Contents

List of figures **xvii**

List of tables **xix**

Preface to the second edition **xxi**

Preface to the first edition **xxiii**

1 Data input and output **1**
1.1 Input . . . 1
1.1.1 Native dataset . . . 1
1.1.2 Fixed format text files . . . 2
1.1.3 Other fixed files . . . 3
1.1.4 Reading more complex text files . . . 3
1.1.5 Comma separated value (CSV) files . . . 4
1.1.6 Read sheets from an Excel file . . . 5
1.1.7 Read data from R into SAS . . . 5
1.1.8 Read data from SAS into R . . . 6
1.1.9 Reading datasets in other formats . . . 6
1.1.10 Reading data with a variable number of words in a field . . . 7
1.1.11 Read a file byte by byte . . . 8
1.1.12 Access data from a URL . . . 9
1.1.13 Read an XML-formatted file . . . 9
1.1.14 Manual data entry . . . 10
1.2 Output . . . 11
1.2.1 Displaying data . . . 11
1.2.2 Number of digits to display . . . 11
1.2.3 Save a native dataset . . . 12
1.2.4 Creating datasets in text format . . . 12
1.2.5 Creating Excel spreadsheets . . . 12
1.2.6 Creating files for use by other packages . . . 13
1.2.7 Creating HTML formatted output . . . 14
1.2.8 Creating XML datasets and output . . . 14
1.3 Further resources . . . 15

2 Data management **17**
2.1 Structure and meta-data . . . 17
2.1.1 Access variables from a dataset . . . 17
2.1.2 Names of variables and their types . . . 17
2.1.3 Values of variables in a dataset . . . 18

2.1.4 Label variables . . . 18
2.1.5 Add comment to a dataset or variable . . . 19
2.2 Derived variables and data manipulation . . . 19
2.2.1 Add derived variable to a dataset . . . 19
2.2.2 Rename variables in a dataset . . . 19
2.2.3 Create string variables from numeric variables . . . 20
2.2.4 Create categorical variables from continuous variables . . . 20
2.2.5 Recode a categorical variable . . . 21
2.2.6 Create a categorical variable using logic . . . 21
2.2.7 Create numeric variables from string variables . . . 22
2.2.8 Extract characters from string variables . . . 23
2.2.9 Length of string variables . . . 23
2.2.10 Concatenate string variables . . . 24
2.2.11 Set operations . . . 24
2.2.12 Find strings within string variables . . . 25
2.2.13 Find approximate strings . . . 25
2.2.14 Replace strings within string variables . . . 26
2.2.15 Split strings into multiple strings . . . 26
2.2.16 Remove spaces around string variables . . . 27
2.2.17 Upper to lower case . . . 27
2.2.18 Lagged variable . . . 28
2.2.19 Formatting values of variables . . . 28
2.2.20 Perl interface . . . 29
2.2.21 Accessing databases using SQL (structured query language) . . . 29
2.3 Merging, combining, and subsetting datasets . . . 29
2.3.1 Subsetting observations . . . 30
2.3.2 Drop or keep variables in a dataset . . . 30
2.3.3 Random sample of a dataset . . . 31
2.3.4 Observation number . . . 32
2.3.5 Keep unique values . . . 32
2.3.6 Identify duplicated values . . . 32
2.3.7 Convert from wide to long (tall) format . . . 33
2.3.8 Convert from long (tall) to wide format . . . 34
2.3.9 Concatenate and stack datasets . . . 35
2.3.10 Sort datasets . . . 35
2.3.11 Merge datasets . . . 35
2.4 Date and time variables . . . 37
2.4.1 Create date variable . . . 37
2.4.2 Extract weekday . . . 38
2.4.3 Extract month . . . 38
2.4.4 Extract year . . . 38
2.4.5 Extract quarter . . . 38
2.4.6 Create time variable . . . 39
2.5 Further resources . . . 39
2.6 Examples . . . 39
2.6.1 Data input and output . . . 39
2.6.2 Data display . . . 43
2.6.3 Derived variables and data manipulation . . . 44
2.6.4 Sorting and subsetting datasets . . . 51

3 Statistical and mathematical functions **53**
3.1 Probability distributions and random number generation 53
3.1.1 Probability density function 53
3.1.2 Quantiles of a probability density function 54
3.1.3 Setting the random number seed 55
3.1.4 Uniform random variables 55
3.1.5 Multinomial random variables 56
3.1.6 Normal random variables 56
3.1.7 Multivariate normal random variables 56
3.1.8 Truncated multivariate normal random variables 58
3.1.9 Exponential random variables 58
3.1.10 Other random variables 58
3.2 Mathematical functions 59
3.2.1 Basic functions 59
3.2.2 Trigonometric functions 60
3.2.3 Special functions 60
3.2.4 Integer functions 60
3.2.5 Comparisons of floating point variables 61
3.2.6 Complex numbers 61
3.2.7 Derivatives 62
3.2.8 Integration 62
3.2.9 Optimization problems 62
3.3 Matrix operations 63
3.3.1 Create matrix from vector 63
3.3.2 Combine vectors or matrices 63
3.3.3 Matrix addition 64
3.3.4 Transpose matrix 64
3.3.5 Find the dimension of a matrix or dataset 64
3.3.6 Matrix multiplication 65
3.3.7 Invert matrix 65
3.3.8 Component-wise multiplication 66
3.3.9 Create submatrix 66
3.3.10 Create a diagonal matrix 66
3.3.11 Create a vector of diagonal elements 67
3.3.12 Create a vector from a matrix 67
3.3.13 Calculate the determinant 67
3.3.14 Find eigenvalues and eigenvectors 67
3.3.15 Find the singular value decomposition 68
3.4 Examples 68
3.4.1 Probability distributions 68

4 Programming and operating system interface **71**
4.1 Control flow, programming, and data generation 71
4.1.1 Looping 71
4.1.2 Conditional execution 72
4.1.3 Sequence of values or patterns 73
4.1.4 Referring to a range of variables 74
4.1.5 Perform an action repeatedly over a set of variables 74
4.1.6 Grid of values 75
4.1.7 Debugging 76
4.1.8 Error recovery 76

4.2 Functions and macros . . . 77
4.2.1 SAS macros . . . 77
4.2.2 R functions . . . 78
4.3 Interactions with the operating system . . . 78
4.3.1 Timing commands . . . 78
4.3.2 Suspend execution for a time interval . . . 79
4.3.3 Execute a command in the operating system . . . 79
4.3.4 Command history . . . 80
4.3.5 Find working directory . . . 80
4.3.6 Change working directory . . . 80
4.3.7 List and access files . . . 81

5 Common statistical procedures 83
5.1 Summary statistics . . . 83
5.1.1 Means and other summary statistics . . . 83
5.1.2 Other moments . . . 84
5.1.3 Trimmed mean . . . 84
5.1.4 Quantiles . . . 85
5.1.5 Centering, normalizing, and scaling . . . 85
5.1.6 Mean and 95% confidence interval . . . 86
5.1.7 Proportion and 95% confidence interval . . . 86
5.1.8 Maximum likelihood estimation of parameters . . . 86
5.2 Bivariate statistics . . . 87
5.2.1 Epidemiologic statistics . . . 87
5.2.2 Test characteristics . . . 87
5.2.3 Correlation . . . 89
5.2.4 Kappa (agreement) . . . 89
5.3 Contingency tables . . . 90
5.3.1 Display cross-classification table . . . 90
5.3.2 Displaying missing value categories in a table . . . 90
5.3.3 Pearson chi-square statistic . . . 91
5.3.4 Cochran–Mantel–Haenszel test . . . 91
5.3.5 Cramér's V . . . 91
5.3.6 Fisher's exact test . . . 92
5.3.7 McNemar's test . . . 92
5.4 Tests for continuous variables . . . 92
5.4.1 Tests for normality . . . 92
5.4.2 Student's t test . . . 93
5.4.3 Test for equal variances . . . 93
5.4.4 Nonparametric tests . . . 94
5.4.5 Permutation test . . . 94
5.4.6 Logrank test . . . 95
5.5 Analytic power and sample size calculations . . . 95
5.6 Further resources . . . 97
5.7 Examples . . . 97
5.7.1 Summary statistics and exploratory data analysis . . . 97
5.7.2 Bivariate relationships . . . 101
5.7.3 Contingency tables . . . 103
5.7.4 Two sample tests of continuous variables . . . 107
5.7.5 Survival analysis: logrank test . . . 111

6 Linear regression and ANOVA **113**
6.1 Model fitting . . . 113
6.1.1 Linear regression . . . 113
6.1.2 Linear regression with categorical covariates . . . 114
6.1.3 Changing the reference category . . . 114
6.1.4 Parameterization of categorical covariates . . . 115
6.1.5 Linear regression with no intercept . . . 116
6.1.6 Linear regression with interactions . . . 117
6.1.7 One-way analysis of variance . . . 117
6.1.8 Analysis of variance with two or more factors . . . 117
6.2 Tests, contrasts, and linear functions of parameters . . . 118
6.2.1 Joint null hypotheses: several parameters equal 0 . . . 118
6.2.2 Joint null hypotheses: sum of parameters . . . 118
6.2.3 Tests of equality of parameters . . . 119
6.2.4 Multiple comparisons . . . 119
6.2.5 Linear combinations of parameters . . . 120
6.3 Model diagnostics . . . 120
6.3.1 Predicted values . . . 120
6.3.2 Residuals . . . 121
6.3.3 Standardized and Studentized residuals . . . 121
6.3.4 Leverage . . . 122
6.3.5 Cook's D . . . 122
6.3.6 DFFITS . . . 123
6.3.7 Diagnostic plots . . . 123
6.3.8 Heteroscedasticity tests . . . 124
6.4 Model parameters and results . . . 124
6.4.1 Parameter estimates . . . 124
6.4.2 Standardized regression coefficients . . . 124
6.4.3 Standard errors of parameter estimates . . . 125
6.4.4 Confidence interval for parameter estimates . . . 125
6.4.5 Confidence limits for the mean . . . 125
6.4.6 Prediction limits . . . 126
6.4.7 R-squared . . . 127
6.4.8 Design and information matrix . . . 127
6.4.9 Covariance matrix of parameter estimates . . . 127
6.4.10 Correlation matrix of parameter estimates . . . 128
6.5 Further resources . . . 128
6.6 Examples . . . 128
6.6.1 Scatterplot with smooth fit . . . 129
6.6.2 Linear regression with interaction . . . 130
6.6.3 Regression diagnostics . . . 135
6.6.4 Fitting the regression model separately for each value of another variable . . . 138
6.6.5 Two-way ANOVA . . . 139
6.6.6 Multiple comparisons . . . 144
6.6.7 Contrasts . . . 146

7 Regression generalizations and modeling 149
7.1 Generalized linear models . . . 149
7.1.1 Logistic regression model . . . 149
7.1.2 Conditional logistic regression model . . . 151
7.1.3 Exact logistic regression . . . 152
7.1.4 Ordered logistic model . . . 152
7.1.5 Generalized logistic model . . . 152
7.1.6 Poisson model . . . 153
7.1.7 Negative binomial model . . . 153
7.1.8 Log-linear model . . . 153
7.2 Further generalizations . . . 154
7.2.1 Zero-inflated Poisson model . . . 154
7.2.2 Zero-inflated negative binomial model . . . 154
7.2.3 Generalized additive model . . . 155
7.2.4 Nonlinear least squares model . . . 155
7.3 Robust methods . . . 156
7.3.1 Quantile regression model . . . 156
7.3.2 Robust regression model . . . 156
7.3.3 Ridge regression model . . . 156
7.4 Models for correlated data . . . 157
7.4.1 Linear models with correlated outcomes . . . 157
7.4.2 Linear mixed models with random intercepts . . . 158
7.4.3 Linear mixed models with random slopes . . . 158
7.4.4 More complex random coefficient models . . . 159
7.4.5 Multilevel models . . . 160
7.4.6 Generalized linear models with correlated outcomes . . . 160
7.4.7 Generalized linear mixed models . . . 161
7.4.8 Generalized estimating equations . . . 161
7.4.9 MANOVA . . . 162
7.4.10 Time series model . . . 162
7.5 Survival analysis . . . 163
7.5.1 Proportional hazards (Cox) regression model . . . 163
7.5.2 Proportional hazards (Cox) model with frailty . . . 163
7.5.3 Nelson–Aalen estimate of cumulative hazard . . . 164
7.5.4 Testing the proportionality of the Cox model . . . 164
7.5.5 Cox model with time-varying predictors . . . 165
7.6 Multivariate statistics and discriminant procedures . . . 166
7.6.1 Cronbach's α . . . 166
7.6.2 Factor analysis . . . 166
7.6.3 Recursive partitioning . . . 166
7.6.4 Linear discriminant analysis . . . 167
7.6.5 Latent class analysis . . . 167
7.6.6 Hierarchical clustering . . . 168
7.7 Complex survey design . . . 168
7.8 Model selection and assessment . . . 169
7.8.1 Compare two models . . . 169
7.8.2 Log-likelihood . . . 170
7.8.3 Akaike Information Criterion (AIC) . . . 170
7.8.4 Bayesian Information Criterion (BIC) . . . 170
7.8.5 LASSO model . . . 170
7.8.6 Hosmer–Lemeshow goodness of fit . . . 171

7.8.7 Goodness of fit for count models . . . 171
7.9 Further resources . . . 172
7.10 Examples . . . 172
7.10.1 Logistic regression . . . 172
7.10.2 Poisson regression . . . 176
7.10.3 Zero-inflated Poisson regression . . . 178
7.10.4 Negative binomial regression . . . 180
7.10.5 Quantile regression . . . 181
7.10.6 Ordered logistic . . . 182
7.10.7 Generalized logistic model . . . 183
7.10.8 Generalized additive model . . . 185
7.10.9 Reshaping a dataset for longitudinal regression . . . 187
7.10.10 Linear model for correlated data . . . 190
7.10.11 Linear mixed (random slope) model . . . 193
7.10.12 Generalized estimating equations . . . 197
7.10.13 Generalized linear mixed model . . . 199
7.10.14 Cox proportional hazards model . . . 200
7.10.15 Cronbach's α . . . 201
7.10.16 Factor analysis . . . 202
7.10.17 Recursive partitioning . . . 205
7.10.18 Linear discriminant analysis . . . 206
7.10.19 Hierarchical clustering . . . 208

8 A graphical compendium 211
8.1 Univariate plots . . . 211
8.1.1 Barplot . . . 211
8.1.2 Stem-and-leaf plot . . . 212
8.1.3 Dotplot . . . 212
8.1.4 Histogram . . . 213
8.1.5 Density plots . . . 213
8.1.6 Empirical cumulative probability density plot . . . 214
8.1.7 Boxplot . . . 214
8.1.8 Violin plots . . . 215
8.2 Univariate plots by grouping variable . . . 215
8.2.1 Side-by-side histograms . . . 215
8.2.2 Side-by-side boxplots . . . 215
8.2.3 Overlaid density plots . . . 216
8.2.4 Bar chart with error bars . . . 216
8.3 Bivariate plots . . . 217
8.3.1 Scatterplot . . . 217
8.3.2 Scatterplot with multiple y values . . . 218
8.3.3 Scatterplot with binning . . . 219
8.3.4 Transparent overplotting scatterplot . . . 219
8.3.5 Bivariate density plot . . . 220
8.3.6 Scatterplot with marginal histograms . . . 220
8.4 Multivariate plots . . . 221
8.4.1 Matrix of scatterplots . . . 221
8.4.2 Conditioning plot . . . 221
8.4.3 Contour plots . . . 222
8.4.4 3-D plots . . . 222
8.5 Special purpose plots . . . 223

8.5.1 Choropleth maps . . . 223
8.5.2 Interaction plots . . . 223
8.5.3 Plots for categorical data . . . 224
8.5.4 Circular plot . . . 224
8.5.5 Plot an arbitrary function . . . 224
8.5.6 Normal quantile-quantile plot . . . 225
8.5.7 Receiver operating characteristic (ROC) curve . . . 225
8.5.8 Plot confidence intervals for the mean . . . 226
8.5.9 Plot prediction limits from a simple linear regression . . . 226
8.5.10 Plot predicted lines for each value of a variable . . . 226
8.5.11 Kaplan–Meier plot . . . 227
8.5.12 Hazard function plotting . . . 228
8.5.13 Mean-difference plots . . . 228
8.6 Further resources . . . 230
8.7 Examples . . . 230
8.7.1 Scatterplot with multiple axes . . . 230
8.7.2 Conditioning plot . . . 232
8.7.3 Scatterplot with marginal histograms . . . 232
8.7.4 Kaplan–Meier plot . . . 234
8.7.5 ROC curve . . . 235
8.7.6 Pairs plot . . . 236
8.7.7 Visualize correlation matrix . . . 238

9 Graphical options and configuration 241
9.1 Adding elements . . . 241
9.1.1 Arbitrary straight line . . . 242
9.1.2 Plot symbols . . . 242
9.1.3 Add points to an existing graphic . . . 243
9.1.4 Jitter points . . . 243
9.1.5 Regression line fit to points . . . 244
9.1.6 Smoothed line . . . 244
9.1.7 Normal density . . . 245
9.1.8 Marginal rug plot . . . 245
9.1.9 Titles . . . 246
9.1.10 Footnotes . . . 246
9.1.11 Text . . . 246
9.1.12 Mathematical symbols . . . 247
9.1.13 Arrows and shapes . . . 247
9.1.14 Add grid . . . 248
9.1.15 Legend . . . 248
9.1.16 Identifying and locating points . . . 249
9.2 Options and parameters . . . 250
9.2.1 Graph size . . . 250
9.2.2 Grid of plots per page . . . 250
9.2.3 More general page layouts . . . 251
9.2.4 Fonts . . . 252
9.2.5 Point and text size . . . 252
9.2.6 Box around plots . . . 252
9.2.7 Size of margins . . . 253
9.2.8 Graphical settings . . . 253
9.2.9 Axis range and style . . . 253

9.2.10 Axis labels, values, and tick marks . . . 254
9.2.11 Line styles . . . 254
9.2.12 Line widths . . . 255
9.2.13 Colors . . . 255
9.2.14 Log scale . . . 255
9.2.15 Omit axes . . . 256
9.3 Saving graphs . . . 256
9.3.1 PDF . . . 256
9.3.2 Postscript . . . 256
9.3.3 RTF . . . 257
9.3.4 JPEG . . . 258
9.3.5 Windows Metafile (WMF) . . . 258
9.3.6 Bitmap image file (BMP) . . . 258
9.3.7 Tagged image file format (TIFF) . . . 259
9.3.8 Portable Network Graphics (PNG) . . . 259
9.3.9 Closing a graphic device . . . 260

10 Simulation **261**
10.1 Generating data . . . 261
10.1.1 Generate categorical data . . . 261
10.1.2 Generate data from a logistic regression . . . 263
10.1.3 Generate data from a generalized linear mixed model . . . 264
10.1.4 Generate correlated binary data . . . 267
10.1.5 Generate data from a Cox model . . . 269
10.1.6 Sampling from a challenging distribution . . . 271
10.2 Simulation applications . . . 274
10.2.1 Simulation study of Student's t test . . . 274
10.2.2 Diploma (or hat-check) problem . . . 276
10.2.3 Monty Hall problem . . . 278
10.3 Further resources . . . 280

11 Special topics **281**
11.1 Processing by group . . . 281
11.2 Simulation-based power calculations . . . 284
11.3 Reproducible analysis and output . . . 287
11.4 Advanced statistical methods . . . 290
11.4.1 Bayesian methods . . . 290
11.4.2 Propensity scores . . . 296
11.4.3 Bootstrapping . . . 303
11.4.4 Missing data . . . 304
11.4.5 Finite mixture models with concomitant variables . . . 311
11.5 Further resources . . . 313

12 Case studies **315**
12.1 Data management and related tasks . . . 315
12.1.1 Finding two closest values in a vector . . . 315
12.1.2 Tabulate binomial probabilities . . . 317
12.1.3 Calculate and plot a running average . . . 318
12.1.4 Create a Fibonacci sequence . . . 320
12.2 Read variable format files . . . 321
12.3 Plotting maps . . . 324

12.3.1 Massachusetts counties, continued . . . 324
12.3.2 Bike ride plot . . . 325
12.3.3 Choropleth maps . . . 327
12.4 Data scraping and visualization . . . 329
12.4.1 Scraping data from HTML files . . . 330
12.4.2 Reading data with two lines per observation . . . 331
12.4.3 Plotting time series data . . . 333
12.4.4 URL APIs and truly random numbers . . . 334
12.5 Manipulating bigger datasets . . . 336
12.6 Constrained optimization: the knapsack problem . . . 337

A Introduction to SAS **341**
A.1 Installation . . . 341
A.2 Running SAS and a sample session . . . 341
A.3 Learning SAS and getting help . . . 346
A.4 Fundamental elements of SAS syntax . . . 347
A.5 Work process: The cognitive style of SAS . . . 349
A.6 Useful SAS background . . . 349
A.6.1 Dataset options . . . 349
A.6.2 Subsetting . . . 350
A.6.3 Formats and informats . . . 350
A.7 Output Delivery System . . . 351
A.7.1 Saving output as datasets and controlling output . . . 351
A.7.2 Output file types and ODS destinations . . . 355
A.8 SAS macro variables . . . 355
A.9 Miscellanea . . . 356

B Introduction to R and RStudio **357**
B.1 Installation . . . 358
B.1.1 Installation under Windows . . . 358
B.1.2 Installation under Mac OS X . . . 359
B.1.3 RStudio . . . 359
B.1.4 Other graphical interfaces . . . 359
B.2 Running R and sample session . . . 360
B.2.1 Replicating examples from the book and sourcing commands . . . 361
B.2.2 Batch mode . . . 362
B.3 Learning R and getting help . . . 362
B.4 Fundamental structures and objects . . . 365
B.4.1 Objects and vectors . . . 365
B.4.2 Indexing . . . 365
B.4.3 Operators . . . 366
B.4.4 Lists . . . 366
B.4.5 Matrices . . . 367
B.4.6 Dataframes . . . 367
B.4.7 Attributes and classes . . . 369
B.4.8 Options . . . 369
B.5 Functions . . . 369
B.5.1 Calling functions . . . 369
B.5.2 The `apply` family of functions . . . 370
B.6 Add-ons: packages . . . 371
B.6.1 Introduction to packages . . . 371

B.6.2 CRAN task views 372
B.6.3 Installed libraries and packages 373
B.6.4 Packages referenced in this book 374
B.6.5 Datasets available with R 377
B.7 Support and bugs 377

C The HELP study dataset **379**
C.1 Background on the HELP study 379
C.2 Roadmap to analyses of the HELP dataset 379
C.3 Detailed description of the dataset 381

References **385**

Subject index **399**

SAS index **419**

R index **431**