Preface

There are many books that are excellent sources of knowledge about individual statistical tools (survival models, general linear models, etc.), but the art of data analysis is about choosing and using multiple tools. In the words of Chatfield [69, p. 420] "...students typically know the technical details of regression for example, but not necessarily when and how to apply it. This argues the need for a better balance in the literature and in statistical teaching between *techniques* and problem solving *strategies*." Whether analyzing risk factors, adjusting for biases in observational studies, or developing predictive models, there are common problems that few regression texts address. For example, there are missing data in the majority of datasets one is likely to encounter (other than those used in textbooks!) but most regression texts do not include methods for dealing with such data effectively, and texts on missing data do not cover regression modeling.

This book links standard regression modeling approaches with

- methods for relaxing linearity assumptions that still allow one to easily obtain predictions and confidence limits for future observations, and to do formal hypothesis tests,
- nonadditive modeling approaches not requiring the assumption that interactions are always linear $\times$ linear,
- methods for imputing missing data and for penalizing variances for incomplete data,

- methods for handling large numbers of predictors without resorting to problematic stepwise variable selection techniques,
- data reduction methods (some of which are based on multivariate psychometric techniques too seldom used in statistics) that help with the problem of "too many variables to analyze and not enough observations" as well as making the model more interpretable when there are predictor variables containing overlapping information,
- methods for quantifying predictive accuracy of a fitted model,
- powerful model validation techniques based on the bootstrap, that allow the analyst to estimate predictive accuracy nearly unbiasedly without holding back data from the model development process, and
- graphical methods for understanding complex models.

On the last point, this text has special emphasis on what could be called "presentation graphics for fitted models" to help make regression analyses more palatable to nonstatisticians. For example, nomograms have long been used to make equations portable, but they are not drawn routinely because doing so is very labor intensive. An S-PLUS function called `nomogram` in the library described below draws nomograms from a regression fit, and these diagrams can be used to communicate modeling results as well as to obtain predicted values manually even in the presence of complex variable transformations.

Most of the methods in this text apply to all regression models, but special emphasis is given to some of the most popular ones: multiple regression using least squares, the binary logistic model, two logistic models for ordinal responses, parametric survival regression models, and the Cox semiparametric survival model. There is also a chapter on nonparametric transform-both-sides regression. Emphasis is given to detailed case studies for these methods as well as for data reduction, imputation, model simplification, and other tasks. The majority of examples are from biomedical research. However, the methods presented here have broad application to other areas including economics, epidemiology, sociology, psychology, engineering, and predicting consumer behavior and other business outcomes.

This text is intended for Masters or PhD level graduate students who have had a general introductory probability and statistics course and who are well versed in ordinary multiple regression and intermediate algebra. The book is also intended to serve as a reference for data analysts and statistical methodologists. Readers without a strong background in applied statistics may wish to first study one of the many introductory applied statistics and regression texts that are available; Katz' small book on multivariable analysis[232] is especially helpful to clinicians and epidemiologists. The paper by Nick and Hardin[325] also provides a good introduction to multivariable modeling and interpretation.

The overall philosophy of this book is summarized by the following statements.

- Satisfaction of model assumptions improves precision and increases statistical power.
- It is more productive to make a model fit step by step (e.g., transformation estimation) than to postulate a simple model and find out what went wrong.
- Graphical methods should be married to formal inference.
- Overfitting occurs frequently, so data reduction and model validation are important.
- In most research projects the cost of data collection far outweighs the cost of data analysis, so it is important to use the most efficient and accurate modeling techniques, to avoid categorizing continuous variables, and to not remove data from the estimation sample just to be able to validate the model.
- The bootstrap is a breakthrough for statistical modeling, and the analyst should use it for many steps of the modeling strategy, including derivation of distribution-free confidence intervals and estimation of optimism in model fit that takes into account variations caused by the modeling strategy.
- Imputation of missing data is better than discarding incomplete observations.
- Variance often dominates bias, so biased methods such as penalized maximum likelihood estimation yield models that have a greater chance of accurately predicting future observations.
- Carefully fitting an improper model is better than badly fitting (and overfitting) a well-chosen one.
- Methods that work for all types of regression models are the most valuable.
- Using the data to guide the data analysis is almost as dangerous as not doing so.
- There are benefits to modeling by deciding how many degrees of freedom (i.e., number of regression parameters) can be "spent," deciding where they should be spent, and then spending them.

On the last point, the author believes that significance tests and P-values are problematic, especially when making modeling decisions. Judging by the increased emphasis on confidence intervals in scientific journals there is reason to believe that hypothesis testing is gradually being deemphasized. Yet the reader will notice that this text contains many P-values. How does that make sense when, for example, the text recommends against simplifying a model when a test of linearity is not significant? First, some readers may wish to emphasize hypothesis testing in general, and some hypotheses have special interest, such as in pharmacology where one may

be interested in whether the effect of a drug is linear in log dose. Second, many of the more interesting hypothesis tests in the text are tests of complexity (nonlinearity, interaction) of the overall model. Null hypotheses of linearity of effects in particular are frequently rejected, providing formal evidence that the analyst's investment of time to use more than simple statistical models was warranted.

The text emphasizes predictive modeling, but as discussed in Chapter 1, developing good predictions goes hand in hand with accurate estimation of effects and with hypothesis testing (when appropriate). Besides emphasis on multivariable modeling, the text includes a chapter (16) introducing survival analysis and methods for analyzing various types of single and multiple events. This book does not provide examples of analyses of one common type of response variable, namely, cost and related measures of resource consumption. However, least squares modeling presented in Chapter 7, the robust rank-based methods presented in Chapters 13 and 19, and the transform-both-sides regression models discussed in Chapter 15 are very applicable and robust for modeling economic outcomes. See [120] and [177] for example analyses of such dependent variables using, respectively, the Cox model and nonparametric additive regression. The central Web site for this book (see the Appendix) has much more material on the use of the Cox model for analyzing costs.

Heavy use is made of the S-Plus statistical software environment from Insightful Corporation (Seattle, Washington). A few small examples using SAS (SAS Institute, Inc., Cary, North Carolina) are also described. S-Plus is the focus because it is an elegant object-oriented system in which it is easy to implement new statistical ideas. Many S-Plus users around the world have done so, and their work has benefitted many of the procedures described here. S-Plus also has a uniform syntax for specifying statistical models (with respect to categorical predictors, interactions, etc.), no matter which type of model is being fitted.[65]

A free, open-source statistical software system called R has become available in the past few years. The R language is similar to the S language on which S-Plus is based. Most of the functions used in this text are expected to be adapted to the R system. See the book's Web site for updated information about software availability.

Readers who don't use S-Plus, R, or any other statistical software environment will still find the statistical methods and case studies in this text useful, and it is hoped that the code that is presented will make the statistical methods more concrete. At the very least, the code demonstrates that all of the methods presented in the text are feasible.

This text does not teach analysts how to use S-Plus or R. For that, the reader may wish to consult Venables and Ripley[434] (which is an excellent companion to this text) as well as texts by Spector,[395] Krause and Olson,[250] and others, along with S-Plus manuals.[308] A free resource is a book by Alzola and Harrell[15] available on this text's Web site. That document teaches general S-Plus concepts as well as how to use add-on libraries described below. The document is also useful for SAS users who are new to S-Plus. See the Appendix for more information.

In addition to powerful features that are built into S-PLUS, this text uses a library of freely available S-PLUS functions called `Design` written by the author. `Design`, so named because of its tracking of modeling details related to the expanded X or design matrix, is a series of over 200 functions for model fitting, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. `Design` includes functions for least squares and penalized least squares multiple regression modeling in addition to functions for binary and ordinal logistic regression and survival analysis that are emphasized in this text. Other freely available miscellaneous S-PLUS functions used in the text are found in the `Hmisc` library also written by the author. Functions in `Hmisc` include facilities for data reduction, imputation, power and sample size calculation, advanced table making, recoding variables, importing and inspecting data, and general graphics. Consult the Appendix for information on obtaining `Hmisc` and `Design`.

The author and his colleagues have written SAS macros for fitting restricted cubic splines and for other basic operations. See the Appendix for more information as well as notes on using SAS procedures for many of the models discussed in this text. It is unfair not to mention some excellent capabilities of other statistical packages such as Stata and SYSTAT, but the extendibility and graphics of S-PLUS and R make them especially attractive for all aspects of the comprehensive modeling strategy presented in this book.

Portions of Chapters 4 and 19 were published as reference [185]. Some of Chapter 13 was published as reference [188].

The author may be contacted by electronic mail at `fharrell@virginia.edu` and would appreciate being informed of unclear points, errors, and omissions in this book. Suggestions for improvements and for future topics are also welcome. As described in the Web site, instructors may contact the author to obtain copies of quizzes and extra assignments (both with answers) related to much of the material in the earlier chapters, and to obtain full solutions (with graphical output) to the majority of assignments in the text.

Acknowledgements

A good deal of the writing of this book was done during my 17 years on the faculty of Duke University. I wish to thank my close colleague Kerry Lee for providing many valuable ideas, fruitful collaborations, and well-organized lecture notes from which I have greatly benefited over the past years. Terry Therneau of Mayo Clinic has given me many of his wonderful ideas for many years, and has written state-of-the-art S-PLUS software for survival analysis that forms the core of survival analysis software in my `Design` library. Michael Symons of the Department of Biostatistics of the University of North Carolina at Chapel Hill and Timothy Morgan of the Biometry Department at Wake Forest University School of Medicine also provided

course materials, some of which motivated portions of this text. My former clinical colleagues in the Cardiology Division at Duke University, Robert Califf, Phillip Harris, Mark Hlatky, Dan Mark, David Pryor, and Robert Rosati for many years provided valuable motivation, feedback, and ideas through our interaction on clinical problems. Besides Kerry Lee, statistical colleagues L. Richard Smith, Lawrence Muhlbaier, and Elizabeth DeLong clarified my thinking and gave me new ideas on numerous occasions. Charlotte Nelson and Carlos Alzola have frequently helped me debug S-PLUS routines when they thought they were just analyzing data.

Former students Bercedis Peterson, James Herndon, Robert McMahon, and Yuan-Li Shen have provided many insights into logistic and survival modeling. Associations with Doug Wagner and William Knaus of the University of Virginia, Ken Offord of Mayo Clinic, David Naftel of the University of Alabama in Birmingham, and Phil Miller of Washington University have provided many valuable ideas and motivations for this work, as have Michael Schemper of Vienna University, Janez Stare of Ljubljana University, Slovenia, and Ewout Steyerberg of Erasmus University, Rotterdam. Richard Goldstein of Qualitas, Inc., along with several anonymous reviewers, provided many helpful criticisms of a previous version of this manuscript that resulted in significant improvements, and critical reading by Bob Edson (VA Cooperative Studies Program, Palo Alto) resulted in many error corrections. Thanks to Brian Ripley of the University of Oxford for providing many helpful software tools and statistical insights that greatly aided in the production of this book, and to Bill Venables of CSIRO Australia for wisdom, both statistical and otherwise. Thanks also to John Kimmel of Springer-Verlag whose ideas and encouragement have been invaluable. This work would also not have been possible without the S environment developed by Rick Becker, John Chambers, Allan Wilks, and many other researchers at Lucent Technologies and several universities, or without the S-PLUS environment developed by many programmers and researchers at Insightful Corporation.

This work was supported by grants and contracts from the following U.S. agencies: Agency for Healthcare Research and Quality; National Library of Medicine; National Heart, Lung and Blood Institute and the National Center for Research Resources of the National Institutes of Health; National Cancer Institute; Robert Wood Johnson Foundation; National Center for Health Statistics; Roche Pharmaceuticals; the Biometry Training Program, Duke University; and the Department of Health Evaluation Sciences, University of Virginia School of Medicine.

Frank E. Harrell, Jr.
University of Virginia
April 2001

Contents

Preface **vii**

Typographical Conventions **xxiii**

1 Introduction **1**

1.1 Hypothesis Testing, Estimation, and Prediction 1

1.2 Examples of Uses of Predictive Multivariable Modeling 3

1.3 Planning for Modeling 4

1.3.1 Emphasizing Continuous Variables 6

1.4 Choice of the Model 6

1.5 Further Reading 8

2 General Aspects of Fitting Regression Models **11**

2.1 Notation for Multivariable Regression Models 11

2.2 Model Formulations 12

2.3 Interpreting Model Parameters 13

2.3.1 Nominal Predictors 14

2.3.2 Interactions 14

2.3.3 Example: Inference for a Simple Model . . . 15
2.4 Relaxing Linearity Assumption for Continuous Predictors . . . 16
2.4.1 Simple Nonlinear Terms . . . 16
2.4.2 Splines for Estimating Shape of Regression Function and Determining Predictor Transformations . . . 18
2.4.3 Cubic Spline Functions . . . 19
2.4.4 Restricted Cubic Splines . . . 20
2.4.5 Choosing Number and Position of Knots . . . 23
2.4.6 Nonparametric Regression . . . 24
2.4.7 Advantages of Regression Splines over Other Methods . . . 26
2.5 Recursive Partitioning: Tree-Based Models . . . 26
2.6 Multiple Degree of Freedom Tests of Association . . . 27
2.7 Assessment of Model Fit . . . 29
2.7.1 Regression Assumptions . . . 29
2.7.2 Modeling and Testing Complex Interactions . . . 32
2.7.3 Fitting Ordinal Predictors . . . 34
2.7.4 Distributional Assumptions . . . 35
2.8 Further Reading . . . 36
2.9 Problems . . . 37

3 Missing Data 41
3.1 Types of Missing Data . . . 41
3.2 Prelude to Modeling . . . 42
3.3 Missing Values for Different Types of Response Variables . . . 43
3.4 Problems with Simple Alternatives to Imputation . . . 43
3.5 Strategies for Developing Imputation Algorithms . . . 44
3.6 Single Conditional Mean Imputation . . . 47
3.7 Multiple Imputation . . . 47
3.8 Summary and Rough Guidelines . . . 48
3.9 Further Reading . . . 50
3.10 Problems . . . 51

4 Multivariable Modeling Strategies 53
4.1 Prespecification of Predictor Complexity Without Later Simplification . . . 53

4.2 Checking Assumptions of Multiple Predictors Simultaneously . . . 56
4.3 Variable Selection . . . 56
4.4 Overfitting and Limits on Number of Predictors . . . 60
4.5 Shrinkage . . . 61
4.6 Collinearity . . . 64
4.7 Data Reduction . . . 66
4.7.1 Variable Clustering . . . 66
4.7.2 Transformation and Scaling Variables Without Using Y . . . 67
4.7.3 Simultaneous Transformation and Imputation . . . 69
4.7.4 Simple Scoring of Variable Clusters . . . 70
4.7.5 Simplifying Cluster Scores . . . 72
4.7.6 How Much Data Reduction Is Necessary? . . . 73
4.8 Overly Influential Observations . . . 74
4.9 Comparing Two Models . . . 77
4.10 Summary: Possible Modeling Strategies . . . 79
4.10.1 Developing Predictive Models . . . 79
4.10.2 Developing Models for Effect Estimation . . . 82
4.10.3 Developing Models for Hypothesis Testing . . . 83
4.11 Further Reading . . . 84

5 Resampling, Validating, Describing, and Simplifying the Model 87
5.1 The Bootstrap . . . 87
5.2 Model Validation . . . 90
5.2.1 Introduction . . . 90
5.2.2 Which Quantities Should Be Used in Validation? . . . 91
5.2.3 Data-Splitting . . . 91
5.2.4 Improvements on Data-Splitting: Resampling . . . 93
5.2.5 Validation Using the Bootstrap . . . 94
5.3 Describing the Fitted Model . . . 97
5.4 Simplifying the Final Model by Approximating It . . . 98
5.4.1 Difficulties Using Full Models . . . 98
5.4.2 Approximating the Full Model . . . 99
5.5 Further Reading . . . 101

6 S-Plus Software 105

6.1 The S Modeling Language . . . 106
6.2 User-Contributed Functions . . . 107
6.3 The `Design` Library . . . 108
6.4 Other Functions . . . 119
6.5 Further Reading . . . 120

7 Case Study in Least Squares Fitting and Interpretation of a Linear Model 121
7.1 Descriptive Statistics . . . 122
7.2 Spending Degrees of Freedom/Specifying Predictor Complexity . . . 127
7.3 Fitting the Model Using Least Squares . . . 128
7.4 Checking Distributional Assumptions . . . 131
7.5 Checking Goodness of Fit . . . 135
7.6 Overly Influential Observations . . . 135
7.7 Test Statistics and Partial R^2 . . . 136
7.8 Interpreting the Model . . . 137
7.9 Problems . . . 142

8 Case Study in Imputation and Data Reduction 147
8.1 Data . . . 147
8.2 How Many Parameters Can Be Estimated? . . . 150
8.3 Variable Clustering . . . 151
8.4 Single Imputation Using Constants or Recursive Partitioning . . . 154
8.5 Transformation and Single Imputation Using `transcan` . . . 157
8.6 Data Reduction Using Principal Components . . . 160
8.7 Detailed Examination of Individual Transformations . . . 168
8.8 Examination of Variable Clusters on Transformed Variables . . . 169
8.9 Transformation Using Nonparametric Smoothers . . . 170
8.10 Multiple Imputation . . . 172
8.11 Further Reading . . . 175
8.12 Problems . . . 176

9 Overview of Maximum Likelihood Estimation 179
9.1 General Notions—Simple Cases . . . 179
9.2 Hypothesis Tests . . . 183

9.2.1 Likelihood Ratio Test . . . 183
9.2.2 Wald Test . . . 184
9.2.3 Score Test . . . 184
9.2.4 Normal Distribution—One Sample . . . 185
9.3 General Case . . . 186
9.3.1 Global Test Statistics . . . 187
9.3.2 Testing a Subset of the Parameters . . . 187
9.3.3 Which Test Statistics to Use When . . . 189
9.3.4 Example: Binomial—Comparing Two Proportions . . . 190
9.4 Iterative ML Estimation . . . 192
9.5 Robust Estimation of the Covariance Matrix . . . 193
9.6 Wald, Score, and Likelihood-Based Confidence Intervals . . . 194
9.7 Bootstrap Confidence Regions . . . 195
9.8 Further Use of the Log Likelihood . . . 202
9.8.1 Rating Two Models, Penalizing for Complexity . . . 202
9.8.2 Testing Whether One Model Is Better than Another . . . 203
9.8.3 Unitless Index of Predictive Ability . . . 203
9.8.4 Unitless Index of Adequacy of a Subset of Predictors . . . 205
9.9 Weighted Maximum Likelihood Estimation . . . 206
9.10 Penalized Maximum Likelihood Estimation . . . 207
9.11 Further Reading . . . 210
9.12 Problems . . . 212

10 Binary Logistic Regression 215
10.1 Model . . . 215
10.1.1 Model Assumptions and Interpretation of Parameters . . . 217
10.1.2 Odds Ratio, Risk Ratio, and Risk Difference . . . 220
10.1.3 Detailed Example . . . 221
10.1.4 Design Formulations . . . 227
10.2 Estimation . . . 228
10.2.1 Maximum Likelihood Estimates . . . 228
10.2.2 Estimation of Odds Ratios and Probabilities . . . 228
10.3 Test Statistics . . . 229
10.4 Residuals . . . 230

10.5 Assessment of Model Fit . . . 230
10.6 Collinearity . . . 244
10.7 Overly Influential Observations . . . 245
10.8 Quantifying Predictive Ability . . . 247
10.9 Validating the Fitted Model . . . 249
10.10 Describing the Fitted Model . . . 253
10.11 S-PLUS Functions . . . 257
10.12 Further Reading . . . 264
10.13 Problems . . . 265

11 Logistic Model Case Study 1: Predicting Cause of Death 269
11.1 Preparation for Modeling . . . 269
11.2 Regression on Principal Components, Cluster Scores, and Pretransformations . . . 271
11.3 Fit and Diagnostics for a Full Model, and Interpreting Pretransformations . . . 276
11.4 Describing Results Using a Reduced Model . . . 285
11.5 Approximating the Full Model Using Recursive Partitioning . . . 291
11.6 Validating the Reduced Model . . . 294

12 Logistic Model Case Study 2: Survival of Titanic Passengers 299
12.1 Descriptive Statistics . . . 300
12.2 Exploring Trends with Nonparametric Regression . . . 305
12.3 Binary Logistic Model With Casewise Deletion of Missing Values . . . 305
12.4 Examining Missing Data Patterns . . . 312
12.5 Single Conditional Mean Imputation . . . 316
12.6 Multiple Imputation . . . 320
12.7 Summarizing the Fitted Model . . . 322
12.8 Problems . . . 326

13 Ordinal Logistic Regression 331
13.1 Background . . . 331
13.2 Ordinality Assumption . . . 332
13.3 Proportional Odds Model . . . 333
13.3.1 Model . . . 333
13.3.2 Assumptions and Interpretation of Parameters . . . 333

13.3.3 Estimation . . . 334
13.3.4 Residuals . . . 334
13.3.5 Assessment of Model Fit . . . 335
13.3.6 Quantifying Predictive Ability . . . 335
13.3.7 Validating the Fitted Model . . . 337
13.3.8 S-PLUS Functions . . . 337
13.4 Continuation Ratio Model . . . 338
13.4.1 Model . . . 338
13.4.2 Assumptions and Interpretation of Parameters . . . 338
13.4.3 Estimation . . . 339
13.4.4 Residuals . . . 339
13.4.5 Assessment of Model Fit . . . 339
13.4.6 Extended CR Model . . . 339
13.4.7 Role of Penalization in Extended CR Model . . . 340
13.4.8 Validating the Fitted Model . . . 341
13.4.9 S-PLUS Functions . . . 341
13.5 Further Reading . . . 342
13.6 Problems . . . 342

14 Case Study in Ordinal Regression, Data Reduction, and Penalization 345

14.1 Response Variable . . . 346
14.2 Variable Clustering . . . 347
14.3 Developing Cluster Summary Scores . . . 349
14.4 Assessing Ordinality of Y for each X, and Unadjusted Checking of PO and CR Assumptions . . . 351
14.5 A Tentative Full Proportional Odds Model . . . 352
14.6 Residual Plots . . . 355
14.7 Graphical Assessment of Fit of CR Model . . . 357
14.8 Extended Continuation Ratio Model . . . 357
14.9 Penalized Estimation . . . 359
14.10 Using Approximations to Simplify the Model . . . 364
14.11 Validating the Model . . . 367
14.12 Summary . . . 369
14.13 Further Reading . . . 371

14.14 Problems . . . 371

15 Models Using Nonparametric Transformations of X and Y 375

15.1 Background . . . 375
15.2 Generalized Additive Models . . . 376
15.3 Nonparametric Estimation of Y-Transformation . . . 376
15.4 Obtaining Estimates on the Original Scale . . . 377
15.5 S-PLUS Functions . . . 378
15.6 Case Study . . . 379

16 Introduction to Survival Analysis 389

16.1 Background . . . 389
16.2 Censoring, Delayed Entry, and Truncation . . . 391
16.3 Notation, Survival, and Hazard Functions . . . 392
16.4 Homogeneous Failure Time Distributions . . . 398
16.5 Nonparametric Estimation of S and Λ . . . 400
16.5.1 Kaplan–Meier Estimator . . . 400
16.5.2 Altschuler–Nelson Estimator . . . 403
16.6 Analysis of Multiple Endpoints . . . 404
16.6.1 Competing Risks . . . 404
16.6.2 Competing Dependent Risks . . . 405
16.6.3 State Transitions and Multiple Types of Nonfatal Events . . . 406
16.6.4 Joint Analysis of Time and Severity of an Event . . . 407
16.6.5 Analysis of Multiple Events . . . 407
16.7 S-PLUS Functions . . . 408
16.8 Further Reading . . . 410
16.9 Problems . . . 411

17 Parametric Survival Models 413

17.1 Homogeneous Models (No Predictors) . . . 413
17.1.1 Specific Models . . . 413
17.1.2 Estimation . . . 414
17.1.3 Assessment of Model Fit . . . 416
17.2 Parametric Proportional Hazards Models . . . 417
17.2.1 Model . . . 417

17.2.2 Model Assumptions and Interpretation of Parameters 418
17.2.3 Hazard Ratio, Risk Ratio, and Risk Difference 419
17.2.4 Specific Models 421
17.2.5 Estimation 422
17.2.6 Assessment of Model Fit 423
17.3 Accelerated Failure Time Models 426
17.3.1 Model 426
17.3.2 Model Assumptions and Interpretation of Parameters 427
17.3.3 Specific Models 427
17.3.4 Estimation 428
17.3.5 Residuals 429
17.3.6 Assessment of Model Fit 430
17.3.7 Validating the Fitted Model 434
17.4 Buckley–James Regression Model 435
17.5 Design Formulations 435
17.6 Test Statistics 435
17.7 Quantifying Predictive Ability 436
17.8 S-Plus Functions 436
17.9 Further Reading 441
17.10 Problems 441

18 Case Study in Parametric Survival Modeling and Model Approximation 443
18.1 Descriptive Statistics 443
18.2 Checking Adequacy of Log-Normal Accelerated Failure Time Model 448
18.3 Summarizing the Fitted Model 454
18.4 Internal Validation of the Fitted Model Using the Bootstrap 454
18.5 Approximating the Full Model 458
18.6 Problems 464

19 Cox Proportional Hazards Regression Model 465
19.1 Model 465
19.1.1 Preliminaries 465
19.1.2 Model Definition 466
19.1.3 Estimation of β 466

19.1.4 Model Assumptions and Interpretation of Parameters 468
19.1.5 Example 468
19.1.6 Design Formulations 470
19.1.7 Extending the Model by Stratification 470
19.2 Estimation of Survival Probability and Secondary Parameters 472
19.3 Test Statistics 474
19.4 Residuals 476
19.5 Assessment of Model Fit 476
19.5.1 Regression Assumptions 477
19.5.2 Proportional Hazards Assumption 483
19.6 What to Do When PH Fails 489
19.7 Collinearity 491
19.8 Overly Influential Observations 492
19.9 Quantifying Predictive Ability 492
19.10 Validating the Fitted Model 493
19.10.1 Validation of Model Calibration 493
19.10.2 Validation of Discrimination and Other Statistical Indexes . 494
19.11 Describing the Fitted Model 496
19.12 S-Plus Functions 499
19.13 Further Reading 506

20 Case Study in Cox Regression **509**
20.1 Choosing the Number of Parameters and Fitting the Model 509
20.2 Checking Proportional Hazards 513
20.3 Testing Interactions 516
20.4 Describing Predictor Effects 517
20.5 Validating the Model 517
20.6 Presenting the Model 519
20.7 Problems 522

Appendix **523**

References **527**

Index **559**