Retaining the same accessible format as the popular first edition, *SAS and R: Data Management, Statistical Analysis, and Graphics, Second Edition* explains how to easily perform an analytical task in both SAS and R, without having to navigate through the extensive, idiosyncratic, and sometimes unwieldy software documentation. The book covers many common tasks, such as data management, descriptive summaries, inferential procedures, regression analysis, and graphics, along with more complex applications.

This edition now covers RStudio, a powerful and easy-to-use interface for R. It incorporates a number of additional topics, including application program interfaces (APIs), database management systems, reproducible analysis tools, Markov chain Monte Carlo (MCMC) methods, and finite mixture models. It also includes extended examples of simulations and many new examples.

Through the extensive indexing and cross-referencing, users can directly find and implement the material they need. SAS users can look up tasks in the SAS index and then find the associated R code while R users can benefit from the R index in a similar manner. Numerous example analyses demonstrate the code in action and facilitate further exploration.

**Features**
- Presents parallel examples in SAS and R to demonstrate how to use the software and derive identical answers regardless of software choice
- Takes users through the process of statistical coding from beginning to end
- Contains worked examples of basic and complex tasks, offering solutions to stumbling blocks often encountered by new users
- Includes an index for each software, allowing users to easily locate procedures
- Shows how RStudio can be used as a powerful, straightforward interface for R
- Covers APIs, reproducible analysis, database management systems, MCMC methods, and finite mixture models
- Incorporates extensive examples of simulations
- Provides the SAS and R example code, datasets, and more online

Ken Kleinman and Nicholas J. Horton
SAS and R

Data Management,
Statistical Analysis,
and Graphics
SECOND EDITION
Contents

List of figures xvii
List of tables xix
Preface to the second edition xxvi
Preface to the first edition xxxi

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
CONTENTS

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
CONTENTS

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

6.1 Model fitting

6.1.1 Linear regression

6.1.2 Linear regression with categorical covariates

6.1.3 Changing the reference category

6.1.4 Parameterization of categorical covariates

6.1.5 Linear regression with no intercept

6.1.6 Linear regression with interactions

6.1.7 One-way analysis of variance

6.1.8 Analysis of variance with two or more factors

6.2 Tests, contrasts, and linear functions of parameters

6.2.1 Joint null hypotheses: several parameters equal 0

6.2.2 Joint null hypotheses: sum of parameters

6.2.3 Tests of equality of parameters

6.2.4 Multiple comparisons

6.2.5 Linear combinations of parameters

6.3 Model diagnostics

6.3.1 Predicted values

6.3.2 Residuals

6.3.3 Standardized and Studentized residuals

6.3.4 Leverage

6.3.5 Cook's D

6.3.6 DFFITS

6.3.7 Diagnostic plots

6.3.8 Heteroscedasticity tests

6.4 Model parameters and results

6.4.1 Parameter estimates

6.4.2 Standardized regression coefficients

6.4.3 Standard errors of parameter estimates

6.4.4 Confidence interval for parameter estimates

6.4.5 Confidence limits for the mean

6.4.6 Prediction limits

6.4.7 R-squared

6.4.8 Design and information matrix

6.4.9 Covariance matrix of parameter estimates

6.4.10 Correlation matrix of parameter estimates

6.5 Further resources

6.6 Examples

6.6.1 Scatterplot with smooth fit

6.6.2 Linear regression with interaction

6.6.3 Regression diagnostics

6.6.4 Fitting the regression model separately for each value of another variable

6.6.5 Two-way ANOVA

6.6.6 Multiple comparisons

6.6.7 Contrasts
7 Regression generalizations and modeling

7.1 Generalized linear models

7.1.1 Logistic regression model
7.1.2 Conditional logistic regression model
7.1.3 Exact logistic regression
7.1.4 Ordered logistic model
7.1.5 Generalized logistic model
7.1.6 Poisson model
7.1.7 Negative binomial model
7.1.8 Log-linear model

7.2 Further generalizations

7.2.1 Zero-inflated Poisson model
7.2.2 Zero-inflated negative binomial model
7.2.3 Generalized additive model
7.2.4 Nonlinear least squares model

7.3 Robust methods

7.3.1 Quantile regression model
7.3.2 Robust regression model
7.3.3 Ridge regression model

7.4 Models for correlated data

7.4.1 Linear models with correlated outcomes
7.4.2 Linear mixed models with random intercepts
7.4.3 Linear mixed models with random slopes
7.4.4 More complex random coefficient models
7.4.5 Multilevel models
7.4.6 Generalized linear models with correlated outcomes
7.4.7 Generalized linear mixed models
7.4.8 Generalized estimating equations
7.4.9 MANOVA
7.4.10 Time series model

7.5 Survival analysis

7.5.1 Proportional hazards (Cox) regression model
7.5.2 Proportional hazards (Cox) model with frailty
7.5.3 Nelson–Aalen estimate of cumulative hazard
7.5.4 Testing the proportionality of the Cox model
7.5.5 Cox model with time-varying predictors

7.6 Multivariate statistics and discriminant procedures

7.6.1 Cronbach’s $\alpha$
7.6.2 Factor analysis
7.6.3 Recursive partitioning
7.6.4 Linear discriminant analysis
7.6.5 Latent class analysis
7.6.6 Hierarchical clustering

7.7 Complex survey design

7.8 Model selection and assessment

7.8.1 Compare two models
7.8.2 Log-likelihood
7.8.3 Akaike Information Criterion (AIC)
7.8.4 Bayesian Information Criterion (BIC)
7.8.5 LASSO model
7.8.6 Hosmer–Lemeshow goodness of fit
CONTENTS

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 $\alpha$ .............................................. 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
CONTENTS

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
## CONTENTS

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
CONTENTS

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
List of Figures

3.1 Comparison of standard normal and \( t \) distribution with 1 degree of freedom (df) ................................................. 69
3.2 Descriptive plot of the normal distribution ................................................. 70
5.1 Density plot of depressive symptom scores (CESD) plus superimposed histogram and normal distribution .......................... 100
5.2 Scatterplot of CESD and MCS for women, with primary substance shown as the plot symbol ................................................. 102
5.3 Graphical display of the table of substance by race/ethnicity ......................... 106
5.4 Density plot of age by gender ......................................................................... 111
6.1 Scatterplot of observed values for age and I1 (plus smoothers by substance) .... 130
6.2 SAS table produced with \texttt{latex} destination in \texttt{ODS} ................................................. 134
6.3 Q-Q plot from SAS, default diagnostics from R ................................................. 137
6.4 Empirical density of residuals, with superimposed normal density .................. 137
6.5 Interaction plot of CESD as a function of substance group and gender .......... 140
6.6 Boxplot of CESD as a function of substance group and gender ......................... 140
6.7 Pairwise comparisons .............................................................................. 146
7.1 Scatterplots of smoothed association of PCS with CESD .............................. 186
7.2 Side-by-side box plots of CESD by treatment and time ..................................... 193
7.3 Recursive partitioning tree from R ................................................................. 206
7.4 Graphical display of assignment probabilities or score functions from linear discriminant analysis by actual homeless status .............. 209
7.5 Results from hierarchical clustering ................................................................. 210
8.1 Plot of InDUC and MCS vs. CESD for female alcohol-involved subjects ...... 231
8.2 Association of MCS and CESD, stratified by substance and report of suicidal thoughts ...................................................................... 233
8.3 Association of MCS and CESD with marginal histograms ......................... 234
8.4 Kaplan–Meier estimate of time to linkage to primary care by randomization group ............................................................................ 236
8.5 Receiver operating characteristic curve for the logistical regression model predicting suicidal thoughts using the CESD as a measure of depressive symptoms (sensitivity = true positive rate; 1-specificity = false positive rate) ...... 237
8.6 Pairsplot of variables from the HELP dataset ............................................. 238
8.7 Visual display of correlations and associations ............................................... 240
10.1 Plot of true and simulated distributions ....................................................... 274
LIST OF FIGURES

11.1 Sample Markdown input file ........................................ 288
11.2 Formatted output from R Markdown example ...................... 289
12.1 Running average for Cauchy and $t$ distributions ............... 320
12.2 Massachusetts counties .............................................. 324
12.3 Bike plot with map background ..................................... 326
12.4 Choropleth map ........................................................ 329
12.5 Sales plot ............................................................... 334
12.6 Number of flights departing Bradley airport on Mondays over time . 338
A.1 SAS Windows interface ................................................ 342
A.2 Running a SAS program ................................................ 343
A.3 Results from proc print .............................................. 344
A.4 Results from proc univariate ....................................... 345
A.5 The SAS window after running the sample session code .......... 346
A.6 The SAS Explorer window ............................................ 347
A.7 Opening the on-line help ............................................. 348
A.8 The SAS Help and Documentation window ......................... 348
B.1 R Windows graphical user interface .................................. 358
B.2 R Mac OS X graphical user interface ................................. 359
B.3 RStudio graphical user interface ..................................... 360
B.4 Sample session in R .................................................... 361
B.5 Documentation on the mean() function ............................ 363
B.6 Display after running RSiteSearch("eta squared anova") ........ 364
List of Tables

3.1 Quantiles, probabilities, and pseudo-random number generation: distributions available in SAS and R ................................. 54
6.1 Formatted results using the xtable package ................................. 134
7.1 Generalized linear model distributions supported by SAS and R ........ 150
11.1 Bayesian modeling functions available within the MCMCpack package 292
12.1 Weights, volume, and values for the knapsack problem .................. 337
B.1 CRAN task views .................................................... 373
C.1 Analyses undertaken using the HELP dataset .............................. 379
C.2 Annotated description of variables in the HELP dataset ................. 381
Preface to the second edition

Software systems evolve, and so do the approaches and expertise of statistical analysts.

After the publication of the first edition of *SAS and R: Data Management, Statistical Analysis, and Graphics*, we began a blog in which we explored many new case studies and applications, ranging from generating a Fibonacci series to fitting finite mixture models with concomitant variables. We also discussed some additions to SAS and new or improved R packages. The blog now has hundreds of entries and (according to Google Analytics) has received hundreds of thousands of visits.

The volume you are holding is nearly 50% longer than the first edition, and much of the new material is adapted from these blog entries, while it also includes other improvements and additions which have emerged in the last few years.

We have extensively reorganized the material in the book and created three new chapters. The first, *Simulation*, includes examples where data are generated from complex models such as mixed effects models and survival models, and from distributions using the Metropolis–Hastings algorithm. We also explore three interesting statistics and probability examples via simulation. The second is *Special topics*, where we describe some key features, such as processing by group, and detail several important areas of statistics, including Bayesian methods, propensity scores, and bootstrapping. The last is *Case studies*, where we demonstrate examples of some data management tasks, read complex files, make and annotate maps, and show how to “scrape” data from web pages.

We also cover some important new tools, including the use of RStudio, a powerful and easy-to-use front end for R that adds innumerable features to R. In our experience, it at least doubles the productivity of R users, and our SAS-using students find it an extremely comfortable interface that bears some similarity to the SAS GUI.

We have added a separate section and examples that describe “reproducible analysis.” This is the notion that code, results, and interpretation should live together in a single place. We used two reproducible analysis systems (SASweave and Sweave) to generate the example code and output in the book. Code extracted from these files is provided on the book web site. In this edition, we provide a detailed discussion of the philosophy and use of these systems. In particular, we feel that the *knitr* and *markdown* packages for R, which are tightly integrated with RStudio, should become a part of every R user’s toolbox. We can’t imagine working on a project without them.

Finally, we’ve reorganized much of the material from the first edition into smaller, more focused chapters. Users will now find separate (and enhanced) chapters on data input and output, data management, statistical and mathematical functions, and programming, rather than a single chapter on “data management.” Graphics are now discussed in two chapters: one on high-level types of plots, such as scatterplots and histograms, and another on customizing the fine details of the plots, such as the number of tick marks and the color of plot symbols.

We’re immensely gratified by the positive response the first edition elicited, and hope the current volume will be as useful to you.
On the web
The book website at http://www.amherst.edu/~nhorton/sasr2 includes the table of contents, the indices, the HELP dataset, example code in SAS and R, a pointer to the blog, and a list of errata.

Acknowledgments
In addition to those acknowledged in the first edition, we would like to thank Kathryn Aloisio, Gregory Call, J.J. Allaire and the RStudio developers, plus the many individuals who have created and shared R packages or SAS macros. Their contributions to SAS, R, or \LaTeX{} programming efforts, comments, guidance, and/or helpful suggestions on drafts of the revision have been extremely helpful. Above all we greatly appreciate Sara and Julia as well as Abby, Alana, Kinari, and Sam, for their patience and support.

Amherst, MA
March 16, 2014
Preface to the first edition

SAS™ (SAS Institute [153]) and R (R development core team [135]) are two statistical software packages used in many fields of research. SAS is commercial software developed by SAS Institute; it includes well-validated statistical algorithms. It can be licensed but not purchased. Paying for a license entitles the licensee to professional customer support. However, licensing is expensive and SAS sometimes incorporates new statistical methods only after a significant lag. In contrast, R is free, open-source software, developed by a large group of people, many of whom are volunteers. It has a large and growing user and developer base. Methodologists often release applications for general use in R shortly after they have been introduced into the literature. Professional customer support is not provided, though there are many resources for users. There are settings in which one of these useful tools is needed, and users who have spent many hours gaining expertise in the other often find it frustrating to make the transition.

We have written this book as a reference text for users of SAS and R. Our primary goal is to provide users with an easy way to learn how to perform an analytic task in both systems, without having to navigate through the extensive, idiosyncratic, and sometimes (often?) unwieldy documentation each provides. We expect the book to function in the same way that an English–French dictionary informs users of both the equivalent nouns and verbs in the two languages as well as the differences in grammar. We include many common tasks, including data management, descriptive summaries, inferential procedures, regression analysis, multivariate methods, and the creation of graphics. We also show some more complex applications. In toto, we hope that the text will allow easier mobility between systems for users of any statistical system.

We do not attempt to exhaustively detail all possible ways available to accomplish a given task in each system. Neither do we claim to provide the most elegant solution. We have tried to provide a simple approach that is easy to understand for a new user, and have supplied several solutions when they seem likely to be helpful. Carrying forward the analogy to an English–French dictionary, we suggest language that will communicate the point effectively, without listing every synonym or providing guidance on native idiom or eloquence.

Who should use this book

Those with an understanding of statistics at the level of multiple-regression analysis will find this book helpful. This group includes professional analysts who use statistical packages almost every day as well as statisticians, epidemiologists, economists, engineers, physicians, sociologists, and others engaged in research or data analysis. We anticipate that this tool will be particularly useful for sophisticated users, those with years of experience in only one system, who need or want to use the other system. However, intermediate-level analysts should reap the same benefit. In addition, the book will bolster the analytic abilities of a relatively new user of either system, by providing a concise reference manual and annotated examples executed in both packages.
Using the book

The book has three indices, in addition to the comprehensive Table of Contents. These include: 1) a detailed topic (subject) index in English; 2) a SAS index, organized by SAS syntax; and 3) an R index, describing R syntax. SAS users can use the SAS index to look up a task for which they know the SAS code and turn to a page with that code as well as the associated R code to carry out that task. R users can use the dictionary in an analogous fashion using the R index.

Extensive example analyses are presented; see Table C.1 for a comprehensive list. These employ a single dataset (from the HELP study), described in Appendix C. Readers are encouraged to download the dataset and code from the book website. The examples demonstrate the code in action and facilitate exploration by the reader.

Differences between SAS and R

SAS and R are so fundamentally distinct that an enumeration of their differences would be counterproductive. However, some differences are important for new users to bear in mind.

SAS includes data management tools that are primarily intended to prepare data for analysis. After preparation, analysis is performed in a distinct step, the implementation of which effectively cannot be changed by the user, though often extensive options are available. R is a programming environment tailored for data analysis. Data management and analysis are integrated. This means, for example, that calculating body mass index (BMI) from weight and height can be treated as a function of the data, and as such is as likely to appear within a data analysis as in making a “new” piece of data to keep.

SAS Institute makes decisions about how to change the software or expand the scope of included analyses. These decisions are based on the needs of the user community and on corporate goals for profitability. For example, when changes are made, backwards compatibility is almost always maintained, and documentation of exceptions is extensive. SAS Institute’s corporate conservatism means that techniques are sometimes not included in SAS until they have been discussed in the peer-reviewed literature for many years. While the R core team controls base functionality, a very large number of users have developed functions for R. Methodologists often release R functions to implement their work concurrently with publication. While this provides great flexibility, it comes at some cost. A user-contributed function may implement a desired methodology, but code quality may be unknown, documentation scarce, and paid support nonexistent. Sometimes a function which once worked may become defunct due to a lack of backwards compatibility and/or the author’s inability to, or lack of interest in, updating it.

Other differences between SAS and R are worth noting. Data management in SAS is undertaken using row by row (observation-level) operations. R is inherently a vector-based language, where columns (variables) are manipulated. R is case sensitive, while SAS is generally not.

Where to begin

We do not anticipate that the book will be read cover to cover. Instead, we hope that the extensive indexing, cross-referencing, and worked examples will make it possible for readers to directly find and then implement what they need. A user new to either SAS or R should begin by reading the appropriate appendix for that software package, which includes a sample session and overview.
PREFACE

On the web

The book website includes the Table of Contents, the indices, the HELP dataset, example code in SAS and R, and a list of errata.

Acknowledgments

We would like to thank Rob Calver, Shashi Kumar, and Sarah Morris for their support and guidance at Informa CRC/Chapman and Hall, the Department of Statistics at the University of Auckland for graciously hosting NH during a sabbatical leave, and the Office of the Provost at Smith College. We also thank Allyson Abrams, Tanya Hakim, Ross Ihaka, Albyn Jones, Russell Lenth, Brian McArdle, Paul Murrell, Alastair Scott, David Schoenfeld, Duncan Temple Lang, Kristin Tyler, Chris Wild, and Alan Zaslavsky for contributions to SAS, R, or \LaTeX programming efforts, comments, guidance, and/or helpful suggestions on drafts of the manuscript.

Above all we greatly appreciate Sara and Julia as well as Abby, Alana, Kinari, and Sam, for their patience and support.

Amherst, MA and Northampton, MA
March 2009
Chapter 1

Data input and output

This chapter reviews data input and output, including reading and writing files in spreadsheet, ASCII file, native, and foreign formats.

1.1 Input

Both SAS and R provide comprehensive support for data input and output. In this section we address aspects of these tasks.

SAS native datasets are rectangular files with data stored in a special format. They have the form filename.sas7bdat or something similar, depending on version. In the following, we assume that files are stored in directories and that the locations of the directories in the operating system can be labeled using Windows syntax (though SAS allows UNIX/Linux/Mac OS X-style forward slash as a directory delimiter on Windows). Other operating systems will use local idioms in describing locations.

R organizes data in dataframes (B.4.6), or connected series of rectangular arrays, which can be saved as platform independent objects. R also allows UNIX-style directory delimiters (forward slash) on Windows.

1.1.1 Native dataset

SAS

```
libname libref "dir_location";
data ds;
   set libref.sasfilename; /* Note: no file extension */
... run;
```

```
or

data ds;
   set "dir_location\sasfilename.sas7bdat"; /* Windows only */
   set "dir_location/sasfilename.sas7bdat";
   /* works on all OS including Windows */
... run;
```

Note: The file sasfilename.sas7bdat is created by using a libref in a data statement; see 1.2.3.
CHAPTER 1. DATA INPUT AND OUTPUT

R

```r
load(file="dir_location/savedfile")  # works on all OS including Windows
load(file="dir_location\savedfile")  # Windows only
```

*Note:* Forward slash is supported as a directory delimiter on all operating systems; a double backslash is supported under Windows. The file `savedfile` is created by `save()` (see 1.2.3). Running the command `print(load(file="dir_location/savedfile"))` will display the objects that are added to the workspace.

1.1.2 Fixed format text files

See 1.1.4 (read more complex fixed files) and 12.2 (read variable format files).

SAS

```sas
data ds;
    infile 'C:\file_location\filename.ext';
    input varname1 ... varnamek;
run;
```

or

```sas
filename filehandle 'file_location/filename.ext';
```

```sas
proc import datafile=filehandle
    out=ds dbms=dlm;
    getnames=yes;
run;
```

*Note:* The `infile` approach allows the user to limit the number of rows read from the data file using the `obs` option. Character variables are noted with a trailing `$`, e.g., use a statement such as `input varname1 varname2 $ varname3` if the second position contains a character variable (see 1.1.4 for examples). The `input` statement allows many options and can be used to read files with variable format (12.2).

In `proc import`, the `getnames=yes` statement is used if the first row of the input file contains variable names (the variable types are detected from the data). If the first row does not contain variable names then the `getnames=no` option should be specified. The `guessingrows` option (not shown) will base the variable formats on other than the default 20 rows. The `proc import` statement will accept an explicit file location rather than a file associated by the `filename` statement as in 7.10.

Note that in Windows installations, SAS accepts either slashes or backslashes to denote directory structures. For Linux, only forward slashes are allowed. Behavior in other operating systems may vary.

In addition to these methods, files can be read by selecting the Import Data option on the file menu in the GUI.

R

```r
ds = read.table("dir_location\file.txt", header=TRUE)  # Windows only
```

or

```r
ds = read.table("dir_location/file.txt", header=TRUE)  # all OS (including Windows)
```

*Note:* Forward slash is supported as a directory delimiter on all operating systems; a double backslash is supported under Windows. If the first row of the file includes the name of the variables, these entries will be used to create appropriate names (reserved characters such as `$` or `’` are changed to `’`) for each of the columns in the dataset. If the first row doesn’t include the names, the `header` option can be left off (or set to `FALSE`), and the variables will be called `V1`, `V2`, ... `Vn`. A limit on the number of lines to be read can be specified