no missing information on the predictors and missing data are mainly in the outcome, or both outcome and predictors are missing (as when a participant drops out of a study). In such cases, we can focus on the outcome variable, for which we are already hypothesizing a distribution, and categorize the missing data mechanisms relatively succinctly.

11.2.1 *Mechanisms for Missing Data*

Because we will want to describe the way in which the data came to be missing, it is worthwhile to consider a formal statistical model and develop some notation.

In that spirit, we envision a “complete” dataset, where all the data are present. We will think of this in the context of a longitudinal cohort study with regularly scheduled observation times, but the ideas apply more generally. Our complete dataset would be one with all outcome and all predictors measured on each person for each visit. Next, consider one of the variables that actually has missing data.

Let R_{it} be 1 if the i th participant has a measured value of the variable with missing data at visit time t and zero if it has “become” missing. So R is a binary indicator of whether a data value is present or not. For each variable that has missing data, we can now classify various missing data mechanisms by how they relate to the probability that $R_{it} = 1$. If factors are unrelated to this probability, then they have no bearing on the missing data process.

A common practice with missing data in a longitudinal study is to look at baseline characteristics of participants who had missing data later in the study. If variables differ significantly between those with and those without missing data (e.g., their age, gender, or baseline value of the outcome) then we can begin to understand what is related to $R_{it} = 1$. For example, [Splieth et al. \(2005\)](#) obtained a baseline oral health assessment of all first- and second-grade schoolchildren in a city in Germany. They compared the oral health of children whose parents did and did not allow them to participate in a cavity prevention program and longitudinal follow-up. They found that the children not participating were older and had poorer dental health compared to the participants. Failure to recognize this selective participation would result in biased estimates of average values. The formal classification scheme we consider next takes the idea of relating missing data to baseline covariates a step further.

11.2.1.1 Missing Completely at Random (MCAR)

There are three common classifications of the missing data process. Data are said to be *missing completely at random* (MCAR) if $P(R_{it} = 1)$ does not depend on any of the variables. For SCUT this would mean, for example, that the probability a logmar value at 3 months was missing was unrelated to the previous, current or future logmar values and also unrelated to visual acuity, ulcer location, ulcer size, or type of infecting organism. If we observed, for example, that participants with very poor logmar values at baseline were less likely to return then we would know that the MCAR scenario would not apply.

With $\mathbf{X}$ representing all the predictors and $\mathbf{Y}$ representing all the outcomes, this can be formally stated as

$$P(R_{it} = 1 | \mathbf{Y}, \mathbf{X}) = P(R_{it} = 1), \quad (11.1)$$

i.e., the probability of the data being missing is not associated with any part of the data. Another way to interpret this is that knowing the values of the outcomes and predictors would not change our estimate of the likelihood that a particular data

value is missing. While a useful conceptual “baseline” definition, MCAR is often not a reasonable assumption. For example, in longitudinal studies where there is missing data, there is almost invariably more missing data later in the study. So, at the very least, the predictor time or visit would be associated with the probability that an observation is missing.

11.2.1.2 Covariate-Dependent Missing Completely at Random (CD-MCAR)

A minor, but important, variation of this definition is *covariate-dependent missing completely at random* (CD-MCAR), which is mainly applicable to missing *outcome* data. In this situation, the probability of the outcome being missing can depend on the predictors which are part of the statistical model but does not depend on the other outcomes. With $\mathbf{X}^{\text{obs}}$ representing all the observed information for predictors which will be included in our model, we would formally write this as

$$P(R_{it} = 1 | \mathbf{Y}, \mathbf{X}) = P(R_{it} = 1 | \mathbf{X}^{\text{obs}}). \quad (11.2)$$

For SCUT this would mean, for example, that the probability a logmar value was missing was unrelated to the 3 weeks, 3 months, or 12 months logmar values but could be related to visit, VA, ulcer location, ulcer size, or type of infecting organism. If we observed, after accounting for differences due to the predictors, that participants with very poor logmar values at 3 weeks were more likely to return at 3 months then we would know that the covariate-dependent MCAR scenario would not apply.

11.2.1.3 Missing at Random (MAR)

A yet more flexible specification is that data are *missing at random* (MAR). This assumption handles a variety of more plausible scenarios. In MAR, the probability of missing data may depend not only on the covariates in the model but also on observed outcomes.

With $\mathbf{Y}^{\text{obs}}$ representing all the observed outcome information, formally this would be written as

$$P(R_{it} = 1 | \mathbf{Y}, \mathbf{X}) = P(R_{it} = 1 | \mathbf{Y}^{\text{obs}}, \mathbf{X}^{\text{obs}}). \quad (11.3)$$

In the SCUT example, the MAR scenario would allow for people with worse VA at 3 weeks or 3 months to be missing more frequently at 12 months and also to depend on visit, baseline logmar, VA, ulcer location, ulcer size, or type of infecting organism. In the HERS example, in Table 11.6, we artificially created data that was MAR.

11.2.1.4 Missing Not at Random (MNAR)

Finally, it may be that the probability a data value is missing depends on unobserved quantities, for example, the outcome we would have measured were it not missing. For instance, consider SCUT patients with identical baseline and 3 week visual acuities. Suppose the ones whose VA did not improve are more likely to make the 3 month visit (to get checked by the doctors). Then the fact that the data are missing would depend on the unobserved 3-month outcome. This scenario is called *missing not at random* or MNAR. In the HERS example, in Table 11.7, we artificially created data that was MNAR.

More formally, simplification of the model for $P(R_{it} = 1 | \mathbf{Y}, \mathbf{X})$ would not be possible as we did in, for example, (11.2). Unfortunately, but perhaps not surprising and because MNAR depends on *unobserved* quantities, we cannot verify or rule out a MNAR process from the observed data alone. Instead, if we suspect the data are MNAR the best we can do is conduct sensitivity analyses. One way to do so is via multiple imputation, described in Sect. 11.5.

Why are these characterizations important? Their utility is that we can now describe more carefully when standard types of analyses can be expected to give answers free of bias due to the missing data. We give more details and caveats beginning in Sect. 11.5 but in essence:

- When the data are MCAR, any method of analysis will give unbiased answers.
- When the outcome data are CD-MCAR, and those covariates are included in the statistical model, any method of analysis will give unbiased answers for regression coefficients and predicted values. Care still needs to be taken with calculations that average over values of the covariates (e.g., an average of the predicted values, or estimation of marginal effects) because those may not have the same distribution of covariate values as in the complete data.
- When the outcome data are MAR, correctly specified, likelihood-based analysis methods (e.g., mixed models) will give unbiased answers, but other methods (e.g., GEEs) may not.
- When the data are MNAR, any standard method of analysis may be biased.

Moving from MCAR to CD-MCAR accommodates the common situation in which missing data depend on measured covariates. Going from CD-MCAR to MAR allows even more elaborate dependence of the missing data—on measured covariates *and* outcomes—and will therefore include missing data mechanisms that have a higher chance of being applicable in practice. This makes likelihood-based methods especially attractive because they can continue to give unbiased answers even if the data are MAR. To reflect this fact, data which are MAR (or the more stringent requirements of MCAR or CD-MCAR) are sometimes called “ignorable”. Notably, although we postulate the MAR condition in terms of (11.3), if we are using likelihood-based methods, we need not specify a explicit statistical model for it, no matter how complicated the dependence might be. Instead, we can focus on developing a model for the complete data. This avoids being distracted by modeling a missingness mechanism which is likely to be imperfectly understood.

11.3 Simple Approaches to Handling Missing Data

We begin our discussion of methods of addressing missing data with a number of simple (and sometimes simplistic) methods that have been used previously. We return to the context of the HERS and SCUT trials.

11.3.1 *Include a Missing Data Category*

A simple approach to completing a dataset with missing values in a categorical *predictor* is to define a separate category for the missing values. In Sect. 4.3, we note that women in the HERS cohort responded to a question about how physically active they considered themselves compared to other women of their age. The five-level response ranged from “much less active” to “much more active”, and was coded in order from 1 to 5. A solution to missing data for this predictor is to define a category designated as “missing”. Physical activity is then analyzed as a categorical variable with six categories with all observations in the sample having a defined value for this variable. This is appealing because it avoids imputing values to incomplete observations but allows all observations to be used in the analysis.

Unfortunately, this can create biased estimates for other regression coefficients in the model, even when the data are MCAR. The reason for this is that, for the subset coded as missing, we are not adjusting for the value of physical activity, whereas for the rest of the data we are. So regression coefficients (for predictors other than physical activity) that are estimated from the model using the six category version of physical activity are a blend of the coefficient before and after adjustment for physical activity. Bias is introduced when the unadjusted and adjusted coefficients differ and there is a sizeable percentage of observations in the missing data category. On the other hand, if the adjusted and unadjusted coefficients are similar and the percentage of observations in the missing data category is small, little bias will be introduced.

11.3.2 *Last Observation or Baseline Carried Forward*

In SCUT, vision tends to improve rapidly in the first month as the infection is treated and has usually stabilized by 3 months. As patients feel better, they are less likely to return to the clinic for follow-up appointments and nearly 30% of 12 month visual measurements are missing due to loss to follow-up.

One approach to handling a missing 12-month outcome value in the SCUT trial is to use (or “carry forward”) a patient’s 3 month VA measure. If the 3-month value is not available the 3-week (or, if that is missing, the baseline value) value would be used. This approach is called *last observation carried forward* (LOCF) because missing values are filled in with the last available value from the same person. This

approach can be used with either outcomes or predictors. The LOCF approach has the appeal of using the most proximate available VA measure to complete the data. It has been argued that this is a conservative method because it assumes no change in measured values for the missing data.

The method has substantial disadvantages. In SCUT, for instance, visual acuity improves substantially from 3 weeks to 3 months. Hence, LOCF would be implausible for such data and almost certainly underestimate VA if values are carried forward, potentially leading to biased estimates. Second, a single value is substituted for the missing value. As with single imputation, if a standard analysis is then applied to the completed data set, this uncertain, filled-in value is treated as if it were an actual measurement and leads to falsely precise analyses. This is a concern whether or not carrying forward values is approximately correct, on average.

Consider a study of people initiating an experimental treatment to reduce hypertension with repeated measures of their blood pressure, subject to missing values. If the missing values are due to study dropout and the participants must discontinue the experimental treatment, then we might reasonably expect that the blood pressure values would return to pretreatment levels. This would be captured by using the baseline value (rather than the last value prior to dropout) to fill in missing values. This approach is termed baseline value carried forward (BCF) and it is very similar in spirit and execution to LOCF except that a baseline value is used to replace the missing value. While imputing using the baseline value might be reasonable for the above example, the immediate return to baseline assumption may not be plausible in other contexts. BCF, like LOCF, under-accounts for the variation due to the single imputed value.

11.3.2.1 Other Single Imputation Approaches

Other approaches use information from the remainder of the data set to infer a single value. Suppose values of a variable like income are missing in a sample. A typical value, such as the mean or median of observed values, could be used. While this can generate a reasonable value for a continuous value, like income, mean values would produce an implausible value for a categorical value, like race. For categorical variables, the method could be adapted to impute the race as the most common answer (e.g., white) if the variable is categorical. The main advantage of all of these “single imputation” approaches is their simplicity in generating the imputation (substituting means, modes, or previously measured values). However, this simplicity may reflect a lack of critical thinking about the relationship of missing data to observed data. In the SCUT trial for example, a better imputation for a missing 3 month VA measure might be to use the 3-week value augmented by the expected change in VA from 3 weeks to 3 months.

With a variable such as income, it is highly possible that the value to be measured contributes to the chance that it will not be observed, which might lead to data that are MNAR. A better approach to imputation might use values of other covariates such as zip code, age, and/or education to predict the missing values of income. If those covariates were able to account for the dependence of missingness on income,

then the data would be MAR. Thus, superior imputations will need to be informed by a model for the data and for the mechanism which underlies the missing values. Methods such as LOCF or BCF skip this crucial step of model development.

Furthermore, any single imputation approach that applies standard analysis methods to the completed data set can seriously underestimate the variation in the data set, giving standard errors that are too small and CIs which are too narrow. These deficiencies can be corrected by applying the method of multiple imputation which we discuss in Sect. 11.5.

11.4 Methods for Handling Missing Data

We now return to more general approaches for handling missing data. The recommended methods depend on both the pattern of missing data (drop out from the study, missing predictors only, etc.) and the missing data mechanism. A key distinction is whether there is missing data for the predictor variables with at least some of those instances having observed values of the outcome. In such a case, we recommend using multiple imputation, described in more detail in Sect. 11.5.

For situations in which the predictors are mostly complete and the issue is data missing in the outcome variable, we divide our presentation and recommendations by the mechanism of the missing data: missing completely at random (MCAR—Sect. 11.7), covariate-dependent missing completely at random (CD-MCAR—Sect. 11.8) or missing at random (MAR—for hierarchical analyses only and in Sect. 11.9). When the data are MCAR or CD-MCAR, relatively simple approaches may suffice. For data that are MAR, several approaches are possible.

11.5 Missing Data in the Predictors and Multiple Imputation

The first distinction in recommended analysis strategies is whether there is missing data in the predictors (even if there is also missing data in the outcomes) and the missing data can be assumed to be MAR. With missing predictor data, we recommend the approach of multiple imputation, which we introduced briefly in Sect. 11.1.2. The basic idea is not only to fill in a reasonable value for the missing data but also to incorporate some random error. While it may seem counterproductive to add in random error, it is a convenient device for properly reflecting the degree of uncertainty due to the missing data. By doing it a number of times (hence the adjective multiple in multiple imputation), we can get valid estimates of standard errors (and hence CIs and p -values), and by averaging the results, not have them unduly affected by the random error. It turns out, perhaps surprisingly, that the process does not need to be repeated very many times. A typical number is five or ten.

Table 11.8 Regression model for imputing glucose

```
. regress glucose bmi csmker white sbp diabetes
```

Source	SS	df	MS	Number of obs =	1355
Model	1019590.52	5	203918.103	F(5, 1349) =	213.11
Residual	1290806.09	1349	956.861448	Prob > F =	0.0000
				R-squared =	0.4413
				Adj R-squared =	0.4392
Total	2310396.61	1354	1706.34905	Root MSE =	30.933

glucose	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
bmi	.4921757	.1568229	3.14	0.002	.1845325 .7998189
csmker	1.183684	2.603247	0.45	0.649	-3.923168 6.290536
white	9.180278	2.863755	3.21	0.001	3.562382 14.79817
sbp	.0342977	.0447849	0.77	0.444	-.0535579 .1221532
diabetes	60.45312	1.977712	30.57	0.000	56.57339 64.33285
_cons	69.66885	8.108395	8.59	0.000	53.76242 85.57528

The steps in multiple imputation are essentially as follows:

- (1) Specify a probabilistic model for how to fill in the missing data.
- (2) Using the model, fill in (impute) the missing data with random error.
- (3) Repeat the imputation a small number of times (e.g., five) so you end up with multiple versions of the data set, each with somewhat different values of the imputed variable(s).
- (4) For each of the imputed data sets, calculate the quantities of interest (e.g., the regression coefficients).
- (5) Average the quantity of interest across the imputed data sets.
- (6) Calculate a standard error based on the average of the model-based variation plus the variation in the calculated quantities of interest across the imputed data sets.

The first step, of specifying the imputation model, is the most difficult and involves building a regression model for the variable with missing data; the subject of this entire book! The remaining steps are relatively automatic and are handled by statistical software.

For the HERS example of Sect. 11.1.2, the variable we imputed was glucose. Our probabilistic model was a linear regression model for glucose with predictors of SBP, BMI, being white, current smoking status, and development of diabetes. The standard assumption of a linear regression model is that the error terms are normally distributed. Table 11.8 gives the output from fitting that regression equation.

Given values of BMI, current smoking status, being white, SBP, and development of diabetes, we can use the regression equation to generate a predicted glucose value for those with missing values. But in multiple imputation we do more. The regression output from the table also gives the value of the root mean square error, 30.933, which quantifies the degree of uncertainty (i.e., the error term) in the regression equation. Under the assumptions of the linear regression model, those

Table 11.9 Stata code for imputing glucose

```

mi set wide
mi register imputed glucose
mi impute reg glucose bmi csmker white sbp diabetes, add(5) rseed(271828) ///
> force
mi estimate: regress sbp glucose white bmi

```

errors are normally distributed, with means zero and standard deviation 30.933. So to impute the missing values of glucose, we calculate the predicted value of glucose and then add a normally distributed error term with standard deviation 30.933.

As an example, one of the HERS participants who had a missing glucose measurement had a BMI of 24.68, was not a current smoker, was white, had a SBP of 130, and was not diabetic. Using the coefficients from Table 11.8, her predicted glucose value is 95.45. To impute a value for her, we would add a normally distributed error term with mean zero and standard deviation 30.933. Using the `rnormal(0,30.9333)` command twice in Stata to generate random normal variables with the correct mean and standard deviation gave the values 42.98 and -13.34 . So her imputed glucose value for the first imputed data set would be $95.45 + 42.98 = 138.43$ and would be $95.45 - 13.34 = 82.11$ for the second. This process is repeated for each of the missing glucose values and each imputed data set.

Next, for each imputed dataset, we perform the regression of SBP on glucose, BMI, and white. Suppose our interest lies in understanding the relationship between SBP and BMI. We would have five estimates of the regression coefficient, each slightly different due to the different imputed values. Averaging those five estimates gives us our multiple imputation estimate and the standard error is calculated both from the model-based standard errors and the variation in the coefficients from imputed data set to imputed data set, which measures the amount of uncertainty due to imputing the values of glucose.

Across the five imputations, the values of the coefficient for BMI were 0.145, 0.149, 0.138, 0.150, and 0.152 with an average of 0.147. The model-based standard errors were 0.083, 0.083, 0.084, 0.083, and 0.084 with an average of 0.083. So the estimated coefficient for BMI from the multiple imputation is 0.147 and the standard error is slightly higher than the average of the model-based standard errors (due to the imputation to imputation variability) and is equal to 0.084.

While no one step in the multiple imputation process is difficult, conducting the multiple imputations and analyses and assembling the results is tedious and could be error-prone if done manually. So programs like Stata automate the process. The results reported in Table 11.3 were generated with the Stata code given in Table 11.9.

11.5.1 Remarks About Using Multiple Imputation

In the HERS example, we used the outcome of the original analysis (SBP) to impute glucose. And then we turned around and used the imputed glucose value in the

regression of SBP on glucose and the other predictors. This may seem like cheating but is actually needed to obtain unbiased estimates (Little 1992). In fact, multiple imputation does not distinguish the roles of outcome and predictor, but instead regards all the variables on an equal footing. So, whenever variables are associated with one another, it becomes important to utilize that information in the imputation model. And it is usually the case that we include predictors (e.g., glucose) in our modeling precisely because we expect them to be associated with the outcome. So if those predictors are missing, it is important to use the original outcome variable in the imputation modeling.

Multiple imputation has a long history of use in sample surveys. Many surveys (like NHANES, the National Health and Nutrition Examination Survey) are extremely comprehensive and are released for public use. In this case, it is difficult to predict which analyses will be undertaken, which variables will be used in analyses, and which will be treated as outcome variables. Because multiple imputation does not distinguish outcomes from predictors, users need not worry about the distinction with regards how the data were imputed. Contrast this with a method like creating a missing data category, which only works for categorical predictors. Furthermore, the imputation algorithms may be quite complicated and difficult to implement, but the analysis of the data is straightforward using routines like those available in Stata. So once a multiple imputation is produced, it can be used in a variety of situations.

Because of the potential for using a multiply imputed data set for many purposes, when imputing missing data it is important to err on the side of flexibility rather than parsimony. If the model for imputation is overly simplistic, those assumptions will be built into the portion of the data that has been imputed. For example in HERS, if the relationship between glucose and BMI were non-linear, but the imputation model assumed it to be linear, then predictions might be biased. Or if we assumed there was no interaction between BMI and race in imputing glucose and if a later analysis searched for interactions, interaction effects would be attenuated.

11.5.2 Approaches to Multiple Imputation

In practice, the pattern of missing data across variables can be quite different. In the HERS example, BMI and current smoking status (yes/no) had missing data in addition to glucose, whereas race (white versus other) had very little missing data. With multiple missing variables, a number of complications arise. First, how should we handle variables with distributions other than normal such as current smoking status, which is binary? Second, if we want to impute glucose using SBP, BMI, race, and current smoking status, what do we do about the missing data for BMI and current smoking status? Third, once the data are filled in, how should we update the parameter estimates? There are three main approaches for dealing with multiple imputation across a number of variables: iterative chained imputation, multivariate normal (MVN) imputation, and Monte Carlo Markov chain. We describe in more detail the first two.

11.5.2.1 Iterative Chained Equations Imputation

Using iterative chained equations (ICEs) imputation, we build regression models for each of the variables with missing data, in turn treating them as outcome variables and using the rest of the variables as possible predictors. For the HERS example, in addition to the model we have already built for glucose, we would need models for BMI and current smoking status. Because current smoking status is a binary variable, a logical imputation model would be to use a logistic regression with predictors of SBP, BMI, race, and glucose. From the logistic regression model, we would get a predicted probability of being a current smoker. We would then generate a random binary outcome with the predicted probability (which could be done using the `rbinomial` command in Stata). Once the value of current smoking was imputed, this could be used as a predictor in a regression model to impute BMI. These regression equations are used to fill in each of the variables in turn. The whole process is repeated a number of times to reach a “steady state” so that the results do not depend on the order in which the variables are imputed.

An important advantage of this approach is the ability to tailor the model for imputing each variable, both with respect to its distribution (e.g., normal, binary, or multiple categories) as well as the inclusion of predictors, possibly with non-linear terms and/or interactions. Currently, Stata allows regression models of the following types via its `mi impute chained` command: linear regression (regular, truncated, and interval), logistic (binary, ordinal, and multinomial), Poisson, and negative binomial. This is also its most important disadvantage: a regression model has to be constructed for each of the variables for which there is a significant percentage of missing data. With, say, 20 variables with missing data, the regression modeling effort increases 20-fold, even though this may not be the scientific focus of the analysis. These regression models need to be built with care so as not to introduce out of range or implausible imputed values.

11.5.2.2 Multivariate Normal Imputation

A simpler to use method available in many statistical software packages is to impute the missing data assuming all the variables follow a joint, normal distribution. While this is invariably an incorrect assumption when there are a number of variables with missing data, it has often been found to perform well in practice. This is because, even though the distributional assumptions may be suspect, imputation assuming a MVN distribution still retains the proper average values and correlations among the variables. When the later analysis depends only on such quantities (as when the ultimate analysis is a linear regression) this method may suffice.

For example, in Sect. 11.5.3 we wish to impute the binary predictor variable race, which is coded as 1 for white and 0 otherwise. Recall that when a variable is coded as 0 and 1, its mean is equal to proportion of observations falling in the category coded as 1. When imputing such a variable, MVN imputation will generate a continuous variable in its place, but one which will have the proper mean (in the sense that the

mean will properly reflect the proportion falling in category 1). Of course, care must be taken when using such a variable in an analysis: since it is no longer categorical it cannot be treated as such in a prediction equation. Software packages such as SAS allow the user to round off to 0 or 1 to recover this aspect of the data.

Although MVN imputation often gives sensible answers, in some cases it may be important to retain more detailed aspects of the distribution (e.g., the proportion exceeding a threshold), and MVN imputation may lead to suspect conclusions. Another situation in which the multivariate normal assumption is not satisfactory is when one or more variable is a nominal categorical variable, e.g., marital status (single and never married, married, divorced).

If most of the variables to be imputed are approximately normally distributed and there are no nominal categorical variables, then it is probably safe to use MVN imputation, which is often easier to implement in practice. However, if there are nominal categorical variables, or the predictors are highly nonnormally distributed, then iterative chained imputation is the recommended approach.

11.5.3 Multiple Imputation for HERS

We demonstrate the use of ICEs imputation and MVN imputation using the HERS dataset and two regression analyses: regression of SBP on glucose, BMI, and race (white or not), which has missing data on two continuous predictors (glucose and BMI) and the regression of SBP on glucose, current smoking status (yes/no), and race, which has missing data on a continuous predictor (glucose) and a binary predictor (smoking status).

Using the ICE methodology, we built linear regression models for glucose and BMI and a logistic regression model for current smoking status. We considered two approaches to modeling: parsimonious and flexible. In the parsimonious approach, we included the other variables in the imputation model as is. So, for example, the parsimonious imputation model for BMI was a linear regression with predictors of SBP, glucose, race, and current smoking status. In the flexible approach, we included all two way interactions and quadratic versions of numerical predictors. So, for example, the flexible imputation model for current smoking status was a logistic regression with predictors of glucose, BMI, and SBP, the squared versions of each of those, race and all the two way interactions such as race by BMI, race by SBP, BMI times SBP, etc.

We compared this to the MVN approach, which assumes that SBP, glucose, BMI, and current smoking status are MVN and imputes the values under that assumption. Table 11.10 lists the sample sizes, regression coefficients, and p -values for a complete case analysis and the three approaches to multiple imputation. Similarly, Table 11.11 lists the values for a regression of SBP on glucose, race, and current smoking status. We might expect the MVN approach to do more poorly for this model since current smoking status is a binary variable.

Table 11.10 HERS model fit comparisons with different multiple imputation strategies: regression of SBP on glucose, race, and BMI

MI method	N	Parameter estimates			p-values		
		Glucose	Race	BMI	Glucose	Race	BMI
Complete case	1385	0.030	−1.54	0.06	0.02	0.36	0.49
ICE parsimony	1871	0.029	−2.52	0.13	0.02	0.09	0.11
ICE flexible	1871	0.028	−2.53	0.14	0.02	0.09	0.09
MVN	1871	0.030	−2.44	0.14	0.02	0.10	0.10

Table 11.11 HERS model fit comparisons with different multiple imputation strategies: regression of SBP on glucose, race, and current smoking status

MI method	N	Parameter estimates			p-values		
		Glucose	Race	Smoke	Glucose	Race	Smoke
Complete case	1370	0.028	−2.04	−1.55	0.02	0.23	0.32
ICE parsimony	1871	0.032	−2.75	−0.96	0.007	0.06	0.49
ICE flexible	1871	0.032	−2.77	−0.94	0.006	0.06	0.49
MVN	1871	0.033	−2.68	−1.01	0.005	0.07	0.46

The imputation analyses differ from the complete case analyses in several important aspects:

- The imputation methods are based on imputed versions of the complete data set with 1,871 observations.
- For a number of the coefficients, the imputations give materially different estimates of the coefficients compared to the complete case analysis, e.g., the coefficient for race.
- The imputations, which use all the observed data, often have smaller p -values than the complete case analysis.

Turning to comparisons among the various imputation methods, we observe that

- The flexible and parsimonious approaches to ICE gave virtually the same answers.
- The MVN approach gave somewhat different answers than the two ICE approaches, but all three imputation approaches gave answers similar to one another and somewhat different than the complete case analysis.
- The MVN approach seemed to do a creditable job even when imputing the binary variable, race.

The example serves to illustrate both the advantages and disadvantages of multiple imputation. It uses all the observed data while properly reflecting the fact some of the data are missing. It may have reduced the bias in some of the regression coefficients. It properly reflects the fact some of the data are missing but allows for reduced standard errors and generally smaller p -values. But it came at the cost of either having to construct a model for each of the original predictor variables (for ICE) or hypothesize a MVN model for all the predictor variables that had substantial missing data and led to a somewhat more complicated overall analysis.

11.6 Deciding Which Missing Data Mechanism May Be Applicable

The key to using multiple imputation is to build regression models to fill in predictors or outcomes that have missing data. When the predictors have missing data, the outcome variables will usually be part of the imputation models. In the next few sections, we consider situations where the main missing data are *outcome* data. Our recommended strategies depend on which missing data mechanism is to be assumed so we give some guidance here as to how to choose.

As noted above, e.g., (11.2), the different missing data mechanisms are distinguished by dependence of the probability of the data being missing on different quantities. In CD-MCAR dependence is on covariates and, in MAR, dependence is on the outcome and possibly also on covariates. Distinguishing between these cases can be done in a descriptive manner or using a more formal statistical model.

For example, in the SCUT trial and considering missing outcome (visual acuity) data at the 3-month visit, we would calculate descriptive statistics for those with and without missing data. If the average ulcer size, the proportion with the ulcer in the center of the eye, or the proportion of gram positive infections (all measured at baseline) differed between those with a missing VA measurement at 3 months and those with it present, then we would know the data could not be considered MCAR, but instead would be at least CD-MCAR. We could formally test the association by conducting a t -test for ulcer size or χ^2 tests for whether the ulcer is in the center of the eye or type of infection across the missingness groups.

Alternatively, we could define an indicator variable R_i , equal to 1 if the 3-month measurement was present and zero otherwise, and conduct a logistic regression to assess the association of missingness with the covariates. If we found that any of the covariates is associated with missingness, it would establish that the data could not be MCAR.

By further considering previously measured outcomes (e.g., the value of VA at 3 weeks), we can check to see if the CD-MCAR assumption is inadequate. If the VA at 3 weeks differed between the groups with 3 month VA data present and absent that would suggest the missing data mechanism to be at least MAR. More rigorously, if VA at 3 weeks was predictive of missing VA at 3 months in a logistic regression model that also contained the covariates that were related to missingness then we would know that the assumption of CD-MCAR was inadequate.

With substantial amounts of missing data, it is invariably good practice to conduct descriptive analyses to understand to what extent the missing data are associated with measured variables. As noted above this can help rule out simpler mechanism such as MCAR (which rarely holds in practice) and CD-MCAR. As noted earlier, because MNAR depends on *unobserved* quantities, we cannot verify or rule out a MNAR process from the observed data alone.

11.7 Missing Outcomes, Missing Completely at Random

We now consider datasets for which there is missing data in the outcomes but where any missing data in the predictor variables is negligible or occurs along with missing data in the outcome (as when a participant drops out of a study). The easiest case to deal with is when the data are MCAR, i.e., the missing data are totally unrelated to either the other outcomes or the predictors. In this case, ignoring the missing data does not cause bias and simply leaving the missing data out of the analysis properly reflects the amount of information available. In this case, complete case analysis using any of the usual statistical analysis strategies (e.g., linear regression or logistic regression) is the recommended strategy. It will automatically be adopted by any of the usual statistical packages, including Stata, if you conduct the usual analysis in the presence of missing data.

We again return to the HERS dataset and we fit a model to predict systolic blood pressure (SBP) from BMI, race (white or not), whether the participant was on medication to control their blood pressure (yes/no) and the interaction of BMI and blood pressure medication. From that model, the coefficient of BMI in the on-medication group was 0.24, with a standard error of 0.06. So with each increase in BMI of one unit, there is an associated increase in SBP of about 0.24. However, in the off-medication group, the coefficient is 0.52 with a standard error of 0.11. This is not surprising as we would expect those on medication to have their blood pressure better controlled and less associated with BMI.

We again artificially create missing data to illustrate the consequences. Using a random mechanism, we dropped 75% of the data and refit the above model, so the missing data mechanism is MCAR. The on-medication BMI coefficient was 0.19 with a standard error of 0.10 and the off-medication coefficient was 0.56 with a standard error of 0.21. So, even though we have dropped 75% of the data, the two coefficients are similar to those obtained from the full dataset, as expected. Using GEEs gave virtually the same results, with coefficients of 0.18 and 0.56, respectively.

11.8 Missing Outcomes, Covariate-Dependent Missing Completely at Random

The next level of missing data occurs with data where missingness may depend on a covariate that is in the analysis model as a predictor, but does not depend on other variables (either other outcomes or variables not in the model), that is, covariate-dependent missing completely at random (CD-MCAR). Under CD-MCAR, a complete case analysis yields unbiased estimates of regression coefficients and predictions for given values of the covariates using any of the regression methods we have described. However, quantities that require averaging over members of the sample may not be correct.

Using the HERS data, as in the previous section, we again randomly dropped 75% of the data, but this time all the dropped data was from the on-medication subgroup, which makes up about 80% of the full dataset. This missing data mechanism would be CD-MCAR because it depends on whether the participant is on hypertension medication or not, but not on other variables. We fit the same model as described in the previous section and obtained an on-medication coefficient for BMI of 0.23 with a standard error of 0.16 and an off-medication coefficient of 0.59 with a standard error of 0.12, with the coefficients again quite similar to the full dataset.

But suppose we were interested in the average increase in SBP associated with a one unit increase in BMI. Since the no-medication participants make up about 80% of the cohort, the average increase is a weighted average of the two coefficients: $0.30 = 0.8(0.24) + 0.2(0.52)$. Being more careful with the calculations, the exact value is actually 0.29. But in the CD-MCAR scenario, the proportion of on-medication participants is only about 24%. And so the average increase will be misestimated as 0.51, because the off-medication participants are weighted too heavily.

The correctly blended average can be calculated using Stata's `margins` command as shown in Table 11.12. In that table, `bmi_ctr` is the centered value of BMI (i.e., it has mean zero) and `sbp_cdmcar` is SBP with values missing due to the CD-MCAR mechanism. The `margins` command estimates the value of SBP at the mean value of BMI and at one unit above the mean. Using just the estimation sample gives an associated increase in SBP of about 0.51 ($= 135.1076 - 134.6008$). However, using the `noesample` option generates an estimate for the entire sample, recovering the proper weighting of the on- and off-medication subgroups, and gives an estimate of about 0.30 ($= 133.2468 - 132.9490$), quite close to the full data estimate.

As in the MAR scenario, for CD-MCAR the particular analysis method makes little difference. Using GEEs gave virtually the same answers as the mixed-model approaches reported above.

11.9 Missing Outcomes for Longitudinal Studies, Missing at Random

Longitudinal studies with a planned observation schedule invariably have at least some missing data. Although attempts are usually made to have participants return for every scheduled visit (e.g., yearly), some drop out of the study, either voluntarily or involuntarily (e.g., death), or miss visits. A consequence is that all data that would have been collected at that visit (either outcomes or predictors) will be missing. So use of analysis strategies to minimize bias due to missing data are essential. For example, the Osteoarthritis Initiative, a well-conducted cohort study, enrolled 4,796 individuals, attempting to collect data yearly. After 1 year, 94% were still

Table 11.12 Using the margins command with CD-MCAR missing data

```
. xtmixed sbp_cdmcar c.bmi_ctr white htnmeds htnmeds#c.bmi_ctr || pptid:

Mixed-effects REML regression                Number of obs    =    2291
Group variable: pptid                      Number of groups   =    972

                                           Obs per group: min =    1
                                           avg   =    2.4
                                           max   =    6

                                           Wald chi2(4)       =    30.08
                                           Prob > chi2        =    0.0000

Log restricted-likelihood = -9527.0924
```

sbp_cdmcar	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
bmi_ctr	.5946715	.1249834	4.76	0.000	.3497085	.8396344
white	.6615583	2.262646	0.29	0.770	-3.773146	5.096263
htnmeds	-2.858138	.9904139	-2.89	0.004	-4.799314	-.9169625
htnmeds# c.bmi_ctr						
1	-.3666237	.1957226	-1.87	0.061	-.750233	.0169856
_cons	134.6577	2.258123	59.63	0.000	130.2318	139.0835

```
-----+-----
Random-effects parameters | Estimate Std. Err. [95% Conf. Interval]
pptid: Identity          +-----+-----
sd(_cons)                | 14.58279 .4669968 13.69562 15.52742
-----+-----
sd(Residual)             | 11.41371 .2233941 10.98415 11.86006
-----+-----
LR test versus linear regression: chibar2(01)=736.07 Prob >= chibar2= 0.0000

. margins, at(bmi_ctr=0) at(bmi_ctr=1)

Predictive margins                Number of obs    =    2291

Expression : Linear prediction, fixed portion, predict()

1._at      : bmi_ctr          =    0
2._at      : bmi_ctr          =    1
```

	Margin	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
_at						
1	134.6008	.5746835	234.22	0.000	133.4745	135.7272
2	135.1076	.5946754	227.20	0.000	133.9421	136.2732

```
-----+-----
. margins, at(bmi_ctr=0) at(bmi_ctr=1) noesample

Predictive margins                Number of obs    =    9157

Expression : Linear prediction, fixed portion, predict()
1._at      : bmi_ctr          =    0
2._at      : bmi_ctr          =    1
```

	Margin	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
_at						
1	132.9490	.6731565	197.50	0.000	131.6297	134.2684
2	133.2468	.6880263	193.67	0.000	131.8983	134.5953

being followed, with 6% dead or lost to follow-up and, after 2 years, 90% were still being followed. If the data are MCAR or CD-MCAR then the analysis strategies suggested above will work for longitudinal data. But in a longitudinal study, it is quite possible that missingness is related to previously measured *outcomes*, making the data MAR. For example, in the OAI, a patient with an MRI (magnetic resonance image) showing advanced osteoarthritis at one visit may be less likely to come in for the next visit, since it would entail lengthy data collection and another MRI.

The situation of MAR represents a reasonable one for a wide variety of missing data problems. This is a middle ground between MCAR and MNAR for which the choice of analysis strategy can make a difference. Three general approaches have been suggested for dealing with MAR data in longitudinal studies: use maximum-likelihood based methods, use inverse weighting methods, or use multiple imputation.

11.9.1 *ML and MAR*

In Sect. 11.1.2, we contrasted the use of generalized estimating equations and linear mixed models in a particular example. Under a MAR situation we showed that the generalized estimating equations approach gave biased results whereas the linear mixed-model analysis did not. This result generalizes to the wider class of models fit by maximum likelihood (see Sect. 5.6 for the definition of maximum likelihood). Namely that a simple strategy for dealing with MAR data is to use approaches wherein the models are fit by the method of maximum likelihood, such as the random effects models described in Sect. 7.5. As long as the model is correct in both its fixed and random components, this fitting technique leads to methods that are not biased. A more detailed explanation as to why maximum likelihood avoids bias with MAR data is given below in Sect. 11.10.

Commonly used approaches which use maximum likelihood include linear mixed-model analyses (Stata `xtmixed` or `xtreg` [with the `mle` option]; SAS Proc MIXED; SPSS linear mixed model routines) and random effects logistic or Poisson regression models (Stata `xtlogit`, `xtmelogit`, `xtpoisson`, `xtmepoisson`, and others; SAS Proc NLMIXED). The primary method for longitudinal data which does *not* use maximum likelihood is GEEs (see Sect. 7.4), which is therefore subject to bias under MAR data.

When maximum-likelihood methods are a natural analysis strategy, we generally recommend them since they obviate the need to model the missingness mechanism. And for studies that are not on a regularly scheduled visit time, it is not clear what data should be imputed. When following a maximum-likelihood analysis strategy and for cases where there is a significant portion of missing outcome data, care should be taken on model diagnostics (e.g., checking for interactions and correct specification of the variance-covariance structure). This is because the ability of maximum likelihood to adjust for missingness depends on specifying a correct or nearly correct model.

11.9.2 Multiple Imputation

In a longitudinal study with MAR missing data, maximum-likelihood methods automatically correct for missing data without having to specify a model for the missingness. But multiple imputation is also a viable method, building a model to impute the missing outcomes based on the covariates and previously measured outcomes.

There are, however, circumstances in which multiple imputation is to be recommended over maximum likelihood. If the preferred analysis strategy is GEEs (or another, non-likelihood-based method) then multiple imputation is an attractive strategy to deal with missing data. This is because it can reduce the bias associated with the use of non-likelihood-based methods under MAR missing data.

So far we have assumed that, when missingness is dependent on the predictors, these are predictors that can be included in the analysis model. This will not always be the case, for example, if drop out in a longitudinal study depends on a mediator. In SCUT, for example, suppose that individuals whose ulcers have cleared by three weeks are less likely to return at 3 months since it is not as urgent for them to visit the clinic. To properly account for missingness in an ML analysis, we would need to include presence of an ulcer at 3 months in the model. But this will also adjust away some of the treatment effect, which we do not wish to do. This would be an example of a situation in which a variable (presence of an ulcer at 3 weeks) is needed to make the MAR assumption plausible but is not useful for the analysis model. This is another situation in which multiple imputation is an attractive approach: we can use the mediator in the imputation model, but leave it out of the analysis model.

11.9.3 Inverse Probability Weighting

Another family of methods which use the MAR assumption are those based on *inverse probability weighting*. The basic idea is to use complete observations to represent incomplete observations, just as we did for potential outcomes in Subject. 9.1.8. For instance, in the SCUT example, suppose that we could make the assumption that the probability of missing visual acuity (VA) at the second (3 month) visit depended only on the distance the patient lives from the clinic and their VA at enrollment.

In that case, for patient i at visit 2

$$P(R_{i2} = 1 | \mathbf{Y}, \mathbf{X}) = P(R_{i2} = 1 | \mathbf{X}) \quad (11.4)$$

or more specifically it is equal to $P(R_{i2} = 1 | x_{1i}, x_{2i})$, where x_{1i} is the distance patient i lives from the clinic and x_{2i} is his or her VA at baseline. This simplification means that we postulate covariate-dependent MCAR for the missing outcomes.

Suppose, for specific values of clinic distance and VA, the probability of observing the visit 2 outcome (11.4) is equal to $1/2$. Then only about half of the patients with these values of x_{1i} and x_{2i} will have VA data at the second visit and it would be reasonable to “double-count” their values to represent the missing values. Similarly, if the probability were $1/3$, we would observe only about $1/3$ of the outcomes and it would be reasonable to “triple-count” the participants for whom we observed the outcome. In general, we would up-weight observed outcomes by one divided by the probability of being measured, hence the name, inverse probability weighting. This is the spirit that underlies inverse weighting methods.

Many statistical packages allow the incorporation of weights, but care must be taken. Often, for example the `_weight` statement in SAS, a weight of 2 would represent 2 actual measured observations with the same value. This is distinct from our situation in which a weight of 2 would mean we are using a single measured value to represent itself and an unmeasured value. The weights that are needed for inverse weighting estimation are sometimes called *probability* or *sampling* weights and are implemented for many of the commands in Stata using the `pweight` option. Using the more standard weighting as in SAS gives the correct estimate, but incorrectly implies there is more actual measured data and hence will give standard errors and p -values that are too small and CIs that are too narrow. For some routines, this can be corrected by using robust standard errors.

11.9.3.1 Comments on Inverse Probability Weighting

Inverse probability estimates require that we specify or estimate the *probability of observing* an outcome at 3 months. We might do this by developing a regression model, like logistic regression, for the probability of a measured value in terms of observed data (much like the propensity score method discussed in Sect. 9.4.3). Other methods discussed in this chapter based on the MAR assumption rely on postulating a correct model for the outcomes. For example, in the SCUT trial, we might postulate a linear mixed-effects model for the VA measures. These approaches use MI- or ML- based estimation and are able to avoid specifying a model for the missing data mechanism but depend on the correctness of the outcome model to adjust for missing data. In contrast, inverse weighting adjusts for missing data through the weighting scheme and does not depend as strongly on the correctness of the outcome model. Inverse weighting has been suggested in situations using analyses such as generalized estimating equations.

We have several concerns about the use of inverse probability methods and cannot recommend them in general. In many situations, the probability of a measured value can be small, leading to large inverse weights. The large weight given to a few observations means that these values significantly influence the results, leading to unstable estimates and loss of efficiency.

If IPW is used, weights should be carefully monitored. And even if weights are not large, inverse weighting can be notably less efficient than an analysis based on a carefully chosen model for the complete data. In many, if not most, situations,

a plausible model for missingness is poorly understood. It is, therefore, often more natural to build a model for the complete data and apply methods based on maximum likelihood.

11.10 Technical Details About Maximum Likelihood and Data Which are Missing at Random

We have stated earlier that methods of fitting models using maximum likelihood give valid estimates even when the data are MAR. In this section, we give some explanation as to why that is so and contrast maximum-likelihood with multiple imputation. The comparison rests on a particular way in which maximum likelihood estimates can be calculated, called the Expectation–Maximization Algorithm, or *EM algorithm* for short, an approach that has often been of utility in missing data problems. The EM algorithm operates by starting with a guess as to the values of the estimates and improves them using an expectation calculation and then a maximization calculation. The expectation and maximization calculations are repeated until the estimates stabilize. This gives the same answer as directly finding the maximum of the likelihood of the observed data.

11.10.1 An Example of the EM Algorithm

Suppose we wanted to estimate the average number of emergency room visits per person in a year for the population of people served by a particular emergency room. But suppose we only had emergency-room (ER) data and did not know the size of the population who might use that emergency room. If we had data for everyone, we would just calculate the average value. But we have a problem since we do not have a record of those who did not visit the emergency room that year, that is, those whose outcome is equal to 0. And clearly calculating the average among those who *were* seen in the emergency room will drastically overestimate the average.

If we had a preliminary estimate of the average and a probabilistic model for how often people visit the emergency room, we could predict how many we would expect to have a zero value. One such model is the Poisson distribution, for which the probability of an individual visiting the ER exactly x times during the year, $P(x)$, is given by the formula

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad (11.5)$$

where λ is the average number of visits per year and $x!$ is “ x -factorial”, e.g., $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$, and, by convention, $0! = 1$. Plugging a 0 in for x

in (11.5), the probability of a person not visiting the ER in a year is $\frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda}$. So, if the average is 0.1 visits per year, the probability of no visits is $e^{-0.1} = 0.905$, and we would predict about 90.5% of the people in the population will have no visits.

Suppose our data consist of 1,232 separate people who visited the ER. Of those people, 1,171 visited 1 time, 57 visited twice, and 4 visited 3 times. So there was a total of $1,171 + 2 \times 57 + 3 \times 4 = 1,297$ visits. We are certain that there many people who visited 0 times, but how many? The EM algorithm works by “filling in” the missing data (the number who visited 0 times) making the problem a simple one.

Suppose we start with an initial guess of 0.25 visits per person per year. Then the probability of zero visits would be $e^{-0.25} = 0.779$ and the probability of at least one visit would be $1 - 0.779 = 0.221$. That is, there should be $0.779/0.221$ or 3.52 as many people we did *not* see compared to how many we did see visit the ER. So we would expect that there are $3.52 \times 1,232 = 4,337$ people with zero visits. This is the expectation step of the EM algorithm.

Next we use our data to find the maximum-likelihood estimate of the average simply by calculating the arithmetic average using the filled in data. The total number of visits was 1,297 and we expect there were $4,337 + 1,232 = 5,569$ people, for an average of $1,297/5,569 = 0.233$. This is the maximization step. So we can see that our initial guess was too high and the average rate has tended lower.

With our new estimate of the average, we can calculate an updated probability of not visiting: $e^{-0.233} = 0.792$. And now we expect that there are $0.792/0.208$ or about 3.81 times as many zero visit people as those we actually saw in the ER for a expected number of $3.81 \times 1,232 = 4,698$. So we can further update our estimate of the average as $1,297/(4,698 + 1,232) = 0.219$. Repeating this process many times, the estimate converges to 0.104. This can easily be calculated using a spreadsheet program such as Excel.

The maximum-likelihood estimate can also be calculated directly. It corresponds to finding the value of λ that maximizes the quantity¹

$$\log L = 1297 \log \lambda - 1232 \lambda - 1232 \log(1 - e^{-\lambda}). \quad (11.6)$$

Numerically calculating the maximum of (11.6) also gives the value 0.104. It is also possible to find the maximum-likelihood estimate graphically using the Stata commands given in Table 11.13. The resulting plot is shown in Fig. 11.1.

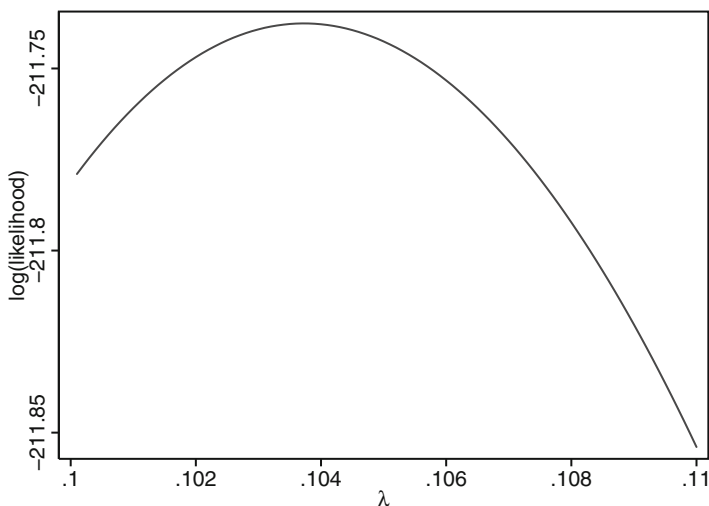
¹Recall that the likelihood is the probability of observing the data. The probability of a specific count for a Poisson model, conditional on being 1 or greater is given by $P(x) = \frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})}$. The product over the entire sample is given by $L = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod x_i! (1 - e^{-\lambda})^n}$, where n is the sample size, and x_i is the count for individual i . It is equivalent and easier to maximize the logarithm of L . We can also ignore the factorial term which does not depend on λ , giving $\log L = \sum x_i \log \lambda - n\lambda - n \log(1 - e^{-\lambda})$.

Table 11.13 Stata commands for plotting the log likelihood

```

clear
set obs 100
gen lambda= n/10000+.1
gen logL=1297*ln(lambda)-1232*lambda-1232*ln(1-exp(-lambda))
twoway line logL lambda, ytitle("log(likelihood)") xtitle("{&lambda}")

```

**Fig. 11.1** Plot of log-likelihood versus the average rate, λ

11.10.2 The EM Algorithm Imputes the Missing Data

The example above illustrates a typical feature of the EM algorithm: using the observed data and the probability model, the EM algorithm fills in the missing data. The analysis of the data then proceeds using the “complete” dataset. The same algorithm can be applied to longitudinal data with missing outcomes. In that case, maximum likelihood is equivalent to filling in the data not observed due to, e.g., drop out or missed visits, using the longitudinal data mixed model. The parameter estimates are then calculated using the complete dataset. As long as the missing data can be reliably predicted from the observed data (which is the case if the longitudinal data model is correct and the MAR assumption holds), the analysis based on the complete dataset is free of bias due to missing data.

Using maximum-likelihood methods with modern computers does not appear to explicitly handle missing data (it just requires the push of a button on a computer). However, when viewed through the lens of the EM algorithm, it is implicitly filling in the missing data based on the assumed probability model being used to fit the data.

11.10.3 ML Versus MI with Missing Outcomes

Maximum likelihood via the EM algorithm may appear to be virtually the same as multiple imputation. Although there are similarities there are also important differences. Perhaps the primary one is that, under MAR, maximum likelihood implicitly selects the right model for filling in the missing data—no model specification is necessary as it is in multiple imputation.

However, because maximum likelihood implicitly assumes a model for “imputation” it cannot be varied. Multiple imputation gives the analyst more options. For example, nonignorable missing data models can be used to check sensitivity to the assumption of MAR. Or violations of the model assumptions can be checked, e.g., what if the assumption of a Poisson distribution was incorrect in the example above? MI also allows the use of techniques other than ML to obtain parameter estimates after the data are imputed.

However, if (a) the model being used for multiple imputation is the same as the one implicitly used by ML, (b) the imputation was performed so many times that the imputation error was negligible, and (c) once imputed, maximum likelihood was used to find parameter estimates, then MI and ML would give the same answers.

11.11 Methods for Data that are Missing Not at Random

We have mentioned previously that standard analysis methods can be biased when the data are MNAR. And, since it is impossible to figure out if the data are MNAR from the observed data, the main strategies to assess the potential impact of MNAR data are sensitivity analyses. Sensitivity analyses proceed by positing a spectrum of MNAR models with checks as to the seriousness of the violation of the MCAR or MAR assumptions required to qualitatively overturn the results of an analysis. If “small” departures from MCAR or MAR lead to different conclusions then the results are taken as tenuous. If “large” departures are required to change the results, then more confidence can be placed in the conclusions. To be convincing, the posited MNAR models and degree of departure from MCAR or MAR need to be defensible in context, which tends to be highly problem specific and so it is difficult to recommend generally applicable strategies. We describe briefly three approaches to MNAR data: pattern mixture models, multiple imputation, and selection models.

11.11.1 Pattern Mixture Models

Consider a study of cognitive decline in which the participants who dropped out had much higher rates of depression at baseline than those with complete data.

We would be concerned that the data was MAR or MNAR and, especially if the rates of decline were quite different, that we might be obtaining biased estimates. What about more detailed comparisons of those with different degrees of missing data?

Our approach to missing data to this point has been what is called a *selection model* approach. We have thought of the observed data as arising from a two-step process. In the first step, the complete data are generated. In the second step (via a process we have described as MCAR, CD-MCAR, MAR, or MNAR), certain of the data are selected for us to observe; the rest is missing.

A very different approach uses what is called a *pattern mixture model*. In this approach, the data are divided into categories according to the pattern of missing data, akin to dividing subjects into those with complete and incomplete data. For example, consider a cohort study where everyone has a baseline observation and there are four planned follow-up visits. Further, suppose the only missing data are because of dropout from the study. Then there are five possible data patterns with regard to presence or absence of data: complete data, missing only visit 5 (i.e., dropout after visit 4), missing visits 4 and 5 (dropout after visit 3), missing visits 3, 4, and 5, and missing visits 2, 3, 4, and 5.

We can now think about dividing up the data according to the missing data pattern and analyzing the data from each pattern separately. The advantage of this approach is that we do not need to think about the missing data mechanism (e.g., MCAR versus MAR). We immediately run in to a problem, however. Returning to the cognitive decline example, what are we to assume about the rate of decline for the participants for whom we only have baseline data? Because we only have a single time point, this group contains no information about the decline over time. If we wish to proceed, we have to make certain assumptions. For example, if we believe the rates of decline are linear over the course of the five year study, we might assume that the rate is the same as that for the group with data for visits 1 and 2 (for which we *can* estimate a linear decline). Or we might assume it is the same as the subgroup with complete data.

If it is reasonable to make simplifying assumptions then the pattern mixture approach is very attractive. Simply by including a categorical predictor for missing data pattern and allowing interactions of key components with that predictor allows the use of standard software packages to accommodate missing data. In the absence of interaction, the analysis gives estimates of the (assumed) common effect. In the presence of interaction, weighted estimates (weighted by the proportion in each missing data pattern) give an estimate of the overall effect.

Unfortunately, it is often the case that there is little guidance in the data as to what models are appropriate and strong assumptions must be made with little opportunity to check them. Further, there are often a multitude of different missing data patterns (it is rarely as simple as described above) which must be grouped subjectively into a manageable, smaller number of categories, each with reasonable sample sizes. These considerations limit the use of pattern mixture models as robust data analysis methods. However, they can still be useful as sensitivity analyses: by varying

the assumptions needed to fit such models, a variety of MNAR missing data mechanisms can be accommodated. See [Little \(1993, 1995\)](#), and [Verbeke and Molenberghs \(2000\)](#) for more in-depth discussion.

11.11.2 Multiple Imputation Under MNAR

Another possible approach to assess sensitivity of results to MNAR missingness is to use multiple imputation but hypothesize an imputation model that allows dependence between the probability that data are missing and the value that would be observed if $R = 1$. [Subak et al. \(2009\)](#) give an example of a trial to encourage weight loss in women with incontinence problems. Their primary analysis imputed end of study values by assuming that women who dropped out of the study, on average, lost no weight, a MNAR mechanism.

11.11.3 Joint Modeling of Outcomes and the Dropout Process

A third strategy is to directly hypothesize a joint model for the complete data and the missing data process and use the observed data to simultaneously estimate the parameters of both models (e.g., [Diggle and Kenward 1994](#)). Not surprisingly, it is difficult to estimate such a model from observed data and they are highly sensitive to the assumed form of the model, something which is not easily checked from the observed data.

11.12 Summary

Missing data are common and many of the simple methods of handling missing data, such as a complete case analysis (the default for most statistical analysis programs), can give misleading results. If it is the predictor variables that are missing in a dataset, we recommend the strategy of multiple imputation. When the main issue is dealing with missing outcomes in a longitudinal study, maximum-likelihood methods are often a good choice. When they are properly specified, they will give valid inference when the data are MAR, whereas generalized estimating equation methods may not. When the analyst needs to exclude important predictors of missingness, in particular mediators, from the outcome model, multiple imputation, and IPW can be useful strategies. Finally, when data are MNAR, pattern mixture models and sensitivity analyses using multiple imputation are recommended.

All techniques for handling missing data require assumptions about how the missing data relate to the observed data. Because the data are missing, these

assumptions cannot be empirically verified. The assumptions are clear in multiple imputation (where we model the missing data), IPW (where we model the probability of missingness), and pattern mixture modeling (where we must make assumptions about covariate effects across missing data patterns). When using maximum-likelihood-based techniques to handle missing-at-random data, the assumptions are inherent and revolve around correct specification of the model, including the variances and correlations in longitudinal data. Because assumptions cannot be verified from the data on hand, it is always a good idea to try a number of techniques of handling missing data to check sensitivity of the conclusions (Hogan et al. 2004).

11.13 Further Notes and References

An attraction of approaching missing data through inverse probability weighting is that it adjusts for missing data through the weighting scheme and does not depend as strongly on the correctness of the outcome model. However, we have noted that it can lead to unstable weights and inefficient analyses. This is an ongoing area of research, with investigations into ways to stabilize the weights, for example, using what is called “robit” regression instead of logistic regression to estimate the probabilities of missingness (Kang and Schafer 2007). Another promising avenue of research is to hedge bets between having to get the outcome or weighting models correct, by using what are known as doubly robust methods (Kang and Schafer 2007). These can correct for missing data when either the model for the inverse weights is correct or the regression model is correct.

The forms of multiple imputation we have illustrated are based on regression models, but there are other alternatives. Scheuren (2005) gives a historical survey of multiple imputation and describes other methods such as “hot deck imputation” (the name comes from a deck of paper “cards” on which data were stored in the early days of the Census Bureau).

Of course, missing data can also occur in situations requiring more complex analyses. For example, there could be missing predictor information in a setting with clustering by facility, physician, and patient. In such a case, just as described in Chap. 7, hierarchical, repeated measures or longitudinal data models must be used to properly impute missing values. Survival analysis is another situation for which imputation of missing predictor information might be required. For survival analysis, the “outcome” consists not only of follow-up time, but also whether censoring has occurred. Both sources of information should be used for imputation, but it is not always clear how to do so. For example, the suggestion to include both the log of the follow-up time and the censoring indicator as predictors in the imputation model can be too simplistic and lead to bias (White and Royston 2009).

11.14 Problems

Problem 11.1. Give an example of a data sampling regime in your research area that is likely to be MAR but not MCAR or CD-MCAR. Briefly explain why.

Problem 11.2. Perform a single imputation for the HERS visit 4 data and verify the results of Table 11.2. Regress glucose on SBP, BMI, ethnicity (white/not white), current smoking status, and diabetes status. Obtain the predicted values for glucose. Create an imputed glucose variable which is equal to the actual glucose value if it is not missing and equal to the predicted value if it is. Using this imputed glucose variable, reproduce the regression of SBP on glucose, white, and BMI given in Table 11.2.

Problem 11.3. How far off are the results when a poor imputation model is used? Singly impute the glucose values (as in Problem 11.2) but using a regression model that contains only current smoking status. How good is this imputation model? Next, compare the estimated effect of glucose on SBP and its statistical significance using this imputation model to the results in Tables 11.1 and 11.2.

Problem 11.4. With the HERS visit 4 data, use the code in Table 11.9 to impute the glucose values. Calculate the SD among the imputed values in glucose to verify that the SD is about 30.9. Hints: the Stata command `egen sd_glu_imp=rowstd(._glucose)` will calculate the standard deviation of the glucose values across the imputed datasets. Summarize those for which the original glucose measurement was missing.

Problem 11.5. What kind of imputation model would you use to impute missing physical activity data in the HERS study? Recall that that variable was a response to a question about how physically active the women considered themselves compared to other women of their age. The five-level response ranged from “much less active” to “much more active,” and was coded in order from 1 to 5. Briefly explain why.

Problems 11.6–11.9 use the data sets `bpmissslong` and `bpmissswide`. The data are based on measurements of SBP in the HERS study. The data set allows us to compare methods of analysis with complete data and under simulated missing data. In the data sets are missing data indicators (`miss_mar` for `bpmissslong` and `miss_mar1` for `bpmissswide`) which have value 1 to flag SBP values which should be dropped to simulate data which displays MAR missingness. In particular, year 1 values from patients with higher baseline SBP are flagged more frequently and hence will be simulated as missing. You can consult the course website for the data sets and more complete documentation and details on Stata code.

Problem 11.6. Using `bpmissswide`,

- (a) Calculate and compare the year 1 SBP (`year1_sbp`) for the complete data and for patients who in the simulated missingness setting would have an available year 1 SBP (i.e., `miss_year` equal to 0).

- (b) Calculate and compare the change in SBP ($\text{year1\_sbp} - \text{base\_sbp}$). What is the mean change in the full sample? What is the mean change restricted among those with available year 1 values in the simulated missingness setting (miss_year equal to 0)?
- (c) Based on (a) and (b) above, how has the simulated missing data mechanism affected estimates of mean of year 1 SBP values and change in SBP from baseline to year 1?

Problem 11.7. Using `bpmisslong`, fit a GEE model with SBP as the outcome and `visit` (`visit`) as the predictor. In Stata, the command would be `xtgee sbp visit, i(pptid) corr(exch)`. Compare a GEE model which uses the full data to one restricted to nonmissing data (miss_year equal to 0). What do you conclude about GEE with MAR missingness?

Problem 11.8. Using `bpmisslong`, fit a mixed linear regression model with SBP as the outcome and `visit` (`visit`) as the predictor. In Stata, the command would be `xtmixed sbp visit || pptid:.` Compare the mixed model which uses the full data to one restricted to nonmissing data (miss_mar equals 0). Compare the results with the GEE results in Problem 11.7. How do you explain the difference in results between the GEE and a linear mixed model with MAR missing data?

Problem 11.9. Using `bpmisswide`,

- (a) Attempt to mimic the effects of multiple imputation by performing imputation to fill in SBP values flagged as missing in the simulated scenario. You may choose the imputation model but it should include baseline SBP, BMI at baseline and year 1 as well as diabetes. In Stata, it will be simplest to perform multivariate normal-based imputations.
- (b) Fit a GEE model (as in Problem 11.7) with multiple imputation. How do the results compare to the results in Problem 11.7? *Note, to fit the GEE model you will need to convert the data from a wide to long format. In Stata, this can be done with the `mi convert` command.*
- (c) Fit a mixed model (as in Problem 11.8) with multiple imputation. How do the results compare to the results in Problem 11.8?

Problem 11.10. The data set `multivisitsbp` extends the HERS SBP data to a series of up to six visits and borrows the set-up used in Problems 11.6–11.9 to simulate missing data through a missing data indicator `miss_mar`.

- (a) To mimic an analysis on complete data, examine a series of models (ignoring the missing data indicators). Fit a GEE model with terms for time (`visit`) and BMI (`bmi`). Then, fit a series of mixed models with fixed effects terms for time and BMI but with varying variance/covariance structures. You might try a random slopes model along with first-, second-, and third-order autoregressive (AR1–AR3). Do you reach similar conclusions about changes in SBP over time (given by the coefficient for `visit`) in these models?

Note: For this data, you can specify the covariance in `xtmixed` with the options `|| ppid: visit, cov(un)` for random slopes and `|| ppid:, residuals(ar 1, t(visit))` for the AR1 model, with AR2 and AR3 defined similarly.

- (b) Repeat the model fits in (a) restricted to available data (`miss_mar` equal to 0) under simulated missingness. Do you reach similar conclusions about changes in SBP over time across these models? How do they compare to the corresponding complete data results in Problem 11.10? Discuss how this might affect choice of variance–covariance structure for mixed models with missing data. Would you prefer a more parsimonious structure (like random intercepts) or a richer one (like third-order autoregressive)? Explain.

11.15 Learning Objectives

- (1) Define the different types of missing data mechanisms (MCAR, CD-MCAR, MAR, MNAR).
- (2) Explain why complete case analysis may lead to biased and/or inefficient analyses.
- (3) Explain the drawbacks of LOCF as an imputation method.
- (4) Identify situations in which ICEs multiple imputation is to be preferred over MVN multiple imputation.
- (5) Use ICEs multiple imputation and MVN multiple imputation to analyze datasets with missing predictor information.
- (6) Explain why maximum-likelihood methods for longitudinal data can be considered methods for handling missing data.
- (7) Explain how multiple imputation can be used as a sensitivity analysis when data are MNAR.
- (8) Use pattern mixture models to analyze datasets with missing outcome data.

Chapter 12

Complex Surveys

Suppose we wanted to estimate the prevalence of diabetes among adults in the US, as well as the effects of diabetes risk factors in this broad target population, both with minimum bias—that is, in such a way that the estimates were truly representative of the target population. Observational cohorts that might be used for these purposes are usually convenience samples, and are often selected from subsets of the population at elevated risk. This would make it difficult to generalize sample diabetes prevalence to the broader target population. We might be more comfortable assuming that sample associations between risk factors and diabetes were valid for the broader population, but the assumption would be hard to check (Problem 12.1).

Observational studies as well as randomized trials use convenience samples for compelling reasons, among them reducing cost and optimizing internal validity. But when unbiased representation of a well-defined target population is of paramount importance, special methods for obtaining and analyzing the sample must be used. Crucial features of such a study are

- All members of the target population must have some chance of being selected for the sample.
- The probability of inclusion can be defined for each element of the sample.

Using data from a sample which meets these two criteria, we could in principle compute unbiased estimates of the number and percent prevalence of diabetes cases in the US adult population, as well as of the effects of measured diabetes risk factors. Surveys implemented by the National Center for Health Statistics (NCHS), including the National Health and Nutrition Examination Survey (NHANES), the National Hospital Discharge Survey (NHDS), and the National Ambulatory Medical Care Survey (NAMCS), are prominent examples of surveys that meet these criteria. Data sets based on these surveys are publicly available on the NCHS website www.cdc.gov/nchs/.

In this chapter, we give only a brief overview of the design and implementation of these surveys, which are complicated and expensive undertakings. Our primary purpose is to provide guidance for secondary analyses using complex survey data.

Fortunately, Stata and other statistical packages make it straightforward and transparent to account properly for the special features of the sampling design in regression analyses using complex survey data.

12.1 Overview of Complex Survey Designs

To provide unbiased estimates of population parameters, complex survey data are *weighted* in inverse proportion to the known probability of inclusion. In addition, to reduce costs, a *complex sampling design* is often used. In many cases, this means initially sampling clusters, known as primary sampling units (PSUs), rather than individuals; only at some later stage are individual study participants selected. This is in contrast to a simple random sample (SRS), in which individuals are directly and independently sampled. Finally, complex samples are often *stratified*, in that the PSUs are sampled within mutually exclusive strata of the target population.

Inverse Probability Weighting

A primary feature of complex surveys is *inverse probability weighting* (IPW). Introduced in Sect. 11.9.3 for dealing with missing data, IPW is the way complex surveys use well-defined probability of inclusion to obtain representative estimates, as we explain below in Sect. 12.2.

One advantage of IPW is that it accommodates *unequal* probability of inclusion in the survey sample. In part, unequal inclusion probabilities arise naturally from variability in the size of primary and secondary clusters. In addition, subgroups of special interest may be sampled at higher rates, so that they comprise a larger proportion of the sample than they do of the target population. The rationale is to ensure adequate precision of estimates both within the subgroup and in contrasting the subgroup to other parts of the larger population, by increasing their numbers in the sample. IPW ensures that overall estimates properly reflect the population proportions comprised by the over-sampled subgroups.

Cluster Sampling

From Chap. 7, it should be clear that the initial sampling of clusters may affect precision, because outcomes for the observations within a cluster are positively correlated in most cases. The change in precision means that for many purposes a larger sample will be required to achieve a given level of statistical certainty. Nonetheless, the complex survey design is cost-effective, because cluster sampling can be implemented in concentrated geographic areas, rather than having to cover the entire area where the target population is found. Moreover, some of the information required to define probability of inclusion need only be obtained for the selected clusters. Especially for nationally representative samples, the savings can be considerable.

In *multistage* designs, there may be several levels of cluster sampling; for example, counties may initially be sampled, and then census tracts within counties, city blocks with census tracts, and households within blocks. Only at the final stage are individual study participants sampled within households. The rationale is again to reduce costs by making the survey easier to implement.

Stratification

An additional feature of many complex surveys is that clusters may be selected from within mutually exclusive and exhaustive *strata*, usually geographic, which cover the entire target population. To the extent that subsets of the target populations are more similar within than across strata, this can increase the precision of estimates of population means and totals.

Example: NHANES

NHANES is a series of complex, multistage probability samples representative of the civilian, noninstitutionalized US population. Interviews and physical exams are used to ascertain a wide range of demographic, risk-factor, laboratory, and disease outcome variables. In NHANES III, conducted between 1988 and 1994, the PSUs were primarily counties. Thirteen large PSUs were selected with certainty, and the remaining 68 were selected with probability proportional to PSU population size, two from each of 34 geographic strata. At the second stage of cluster sampling in NHANES III, area segments, often composed of city or suburban blocks, were selected. In the first half of the survey, special segments were defined for new housing built since the 1980 census, so that no portion of the target population would be systematically excluded; in the second half, more recent information from the 1990 census made this unnecessary. The third stage of sampling was households, which were carefully enumerated within the area segments. At the fourth and final stage, survey participants were selected from within households.

At each stage, sampling rates were controlled so that the probability of inclusion for each participant could be precisely estimated. Children and people over 65 as well as African Americans and Mexican Americans were over-sampled. Almost 34,000 people were interviewed and of these roughly 31,000 participated in the physical exam. Data from NHANES III have been used in many epidemiologic and clinical investigations.

12.2 Inverse Probability Weighting

We pointed out that in selecting a representative sample, every member of the target population has to have some chance of being included in the sample. To put it another way, no part of the target population can be systematically excluded.

In addition, we said that for every element of the sample, the probability of inclusion must be known. Essentially this is what is meant by a so-called *probability sample*. Analysis of such samples makes use of information about probability of inclusion to produce unbiased estimates of the parameters of the target population.

To see how this works, consider a SRS of size 100, drawn at random from a target population of size 100,000. In this simple case, each member of the sample had a one-in-a-thousand chance of being included in the sample. The so-called *sampling fraction*, another term for the probability of inclusion, would be 0.001 for this sample, and constant across observations. Furthermore, we could think of each member of the sample as representing 1,000 members of the target population. If we wanted to estimate the percent prevalence of diabetes in the target population, the proportion with diabetes in the sample would work fine in this case, for reasons that we explain below. Likewise, the average age of the sample would be an unbiased estimate of mean age in the population.

Now consider the more interesting case of estimating the *number* of diabetics in the population. Suppose there were five diabetics in the sample. Since each represents 1,000 members of the target population, an unbiased (though obviously noisy) estimate of the population number of diabetics would be 5,000. Essentially this would be a *weighted sum* of the number of the diabetics in the sample, where each gets weight 1,000, or the number in the population that each sample participant represents. Formally, the weight is the reciprocal of the sampling fraction of 0.001. Note that the overall sum of these sample *inverse probability weights* equals the population size.

Definition: Inverse probability weights are the reciprocal of the probability of inclusion, and are interpretable as the number of elements in the target population which each sampled observation represents.

Next, consider the more typical case where the probability of inclusion varies across participants. To make this concrete, suppose that women and men both number 100,000 in the target population, but that the sample includes 200 women and 100 men, for sampling fractions of 0.002 and 0.001, respectively. In this sample, each man represents 1,000 men in the population, but each woman represents only 500 women. Thus the IPW for each man in the sample would be 1,000, and for each woman, 500.

In this case, to estimate means for the whole target population, we would need to use *weighted* sample averages. These would no longer equal their unweighted counterparts, in which men would be under represented. The formula for the weighted average is

$$E_w[Y] = \frac{\sum_i w_i y_i}{\sum_i w_i}, \quad (12.1)$$

where $E_w[Y]$ denotes the weighted average of the outcome variable Y , y_i is the value of Y for participant i , and w_i is the corresponding probability weight.

Furthermore, if Y were a binary indicator variable coded 1 = diabetic and 0 = nondiabetic, then (12.1) also holds for estimating the population proportion

with diabetes. As we pointed out in Sect. 4.3, this equivalence between averages and proportions only holds with the 0–1 indicator coding of Y . In addition, with this coding of Y , the weighted estimate of the total number in the population with diabetes is simply $\sum w_i y_i$ —the sum of the weights for the diabetics in the sample.

12.2.1 Accounting for Inverse Probability Weights in the Analysis

Taking account of the IPWs, which are included in the NHANES, NHDS, NAMCS, and other NCHS datasets, is essential for obtaining unbiased estimates. The differences between the weighted and unweighted estimates can be considerable. For example, the unweighted proportion with diabetes among adult respondents in NHANES III is 7.4%, but the weighted proportion is 4.8%. The corresponding unweighted estimate of the number of adult diabetics at the time of NHANES III was 12.5 million, as compared to a weighted estimate of 8.1 million—not a trivial difference for estimating the burden of disease and health services needs. All statistical packages for complex surveys accommodate IPWs.

12.2.2 Inverse Probability Weights and Missing Data

Estimation of population parameters, in particular totals, means, and proportions, can be quite vulnerable to missing data. The potential for bias arises because the non-responders usually differ systematically from responders, especially when the response of interest is sensitive. The nonresponders are not *missing completely at random* (MCAR). However, we might be willing to assume that the data are MCAR within relatively homogeneous demographic subgroups defined by measured covariates. In the framework of Chap. 11, the data are assumed to be *CD-MCAR*.

12.2.2.1 Adjustment of IPWs to Account for Unit Non-Response

In NHANES as in many complex surveys, the inverse probability weights are adjusted to account for missing observations—so-called *unit non-response*—in such a way as to minimize the potential for bias. Under the CD-MCAR assumption, the inverse probability weights are adjusted within relatively homogeneous demographic subgroups. Specifically, for each such subgroup, the weights for the responders are inflated by a fixed factor, determined so that the adjusted weights for the responders sum to the total of the original inverse probability weights for both responders and non-responders. In short, the responders in the subgroup are made to stand in for the non-responders.

In many complex surveys, a second so-called *poststratification* adjustment is made to ensure that the IPWs sum to regional totals for the target population, which are known from the US Census.

12.2.2.2 Multiple Imputation to Account for Item Non-Response

In addition to unit non-response, we also need to be concerned about *item* non-response, or missing responses on particular questions by study participants. The recommended approach to item non-response in complex surveys is *multiple imputation* ([Rubin 1987, 1996](#)); see Sect. 11.5.

12.3 Clustering and Stratification

In contrast to accounting for the inverse probability weights, which is required mainly to avoid bias, taking account of the stratification and clustering of observations due to the complex sampling design is required solely to get the standard errors, CIs, and *P*-values right, and has no effect on the point estimates. Unlike the point estimates, standard errors accounting for the special characteristics of a complex survey do differ from what would be obtained in standard weighted regression routines, sometimes in ways that are crucial to the conclusions of the analysis.

The default standard errors, CIs, and *P*-values provided by most survey packages including Stata are calculated using so-called *linearization*. These are closely related to the robust standard errors available with many Stata regression commands, and thus account, as with longitudinal and hierarchical data, for clustering. In Stata, the main difference is that in testing whether each estimated regression coefficient differs from zero, the survey routines use a *t*-test with degrees of freedom equal to the number of PSUs minus the number of strata, rather than the asymptotic *Z*-test used in GEE. In addition, stratification is taken into account.

Of note, these methods for analyzing survey data do not directly extend to random effects models, introduced in Chap. 7, which represent a different approach to clustered data. [Rabe-Hesketh and Skrondal \(2006\)](#) propose a pseudo-likelihood approach to analyzing multi-level data with a binary outcome, which is implemented in the downloadable `gllamm` package for Stata.

12.3.1 Design Effects

Because of positive correlation within clusters, the standard errors of parameter estimates from a complex survey are often (but not always) inflated as compared

to estimates from a SRS of the same size. This inflation can be summarized by a *design effect*:

Definition: The *design effect* is the ratio of the true variance of a parameter estimate from a complex survey to the variance of the estimate if it were based on data from a simple random sample.

Note that design effects can vary for different parameters estimated in the same survey, because some predictors may be more highly concentrated and some outcomes more highly correlated within clusters than others. Furthermore, design effects in regression may vary with the degree to which the regression effect is estimated by contrasting observations within as opposed to between clusters.

12.4 Example: Diabetes in NHANES

Stata makes it easy to run a regression analysis taking account of the special features of a complex survey. Variables identifying the PSU, IPW, and stratum for each observation are first specified using the `svyset` command. For our NHANES example, the `svyset` command takes the form

```
svyset sdppsuh [pweight = wtpfqx6], strata(sdpstra6)
```

The regression is then run using the usual Stata commands, in conjunction with the `svy:` command prefix.

Table 12.1 shows three logistic models for prevalent diabetes estimated using data from NHANES III. The predictors are age (per 10 years), ethnicity, and sex. The reference group for ethnicity is whites. The odds-ratio estimates given by unweighted logistic regression (Model 1) differ both quantitatively and qualitatively from the results of the weighted and survey analyses (Models 2 and 3), which are identical. In the unweighted model, women appear to be at about 20% higher risk, but this does not hold up after accounting for probability of inclusion; similarly, the apparently increased risk among African Americans and Mexican Americans is smaller after accounting for the weights.

In addition, the standard errors differ across all three models, in part because the survey model takes proper account of stratification as well as clustering within PSUs. Note that in accommodating IPWs in Model 2, Stata by default uses robust standard errors, which are similar to the Linearized Std. Err. estimates given for Model 3.

We can obtain the design effect for each parameter estimate using the Stata `postestimation` command `estat effects, deff`. In the survey logistic model for prevalent diabetes shown in Table 12.1, the design effects are 2.7 for age, 0.93 for African American, 0.41 for Mexican American, 2.0 for other ethnicity, and 1.7 for sex. The increase in precision for the coefficient for Mexican Americans results from the strong concentration of this subgroup in a few PSUs, so that the comparison with whites rests primarily on within-cluster contrasts. In contrast,

Table 12.1 Unweighted, weighted, and survey logistic models for diabetes

. * Model 1: Unweighted logistic model ignoring weights and clustering						
. logit diabetes age10 aframer mexamer othereth female, or nolog						
Logistic regression			Number of obs		= 18140	
			LR chi2(5)		= 1148.81	
			Prob > chi2		= 0.0000	
Log-likelihood = -4206.1375			Pseudo R2		= 0.1202	
Diabetes	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age10	1.679618	.0284107	30.66	0.000	1.624847	1.736235
aframer	2.160196	.1651839	10.07	0.000	1.859534	2.50947
mexamer	2.784521	.2125535	13.42	0.000	2.39759	3.233896
othereth	1.25516	.2297557	1.24	0.214	.8767735	1.796845
female	1.200066	.0713788	3.07	0.002	1.068013	1.348447
. * Model 2: Weighted logistic model still ignoring clustering						
. logit diabetes age10 aframer mexamer othereth female [pweight = wtpfqx6], ///						
or nolog						
Logistic regression			Number of obs		= 18140	
			Wald chi2(5)		= 523.98	
			Prob > chi2		= 0.0000	
Log-pseudolikelihood = -28717819			Pseudo R2		= 0.1124	
Diabetes	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
age10	1.704453	.0420649	21.61	0.000	1.62397	1.788925
aframer	1.823747	.1727191	6.34	0.000	1.514785	2.195726
mexamer	1.915197	.2029156	6.13	0.000	1.556068	2.357211
othereth	1.031416	.2386775	0.13	0.894	.6553287	1.623335
female	.9805769	.0992109	-0.19	0.846	.8041933	1.195647
. * Model 3: Survey model accounting for weights, stratification, and clustering						
. svy: logit diabetes age10 aframer mexamer othereth female, or nolog						
(running logit on estimation sample)						
Survey: Logistic regression						
Number of strata		= 49	Number of obs		= 18140	
Number of PSUs		= 98	Population size		= 168471391	
			Design df		= 49	
			F(5, 45)		= 80.86	
			Prob > F		= 0.0000	
Diabetes	Odds Ratio	Linearized Std. Err.	t	P> t	[95% Conf. Interval]	
age10	1.704453	.0479719	18.95	0.000	1.610726	1.803635
aframer	1.823747	.1840181	5.96	0.000	1.48903	2.233705
mexamer	1.915197	.1934747	6.43	0.000	1.56332	2.346276
othereth	1.031416	.225949	0.14	0.888	.6641157	1.601857
female	.9805769	.0921775	-0.21	0.836	.811784	1.184467
. estat effects, deff						
Diabetes	Coef.	Linearized Std. Err.	DEFF			
age10	.5332443	.028145	2.72072			
aframer	.6008932	.1009011	.933096			
mexamer	.6498207	.1010208	.415208			
othereth	.0309323	.2190668	1.98449			
female	-.0196142	.0940033	1.67026			
_cons	-5.798575	.2023545	3.05472			

women are about half of respondents in all PSUs, so that more of the information for the comparison with men comes from less efficient between-PSU contrasts (Problems 12.3 and 12.4).

In summary, accounting for IPW mainly affects the point estimates and secondarily the standard errors, while accounting for stratification and clustering only affects the latter.

12.5 Some Details

12.5.1 Ignoring Secondary Levels of Clustering

We pointed out earlier that NHANES is a *multistage* complex survey, meaning that area segments are selected within PSUs, then blocks with segments and households within blocks, before individuals are finally selected. Thus clusters are nested within clusters. For the NCHS surveys, multistage design is typical.

SUDAAN and recent versions of Stata make it possible to account more completely for the effects of multistage cluster sampling, by specifying identifiers for *secondary sampling units* (SSUs). They also accommodate so-called *finite population correction factors* to account for the fact that both PSUs and SSUs are sampled without replacement from relatively small “populations” of PSUs and SSUs.

However, only the stratum and PSU identifiers are provided with the NHANES data; to protect the confidentiality of survey respondents, no information is provided about area segment or block—the SSUs. Fortunately, in large samples like NHANES, the robust *sandwich* standard error calculations used in `svy` regression commands will properly reflect differences in the degree of correlation between observations sampled from the same or different SSUs within each PSU.

12.5.2 Other Methods of Variance Estimation

NHANES 2000, next in the series after NHANES III, began collecting data in 1999 and continues to sample yearly, using a similar complex multistage design. A nationally representative sample of approximately 5,000 participants is obtained each year. Data for the first two years were available in mid-2003. Although stratum and (psuedo) PSU identifiers have since been made available, they were not provided in 2003, to protect the confidentiality of study participants. Other surveys that do not provide stratum and PSU identifiers include the NHDS, and until recently, the National Ambulatory Medical Care Survey (NAMCS).

12.5.2.1 Relative Standard Errors

For the NHDS, constants for computing *relative standard errors* are provided with the documentation, so that approximate CIs for means and proportions can be calculated, but regression analysis is not possible.

12.5.2.2 Jackknife and Balanced Repeated Replication

Two other methods of variance estimation are implemented in Stata as well as the SUDAAN and WESVAR packages, and are compatible with regression analyses. The *jackknife* method uses a resampling procedure to estimate variability. The complete sample is split into K groups in such a way as to reflect the complex sampling structure but obscure geographic location, and a set of jackknife weights corresponding to each group is provided. In the k th set, the weights for group k are set to zero, and adjusted for the remaining groups, using adjustment methods already described for dealing with nonresponse. The analysis is then carried out $K + 1$ times, once with the original weights and once with each of the K sets of jackknife weights. It should be clear that the group with jackknife weights equal to zero will be omitted from that analysis. Then the variance of the overall estimates is estimated by variability among the jackknife estimates, appropriately scaled (Rust 1985; Rust and Rao 1996).

A related method for variance estimation called *balanced repeated replication* (BRR) is also implemented in Stata as well as SUDAAN and WESVAR, but is beyond the scope of this chapter.

12.5.3 Model Checking

In addition to accounting for clustering, stratification, and inverse probability weighting, we need to do standard model checking in regression analyses using complex survey data. These should include checks for linearity of the effects of continuous predictors, possibly using restricted cubic splines, and for omitted interactions. One useful tool is the Stata `postestimation` command `estat gof`, which extends the Hosmer–Lemeshow goodness of fit test to logistic and probit models for binary responses in survey data.

12.5.4 Postestimation Capabilities in Stata

Other useful `postestimation` commands, including `margins`, for obtaining average causal effects (Sect. 9.3.4), are also available. We also note that the factor

notation used to include categorical variables, quadratic terms, and interactions (Sects. 4.3 and 4.6) carries over without change to the Stata survey regression commands.

12.5.5 Other Statistical Packages for Complex Surveys

In addition to Stata, three other software packages make it straightforward to carry out descriptive as well as regression analyses using complex survey data. These packages include

- SUDAAN (Research Triangle Institute, Research Triangle Park, NC; www.rti.org),
- SAS (SAS Institute, Cary, NC; www.sas.com),
- WESVAR (Westat, Inc., Rockville MD; www.westat.com).

12.6 Summary

Complex surveys, unlike many convenience samples, can provide representative estimates of the parameters of a target population. However, to obtain these estimates and compute valid standard errors, CIs, and P -values, such surveys have to be analyzed using methods that take account of the special features of the design, including multistage cluster sampling, stratification, and the fact that not all members of the population have an equal chance of being included in the sample. A number of software packages make it straightforward to carry out multi-predictor regression analyses using complex survey data.

12.7 Further Notes and References

For in-depth treatments of the many topics not covered in our brief overview focusing on regression analyses, leading books about complex surveys include [Levy and Lemeshow \(1999\)](#), [Korn and Graubard \(1999\)](#), [Scheaffer \(1996\)](#), [Kish \(1995\)](#), and [Cochran \(1977\)](#). These books deal comprehensively with the design of complex surveys and the underlying statistical theory. They also cover more specific topics including ratio estimators, variance estimation for subpopulations, and analysis of longitudinal surveys and using multiple surveys.

12.8 Problems

Problem 12.1. Taking HIV infection as an example, explain why it might be more problematic to generalize estimates of prevalence from a convenience sample than to generalize estimates of risk factor effects. For the latter, we essentially have to assume that there is little or no interaction between the risk factor and being represented in the sample. Does this make sense?

Problem 12.2. Show that (12.1) reduces to the unweighted average $\sum y_i/n$ when $w_i \equiv w$.

Problem 12.3. Judging from the logistic model shown in Table 12.1, which was used to assess risk factors for diabetes, design effects greater than 1.0 appear to be more common than design effects less than 1.0. Describe what would happen in these two cases to model standard errors, CIs, and P -values, if we were to analyze the survey data incorrectly, ignoring the clustering. In which case would we be more likely to make a type-I error? In which case would we be likely to dismiss an important risk factor? Can we reliably predict whether the design effect will be greater or less than 1.0?

Problem 12.4. In contrast to the design effects in regression analyses, design effects for means, proportions, and totals are almost always greater than 1.0. Explain why this should be the case.

12.9 Learning Objectives

- (1) Describe the rationale for and special features of a complex survey.
- (2) Identify what can go wrong if the analysis of a complex survey ignores inverse probability weights, strata, and cluster sampling.
- (3) Know how to use data from NHANES III or a similar complex survey to validly estimate the parameters of multi-predictor linear and logistic regression models, with standard errors, CIs, and P -values that properly reflect the complex survey design.

Chapter 13

Summary

13.1 Introduction

Our goal in writing this book was to provide researchers and students with a practical guide to the analysis of data from research studies focusing on the relationship between outcomes and multiple predictor variables. Through our experience as coinvestigators and instructors at the University of California, San Francisco, we have observed that students and researchers from many fields can benefit greatly from being able to conduct their own data analyses. Mastering these skills promotes better study designs, clearer and more informative papers and presentations, and more focused and productive interactions with professional statisticians concerning more advanced topics.

Despite the fundamentally mathematical foundations of statistics, the prerequisites needed to acquire adequate data analysis skills are surprisingly nontechnical. Perhaps the most important one is critical thinking. As is true with many technical fields, the key ideas underlying the methods presented here become much clearer when applied in actual data analyses. All of them are characterized by a common structure that mirrors the majority of research questions arising in clinical research: the relationship between an outcome and measured explanatory variables.

In this chapter, we provide a brief review of the general approach to data analysis developed in this book, and provide guidance on how to use it as a resource to address particular analytical issues. We also briefly discuss a number of topics relevant to investigators undertaking their own data analyses, including development of analysis plans and finding help with technical questions. Finally, we discuss briefly some advanced topics that are not covered extensively in this book, and represent areas of current research that are relevant to many modern applications of regression methods.

13.2 Selecting Appropriate Statistical Methods

Selection of the right statistical tool to apply in addressing a research question is not always easy. Despite a number of unsuccessful attempts to use concepts from artificial intelligence in the development of algorithms to automate this process, common sense and experience remain most important for choosing an appropriate analysis method. In this section, we provide some general guidelines on selecting statistical methods, with references to appropriate chapters and sections in the book. In keeping with our overall theme, we assume that the research question and available data involve investigating the relationship between a specified outcome and one or multiple measured predictor variables.

The first step in most data analyses is to define clearly the candidate outcome and predictor variable(s) and choose an appropriate analytic approach. As described in Sect. 1.1, outcomes can generally be classified as being either numeric (e.g., measured characteristics such as cholesterol level or body weight) or categorical (e.g., disease status indicators). Table 13.1 uses this classification to distinguish the main types of outcomes considered in the book (that subsume the majority considered in health research applications), along with the standard regression approaches for each, and the chapters in which they are discussed. Clearly many outcomes do not fit cleanly into the categories provided in the table. For example, the severity score in the back pain example introduced in Chap. 1 could be considered as either continuous or as a categorical variable with ordinal categories. In many such cases, the decision of how to consider such variables for the purpose of analysis will be driven by practicality (e.g., available software) and/or convention. In cases where multiple approaches are available, it is often a good idea to try more than one to insure that results are not sensitive to the choice.

Although the type of outcome usually dictates the choice of which regression model to consider, further consideration of how the outcome is observed and measured is necessary before settling on an analysis approach. A fundamental consideration is whether individual outcomes can be viewed as independent or not. Examples of studies with independent outcomes include diagnosis of CHD in participants in the WCGS study (used for examples in Chaps. 2–5) and baseline glucose levels in women participating in the HERS study (Sect. 4.2). Dependence between outcomes can arise in a number of ways detailed in Chap. 7. These include repeated measures of outcomes measured in the same individuals, or outcomes

Table 13.1 Outcome, regression model, and chapter reference

Outcome classification	Outcome type	Regression model	Chapter reference
Numerical	Continuous	Linear	4
	Count	Poisson model	8
	Time-to-event	Proportional hazards	6
Categorical	Binary	Logistic	5
	Ordinal	Proportional odds	5
	Nominal	Polytomous logistic	5

on different individuals that are associated via a shared environment or genetic relationship (e.g., disease outcomes among members of the same family). Examples include repeated measures of fat content of feces (Sect. 7.1) and birthweights of first- and last-born infants from the same mothers (Sect. 7.3). As described in Chap. 7, most of the regression approaches for independent outcomes have direct analogs applicable in the dependent outcome setting.

In addition to dependence between individual outcomes, it is also important to consider how individuals were selected for inclusion in the study being analyzed. Although for many studies, it is reasonable to assume that study participants in a defined population had equal chances of being selected, in some cases these chances are controlled by the investigator to obtain a sample with desired properties. Examples include case-control studies for binary outcomes and complex sample surveys. As illustrated in Sect. 5.3 and Chap. 12, regression methods for such studies generally involve minor modifications of techniques applicable for independent samples.

Finally, we want to stress that despite the large number of outcome types and corresponding approaches to regression modeling covered here, the tools used for model fitting and evaluation are quite similar in most cases. Key concepts and techniques in model construction and interpretation such as accounting for confounding, mediation, and interaction and non-linearity are shared across approaches as well. Experience with regression modeling for different types of outcomes and study designs will surely reinforce these points.

13.3 Planning and Executing a Data Analysis

Data analyses are usually complex and benefit from careful planning in order to proceed in a timely and organized fashion. In our experience, few analyses are limited to straightforward application of textbook procedures. Invariably, technical questions arise related to data structure and/or quality, application of particular techniques, use of software programs, and interpretation of results. In this section, we provide some advice on several topics related to conducting an efficient analysis.

13.3.1 *Analysis Plans*

Before beginning a data analysis, it is useful to formulate a plan for how the work will proceed. For randomized controlled trials, analysis plans are generally specified in advance by the study protocol. For observational and clinical studies, preliminary plans are often formulated at the proposal stage. However, even when existing plans are not available to guide analyses, a clear outline of the important issues and tasks

can aid in organizing the process. A detailed plan should include a summary of the study design, statements of the research hypotheses, descriptions of each stage of analysis, and clear procedures for record-keeping, data distribution, and security.

13.3.2 Choice of Software

Fortunately, there are a number of excellent software packages available that implement the majority of techniques discussed here. Although we have used Stata in our examples, SAS, S-PLUS, and SPSS all provide commercial alternatives that offer many of the same facilities and run on a variety of computer platforms and operating systems. Also, the R language for statistical computing and graphics ([R Development Core Team 2004](#)) is freely available and includes most of the procedures presented here. Finally, there are a number of special-purpose programs providing methods not well-represented in the major packages, including StatXact and LogXact (exact inference for contingency tables and logistic regression), and SUDAAN (analysis of data from complex surveys). Frequently, multiple programs will be used for a given analysis. For example, SAS may be used in preparation of analysis data sets, and specific analyses conducted in Stata or R. Fortunately, there are programs such as StatTransfer that translate data sets between common formats used by different analysis packages, preserving important features such as variable labels and formats.

13.3.3 Data Preparation

Perhaps the single most time consuming phase of any data analysis is preparation of analysis-ready data sets from source data. Source data frequently reside in relational databases or proprietary formats and must be exported and re-formatted for specific analyses. Since particular analytic procedures rely on specific data structures and variable definitions, sufficient time and resources should be allocated for proper preparation and checking of analysis data sets prior to conducting statistical analyses.

13.3.4 Record Keeping and Reproducibility of Results

An important part of a complete data analysis includes keeping files of relevant commands and procedures used in each of the stages above. Adding comments and explanatory text to programs and keeping text files outlining the analysis procedures

and cataloging the important files are very useful in this regard. This information should be kept in an identifiable place (preferably organized with other project-specific materials) and backed up in a secure location for disaster recovery.

Because a typical data analysis involves a large number of steps, having all files necessary to recreate results from source data can save work for revision of research publications, and is critical in demonstrating that the results are reproducible. The merits of making this material, including source data, public are a topic of current debate in the scientific literature. See [Sedransk et al. \(2010\)](#) for a discussion of relevant issues from the statistician's perspective.

13.3.5 Data Security

Records from research studies often contain sensitive patient information and must be protected from unauthorized access. Although studies generally have data security measures in place to protect primary data sources, data analyses often involve creation of multiple datasets that may be distributed between investigators. As a general rule, it is a good practice to keep analysis datasets physically separate from source data, with any variables that can be linked to participant identities removed. Make sure that all analysis and data distribution procedures conform to current government, institutional, and study-specific guidelines on data security and protected health information.

13.3.6 Consulting a Statistician

As we have noted frequently in the text, there are many instances where analysis issues arise that do not fall in the neat categories typical of many of the examples. Complex sampling schemes, extensive missing data, unusual patterns of censoring, misclassification in measured outcomes and predictors, causal inferences in longitudinal observational studies subject to time-dependent confounding—all are examples of situations where standard methods and attendant assumptions may not apply without modification. Being able to recognize these circumstances is an important step in addressing these issues. When faced with an analysis problem that appears to fall outside of the range of techniques covered here, having access to a professional statistician is a valuable resource. For investigators at research institutions, the best way to insure the availability of sound statistical support is to include a statistician as a consultant or coinvestigator in proposals. Participating in courses or workshops on specialized statistical methods is another way to gain access to expert advice on advanced topics.

13.3.7 Use of Internet Resources

The Internet provides a vast and very valuable resource to assist in selection of statistical methods and planning data analyses. Frequently, answers to questions about particular applications and methods can quickly be found via a search using one of the available Web search engines. Unfortunately, even judicious searches often yield too many results to review completely. Also, the relevance of returned results is frequently influenced by factors completely unrelated to their scientific value. For these reasons, beginning with searches of established research resources such as the PubMed interface to the MEDLINE index and the Current Index to Statistics will often yield more focused searches. Many educational institutions and private companies provide free online access to electronic scientific journals. Also, statistical software sites frequently have online documentation and message lists that can provide useful information on the use of particular methods. Finally, message boards related to particular software programs and academic interests can frequently be a good way to get answers to analysis questions. Of course, unless the qualifications of individuals posting are known, blindly following advice can be dangerous.

13.4 Further Notes and References

13.4.1 Multiple Hypothesis Tests

The majority of the examples and applications considered in this book can be characterized by single outcome variables and their relationship to one or multiple predictors. While these are representative of many of the research questions that arise in epidemiological and medical research, we have largely ignored issues that arise when analyses include testing multiple hypotheses. These can arise in many contexts, including genomic studies that seek to identify important predictors of a primary disease outcome from a potentially very large pool of candidates, and in clinical studies investigating the effect of a treatment on multiple disease outcomes. The primary concern in these examples is the inflation of type-I error resulting from the occurrence of false-positive results arising from multiple hypothesis tests. Valid inferences in these situations generally involve adjustment of P -values from individual tests to control family-wise error rate (FER) to desired levels.

Consider a study of the use of gene expression data in the classification of two types of acute leukemia (myeloid and lymphoblastic) (Golub et al. 1999). RNA from bone marrow samples from 38 patients (27 lymphoblastic and 11 myeloid) was hybridized to oligonucleotide microarrays, each containing probes for 6,817 genes. The research questions centered on the use of genes as predictors for leukemia type. Although some form of binary regression model relating the disease outcome to predictors is clearly appropriate in this example, the fact that the number of

candidate predictors greatly outnumber the observations, and that the correlation between predictors may be quite complex (reflecting functional relationships between genes) raises a number of difficult computational and inferential issues. Clearly, an analysis that screened for candidate genes via independent hypothesis tests of each would potentially yield many false-positive results if the type-I error was fixed at the conventional 5% level.

Conventional procedures for controlling FER such as the Bonferroni correction outlined in Sects. 3.1.5 and 4.3.4 may be quite stringent in this example, resulting in significance levels that may rule out even associations of interest as potential false-positive results. These concerns have led to development of multiple testing procedures designed to control the *false discovery rate* (FDR), defined as the number of false-positives relative to the total number of positives, rather than focus solely on the former. In the example, the choice of an FDR of 5% implies that on average, 5% of genes selected as positively associated with the leukemia outcome would represent false-positive results. This approach generally results in improved power relative to procedures designed to minimize FER, at the expense of an increased likelihood of type-I errors. We refer readers to the seminal papers by [Benjamini and Hochberg \(1995\)](#) and [Storey \(2002\)](#) for further information about FDR procedures.

Multiple testing problems also arise in studies involving the effect of a predictor of interest on multiple outcomes. For example, randomized trials may consider more than one primary outcome in addition to a number of secondary outcomes. This is common in fields such as psychiatry, where treatments may influence a number of behavioral characteristics, many of which are related. Similar issues arise in *subgroup analyses*, which repeat the primary outcome in groups of individuals defined by enrollment characteristics in an effort to identify factors that may influence treatment efficacy. They are also a concern in safety analyses, in which rates of occurrence of adverse events are compared between arms. In all these situations, conventional hypothesis testing with no adjustment for multiple testing can lead to potentially misleading conclusions about results. Results from the Bonferroni adjustment in these situations is expected to be fairly conservative, both because it ignores correlations between outcomes and also gives equal weight to each. Alternative procedures tailored to prespecified ordering of hypotheses about primary and secondary endpoints are sometimes appropriate. These issues are discussed further in [Dmitrienko et al. \(2009\)](#) and [Piantadosi \(2005\)](#).

As discussed in Sect. 10.3, multiple comparison issues are also a concern in regression analyses targeting the relationship between an outcome and multiple predictors, where the primary goal is to identify important predictors and characterize their relationship to the outcome rather than construct a model that provides accurate outcome prediction. In this case, use of formal adjustment methods is debatable.

13.4.2 Statistical Learning

In Sect. 10.1, we considered the application of regression methods in developing clinical prediction models. These problems are typically characterized by using a

potentially large collection of predictor variables to develop a regression model for predicting individual patient outcomes with the aim of minimizing prediction error. Regression methods represent just one approach in a large class of *statistical learning* methods for such addressing such problems. Many of these methods are computationally intensive, and depart radically from the familiar additive linear structure familiar from the models presented here. We refer readers to [Hastie et al. \(2009\)](#) for a book-length overview of some modern approaches being applied in this area.

References

- Aalen, O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, **8**(8), 907–925.
- Ades, A., Parker, S., Walker, J., Edginton, M., Taylor, G. P. and Weber, J. N. (2000). Human t cell leukaemia/lymphoma virus infection in pregnant women in the United Kingdom: population study. *British Medical Journal*, **320**, 1497–1501.
- Allen, D. M. and Cady, F. B. (1982). *Analyzing Experimental Data by Regression*. Wadsworth, Belmont, CA.
- Altman, D. G. and Andersen, P. K. (1989). Bootstrap investigation of the stability of the Cox regression model. *Statistics in Medicine*, **8**, 771–783.
- Altman, D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, **19**, 453–473.
- Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology*, **26**, 1323–1333.
- Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91**(434), 444–455.
- Angrist, J. D. and Krueger, A. (1992). Estimating the payoff to schooling using the Vietnam era draft lottery. *National Bureau of Economic Research, Working Paper 4067*.
- Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press, Princeton.
- Antman, E. M., Cohen, M., Bernink, P. J. L. M., McCabe, C. H., Horaceck, T., Papuchis, G., Mautner, B., Corbalan, R., Radley, D. and Braunwald, E. (2000). The TIMI Risk Score for unstable angina/non-ST elevation MI. *Journal of the American Medical Association*, **284**(7), 835–842.
- Aurora, P., Whitehead, B. and Wade, A. (1999). Lung transplantation and life extension in children with cystic fibrosis. *Lancet*, **354**, 1591–1593.
- Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*, **26**, 3078–3094.
- Austin, P. C. (2009). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, **29**, 661–677.
- Austin, P. C., Grootendorst, P. and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*, **26**, 734–753.
- Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**(6), 1173–1182.

- Baum, C., Schaffer, M. and Stillman, S. (2003). Instrumental variables and gmm: estimation and testing. *The Stata Journal*, **3**(1), 1–31.
- Beach, M. L. and Meier, P. (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials*, **10**, 161S–175S.
- Begg, M. D. and Lagakos, S. (1993). Loss in efficiency caused by omitted covariates and misspecifying exposure in logistic regression models. *Journal of the American Statistical Association*, **88**(421), 166–170.
- Bellamy, S. L., Lin, J. Y. and Ten Have, T. R. (2007). An introduction to causal modeling in clinical trials. *Clinical Trials*, **4**, 58–73.
- Belsey, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*. John Wiley & Sons, New York, Chichester.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, **57**(1), 289–300.
- Bernardo, M. V. P., Lipsitz, S. R., Harrington, D. P. and Catalano, P. J. (2000). Sample size calculations for failure time random variables in non-randomized studies. *Journal of the Royal Statistical Society (Series D): The Statistician*, **49**, 31–40.
- Black, D., Cummings, S., Karpf, D., Cauley, J., Thompson, D., Nevitt, M., Bauer, D., Genant, H., Haskell, W., Marcus, R., Ott, S., Torner, J., Quandt, S., Reiss, T. and Group, K. E. K. F. I. T. R. (1996a). Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. *Lancet*, **348**, 1535–1541.
- Black, D., Cummings, S., Karpf, D., Cauley, J., Thompson, D., Nevitt, M., Bauer, D., Genant, H., Haskell, W., Marcus, R. et al. (1996b). Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. *The Lancet*, **348**(9041), 1535–1541.
- Bradu, D. and Mundlak, Y. (1970). Estimation in lognormal linear models. *Journal of the American Statistical Association*, **65**, 198–211.
- Brant, L. J., Sheng, S. L., Morrell, C. H., Verbeke, G. N., Lesaffre, E. and Carter, H. B. (2003). Screening for prostate cancer by using random-effects models. *Journal of the Royal Statistical Society: Series A*, **166**, 51–62.
- Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, **16**(3), 199–231.
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth Publishing Co., Inc, Belmont, CA.
- Breslow, N. E. and Day, N. E. (1984). *Statistical Methods in Cancer Research Volume I: The Analysis of Case-Control Studies*. Oxford University Press, Lyon.
- Brookes, S. T., Whitley, E., Peters, T. J., Mulheran, P. A., Egger, M. and Smith, G. D. (2001). *Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives*. The National Coordinating Centre for Health Technology Assessment, University of Southampton, Southampton, UK.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Stürmer, T. (2006a). Variable selection for propensity score models. *American Journal of Epidemiology*, **163**(12), 1149–1156.
- Brookhart, M. A., Wang, P. S., Solomon, D. H. and Schneeweiss, S. (2006b). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*, **17**(3), 268–275.
- Brown, J., Vittinghoff, E., Wyman, J. F., Stone, K. L., Nevitt, M. C., Ensrud, K. E. and Grady, D. (2000). Urinary incontinence: does it increase risk for falls and fractures? Study of Osteoporotic Fractures Research Group. *Journal of the American Geriatric Society*, **B48**, 721–725.
- Buchbinder, S. P., Douglas, J. M., McKirnan, D. J., Judson, F. N., Katz, M. H. and MacQueen, K. M. (1996). Feasibility of human immunodeficiency virus vaccine trials in homosexual men in the United States: risk behavior, seroincidence, and willingness to participate. *Journal of Infectious Diseases*, **174**(5), 954–961.
- Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, **53**, 603–618.

- Carey, V., Zeger, S. L. and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517–526.
- Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall/CRC, London, New York.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, **158**, 419–466.
- Chib, S. and Hamilton, B. H. (2002). Semiparametric bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, **110**, 67–89.
- Clark, L., Jr., G. C., Turnbull, B., Slate, E., Chalker, D., Chow, J., Davis, L., Glover, R., Grahm, G., Gross, E., Krongrad, A., Leshner, J., Park, H., Jr., B. S., Smith, C. and Taylor, J. (1996). Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin: a randomized controlled trial. *Journal of the American Medical Association*, **276**(24), 1957–1963.
- Clayton, D. and Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.
- Cleveland, W. S. (1985). *The Elements of Graphing Data*. Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York, Chichester, 3rd ed.
- Cole, S. R. and Hernán, M. A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology*, **31**, 163–165.
- Cole, S. R., Hernán, M. A., Robins, J. M., Anastos, K., Chmiel, J., Detels, R., Ervin, C., Feldman, J., Greenblatt, R., Kingsley, L., Lai, S., Young, M., Cohen, M. and noz, A. M. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology*, **158**, 687–694.
- Collett, D. (2003). *Modelling Binary Data*. Chapman & Hall/CRC, London, New York.
- Concato, J., Peduzzi, P. and Holford, T. R. (1995). Importance of events per independent variable in proportional hazards analysis i. background, goals, and general strategy. *Journal of Clinical Epidemiology*, **48**, 1495–1501.
- Cook, N. R., Buring, J. E. and Ridker, P. M. (2006). The effective of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine*, **145**, 21–29.
- Crager, M. R. (1987). Analysis of covariance in parallel-group clinical trials with pretreatment baselines. *Biometrics*, **43**(4), 895–901.
- D'Agostino, R., Lee, M., Belanger, A., Cupples, L., Anderson, K. and Kannel, W. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine*, **9**(12), 1501–1515.
- D'Agostino, R. B., Russell, M. W., Huse, D. M., Ellison, C., Silberhatz, H., Wilson, P. W. F. and Hartz, S. C. (2000). Primary and subsequent coronary risk appraisal: new results from the Framingham Study. *American Heart Journal*, **139**, 272–281.
- de Bruyn, G., Shiboski, S., van der Straten, A., Blanchard, K., Chipato, T., Ramjee, G., Montgomery, E. and Padian, N. (2011). The effect of the vaginal diaphragm and lubricant gel on acquisition of hsv-2. *Sexually transmitted infections*, **87**(4), 301–5.
- DeGruttola, V. and Tu, X. M. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003–1014.
- Devore, J. and Peck, R. (1986). *Statistics, the Exploration and Analysis of Data*. West Publishing Co., St. Paul, MN.
- Dickson, E. R., Grambsch, P. M. and Fleming, T. R. (1989). Prognosis in primary biliary-cirrhosis - model for decision-making. *Hepatology*, **10**, 1–7.
- Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, 2nd ed.
- Diggle, P. and Kenward, M. (1994). Informative drop-out in longitudinal data analysis (Disc: p73-93). *Applied Statistics*, **43**, 49–73.
- Dmitrienko, A., Tamhane, A. and Bretz, F. (2009). *Multiple testing problems in pharmaceutical statistics*. Chapman & Hall, New York.
- Dobson, A. J. (2001). *An Introduction to Generalized Linear Models*. Chapman & Hall Ltd, London, 2nd ed.

- Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. John Wiley & Sons, New York, Chichester.
- Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, **86**(413), 9–17.
- Efron, B. and Tibshirani, R. (1986). Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, **1**, 54–77.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall Ltd, London, New York.
- Ehrenberg, A. S. C. (1981). The problem of numeracy. *The American Statistician*, **35**, 67–71.
- Fewell, Z., Hernán, M. A., Wolfe, F., Tilling, K., Choi, H. and Sterne, J. A. C. (2004). Controlling for time-dependent confounding using marginal structural models. *Stata Journal*, **4**(4), 402–420.
- Fine, J. and Gray, R. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**(446), 496–497.
- Firth, D. (1993). Bias reduction in maximum likelihood estimates. *Biometrika*, **80**(1), 27–38.
- Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Data Analysis*. John Wiley & Sons, New York.
- Fleiss, J. L. (1988). One-tailed versus two-tailed tests: rebuttal. *Controlled Clinical Trials*, **10**, 227–228.
- Fleiss, J. L., Levin, B. and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*, 3rd Edition. John Wiley & Sons, New York, Chichester, 4th ed.
- Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, **58**, 21–29.
- Freedman, D., Pisani, R., Purves, R. and Adhikari, A. (1991). *Statistics*. W. W. Norton & Co, Inc., New York.
- Freedman, D. A. and Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, **32**(4), 392–409.
- Freedman, L. S., Graubard, B. I. and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167–178.
- Freireich, E. J., Gehan, E., Frei, E. I., Schroeder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., Moon, J. H., Gendel, B. R., Spurr, C. L. and Storrs, R. (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for the evaluation of other potentially useful therapy. *Blood*, **21**, 699–716.
- Friedman, L. M., Furberg, C. D. and Demets, D. L. (1998). *Fundamentals of Clinical Trials*. Springer, New York, 3rd ed.
- Frost, C. and Thompson, S. G. (2000). Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society, Series A, General*, **163**(2), 173–189.
- Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, **41**, 361–372.
- Gail, M. H., Tan, W. Y. and Piantodosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika*, **75**, 57–64.
- Gail, M. H., Wieand, S. and Piantodosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431–444.
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York.
- Glidden, D. V. and Vittinghoff, E. (2004). Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, **23**, 369–388.
- Glymour, M. M., Weuve, J., Berkman, L. F., Kawachi, I. and Robins, J. M. (2005). When is baseline adjustment useful in analyses of change? an example with education and cognitive change. *American Journal of Epidemiology*, **163**(3), 267–278.
- Goldberger, A. S. (1968). The interpretation and estimation of Cobb-Douglas functions. *Econometrica*, **36**, 464–472.

- Goldman, L., Cook, E. F., Johnson, P. A., Brand, D. A., Ronan, G. W. and Lee, T. H. (1996). Prediction of the need for intensive care in patients who come to the emergency departments with acute chest pain. *New England Journal of Medicine*, **334**(23), 1498–1504.
- Goldstein, H. (2003). *Multilevel Statistical Models*. Hodder Arnold, London, 3rd ed.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Gordon, W., Polansky, J., Boscardin, W., Fung, K. and Steinman, M. (2010). Coronary risk assessment by point-based and equation-based Framingham models: significant implications for clinical care. *Journal of General Internal Medicine*, **25**(11), 1145–51.
- Grady, D., Wenger, N. K., Herrington, D., Khan, S., Furberg, C., Hunninghake, D., Vittinghoff, E. and Hulley, S. (2000). Postmenopausal hormone therapy increases risk of venous thromboembolic disease. The Heart and Estrogen/progestin Replacement Study. *Annals of Internal Medicine*, **132**(9), 689–696.
- Greene, W. H. (1998). Gender economics courses in liberal arts colleges: further results. *Journal of Economic Education*, **29**(4), 291–300.
- Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*, **79**(3), 340–349.
- Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine*, **13**, 1665–1677.
- Greenland, S. (2000). An introduction of instrumental variables for epidemiologists. *International Journal of Epidemiology*, **29**, 722–729.
- Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, **14**(3), 300–306.
- Greenland, S. and Brumback, B. (2002). An overview of relations among causal modeling methods. *International Journal of Epidemiology*, **31**(5), 1030–1037.
- Greenland, S., Pearl, J. and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.
- Grodstein, F., Manson, J. E. and Stampfer, M. J. (2001). Postmenopausal hormone use and secondary prevention of coronary events in the Nurses' Health Study. *Annals of Internal Medicine*, **135**, 1–8.
- Haas, J., Phillips, K., Sonneborn, D., McCulloch, C., Baker, L., Kaplan, C. and Liang, E. P.-S. S.-Y. (2004). Variation in access to health care for different racial/ethnic groups by the racial/ethnic composition of an individual's county of residence. *Medical Care*, **42**, 707–714.
- Harrell, F., Lee, K. and Mark, D. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.
- Harrell, F. E. (2005). *Regression Modeling Strategies*. Springer, New York.
- Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, **3**, 143–152.
- Hastie, T. and Tibshirani, R. (1999). *Generalized Additive Models*. Chapman & Hall Ltd, London, New York.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag, New York.
- Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models (with discussion). *Statistical Science*, **1**, 297–318.
- Hauck, W. W., Anderson, S. and Marcus, S. M. (1998). Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials*, **19**, 249–256.
- Hearst, N., Newman, T. and Hulley, S. (1986). Delayed effects of the military draft on mortality: a randomized natural experiment. *New England Journal of Medicine*, **314**, 620–624.
- Heckman, J. (1997). Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources*, **32**(3), 441–462.

- Henderson, R. and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society, Series B, Methodological*, **61**, 367–379.
- Hernán, M. and Robins, J. (2011). *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL.
- Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E. and Robins, J. M. (2008). Observational studies analyzed like randomized experiments. *Epidemiology*, **19**(6), 766–779.
- Hernán, M. A., Brumback, B. and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, **11**, 561–570.
- Hernán, M. A., Brumback, B. and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, **96**(454), 440–448.
- Hernán, M. A., Cole, S. R., Margolick, J., Cohen, M. and Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*, **14**(7), 477–491.
- Hernán, M. A., Hernández-Díaz, S. and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, **15**(5), 615–625.
- Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, **17**(4), 360–372.
- Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, **55**(1), 19–24.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimates for nonorthogonal problems. *Technometrics*, **12**, 55–67.
- Hofer, T., Hayward, R., Greenfield, S., Wagner, E., Kaplan, S. and Manning, W. (1999). The unreliability of individual physician “report cards” for assessing the costs and quality of care of a chronic disease. *Journal of the American Medical Association*, **281**(22), 2098–2105.
- Hogan, J., Roy, J. and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine*, **23**(9), 1455–1497.
- Holcomb, W. L. J., Chaiworapongsa, T., Luke, D. A. and Burgdorf, K. D. (2001). An odd measure of risk: use and misuse of the odds ratio. *Obstetrics and Gynecology*, **98**, 685–688.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **86**(396), 945–960.
- Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons, New York, Chichester.
- Hsieh, F. Y., Bloch, D. A. and Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, **17**, 541–557.
- Hsieh, F. Y. and Lavori, P. W. (2000). Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials*, **21**, 552–560.
- Huber, P. J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In: *The Fifth Berkeley Symposium in Mathematical Statistics and Probability* (edited by L. Le Cam and J. Neyman). University of California Press, Berkeley.
- Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B. and Vittinghoff, E. (1998). Randomized trial of estrogen plus progestin for secondary prevention of heart disease in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. *Journal of the American Medical Association*, **280**(7), 605–613.
- Jewell, N. P. (2004). *Statistics for Epidemiology*. Chapman & Hall/CRC, Boca Raton, FL.
- Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, **103**(481), 101–111.
- Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.
- Kanaya, A., Vittinghoff, E., Shlipak, M. G., Resnick, H. E., Visser, M., Grady, D. and Barrett-Connor, E. (2004). Association of total and central obesity with mortality in postmenopausal women with coronary heart disease. *American Journal of Epidemiology*, **158**(12), 1161–1170.

- Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**(4), 523–539.
- Katan, M. B. (1986). Apolipoprotein e isoforms, serum cholesterol, and cancer. *Lancet*, **1**, 507–508.
- Keiding, N., Filiberti, M., Esbjerg, S., Robins, J. M. and Jacobsen, N. (1999). The graft versus leukemia effect after bone marrow transplantation: a case study using structural nested failure time models. *Biometrics*, **55**, 23–28.
- Kish, L. (1995). *Survey Sampling*. John Wiley & Sons, New York, Chichester.
- Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York.
- Kleinbaum, D. G. (2002). *Logistic Regression: a Self-Learning Text*. Springer-Verlag Inc, New York.
- Ko, H., Hogan, J. W. and Mayer, K. H. (2003). Estimating causal treatment effects using longitudinal natural history studies using marginal structural models. *Biometrics*, **59**, 152–162.
- Korff, M., Barlow, W., Cherkin, D. and Deyo, R. (1994). Effects of practice style in managing back pain. *Annals of Internal Medicine*, **121**, 187–195.
- Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys*. John Wiley & Sons, New York, Chichester.
- Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K. and Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, **163**(3), 262–270.
- Kuzniewicz, M., Escobar, G., Wi, S., Liljestrand, P., McCulloch, C. and Newman, T. (2008). Risk factors for severe hyperbilirubinemia among infants with borderline bilirubin levels. *Journal of Pediatrics*, **153**(2), 234–240.
- Lachenbruch, P. A. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research*, **11**, 297–302.
- Lagakos, S. W. and Schoenfeld, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics*, **40**, 1037–1048.
- Lane, P. W. and Nelder, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics*, **1982**(38), 613–621.
- Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, **41**, 191–201.
- Levy, P. and Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*. Wiley, New York, 3rd ed.
- Li, Z., Meredith, M. P. and Hoseyni, M. S. (2001). A method to assess the proportion of treatment effect explained by a surrogate endpoint. *Statistics in Medicine*, **20**, 3175–3188.
- Linhart, H. and Zucchini, W. (1986). *Model Selection*. John Wiley & Sons, New York, Chichester.
- Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. (1996). *SAS System for Mixed Models*. SAS Publishing, Cary, NC.
- Little, R. J. and Rubin, D. B. (2000). Causal effects in clinical and epidemiologic studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, **21**, 121–145.
- Little, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association*, **83**, 1227–1237.
- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.
- Little, R. J. A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–1121.
- Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, **54**, 217 – 224.

- Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**, 2937–2960.
- MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, **29**, 53–57.
- Magder, L. S. and Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, **146**, 195–203.
- Maldonado, G. and Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology*, **138**, 923–936.
- Mark, S. D. and Robins, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding using a rank preserving structural failure time model. *Statistics in Medicine*, **12**, 1605–1628.
- Martens, E. P., Pestman, W. R., do Boer, A., Belitser, S. V. and Klungel, O. H. (2006). Instrumental variables - application and limitations. *Epidemiology*, **17**(3), 260–267.
- Marubini, E. and Valsecchi, M. G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley & Sons, New York, Chichester.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall Ltd, New York, 2nd ed.
- McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models, 2nd Ed.* John Wiley & Sons, New York, Chichester.
- McNutt, L., Wu, C., Xue, X. and P., H. J. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology*, **157**, 940–943.
- Meier, P., Ferguson, D. J. and Karrison, T. (1985). A controlled trial of extended radical mastectomy. *Cancer*, **55**, 880–891.
- Merwin, E., Stern, S., Jordan, L. M. and Bucci, M. (2009). New estimates for crna vacancies. *AANA Journal*, **77**, 121–129.
- Mickey, R. M. and Greenland, S. (1989). The impact of confounder selection on effect estimation. *American Journal of Epidemiology*, **129**(1), 125–137.
- Miller, A. J. (1990). *Subset Selection in Regression*. Chapman & Hall Ltd, London, New York.
- Miller, R. G., Gong, G. and Munoz, A. (1981). *Survival Analysis*. John Wiley & Sons, New York, Chichester.
- Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T. and Alonso, A. (2002). Statistical challenges in evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*, **23**, 607–625.
- Molinaro, A. and van der Laan, M. J. (2004). Deletion/substitution/addition algorithm for partitioning the covariate space in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 162*.
- Mortimer, K. M., Neugebauer, R., van der Laan, M. and Tager, I. B. (2005). An application of model-fitting procedures for marginal structural models. *American Journal of Epidemiology*, **162**, 382–388.
- Neuhaus, J. (1998). Estimation efficiency with omitted covariates in generalized linear models. *Journal of the American Statistical Association*, **93**, 1124–1129.
- Neuhaus, J. and Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, **80**, 807–815.
- Newman, T. B., Kuzniewicz, M. W., Liljestrand, P., Wi, S. and McCulloch, C. (2009). Numbers needed to treat with phototherapy according to American Academy of Pediatrics Guidelines. *Pediatrics*, **123**, 1352–1359.
- O'Brien, T. R., Busch, M. P., Donegan, E., Ward, J. W., Wong, L., Samson, S. M., Perkins, H. A., Altman, R., Stoneburner, R. L. and Holmberg, S. D. (1994). Heterosexual transmission of human immunodeficiency virus type i from transfusion recipients to their sexual partners. *Journal of AIDS*, **7**, 705–710.
- Orwoll, E., Bauer, D. C., Vogt, T. M. and Fox, K. M. (1996). Axial bone mass in older women. *Annals of Internal Medicine*, **124**(2), 185–197.

- Orwoll, E., Blank, J., Barrett-Connor, E., Cauley, J., Cummings, S., Ensrud, K., Lewis, C., Cawthon, P., Marcus, R., Marshall, L. et al. (2005). Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study—a large observational study of the determinants of fracture in older men. *Contemporary Clinical Trials*, **26**(5), 569–585.
- Padian, N., van der Straten, A., Ramjee, G., Chipato, T., de Bruyn, G., Blanchard, K., Shiboski, S., Montgomery, E., Fancher, H., Cheng, H. et al. (2007). Diaphragm and lubricant gel for prevention of HIV acquisition in southern African women: a randomised controlled trial. *The Lancet*, **370**(9583), 251–261.
- Pagano, M. and Gavreau, K. (1993). *Principles of Biostatistics*. Wadsworth Publishing Co., Belmont, CA.
- Parzen, M. and Lipsitz, S. R. (1999). A global goodness-of-fit statistic for Cox regression models. *Biometrics*, **55**, 580–584.
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669–688.
- Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, **3**, 96–146.
- Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. 2nd ed.. Cambridge University Press, New York.
- Peduzzi, P., Concato, J. and Feinstein, A. R. (1995). Importance of events per independent variable in proportional hazards regression analysis ii. accuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, **48**, 1503–1510.
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, **49**, 1373–1379.
- Pencina, M. J., D’Agostino Sr, R. B., D’Agostino Jr, R. B. and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, **27**, 157–172.
- Permutt, T. H. and Hebel, J. R. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth-weight. *Biometrics*, **45**, 619–622.
- Petersen, M., Sinisi, S. and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology-Baltimore*, **17**(3), 276–284.
- Petersen, M. L., Porter, K., Gruber, S., Want, Y. and van der Laan, M. J. (2010). Diagnosing and responding to violations in the positivity assumption. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (269).
- Piantadosi, S. (2005). *Clinical trials: a methodologic perspective*, vol. 593. Wiley-Blackwell, New Jersey.
- R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.
- Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A*, **169**(4), 805–827.
- Raudenbush, S. W. and Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods (Advanced Quantitative Techniques in the Social Sciences)*. Sage, Newbury Park, CA.
- Ray, W. A. (2003). Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology*, **158**(9), 915–920.
- Robins, J. M., Blevins, D., Ritter, G. and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients. *Epidemiology*, **3**, 319–336.
- Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.
- Robins, J. M. and Greenland, S. (1994). Adjusting for differential rates of pcp prophylaxis in high- versus low-dose azt treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association*, **89**, 737–749.
- Robins, J. M., Greenland, S. and Hu, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, **94**(447), 687–700.

- Robins, J. M., Hernán, M. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.
- Rosenblum, M. and van der Laan, M. (2010). Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, Berlin **6**(2).
- Rosenman, R. H., Friedman, M., Straus, R., Wurm, M., Kositchek, R., Hahn, W. and Werthessen, N. T. (1964). A predictive study of coronary heart disease: the western collaborative group study. *Journal of the American Medical Association*, **189**, 113–120.
- Rothman, K. J. and Greenland, S. (1998). *Modern Epidemiology*. Lippincott Williams & Wilkins Publishers, Philadelphia, PA, 2nd ed.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, Chichester.
- Rubin, D. B. (1996). Multiple imputation after 18 + years. *Journal of the American Statistical Association*, **91**, 473–489.
- Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, **127**, 757–763.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to tobacco litigation. *Health Services and Outcomes Research Methodology*, **2**, 169–188.
- Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, **1**(4), 381–397.
- Rust, K. and Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, **5**, 283–310.
- Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, **13**(4), 279–313.
- Scheaffer, R. L. (1996). *Elementary Survey Sampling*. Duxbury, Boston, 5th ed.
- Scheuren, F. (2005). Multiple imputation: How it began and continues. *The American Statistician*, **59**, 315–319.
- Schmoor, C., Sauerbrei, W. and Schumacher, M. (2000). Sample size considerations for the evaluation of prognostic factors in survival analysis. *Statistics in Medicine*, **19**, 441–452.
- Schmoor, C. and Schumacher, M. (1997). Effects of covariate omission and categorization when analysing randomized trials with the Cox model. *Statistics in Medicine*, **16**, 225–237.
- Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H. and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, **20**(4), 512–522.
- Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, **67**, 145–153.
- Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84**, 57–71.
- Sedransk, N., Young, L., Kelner, K., Moffitt, R., Thakar, A., Raddick, J., Ungvarsky, E., Carlson, R., Apweiler, R., Cox, L. et al. (2010). Make research data public? not always so simple: A dialogue for statisticians and science editors. *Statistical Science*, **25**(1), 41–50.
- Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In: *AIDS Epidemiology: Methodological Issues* (edited by N. Jewell, K. Dietz and V. Farewell). Birkhauser, Boston.
- Self, S. G. and Mauritsen, R. H. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, **48**, 31–39.
- Smith, G. D. and Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, **33**, 30–42.
- Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods*. Iowa State University Press, Ames, IA, eighth ed.
- Snow, J. (1855). *On the Model of Communication of Cholera*. John Churchill, London.

- Sommer, A. and Zeger, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine*, **10**, 45–52.
- Splieth, C., Steffen, H., Welk, A. and Schwahn, C. (2005). Responder and nonresponder analysis for a caries prevention program. *Caries Research*, **39**, 269–272.
- Sterne, J. A. C. and Tilling, K. (2002). G-estimation of causal effects, allowing for time-varying confounding. *Stata Journal*, **2**(2), 164–182.
- Steyerberg, E. W. (2009). *Clinical Prediction Models*. Springer, New York.
- Stone, C. J. (1986). Comment: Generalized additive models. *Statistical Science*, **47**, 312–314.
- Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, **64**(3), 479–498.
- Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, **21**(153), 65–66.
- Subak, L., Wing, R., West, D., Franklin, F., Vittinghoff, E., Creasman, J., Richter, H., Myers, D., Burgo, K., Gorin, A., Macer, J., Kusek, J. and Investigators., D. G. P. (2009). Weight loss to treat urinary incontinence in overweight and obese women. *New England Journal of Medicine*, **360**(5), 481–490.
- Suissa, S. (2008). Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology*, **167**(4), 492–499.
- Sun, G. W., Shook, T. L. and Kay, G. L. (1999). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*, **49**, 907–916.
- Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.
- Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385–395.
- Tilling, K., Sterne, J. A. and Szklo, M. (2002). Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using G-estimation: the Atherosclerosis Risk in Communities Study. *American Journal of Epidemiology*, **155**, 710–718.
- Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data. *Statistica Sinica*, **14**(3), 809–834.
- Tung, P., Kopelnik, A., Banki, N., Ong, K., Ko, N., Lawton, M. T., Gress, D., Drew, B. J., Foster, E., Parmley, W. W. and Zaroff, J. G. (2004). Predictors of neurocardiogenic injury after subarachnoid hemorrhage. *Stroke*, **35**(2), 548–551.
- Van Der Laan, M. and Petersen, M. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, **3**(1).
- van Houwelingen, H. C. (2000). Validation, calibration, revision, and combination of prognostic survival models. *Statistics in Medicine*, **19**, 3401–3415.
- Vanderpump, M. P., Tunbridge, W. M., French, J. M., Appleton, D., Bates, D., Clark, F., Grimley Evans, J., Rodgers, H., Tunbridge, F. and Young, E. T. (1996). The development of ischemic heart disease in relation to autoimmune thyroid disease in a 20-year follow-up study of an English community. *Thyroid*, **6**, 155–160.
- VanderWeele, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology-Baltimore*, **20**(1), 18–26.
- Vansteelandt, S. (2009). Estimating direct effects in cohort and case-control studies. *Epidemiology*, **20**(6), 851–860.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.
- Verweij, P. J. M. and Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, **13**, 2427–2436.
- Vittinghoff, E. and McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, **165**, 710–718.
- Vittinghoff, E., Sen, S. and McCulloch, C. E. (2009). Sample size calculations for evaluating mediation. *Statistics in Medicine*, **28**(4), 1623–1634.
- Vittinghoff, E., Shlipak, M. G., Varosy, P. D., Furberg, C. D., Ireland, C. C., Khan, S. S., Blumenthal, R., Barrett-Connor, E. and Hulley, S. (2003). Risk factors and secondary prevention

- in women with heart disease: The Heart and Estrogen/progestin Replacement Study. *Annals of Internal Medicine*, **138**(2), 81–89.
- Volberding, P. A., Lagakos, S. W. and Koch, M. A. (1990). Zidovudine in asymptomatic human-immunodeficiency-virus infection – a controlled trial in persons with fewer than 500 cd4-positive cells per cubic millimeter. *The New England Journal of Medicine*, **322**(14), 941–949.
- Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**(2), 307–333.
- Walter, L. C., Brand, R. J., Counsell, S. R., Palmer, R. M., Landefeld, C. S., Fortinsky, R. H. and Covinsky, K. E. (2001). Development and validation of a prognostic index for 1-year mortality in older adults after hospitalization. *Journal of the American Medical Association*, **285**(23), 2987–2994.
- Wei, L. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, **11**(14–15), 1871–1879.
- Wei, L. J. and Glidden, D. V. (1997). An overview of statistical methods for multiple failure time data in clinical trials (with discussion). *Statistics in Medicine*, **16**(8), 833–839.
- Weisberg, S. (1985). *Applied Linear Regression*. John Wiley & Sons, New York, Chichester.
- Whelan, T., Levine, M., Willan, A., Gafni, A., Sanders, K., Mirsky, D., Chambers, S., O'Brien, M. A., Reid, S. and Dubois, S. (2004). Effect of a decision aid on knowledge and treatment decision making for breast cancer surgery: a randomized trial. *Journal of the American Medical Association*, **292**, 435–441.
- White, I. R. and Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, **28**, 1982–1998.
- Whooley, M., de Jonge, P., Vittinghoff, E., Otte, C., Moos, R., Carney, R., Ali, S., Carney, R., Na, B., Feldman, M., Schiller, N. and Browner, W. (2008). Depressive symptoms, health behaviors, and risk of cardiovascular events in patients with coronary heart disease. *Journal of the American Medical Association*, **300**(20), 2379–2388.
- Wittelman, J. C. M., D'Agostino, R. B., Stijnen, T., Kannel, W. B., Cobb, J. C., de Ridder, M. A. J., Hofman, A. and Robins, J. M. (1998). G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. *American Journal of Epidemiology*, **148**(4), 390–401.
- Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.
- Yelland, L., Salter, A. and Ryan, P. (2011). Relative risk estimation in randomized controlled trials: A comparison of methods for independent observations. *The International Journal of Biostatistics*, **7**(1), 5.
- Zhang, J. and Yu, K. F. (1998). What's the relative risk? a method for correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*, **280**, 1690–1691.
- Zou, G. (2004). A modified poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, **159**(7), 702.

Index

A

ACE, *see* average causal effect
 Additive model, 105, 165, 208
 Additive risk model, 140, 165
 Adjusted R^2 , 400
 Adjustment, 70–72, 74, 89–94, 156–160, 218–219, 222–224, 228–230
 AIC, *see* Akaike Information Criterion
 Akaike Information Criterion, 400
 Allen–Cady procedure, 420
 Alternative hypothesis, 28, 30, 46, 60–61, 81
 Analysis of covariance, 31
 Analysis of variance, 30–33, 262
 multi-way, 31
 one-way, 30, 78
 two-way, 263
 ANCOVA, *see* analysis of covariance
 ANOVA, *see* analysis of variance
 Area under the curve, 270
 As-treated comparisons, 384
 Asymptotics, *see* large sample behavior
 ATT, *see* average treatment effect in the treated
 Attenuation, 74, 90, 94–99, 218–417
 Attributable risk, 44
 AUC, *see* area under the curve
 Average causal effect, 332
 Average treatment effect in the treated, 339, 360, 361

B

Balanced repeated replication, 478
 Bandwidth, 20, 109, 173
 Baseline outcome
 as a covariate, 273
 Baseline predictor, 289
 Bayesian Information Criterion, 400

Best subsets, 420, 424
 Between-cluster predictor, 301
 Bias–variance trade-off, 397
 BIC, *see* Bayesian Information Criterion
 Bonferroni procedure, 30, 82, 487
 Bootstrap confidence intervals, 62–63, 99, 182, 247, 298, 349
 Borrow strength, 229, 237, 268
 Boxplot, 12, 116, 125
 BRR, *see* balanced repeated replication

C

CACE, *see* complier averaged causal effect
 Case-control studies, 46, 168–173
 Categorical variable, 8
 Causal
 diagram, 160, 418, 428
 inference, fundamental problem of, 333
 interpretation, 72, 89, 94, 99, 100, 102, 104, 157, 169, 418
 Cause-specific hazard, 241, 244–247
 CD-MCAR, *see* covariate-dependent missing completely at random 440, 444, 452
 Ceiling effect, 109, 111
 Censoring, 241
 dependent, 249
 independent, 57, 249
 interval, 249
 reasons for, 204
 right, 55, 203
 Centering, 36, 107, 154, 165, 174, 208, 223, 230, 407
 Change score, 106, 269, 273
 $\bar{\chi}^2$ test, 288
 χ^2 test, 46–48, 146, 154, 212–213, 215–217
 Classification and regression trees, 168, 190

- Cluster resampling, 299
 - Cluster sampling, 479
 - Clustered data, 37, 257, 261
 - Coefficient of determination, 42, 75
 - Collapsibility, 52, 96
 - Collinearity, 107, 174, 407, 418, 421–422
 - Competing risks, 239–247
 - Complete case analysis, 431
 - Complete null hypothesis, 82
 - Complex surveys, 37, 469–479
 - Complier averaged causal effect, 386
 - Component plus residual plot, 109–111, 113, 174
 - Conditional
 - effects, 344
 - independence, 334
 - logistic regression model, 171
 - mean, 335
 - model, 285, 295, 417
 - odds-ratio, 345
 - risk difference, 345
 - Confidence intervals
 - bootstrap, 62–63, 99, 182, 247, 298, 349
 - complex surveys, 474, 479
 - complimentary log-log model, 183
 - Cox proportional hazards model, 213
 - linear regression model, 74–75
 - logistic regression model, 145, 152, 154, 162, 170
 - nonparametric binary models, 190
 - relationship to hypothesis tests, 41
 - simple linear model, 40
 - Confounding, 52, 70–72, 74, 89–94, 156–160, 218–219, 222–224, 228–230, 248, 291, 407–409, 416, 418–422
 - by indication, 100, 227
 - negative, 90, 424
 - no unmeasured, 338
 - patterns, 90
 - Constant variance, 32, 36, 70–73, 119–123, 123
 - test for, 121
 - Contingency table methods, 42–53
 - Continuation ratio model, 119, 191
 - Continuous variable, 8
 - Contrast, linear, 78, 84, 215
 - Controlled direct effect, 371
 - Convenience sampling, 469
 - Correlation, 265
 - coefficient, 18, 33–35, 42, 74–75
 - multiple, 75
 - relationship to regression coefficient, 42
 - Spearman, 34
 - intraclass, 266
 - matrix, 22
 - structure, 279, 280, 473, 477
 - autoregressive, 280
 - exchangeable, 280, 477
 - nonstationary, 280
 - stationary, 280
 - unstructured, 280
 - working, 281, 473
 - within-cluster, 470, 474
 - Count data, 309
 - Counterfactuals, 428
 - Covariance, 33, 265
 - Covariate, *see* predictor
 - balance, 353
 - overlap, 338
 - Covariate dependent MCAR, 440
 - Cox proportional hazards model, 61, 207–239, 422, 423, 473
 - CPR plot, *see* component plus residual plot
 - Cross-validation, 168, 399
 - development set/validation set, 399
 - h-fold, 399
 - Cumulative
 - event function, 59
 - Cumulative event function, 204
 - Cumulative incidence function, 241
 - Cutpoints, 114, 232
- D**
- Data
 - checking, 7
 - count, 309
 - errors, 7
 - Deciles, 22
 - Degrees of freedom, 32, 40, 74, 107, 146, 155, 212, 215, 474
 - Dependent censoring, 249
 - Derived variable, 268, 269
 - Design effect, 301, 474
 - Detectable effects, 130–135, 252–255, 303, 325–327
 - Development set/validation set, 399
 - DFBETAs, 125–127, 175, 236
 - Difference score, 106, 273
 - Direct effect
 - controlled, 371
 - natural, 373
 - Discrete variable, 8
 - Distribution
 - binomial, 143
 - exponential, 208
 - gamma, 315

heavy-tailed, 13
 light-tailed, 13
 non-normal, 312
 normal, 13
 Poisson, 312
 Weibull, 208
 Dummy variable, *see* indicator variable
 Duncan procedure, 84
 Dunnett's test, 83

E

Effect size, 303
 EM algorithm, *see* Expectation-Maximization Algorithm 458
 Error, 36, 73
 family-wise rate, 30, 82, 215, 486
 in predictors, 37
 prediction, 396
 Events per predictor, 180, 222
 Exact
 χ^2 test, 47
 Fisher's test, 47
 logistic regression, 188
 Excess risk, 43–46, 48
 Excess zeros, 318
 Expectation-Maximization Algorithm, 458
 Exponential model, 208
 Extrapolation, 339

F

F -test, 30–33, 79–82, 84–89, 113, 421
 Face validity, 396, 408, 420, 423
 Factor
 fixed, 286
 random, 286
 False discovery rate, 487
 False-negative rate, 166
 False-positive rate, 166
 Family-wise error rate, 30, 82, 215, 486
 FER, *see* family-wise error rate
 Fisher's
 exact test, 47
 least significant difference, 82
 Fitted values, 39, 74, 75, 119, 166
 Fixed factor, 286
 Floor effect, 109, 111

G

G-computation, 388
 G-estimation, 351, 389
 Gamma distribution, 315

GEE, *see* generalized estimation equations 276
 GEEs, 328
 Generalized additive models, 136, 189
 Generalized estimating equations, 276, 281, 455, 473, 474
 Generalized linear models, 119, 122, 309
 choice of distribution, 312, 315, 316, 323
 interpretation of parameters, 312, 316
 link function, 316
 mean-to-variance relationship, 323
 model for mean response, 311, 315, 316
 repeated measures, 328
 Goodness of fit test, 178

H

Hazard, 204–205, 241, 244
 baseline, 208–210, 213, 228
 Breslow estimator, 209, 223
 ratio, 205–206, 235
 Heavy-tailed distribution, 13
 Heteroscedasticity, 32, 36, 70–73, 119–123
 Hierarchical data, 257, 261, 267
 High leverage points, 124–125
 Histogram, 10, 116
 Homoscedasticity, 32, 36, 70–73, 119–123
 Hosmer–Lemeshow test, 178
 Hurdle model, 319
 Hypothesis tests, relationship to confidence intervals, 41

I

Identity link, 316
 Ignorable missing data, 441
 Imputation, 200
 chained, 448
 multiple, 432, 434, 474
 multivariate normal, 448
 single, 434
 Incidence proportion, 45
 Inclusion criterion, 396, 408
 Independence, 29, 36, 73, 143, 192
 conditional, 334
 Independent censoring, 57, 249
 Indicator variable, 76–77, 100, 209, 226
 Infectious disease transmission models, 181
 Inferential goals, 396
 evaluating a predictor of primary interest, 396, 407–418
 identifying multiple important predictors, 396, 418–420
 prediction, 396
 Influential points, 38, 124–128, 175–176, 236

Instrumental variables, 373–382, 390
 Intention to treat, 382
 Interaction, 23, 52, 94, 99–108, 160–165, 190,
 220–222, 229, 238–239, 271, 408,
 409, 418–420
 qualitative, 107
 term, 100, 107
 Interval censoring, 249
 Intraclass correlation, 266
 Inverse probability weighting, 337, 432, 456
 IPW, *see* inverse probability weighting 456
 Iterative chained imputation, 448
 ITT, *see* intention to treat

J

Jackknife, 478

K

Kaplan–Meier estimator, 55–59, 222, 230, 233,
 248
 Kendall's τ , 34
 Kruskal–Wallis test, 32

L

Large sample behavior, 32, 38, 41, 116, 156,
 212, 213
 LASSO, *see* least absolute shrinkage and
 selection operator
 LATE, *see* local average treatment effect
 Learning set/test set, 168
 Least absolute shrinkage and selection
 operator, 403, 419
 Left truncation, 250
 Left-skewed, 13
 Legression coefficient
 interpretation, 73
 Leverage, 124
 Light-tailed distribution, 13
 Likelihood, 145, 149, 192–194, 212
 Likelihood ratio test, 145, 147, 154–156, 171,
 174, 194, 212–213, 215, 217, 423
 Line of means, 35, 109
 Linear
 contrast, 78, 84, 215
 predictor, 73, 207–208, 248, 325
 trend, 109
 test for, 149, 215–216
 tests for, 84–89
 Linear predictor, 142, 248
 Linear regression model, 473
 adjustment, 70–72, 74, 89–94

attenuation, 74, 94–99
 bootstrap confidence intervals, 99
 confidence intervals, 74–75
 confounding, 70–72, 74, 89–94
 hypothesis tests, 74–75
 interaction, 94, 99–108
 interpretation of regression coefficients,
 73
 mediation, 94–99
 model checking, 108–128
 single predictor, 35–41, 70

Linearity, 109–115

log, 173–175, 231

Linearization, Taylor series, 474

Link

identity, 282, 316
 log, 311, 312, 315
 logit, 143, 316
 specification test, 177

Link function, 316

Local average treatment effect, 382

Log-likelihood, *see* likelihood

Log-linearity, 173–175, 231

Logistic regression model, 44, 119, 122, 294,
 316, 319, 422, 423, 473

adjustment, 156–160

bootstrap confidence intervals, 182

conditional, 171

confidence intervals, 162

confounding, 156–160

excess zeros, 319

for matched case-control studies, 171

interaction, 160–165

mediation, 157, 158

repeated measures, 284, 294

Logit link, 143, 316

Logrank test, 60–61, 215

Longitudinal, 270

LOWESS, 18, 109–111, 122, 173, 189,
 205–206, 236

LR, *see* likelihood ratio

LS/TS, *see* development set/validation set

LSD, *see* Fisher's least significant difference

M

Mallow's C_p , 400

Mantel–Haenszel

combined odds ratio, 50

test of homogeneity, 51

MAR, *see* missing at random 451

Marginal

effects, 344–351

mean, 335

- model, 285, 417
 - odds-ratio, 345
 - risk difference, 346
 - structural model, 333, 345
- Masking, 90
- Matching
 - in case-control studies, 171
 - on propensity scores, 361
- Matrix plot, 278
- Maximum likelihood, 192–194, 455
 - estimation, 192, 194
- MCAR, *see* missing completely at random 451
- Mediation, 53, 94–99, 157–160, 219–220, 226, 370–373, 390, 418, 420
- Missing at random, 440, 455, 473
- Missing completely at random, 439, 461, 473
 - covariate dependent, 440
- Missing data, 152, 200, 431, 473, 474
 - at random, 431, 440, 455
 - completely at random, 439, 461
 - generalized estimating equations, 455
 - ignorable, 441
 - informative, 431
 - maximum likelihood, 455
 - not at random, 441, 461
 - pattern mixture models, 462
- Missing not at random, 441, 461
- Mixed model, 286
- MNAR, *see* missing not at random 451
- Model
 - additive, 208
 - checking, 108–128, 173–179, 231–239
 - conditional, 285, 417
 - generalized additive, 136, 189
 - generalized linear, 309
 - marginal, 285, 417
 - marginal structural, 333, 345
 - multiplicative, 208–210, 310
 - nested, 113, 212
 - nonlinear, 324
 - population-averaged, 417
 - size, 222
 - specification, 338, 353
 - structural nested, 351, 389
 - subject-specific, 417
 - sum of squares, 39, 42, 75
- Model size, 180
- Multi-stage sampling, 470, 477
- Multinomial logistic model, 191
- Multiple
 - comparisons, 30, 61, 81, 83, 419, 486, 487
 - hypothesis tests, 486, 487
 - imputation, 432, 434, 474
- Multiplicative model, 105, 165, 208–210, 310
 - risk, 143, 165
- Multivariate normal imputation, 448
- N**
 - Natural direct effect, 373
 - Negative binomial, 122, 318
 - zero-inflated, 320
 - Negative confounding, 90, 424
 - Negative findings, 64, 323
 - Nested models, 113, 154, 156, 194, 212
 - New user
 - design, 369, 370
 - nested cohorts, 370
 - Nominal variable, 8
 - Non-response, 473, 474
 - item, 474
 - unit, 474
 - Nonlinear model, 324
 - Nonparametric, 32, 61, 109, 116, 209, 235
 - Normal distribution, 13, 31, 32, 36, 37, 40, 41, 73, 116–119, 141, 177
 - tests for, 117
 - Null hypothesis, 28–32, 40, 46, 49, 51, 58, 60–61, 74, 76, 81, 117, 212, 417
 - complete, 82
 - multiple, 30, 486
 - partial, 82
 - Number of predictors, 180, 222
 - Numeric variable, 37, 77
- O**
 - Odds ratio, 43–46, 48, 140, 144, 145, 147, 151, 169, 179
 - combined, 50
 - Offset, 182, 311
 - OLS, *see* ordinary least squares
 - One-sided tests, 28–29
 - Ordinal variable, 8
 - Ordinary least squares, 38, 116
 - Outliers, 12, 15, 38, 124–125, 175
 - Overdispersion, 313, 317
 - Oversampling, 107, 470, 471
- P**
 - Paired *t*-test, 29
 - Parallel lines assumption, 104, 191, 209
 - Parsimonious models, 419, 423
 - Partial null hypothesis, 82
 - Pattern mixture models, 462
 - PE, *see* prediction error

- Penalized estimation, 403
 - Percent change, 106
 - Plots
 - adjusted survival curves, 222–224
 - box, 12, 116, 125
 - component plus residual, 109–111, 113, 174
 - histogram, 116
 - Kaplan–Meier, 55–59, 59, 222, 230, 233, 248
 - log minus log survival, 232–235
 - Q-Q, 13, 116
 - residual vs. predictor, 109, 119
 - ROC, 167
 - scatterplot matrix, 23, 278
 - smoothed hazard ratio, 235–236
 - stratified survival curves, 230
 - Poisson
 - distribution, 312
 - model, 122
 - regression model, 316, 317
 - zero-inflated, 320
 - Polytomous logistic model, 191
 - Pooled
 - logistic regression, 183
 - Pooled logistic regression, 183
 - Population-averaged, 285, 295
 - model, 417
 - Positivity, 354, 366
 - assumption, 338
 - restriction, 339, 359
 - Potential outcomes, 332–337
 - cumulative risk estimation, 351
 - estimation, 336, 344–351, 360, 388
 - survival models, 351
 - trials with incomplete adherence, 382
 - Power, 130–135, 252–255, 301–304, 325–327
 - Prediction, 165–168, 248, 267, 293, 396
 - error, 396
 - Predictor
 - events per, 180, 222
 - number of, 180, 222
 - assumptions about, 37
 - baseline, 289
 - binary, 76–77, 213
 - categorical, 48–53, 76–89, 107, 213–216, 234
 - continuous, 35, 37, 108, 119, 217–218
 - events per, 422–423
 - measurement error, 37
 - multiple important, 418–420, 422
 - number of, 422–423
 - of primary interest, 228, 237, 407–418, 421, 423
 - selection, 395–396
 - Allen–Cady procedure, 420
 - backward, 396, 408, 420, 423, 424
 - best subsets, 420, 424
 - forward, 408, 424
 - number of predictors, 341, 422–423
 - stepwise, 408, 420, 424
 - time-dependent, 225–227, 238–239
 - time-invariant, 271, 289
 - time-varying, 289
 - Prevalence, 45
 - Primary sampling unit, 470–474
 - Principal stratification, 385
 - Probability
 - of inclusion, 107, 469–472, 474, 479
 - unequal, 470–472
 - sample, 471
 - weights, 457, 471–474
 - Product limit, *see* Kaplan–Meier estimator
 - Product term, *see* interaction term, 220–222, 229, 238
 - Propensity scores, 352–363, 389
 - advantages and limitations, 363
 - interactions with exposure, 358
 - inverse probability weights, 356
 - matching, 361
 - positivity violations, 359
 - potential outcomes estimation, 360
 - quintiles, 355
 - recommendations, 362
 - restricted cubic splines, 355
 - standardized mortality ratio weights, 361
 - Proportional hazards, 207–210
 - checking, 232–239, 366
 - parametric models, 208, 257
 - Schoenfeld test, 236–237
 - Proportional odds model, 119, 190
 - Pseudo- R^2 , 146
 - PSU, *see* primary sampling unit
- Q**
- Q-Q plot, 13, 116
 - Quadratic term, 111, 174
 - Quartiles, 22
 - Quintiles, 22
- R**
- R^2 , 42, 75, 111–113, 146, 398
 - adjusted, 400
 - pseudo, 146
 - Random effects, 286
 - predicted, 293

- Random factor, 286
- Randomization assumption, 225, 334
- Rank-based methods, 32–35, 61
- Receiver operator characteristic curve, 167
- Reference group, 78, 213–215
- Regression coefficient
 - change in, 125
 - interpretation, 36, 144, 151, 153, 160, 186, 312, 316
 - standardized, 290
 - variance, 74, 423
- Regression dilution bias, 37
- Regression line, 35, 39, 72, 109
- Regression standardization, 388
- Relative hazard, *see* hazard ratio
- Relative risk, 43–46, 48, 140, 169–171, 179
 - model, 180, 186–188
- Relative risk ratio, 191
- Repeated measures
 - data, 261
 - models
 - potential outcomes estimation, 350
 - time-dependent treatments, 367
- Repeated measures models
 - missing data, 431
- Representative sampling, 469, 470
- Reproducibility, 485
- Resample, 299
- Residual
 - sum of squares, 39
 - variance, 74, 422
 - vs predictor plot, 109
 - vs. predictor plot, 119
- Residuals, 39, 231
 - Schoenfeld, 236
 - standardized Pearson, 175
- Restriction, 339
- Ridge regression, 403
- Right truncation, 251
- Right-skewed, 13, 116
- Risk
 - difference, 43–46, 140, 169, 371
 - model, 165, 180, 186–188
 - ratio, 43–46
- Risk difference, 323
- Risk score, 248
- Robust standard error, 281, 474
- Robustness, 31, 32, 59, 117, 209, 257
- ROC curve
 - see* receiver operator characteristic curve, 167
- RSS, *see* residual sum of squares
- RVP plot, *see* residual vs. predictor plot
- S**
 - Sample size, 130, 135, 252, 255, 301, 304, 325, 327, 396, 418, 422
 - adjusting for covariate, 303, 304
 - adjusting for covariates, 130, 132, 252
 - number of predictors, 180, 222, 341
 - Sampling
 - case-control, 169, 172
 - cluster, 470, 474, 479
 - complex, 470, 474
 - convenience, 469
 - fraction, 472
 - multi-stage, 470, 477
 - probability, 471
 - representative, 469, 470
 - weight, 457
 - Scale parameter, 313
 - Scatterplot matrix plot, 23
 - Scatterplot smoother, *see* smoothing, LOWESS
 - Scheffé procedure, 30, 82
 - Schoenfeld
 - residuals, 236
 - test for proportional hazards, 236–237
 - Selection model, 462
 - Semi-parametric, 209
 - Sensitivity, 167
 - Shrinkage estimator, 293, 403
 - Sidak procedure, 30, 82
 - Simple random sample, 470
 - Simpson's paradox, 51
 - Single imputation, 434
 - Skewness, 13, 116, 309
 - Smoothing, 18, 109, 111, 122, 173, 189, 205–206, 236
 - hazard ratio, 236
 - Spearman correlation coefficient, 34
 - Specificity, 167
 - Splines, 174, 184, 234, 271, 289
 - restricted cubic, 174, 184, 234, 355
 - SRS, *see* simple random sample
 - Standard errors, 40, 74, 145
 - complex surveys, 474, 479
 - relative, 477
 - robust, 281, 474
 - Standardized regression coefficient, 75–76, 290
 - Statistical significance
 - lack of, 323
 - Step function, 114, 141, 239
 - Step-down procedure, 84
 - Step-up procedure, 84
 - Stratification in complex surveys, 471–474
 - Stratified Cox model, 228–230, 237

- Student-Newman-Keuls procedure, 84
 - Subgroup analysis, 107, 268
 - Subject-specific, 285, 295
 - model, 417
 - Sum of squares, 39, 74
 - model, 39, 42, 75
 - residual, 39
 - total, 39, 42, 75
 - Survival
 - function, 55–59
 - Kaplan–Meier estimate, 55–59
 - parametric, 59
 - models
 - potential outcomes estimation, 351
 - time-dependent treatments, 367
 - time
 - mean, 59
 - median, 58
 - quantiles, 59
 - Survival function, 204
 - adjusted estimate, 222–224, 230
 - baseline, 223–224
 - Kaplan–Meier estimate, 222, 230, 233, 248
 - stratified estimate, 230
 - Survival models
 - parametric, 257
 - Survival time
 - mean, 224, 248
 - predicted, 224, 248
 - Survivor function, *see* survival function
- T**
- t -distribution, 40
 - t -statistic, 28, 40
 - t -test, 27–33, 40, 74, 76, 78, 113, 421, 474
 - paired, 273
 - t -test
 - paired, 29
 - unequal variance, 32
 - Target population, 192, 469, 470
 - Taylor series linearization, 474
 - Tertiles, 22
 - Test
 - $\bar{\chi}^2$, 288
 - χ^2 , 46, 48, 212, 213, 215, 216
 - F , 30, 33, 79, 82, 84, 89, 113
 - Fisher's exact, 47
 - for trend, 49, 84, 89, 149, 215, 216
 - goodness of fit, 178
 - Hosmer–Lemeshow, 178
 - Kruskal–Wallis, 32
 - likelihood ratio, 145, 147, 154, 156, 171, 174, 194, 212, 213, 215
 - link specification, 177
 - logrank, 60, 61, 215
 - Mantel–Haenszel, 51
 - multiple stage, 84
 - of association, 46, 48
 - of homogeneity, 49, 51
 - t , 27, 33, 40, 74, 76, 78, 89, 113, 474
 - Vuong's, 113
 - Wald, 145, 149, 154, 156, 212, 213, 215, 216, 282
 - Wilcoxon, 32
 - censored, 61
 - Z , 212, 213, 222, 474
 - Time origin, 225, 250
 - Time-dependent
 - covariates, 225–227, 238–239, 364
 - treatments, 364–369, 389
 - Time-invariant predictor, 289
 - Time-varying predictor, 289
 - Total sum of squares, 39, 42, 75
 - Transformations, 15, 32, 111–115, 310
 - back, 128, 217, 314
 - outcome
 - log, 15, 117, 128
 - normalizing, 117–119, 177
 - power, 118
 - rank, 118
 - variance-stabilizing, 121–123, 123
 - predictor
 - categorical, 114, 115
 - linearizing, 111–115, 178
 - log, 15, 111, 128
 - polynomial, 111
 - restricted cubic splines, 113
 - square root, 111
 - smooth, 111–114
 - restricted cubic splines, 114
 - Tree-based methods, 168, 190
 - Trend
 - test for, 84–89, 149, 215–216
 - trend
 - test for, 49, 61
 - Truncated, 319
 - Truncation, 250
 - left, 250
 - right, 251
 - TSS, *see* total sum of squares
 - Two-part model, 319
 - Two-sided tests, 28–29
 - Type-I error, 30–32, 107, 108, 215, 396, 420, 425

U

Unbalanced data, 270
Unbiased estimation, 38, 100, 102, 173, 174
Unequal probability of inclusion, 470
Unequal variance, 32, 36, 70–73, 119–123

V

Variable, 8
 categorical, 8
 continuous, 8
 continuous versus discrete, 8
 dependent, *see* variable, outcome
 derived, 268, 269
 discrete, 8
 independent, *see* predictor
 nominal, 8
 numeric, 8
 ordinal, 8
 outcome, 17
 predictor, 17
 response, 17
 transformations, 15
Variance
 estimation, 74, 474–478
 inflation factor, 74, 416
 predictor, 74

 regression coefficient, 74, 423
 residual, 40, 74, 422
Vuong's test, 113

W

Wald test, 145, 149, 154, 156, 212–213, 215–217, 282, 423
Weibull model, 208
Weight
 inverse probability of censoring, 249
 inverse probability, 337, 356, 365, 456
 of censoring, 365
 probability, 457, 471–474
 sampling, 457
 stabilized, 365
 standardized mortality ratio, 361
 time-dependent, 365
Wilcoxon test, 32
 censored, 61
Winsorization, 118
Within-cluster predictor, 301
Within-subject effects, 291

Z

Z-test, 212–213, 222, 474
Zero-inflated, 320
Zero-truncated, 319