Statistical Modeling for Biomedical Researchers
Contents

Preface page xvii

1 Introduction 1

1.1 Algebraic notation 1

1.2 Descriptive statistics 3

1.2.1 Dot plot 3

1.2.2 Sample mean 4

1.2.3 Residual 4

1.2.4 Sample variance 4

1.2.5 Sample standard deviation 5

1.2.6 Percentile and median 5

1.2.7 Box plot 5

1.2.8 Histogram 6

1.2.9 Scatter plot 6

1.3 The Stata Statistical Software Package 7

1.3.1 Downloading data from my website 8

1.3.2 Creating histograms with Stata 9

1.3.3 Stata command syntax 12

1.3.4 Obtaining interactive help from Stata 13

1.3.5 Stata log files 14

1.3.6 Stata graphics and schemes 14

1.3.7 Stata do files 15

1.3.8 Stata pulldown menus 15

1.3.9 Displaying other descriptive statistics with Stata 20

1.4 Inferential statistics 22

1.4.1 Probability density function 23

1.4.2 Mean, variance, and standard deviation 24

1.4.3 Normal distribution 24

1.4.4 Expected value 25

1.4.5 Standard error 25
2 Simple linear regression

2.1 Sample covariance 45
2.2 Sample correlation coefficient 47
2.3 Population covariance and correlation coefficient 47
2.4 Conditional expectation 48
2.5 Simple linear regression model 49
2.6 Fitting the linear regression model 50
2.7 Historical trivia: origin of the term regression 52
2.8 Determining the accuracy of linear regression estimates 53
2.9 Ethylene glycol poisoning example 54
2.10 95% confidence interval for $y[x] = \alpha + \beta x$ evaluated at $x$ 55
2.11 95% prediction interval for the response of a new patient 56
2.12 Simple linear regression with Stata 57
2.13 Lowess regression 64
2.14 Plotting a lowess regression curve in Stata 65
2.15 Residual analyses 66
2.16 Studentized residual analysis using Stata 69
2.17 Transforming the $x$ and $y$ variables 70
  2.17.1 Stabilizing the variance 70
  2.17.2 Correcting for non-linearity 71
3 Multiple linear regression

3.1 The model

3.2 Confounding variables

3.3 Estimating the parameters for a multiple linear regression model

3.4 $R^2$ statistic for multiple regression models

3.5 Expected response in the multiple regression model

3.6 The accuracy of multiple regression parameter estimates

3.7 Hypothesis tests

3.8 Leverage

3.9 95% confidence interval for $\hat{y}_i$

3.10 95% prediction intervals

3.11 Example: the Framingham Heart Study
    3.11.1 Preliminary univariate analyses

3.12 Scatter plot matrix graphs
    3.12.1 Producing scatter plot matrix graphs with Stata

3.13 Modeling interaction in multiple linear regression
    3.13.1 The Framingham example

3.14 Multiple regression modeling of the Framingham data

3.15 Intuitive understanding of a multiple regression model
    3.15.1 The Framingham example

3.16 Calculating 95% confidence and prediction intervals

3.17 Multiple linear regression with Stata

3.18 Automatic methods of model selection
    3.18.1 Forward selection using Stata
    3.18.2 Backward selection
    3.18.3 Forward stepwise selection
### 3 Backward stepwise selection
- 3.18.4 Backward stepwise selection 123
- 3.18.5 Pros and cons of automated model selection 124

### 3.19 Collinearity

### 3.20 Residual analyses

### 3.21 Influence
- 3.21.1 $\hat{\Delta} \hat{\beta}$ influence statistic 127
- 3.21.2 Cook's distance 127
- 3.21.3 The Framingham example 128

### 3.22 Residual and influence analyses using Stata 129

### 3.23 Using multiple linear regression for non-linear models 133

### 3.24 Building non-linear models with restricted cubic splines 134
- 3.24.1 Choosing the knots for a restricted cubic spline model 137
- 3.25 The SUPPORT Study of hospitalized patients 138
- 3.25.1 Modeling length-of-stay and MAP using restricted cubic splines 138
- 3.25.2 Using Stata for non-linear models with restricted cubic splines 142

### 3.26 Additional reading 154

### 3.27 Exercises 155

### 4 Simple logistic regression 159

- 4.1 Example: APACHE score and mortality in patients with sepsis 159
- 4.2 Sigmoidal family of logistic regression curves 159
- 4.3 The log odds of death given a logistic probability function 161
- 4.4 The binomial distribution 162
- 4.5 Simple logistic regression model 163
- 4.6 Generalized linear model 163
- 4.7 Contrast between logistic and linear regression 164
- 4.8 Maximum likelihood estimation 164
- 4.8.1 Variance of maximum likelihood parameter estimates 165
- 4.9 Statistical tests and confidence intervals 166
- 4.9.1 Likelihood ratio tests 166
- 4.9.2 Quadratic approximations to the log likelihood ratio function 167
- 4.9.3 Score tests 168
Contents

4.9.4 Wald tests and confidence intervals 168
4.9.5 Which test should you use? 169
4.10 Sepsis example 170
4.11 Logistic regression with Stata 171
4.12 Odds ratios and the logistic regression model 174
4.13 95% confidence interval for the odds ratio associated with a unit increase in $x$ 175
4.13.1 Calculating this odds ratio with Stata 175
4.14 Logistic regression with grouped response data 176
4.15 95% confidence interval for $\pi [x]$ 176
4.16 Exact 100(1 − $\alpha$)% confidence intervals for proportions 177
4.17 Example: the Ibuprofen in Sepsis Study 178
4.18 Logistic regression with grouped data using Stata 181
4.19 Simple 2 × 2 case-control studies 187
4.19.1 Example: the Ille-et-Vilaine study of esophageal cancer and alcohol 187
4.19.2 Review of classical case-control theory 188
4.19.3 95% confidence interval for the odds ratio: Woolf’s method 189
4.19.4 Test of the null hypothesis that the odds ratio equals one 190
4.19.5 Test of the null hypothesis that two proportions are equal 190
4.20 Logistic regression models for 2 × 2 contingency tables 191
4.20.1 Nuisance parameters 191
4.20.2 95% confidence interval for the odds ratio: logistic regression 191
4.21 Creating a Stata data file 192
4.22 Analyzing case–control data with Stata 195
4.23 Regressing disease against exposure 197
4.24 Additional reading 198
4.25 Exercises 199

5 Multiple logistic regression 201
5.1 Mantel–Haenszel estimate of an age-adjusted odds ratio 201
5.2 Mantel–Haenszel $\chi^2$ statistic for multiple 2 × 2 tables 203
5.3 95% confidence interval for the age-adjusted odds ratio 204
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>Breslow–Day–Tarone test for homogeneity</td>
<td>204</td>
</tr>
<tr>
<td>5.5</td>
<td>Calculating the Mantel–Haenszel odds ratio using Stata</td>
<td>206</td>
</tr>
<tr>
<td>5.6</td>
<td>Multiple logistic regression model</td>
<td>210</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Likelihood ratio test of the influence of the covariates on the response variable</td>
<td>211</td>
</tr>
<tr>
<td>5.7</td>
<td>95% confidence interval for an adjusted odds ratio</td>
<td>211</td>
</tr>
<tr>
<td>5.8</td>
<td>Logistic regression for multiple $2 \times 2$ contingency tables</td>
<td>212</td>
</tr>
<tr>
<td>5.9</td>
<td>Analyzing multiple $2 \times 2$ tables with Stata</td>
<td>214</td>
</tr>
<tr>
<td>5.10</td>
<td>Handling categorical variables in Stata</td>
<td>216</td>
</tr>
<tr>
<td>5.11</td>
<td>Effect of dose of alcohol on esophageal cancer risk</td>
<td>217</td>
</tr>
<tr>
<td>5.11.1</td>
<td>Analyzing model (5.25) with Stata</td>
<td>219</td>
</tr>
<tr>
<td>5.12</td>
<td>Effect of dose of tobacco on esophageal cancer risk</td>
<td>221</td>
</tr>
<tr>
<td>5.13</td>
<td>Deriving odds ratios from multiple parameters</td>
<td>221</td>
</tr>
<tr>
<td>5.14</td>
<td>The standard error of a weighted sum of regression coefficients</td>
<td>222</td>
</tr>
<tr>
<td>5.15</td>
<td>Confidence intervals for weighted sums of coefficients</td>
<td>222</td>
</tr>
<tr>
<td>5.16</td>
<td>Hypothesis tests for weighted sums of coefficients</td>
<td>223</td>
</tr>
<tr>
<td>5.17</td>
<td>The estimated variance–covariance matrix</td>
<td>223</td>
</tr>
<tr>
<td>5.18</td>
<td>Multiplicative models of two risk factors</td>
<td>224</td>
</tr>
<tr>
<td>5.19</td>
<td>Multiplicative model of smoking, alcohol, and esophageal cancer</td>
<td>225</td>
</tr>
<tr>
<td>5.20</td>
<td>Fitting a multiplicative model with Stata</td>
<td>227</td>
</tr>
<tr>
<td>5.21</td>
<td>Model of two risk factors with interaction</td>
<td>231</td>
</tr>
<tr>
<td>5.22</td>
<td>Model of alcohol, tobacco, and esophageal cancer with interaction terms</td>
<td>233</td>
</tr>
<tr>
<td>5.23</td>
<td>Fitting a model with interaction using Stata</td>
<td>234</td>
</tr>
<tr>
<td>5.24</td>
<td>Model fitting: nested models and model deviance</td>
<td>238</td>
</tr>
<tr>
<td>5.25</td>
<td>Effect modifiers and confounding variables</td>
<td>240</td>
</tr>
<tr>
<td>5.26</td>
<td>Goodness-of-fit tests</td>
<td>240</td>
</tr>
<tr>
<td>5.26.1</td>
<td>The Pearson $\chi^2$ goodness-of-fit statistic</td>
<td>241</td>
</tr>
<tr>
<td>5.27</td>
<td>Hosmer–Lemeshow goodness-of-fit test</td>
<td>242</td>
</tr>
<tr>
<td>5.27.1</td>
<td>An example: the Ille-et-Vilaine cancer data set</td>
<td>242</td>
</tr>
<tr>
<td>5.28</td>
<td>Residual and influence analysis</td>
<td>244</td>
</tr>
<tr>
<td>5.28.1</td>
<td>Standardized Pearson residual</td>
<td>245</td>
</tr>
<tr>
<td>5.28.2</td>
<td>$\Delta \hat{\beta}_j$ influence statistic</td>
<td>245</td>
</tr>
<tr>
<td>5.28.3</td>
<td>Residual plots of the Ille-et-Vilaine data on esophageal cancer</td>
<td>246</td>
</tr>
<tr>
<td>5.29</td>
<td>Using Stata for goodness-of-fit tests and residual analyses</td>
<td>248</td>
</tr>
<tr>
<td>5.30</td>
<td>Frequency matched case–control studies</td>
<td>258</td>
</tr>
<tr>
<td>5.31</td>
<td>Conditional logistic regression</td>
<td>258</td>
</tr>
</tbody>
</table>
5.32 Analyzing data with missing values 258
  5.32.1 Imputing data that is missing at random 259
  5.32.2 Cardiac output in the Ibuprofen in Sepsis Study 260
  5.32.3 Modeling missing values with Stata 263
5.33 Logistic regression using restricted cubic splines 265
  5.33.1 Odds ratios from restricted cubic spline models 266
  5.33.2 95% confidence intervals for $\psi(x)$ 267
5.34 Modeling hospital mortality in the SUPPORT Study 267
5.35 Using Stata for logistic regression with restricted cubic splines 271
5.36 Regression methods with a categorical response variable 278
  5.36.1 Proportional odds logistic regression 278
  5.36.2 Polytomous logistic regression 279
5.37 Additional reading 282
5.38 Exercises 283

6 Introduction to survival analysis 287
  6.1 Survival and cumulative mortality functions 287
  6.2 Right censored data 289
  6.3 Kaplan–Meier survival curves 290
  6.4 An example: genetic risk of recurrent intracerebral hemorrhage 291
  6.5 95% confidence intervals for survival functions 293
  6.6 Cumulative mortality function 295
  6.7 Censoring and bias 296
  6.8 Log-rank test 296
  6.9 Using Stata to derive survival functions and the log-rank test 298
  6.10 Log-rank test for multiple patient groups 305
  6.11 Hazard functions 306
  6.12 Proportional hazards 306
  6.13 Relative risks and hazard ratios 307
  6.14 Proportional hazards regression analysis 309
  6.15 Hazard regression analysis of the intracerebral hemorrhage data 310
  6.16 Proportional hazards regression analysis with Stata 310
  6.17 Tied failure times 311
  6.18 Additional reading 312
  6.19 Exercises 312

© Cambridge University Press
7 Hazard regression analysis

7.1 Proportional hazards model 315
7.2 Relative risks and hazard ratios 315
7.3 95% confidence intervals and hypothesis tests 317
7.4 Nested models and model deviance 317
7.5 An example: the Framingham Heart Study 317
  7.5.1 Kaplan–Meier survival curves for DBP 317
  7.5.2 Simple hazard regression model for CHD risk and DBP 318
  7.5.3 Restricted cubic spline model of CHD risk and DBP 320
  7.5.4 Categorical hazard regression model of CHD risk and DBP 323
  7.5.5 Simple hazard regression model of CHD risk and gender 324
  7.5.6 Multiplicative model of DBP and gender on risk of CHD 325
  7.5.7 Using interaction terms to model the effects of gender and DBP on CHD 326
  7.5.8 Adjusting for confounding variables 327
  7.5.9 Interpretation 329
  7.5.10 Alternative models 330
7.6 Proportional hazards regression analysis using Stata 331
7.7 Stratified proportional hazards models 348
7.8 Survival analysis with ragged study entry 349
  7.8.1 Kaplan–Meier survival curve and the log-rank test with ragged entry 350
  7.8.2 Age, sex, and CHD in the Framingham Heart Study 350
  7.8.3 Proportional hazards regression analysis with ragged entry 351
  7.8.4 Survival analysis with ragged entry using Stata 351
7.9 Predicted survival, log–log plots, and the proportional hazards assumption 354
  7.9.1 Evaluating the proportional hazards assumption with Stata 357
7.10 Hazard regression models with time-dependent covariates 359
  7.10.1 Testing the proportional hazards assumption 361
8

Introduction to Poisson regression: inferences on morbidity and mortality rates

8.1 Elementary statistics involving rates
8.2 Calculating relative risks from incidence data using Stata
8.3 The binomial and Poisson distributions
8.4 Simple Poisson regression for $2 \times 2$ tables
8.5 Poisson regression and the generalized linear model
8.6 Contrast between Poisson, logistic, and linear regression
8.7 Simple Poisson regression with Stata
8.8 Poisson regression and survival analysis
  8.8.1 Recoding survival data on patients as patient–year data
  8.8.2 Converting survival records to person–years of follow-up using Stata
8.9 Converting the Framingham survival data set to person–time data
8.10 Simple Poisson regression with multiple data records
8.11 Poisson regression with a classification variable
8.12 Applying simple Poisson regression to the Framingham data
8.13 Additional reading
8.14 Exercises

9

Multiple Poisson regression

9.1 Multiple Poisson regression model
9.2 An example: the Framingham Heart Study
  9.2.1 A multiplicative model of gender, age, and coronary heart disease
  9.2.2 A model of age, gender, and CHD with interaction terms
  9.2.3 Adding confounding variables to the model
9.3 Using Stata to perform Poisson regression
9.4 Residual analyses for Poisson regression models
  9.4.1 Deviance residuals
<table>
<thead>
<tr>
<th>Contents</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9.5 Residual analysis of Poisson regression models using Stata</td>
<td>424</td>
</tr>
<tr>
<td>9.6 Additional reading</td>
<td>427</td>
</tr>
<tr>
<td>9.7 Exercises</td>
<td>427</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>10</strong> Fixed effects analysis of variance</td>
<td>429</td>
</tr>
<tr>
<td>10.1 One-way analysis of variance</td>
<td>429</td>
</tr>
<tr>
<td>10.2 Multiple comparisons</td>
<td>431</td>
</tr>
<tr>
<td>10.3 Reformulating analysis of variance as a linear regression model</td>
<td>433</td>
</tr>
<tr>
<td>10.4 Non-parametric methods</td>
<td>434</td>
</tr>
<tr>
<td>10.5 Kruskal–Wallis test</td>
<td>435</td>
</tr>
<tr>
<td>10.6 Example: a polymorphism in the estrogen receptor gene</td>
<td>435</td>
</tr>
<tr>
<td>10.7 User contributed software in Stata</td>
<td>438</td>
</tr>
<tr>
<td>10.8 One-way analyses of variance using Stata</td>
<td>439</td>
</tr>
<tr>
<td>10.9 Two-way analysis of variance, analysis of covariance, and other models</td>
<td>446</td>
</tr>
<tr>
<td>10.10 Additional reading</td>
<td>448</td>
</tr>
<tr>
<td>10.11 Exercises</td>
<td>448</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>11</strong> Repeated-measures analysis of variance</td>
<td>451</td>
</tr>
<tr>
<td>11.1 Example: effect of race and dose of isoproterenol on blood flow</td>
<td>451</td>
</tr>
<tr>
<td>11.2 Exploratory analysis of repeated measures data using Stata</td>
<td>453</td>
</tr>
<tr>
<td>11.3 Response feature analysis</td>
<td>459</td>
</tr>
<tr>
<td>11.4 Example: the isoproterenol data set</td>
<td>460</td>
</tr>
<tr>
<td>11.5 Response feature analysis using Stata</td>
<td>463</td>
</tr>
<tr>
<td>11.6 The area-under-the-curve response feature</td>
<td>468</td>
</tr>
<tr>
<td>11.7 Generalized estimating equations</td>
<td>470</td>
</tr>
<tr>
<td>11.8 Common correlation structures</td>
<td>470</td>
</tr>
<tr>
<td>11.9 GEE analysis and the Huber–White sandwich estimator</td>
<td>472</td>
</tr>
<tr>
<td>11.10 Example: analyzing the isoproterenol data with GEE</td>
<td>473</td>
</tr>
<tr>
<td>11.11 Using Stata to analyze the isoproterenol data set using GEE</td>
<td>476</td>
</tr>
<tr>
<td>11.12 GEE analyses with logistic or Poisson models</td>
<td>481</td>
</tr>
<tr>
<td>11.13 Additional reading</td>
<td>481</td>
</tr>
<tr>
<td>11.14 Exercises</td>
<td>482</td>
</tr>
</tbody>
</table>
Appendices

A Summary of statistical models discussed in this text

A.1 Models for continuous response variables with one response per patient 485
A.2 Models for dichotomous or categorical response variables with one response per patient 486
A.3 Models for survival data (follow-up time plus fate at exit observed on each patient) 487
A.4 Models for response variables that are event rates or the number of events during a specified number of patient–years of follow-up. The event must be rare 489
A.5 Models with multiple observations per patient or matched or clustered patients 489

B Summary of Stata commands used in this text

B.1 Data manipulation and description 491
B.2 Analysis commands 493
B.3 Graph commands 497
B.4 Common options for graph commands (insert after comma) 500
B.5 Post-estimation commands (affected by preceding regression-type command) 502
B.6 Command prefixes 504
B.7 Command qualifiers (insert before comma) 504
B.8 Logical and relational operators and system variables (see Stata User’s Guide) 505
B.9 Functions (see Stata Data Management Manual) 506

References 507
Index 513
Preface

The purpose of this text is to enable biomedical researchers to use a number of advanced statistical methods that have proven valuable in medical research. The past forty years have seen an explosive growth in the development of biostatistics. As with so many aspects of our world, this growth has been strongly influenced by the development of inexpensive, powerful computers and the sophisticated software that has been written to run them. This has allowed the development of computationally intensive methods that can effectively model complex biomedical data sets. It has also made it easy to explore these data sets, to discover how variables are interrelated, and to select appropriate statistical models for analysis. Indeed, just as the microscope revealed new worlds to the eighteenth century, modern statistical software permits us to see interrelationships in large complex data sets that would have been missed in previous eras. Also, modern statistical software has made it vastly easier for investigators to perform their own statistical analyses. Although very sophisticated mathematics underlies modern statistics, it is not necessary to understand this mathematics to properly analyze your data with modern statistical software. What is necessary is to understand the assumptions required by each method, how to determine whether these assumptions are adequately met for your data, how to select the best model, and how to interpret the results of your analyses. The goal of this text is to allow investigators to effectively use some of the most valuable multivariate methods without requiring a prior understanding of more than high school algebra. Much mathematical detail is avoided by focusing on the use of a specific statistical software package.

This text grew out of my second semester course in biostatistics that I teach in our Master of Public Health program at the Vanderbilt University Medical School. All of the students take introductory courses in biostatistics and epidemiology prior to mine. Although this text is self-contained, I strongly recommend that readers acquire good introductory texts in biostatistics and epidemiology as companions to this one. Many excellent texts are available on these topics. At Vanderbilt we are currently using Katz (2006) for biostatistics and Gordis (2004) for epidemiology. The statistical
software used in this text is Stata, version 10 (StataCorp, 2007). It was chosen for the breadth and depth of its statistical methods, for its ease of use, excellent graphics and excellent documentation. There are several other excellent packages available on the market. However, the aim of this text is to teach biostatistics through a specific software package, and length restrictions make it impractical to use more than one package. If you have not yet invested a lot of time learning a different package, Stata is an excellent choice for you to consider. If you are already attached to a different package, you may still find it easier to learn Stata than to master or teach the material covered here from other textbooks. The topics covered in this text are linear regression, logistic regression, Poisson regression, survival analysis, and analysis of variance. Each topic is covered in two chapters: one introduces the topic with simple univariate examples and the other covers more complex multivariate models. The text makes extensive use of a number of real data sets. They all may be downloaded from my web site at biostat.mc.vanderbilt.edu/dupontwd/wddtext/. This site also contains complete log files of all analyses discussed in this text.

Changes in the second edition

I have made extensive modifications and additions to the second edition of this text. These can be summarized as follows.

• Since I wrote the first edition, Stata has undergone major improvements that make it much easier to use and enable more powerful graphics. The examples in this text take advantage of these improvements and comply with Stata’s version 10 syntax.

• Stata now has easy-to-use point-and-click commands that may be used as an alternative to Stata’s character-based commands. I have provided documentation for both the point-and-click and character-based versions of all commands discussed in this text.

• Appendix A summarizes the types of data discussed in this text and indicates which statistical methods are most appropriate for each type of data.

• Restricted cubic splines are used to analyze non-linear regression models. This is a simple but powerful approach that can be used to extend logistic and proportional hazards regression models as well as linear regression models.

• Density-distribution sunflower plots are used for the exploratory analysis of dense bivariate data.
The Breslow–Day–Tarone test is used to test the equality of odds ratios across multiple $2 \times 2$ tables.

Likelihood ratio tests of nested models are used extensively.

I have added a brief discussion of proportional odds and polytomous logistic regression.

Predicted survival and log–log plots are used to evaluate the adequacy of the proportional hazards model of survival data.

Additional exercises have been added to several chapters.

Acknowledgements

I would like to thank Gordon R. Bernard, Jeffrey Brent, Norman E. Breslow, Graeme Eisenhofer, Cary P. Gross, Frank E. Harrell, Daniel Levy, Steven M. Greenberg, Fritz F. Parl, Paul Sorlie, Wayne A. Ray, and Alastair J. J. Wood for allowing me to use their data to illustrate the methods described in this text. I am grateful to William Gould and the employees of Stata Corporation for publishing their elegant and powerful statistical software and for providing excellent documentation. I would also like to thank the students in our Master of Public Health program who have taken my course. Their energy, intelligence and enthusiasm have greatly enhanced my enjoyment in preparing this material. Their criticisms and suggestions have profoundly influenced this work. I am grateful to David L. Page, my friend and colleague of 31 years, with whom I have learnt much about the art of teaching epidemiology and biostatistics to clinicians. My appreciation goes to Sarah K. Meredith for introducing me to Cambridge University Press; to William Schaffner and Frank E. Harrell, my chairmen during the writing of the first and second editions, respectively, who enabled my spending the time needed to complete this work; to W. Dale Plummer for programing and technical support with Stata and \texttt{\LaTeX}; to Nicholas J. Cox for proof-reading this text, for his valuable advice, and for writing the \texttt{stripplot} program; to William R. Rising, Patrick G. Arbogast, and Gregory D. Ayers for proof-reading this book and for their valuable suggestions; to Jeffrey S. Pitblado for writing the Stata 8 version of the \texttt{sunflower} program and for allowing me to adapt his \texttt{\LaTeX} style files for use in this book; to Kristin MacDonald for writing the restricted cubic spline module of the \texttt{mkspline} program; to Tebeb Gebretsadik and Knut M. Wittkowski for their helpful suggestions; to Frances Nex for her careful copy editing, to Charlotte Broom, Laura Wood, Richard Marley, and other colleagues at Cambridge University Press, and
Jagdamba Prasad at Aptara Corporation, for producing this beautiful book; and to my mother and sisters for their support during six critical months of this project. Finally, I am especially grateful to my wife, Susan, and sons, Thomas and Peter, for their love and support, and for their cheerful tolerance of the countless hours that I spent on this project.

**Disclaimer:** The opinions expressed in this text are my own and do not necessarily reflect those of the authors acknowledged in this preface, their employers or funding institutions. This includes the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, USA.