Using R and RStudio for Data Management, Statistical Analysis and Graphics

Second Edition

Nicholas J. Horton
Department of Mathematics and Statistics
Amherst College
Massachusetts, U.S.A.

Ken Kleinman
Department of Population Medicine
Harvard Medical School and
Harvard Pilgrim Health Care Institute
Boston, Massachusetts, U.S.A.
# Contents

List of Tables xvi

List of Figures xix

Preface to the second edition xx

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 .................. 1

1.1.3 Other fixed files ......................... 2

1.1.4 Comma-separated value (CSV) files ....... 2

1.1.5 Read sheets from an Excel file .......... 2

1.1.6 Read data from R into SAS ............... 2

1.1.7 Read data from SAS into R ................ 3

1.1.8 Reading datasets in other formats ..... 3

1.1.9 Reading more complex text files ......... 3

1.1.10 Reading data with a variable number of words in a field .... 4

1.1.11 Read a file byte by byte ................ 5

1.1.12 Access data from a URL ................. 5

1.1.13 Read an XML-formatted file .......... 6

1.1.14 Read an HTML table ................... 6

1.1.15 Manual data entry ..................... 7

1.2 Output ...................................... 7

1.2.1 Displaying data ......................... 7

1.2.2 Number of digits to display ............ 7

1.2.3 Save a native dataset ................... 8

1.2.4 Creating datasets in text format ....... 8

1.2.5 Creating Excel spreadsheets .......... 8

1.2.6 Creating files for use by other packages .... 8

1.2.7 Creating HTML formatted output ....... 8

1.2.8 Creating XML datasets and output ..... 9

1.3 Further resources ........................ 9
## CONTENTS

2 Data management ........................................ 11
  2.1 Structure and metadata ................................ 11
    2.1.1 Access variables from a dataset ............... 11
    2.1.2 Names of variables and their types .......... 11
    2.1.3 Values of variables in a dataset .......... 12
    2.1.4 Label variables ................................ 12
    2.1.5 Add comment to a dataset or variable .... 12
  2.2 Derived variables and data manipulation .......... 12
    2.2.1 Add derived variable to a dataset .......... 13
    2.2.2 Rename variables in a dataset .............. 13
    2.2.3 Create string variables from numeric variables . 13
    2.2.4 Create categorical variables from continuous variables . 13
    2.2.5 Recode a categorical variable ............... 14
    2.2.6 Create a categorical variable using logic ... 14
    2.2.7 Create numeric variables from string variables . 15
    2.2.8 Extract characters from string variables .... 15
    2.2.9 Length of string variables .................... 15
    2.2.10 Concatenate string variables .................. 15
    2.2.11 Set operations ................................ 16
    2.2.12 Find strings within string variables ....... 16
    2.2.13 Find approximate strings .................... 16
    2.2.14 Replace strings within string variables ...... 17
    2.2.15 Split strings into multiple strings .......... 17
    2.2.16 Remove spaces around string variables ...... 17
    2.2.17 Convert strings from upper to lower case .... 17
    2.2.18 Create lagged variable ....................... 17
    2.2.19 Formatting values of variables ............... 18
    2.2.20 Perl interface ................................ 18
    2.2.21 Accessing databases using SQL ............... 18
  2.3 Merging, combining, and subsetting datasets ..... 19
    2.3.1 Subsetting observations ....................... 19
    2.3.2 Drop or keep variables in a dataset .......... 19
    2.3.3 Random sample of a dataset ................... 20
    2.3.4 Observation number ............................... 20
    2.3.5 Keep unique values ............................. 20
    2.3.6 Identify duplicated values ..................... 20
    2.3.7 Convert from wide to long (tall) format .... 21
    2.3.8 Convert from long (tall) to wide format ...... 21
    2.3.9 Concatenate and stack datasets ............... 22
    2.3.10 Sort datasets .................................. 22
    2.3.11 Merge datasets ................................ 22
  2.4 Date and time variables .............................. 23
    2.4.1 Create date variable ......................... 23
    2.4.2 Extract weekday ............................... 24
    2.4.3 Extract month .................................. 24
    2.4.4 Extract year .................................. 24
    2.4.5 Extract quarter ................................ 24
    2.4.6 Create time variable ......................... 24
  2.5 Further resources ..................................... 25
  2.6 Examples ............................................. 25
    2.6.1 Data input and output ......................... 25
CONTENTS

2.6.2 Data display ................................................. 27
2.6.3 Derived variables and data manipulation .................. 27
2.6.4 Sorting and subsetting datasets ........................... 31

3 Statistical and mathematical functions ......................... 33
  3.1 Probability distributions and random number generation ... 33
     3.1.1 Probability density function ......................... 33
     3.1.2 Quantiles of a probability density function ......... 33
     3.1.3 Setting the random number seed ........................ 34
     3.1.4 Uniform random variables ............................. 34
     3.1.5 Multinomial random variables ......................... 35
     3.1.6 Normal random variables .............................. 35
     3.1.7 Multivariate normal random variables ................. 35
     3.1.8 Truncated multivariate normal random variables .... 36
     3.1.9 Exponential random variables ......................... 36
     3.1.10 Other random variables .............................. 36
  3.2 Mathematical functions ........................................ 36
     3.2.1 Basic functions ......................................... 36
     3.2.2 Trigonometric functions ............................... 37
     3.2.3 Special functions ....................................... 37
     3.2.4 Integer functions ...................................... 37
     3.2.5 Comparisons of floating-point variables .............. 38
     3.2.6 Complex numbers ....................................... 38
     3.2.7 Derivatives ............................................. 38
     3.2.8 Integration ............................................. 38
     3.2.9 Optimization problems ................................ 39
  3.3 Matrix operations ............................................. 39
     3.3.1 Create matrix from vector ............................. 39
     3.3.2 Combine vectors or matrices ........................... 39
     3.3.3 Matrix addition ........................................ 39
     3.3.4 Transpose matrix ........................................ 40
     3.3.5 Find the dimension of a matrix or dataset ............ 40
     3.3.6 Matrix multiplication ................................... 40
     3.3.7 Finding the inverse of a matrix ....................... 40
     3.3.8 Component-wise multiplication ........................ 40
     3.3.9 Create a submatrix ..................................... 40
     3.3.10 Create a diagonal matrix .............................. 40
     3.3.11 Create a vector of diagonal elements ................. 41
     3.3.12 Create a vector from a matrix ....................... 41
     3.3.13 Calculate the determinant ............................ 41
     3.3.14 Find eigenvalues and eigenvectors .................... 41
     3.3.15 Find the singular value decomposition ............... 41
  3.4 Examples ..................................................... 42
     3.4.1 Probability distributions ............................... 42

4 Programming and operating system interface .................... 45
  4.1 Control flow, programming, and data generation ........... 45
     4.1.1 Looping ................................................ 45
     4.1.2 Conditional execution .................................. 45
     4.1.3 Sequence of values or patterns ....................... 46
     4.1.4 Perform an action repeatedly over a set of variables . 46
CONTENTS

4.1.5 Grid of values ................................................. 47
4.1.6 Debugging ......................................................... 47
4.1.7 Error recovery .................................................. 47

4.2 Functions ............................................................ 48

4.3 Interactions with the operating system ............................................. 49
4.3.1 Timing commands ................................................. 49
4.3.2 Suspend execution for a time interval ......................................... 49
4.3.3 Execute a command in the operating system ..................................... 49
4.3.4 Command history .................................................. 49
4.3.5 Find working directory .............................................. 49
4.3.6 Change working directory ............................................ 50
4.3.7 List and access files ............................................... 50
4.3.8 Create temporary file ............................................... 50
4.3.9 Redirect output ..................................................... 50

5 Common statistical procedures ................................................. 51

5.1 Summary statistics ..................................................... 51
5.1.1 Means and other summary statistics ......................................... 51
5.1.2 Weighted means and other statistics ......................................... 51
5.1.3 Other moments ....................................................... 52
5.1.4 Trimmed mean ....................................................... 52
5.1.5 Quantiles ............................................................ 52
5.1.6 Centering, normalizing, and scaling ......................................... 52
5.1.7 Mean and 95% confidence interval ......................................... 52
5.1.8 Proportion and 95% confidence interval ....................................... 53
5.1.9 Maximum likelihood estimation of parameters .................................... 53

5.2 Bivariate statistics ...................................................... 53
5.2.1 Epidemiologic statistics .............................................. 53
5.2.2 Test characteristics .................................................. 54
5.2.3 Correlation .......................................................... 54
5.2.4 Kappa (agreement) ................................................... 54

5.3 Contingency tables ....................................................... 55
5.3.1 Display cross-classification table ........................................... 55
5.3.2 Displaying missing value categories in a table ................................... 55
5.3.3 Pearson chi-square statistic ............................................ 55
5.3.4 Cochran–Mantel–Haenszel test .......................................... 55
5.3.5 Cramér’s V .......................................................... 56
5.3.6 Fisher’s exact test ..................................................... 56
5.3.7 McNemar’s test ....................................................... 56

5.4 Tests for continuous variables ............................................ 56
5.4.1 Tests for normality ................................................... 56
5.4.2 Student’s t-test ....................................................... 56
5.4.3 Test for equal variances ............................................... 57
5.4.4 Nonparametric tests .................................................. 57
5.4.5 Permutation test ...................................................... 57
5.4.6 Logrank test ........................................................ 58

5.5 Analytic power and sample size calculations ..................................... 58

5.6 Further resources ........................................................ 59

5.7 Examples .............................................................. 59
5.7.1 Summary statistics and exploratory data analysis .................................. 59
5.7.2 Bivariate relationships ............................................... 60
CONTENTS

5.7.3 Contingency tables ............................................. 61
5.7.4 Two sample tests of continuous variables ................. 64
5.7.5 Survival analysis: logrank test ............................. 65

6 Linear regression and ANOVA ................................. 67
6.1 Model fitting .......................................................... 67
   6.1.1 Linear regression ............................................ 67
   6.1.2 Linear regression with categorical covariates ........ 68
   6.1.3 Changing the reference category ......................... 68
   6.1.4 Parameterization of categorical covariates ............ 68
   6.1.5 Linear regression with no intercept .................... 69
   6.1.6 Linear regression with interactions ..................... 69
   6.1.7 Linear regression with big data ......................... 69
   6.1.8 One-way analysis of variance ............................ 70
   6.1.9 Analysis of variance with two or more factors ......... 70
6.2 Tests, contrasts, and linear functions of parameters ....... 70
   6.2.1 Joint null hypotheses: several parameters equal 0 .... 70
   6.2.2 Joint null hypotheses: sum of parameters ............. 70
   6.2.3 Tests of equality of parameters .......................... 70
   6.2.4 Multiple comparisons ...................................... 71
   6.2.5 Linear combinations of parameters ...................... 71
6.3 Model results and diagnostics ................................ 71
   6.3.1 Predicted values ............................................ 72
   6.3.2 Residuals ..................................................... 72
   6.3.3 Standardized and Studentized residuals ............... 72
   6.3.4 Leverage ..................................................... 72
   6.3.5 Cook’s distance ............................................. 72
   6.3.6 DFFITs ....................................................... 73
   6.3.7 Diagnostic plots ............................................ 73
   6.3.8 Heteroscedasticity tests ................................. 73
6.4 Model parameters and results ................................ 73
   6.4.1 Parameter estimates ....................................... 73
   6.4.2 Standardized regression coefficients ..................... 73
   6.4.3 Coefficient plot ............................................. 74
   6.4.4 Standard errors of parameter estimates ................. 74
   6.4.5 Confidence interval for parameter estimates .......... 74
   6.4.6 Confidence limits for the mean .......................... 74
   6.4.7 Prediction limits ........................................... 75
   6.4.8 R-squared ................................................... 75
   6.4.9 Design and information matrix ......................... 75
   6.4.10 Covariance matrix of parameter estimates ............ 75
   6.4.11 Correlation matrix of parameter estimates ............ 76
6.5 Further resources .................................................. 76
6.6 Examples ............................................................ 76
   6.6.1 Scatterplot with smooth fit ................................ 76
   6.6.2 Linear regression with interaction ....................... 77
   6.6.3 Regression coefficient plot ............................... 81
   6.6.4 Regression diagnostics ................................... 81
   6.6.5 Fitting a regression model separately for each value of
        another variable ............................................ 83
   6.6.6 Two-way ANOVA .............................................. 84
   6.6.7 Multiple comparisons ..................................... 87
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6.8 Contrasts</td>
<td>88</td>
</tr>
<tr>
<td>7 Regression generalizations and modeling</td>
<td>91</td>
</tr>
<tr>
<td>7.1 Generalized linear models</td>
<td>91</td>
</tr>
<tr>
<td>7.1.1 Logistic regression model</td>
<td>91</td>
</tr>
<tr>
<td>7.1.2 Conditional logistic regression model</td>
<td>91</td>
</tr>
<tr>
<td>7.1.3 Exact logistic regression</td>
<td>92</td>
</tr>
<tr>
<td>7.1.4 Ordered logistic model</td>
<td>92</td>
</tr>
<tr>
<td>7.1.5 Generalized logistic model</td>
<td>93</td>
</tr>
<tr>
<td>7.1.6 Poisson model</td>
<td>93</td>
</tr>
<tr>
<td>7.1.7 Negative binomial model</td>
<td>93</td>
</tr>
<tr>
<td>7.1.8 Log-linear model</td>
<td>93</td>
</tr>
<tr>
<td>7.2 Further generalizations</td>
<td>93</td>
</tr>
<tr>
<td>7.2.1 Zero-inflated Poisson model</td>
<td>93</td>
</tr>
<tr>
<td>7.2.2 Zero-inflated negative binomial model</td>
<td>94</td>
</tr>
<tr>
<td>7.2.3 Generalized additive model</td>
<td>94</td>
</tr>
<tr>
<td>7.2.4 Nonlinear least squares model</td>
<td>94</td>
</tr>
<tr>
<td>7.3 Robust methods</td>
<td>95</td>
</tr>
<tr>
<td>7.3.1 Quantile regression model</td>
<td>95</td>
</tr>
<tr>
<td>7.3.2 Robust regression model</td>
<td>95</td>
</tr>
<tr>
<td>7.3.3 Ridge regression model</td>
<td>95</td>
</tr>
<tr>
<td>7.4 Models for correlated data</td>
<td>95</td>
</tr>
<tr>
<td>7.4.1 Linear models with correlated outcomes</td>
<td>96</td>
</tr>
<tr>
<td>7.4.2 Linear mixed models with random intercepts</td>
<td>96</td>
</tr>
<tr>
<td>7.4.3 Linear mixed models with random slopes</td>
<td>96</td>
</tr>
<tr>
<td>7.4.4 More complex random coefficient models</td>
<td>97</td>
</tr>
<tr>
<td>7.4.5 Multilevel models</td>
<td>97</td>
</tr>
<tr>
<td>7.4.6 Generalized linear mixed models</td>
<td>97</td>
</tr>
<tr>
<td>7.4.7 Generalized estimating equations</td>
<td>97</td>
</tr>
<tr>
<td>7.4.8 MANOVA</td>
<td>98</td>
</tr>
<tr>
<td>7.4.9 Time series model</td>
<td>98</td>
</tr>
<tr>
<td>7.5 Survival analysis</td>
<td>98</td>
</tr>
<tr>
<td>7.5.1 Proportional hazards (Cox) regression model</td>
<td>98</td>
</tr>
<tr>
<td>7.5.2 Proportional hazards (Cox) model with frailty</td>
<td>99</td>
</tr>
<tr>
<td>7.5.3 Nelson–Aalen estimate of cumulative hazard</td>
<td>99</td>
</tr>
<tr>
<td>7.5.4 Testing the proportionality of the Cox model</td>
<td>99</td>
</tr>
<tr>
<td>7.5.5 Cox model with time-varying predictors</td>
<td>100</td>
</tr>
<tr>
<td>7.6 Multivariate statistics and discriminant procedures</td>
<td>100</td>
</tr>
<tr>
<td>7.6.1 Cronbach’s $\alpha$</td>
<td>100</td>
</tr>
<tr>
<td>7.6.2 Factor analysis</td>
<td>100</td>
</tr>
<tr>
<td>7.6.3 Recursive partitioning</td>
<td>100</td>
</tr>
<tr>
<td>7.6.4 Linear discriminant analysis</td>
<td>100</td>
</tr>
<tr>
<td>7.6.5 Latent class analysis</td>
<td>101</td>
</tr>
<tr>
<td>7.6.6 Hierarchical clustering</td>
<td>101</td>
</tr>
<tr>
<td>7.7 Complex survey design</td>
<td>101</td>
</tr>
<tr>
<td>7.8 Model selection and assessment</td>
<td>102</td>
</tr>
<tr>
<td>7.8.1 Compare two models</td>
<td>102</td>
</tr>
<tr>
<td>7.8.2 Log-likelihood</td>
<td>102</td>
</tr>
<tr>
<td>7.8.3 Akaike Information Criterion (AIC)</td>
<td>102</td>
</tr>
<tr>
<td>7.8.4 Bayesian Information Criterion (BIC)</td>
<td>102</td>
</tr>
<tr>
<td>7.8.5 LASSO model</td>
<td>102</td>
</tr>
</tbody>
</table>
CONTENTS

7.8.6 Hosmer–Lemeshow goodness of fit ............................................. 103
7.8.7 Goodness of fit for count models .................................................. 103
7.9 Further resources .............................................................................. 103
7.10 Examples .......................................................................................... 104
    7.10.1 Logistic regression ....................................................................... 104
    7.10.2 Poisson regression ...................................................................... 105
    7.10.3 Zero-inflated Poisson regression .................................................. 106
    7.10.4 Negative binomial regression ....................................................... 107
    7.10.5 Quantile regression ..................................................................... 107
    7.10.6 Ordered logistic . ........................................................................ 108
    7.10.7 Generalized logistic model ........................................................... 108
    7.10.8 Generalized additive model .......................................................... 109
    7.10.9 Reshaping a dataset for longitudinal regression ............................ 110
    7.10.10 Linear model for correlated data .................................................. 112
    7.10.11 Linear mixed (random slope) model ............................................. 113
    7.10.12 Generalized estimating equations ............................................... 115
    7.10.13 Generalized linear mixed model .................................................. 116
    7.10.14 Cox proportional hazards model ................................................. 117
    7.10.15 Cronbach's α ............................................................................ 117
    7.10.16 Factor analysis .......................................................................... 118
    7.10.17 Recursive partitioning ................................................................ 119
    7.10.18 Linear discriminant analysis ...................................................... 120
    7.10.19 Hierarchical clustering ................................................................ 121

8 A graphical compendium .................................................................... 123
    8.1 Univariate plots .............................................................................. 123
        8.1.1 Barplot ..................................................................................... 123
        8.1.2 Stem-and-leaf plot .................................................................. 124
        8.1.3 Dotplot ................................................................................... 124
        8.1.4 Histogram ............................................................................... 124
        8.1.5 Density plot ............................................................................ 124
        8.1.6 Empirical cumulative probability density plot ................................ 125
        8.1.7 Boxplot ................................................................................... 125
        8.1.8 Violin plots .............................................................................. 125
    8.2 Univariate plots by grouping variable .............................................. 125
        8.2.1 Side-by-side histograms ............................................................... 125
        8.2.2 Side-by-side boxplots ................................................................. 125
        8.2.3 Overlaid density plots ................................................................. 126
        8.2.4 Bar chart with error bars .............................................................. 126
    8.3 Bivariate plots .................................................................................. 127
        8.3.1 Scatterplot ................................................................................ 127
        8.3.2 Scatterplot with multiple y values .............................................. 127
        8.3.3 Scatterplot with binning ............................................................... 128
        8.3.4 Transparent overplotting scatterplot ........................................... 128
        8.3.5 Bivariate density plot ................................................................. 128
        8.3.6 Scatterplot with marginal histograms ........................................... 129
    8.4 Multivariate plots .............................................................................. 129
        8.4.1 Matrix of scatterplots ................................................................. 129
        8.4.2 Conditioning plot ..................................................................... 129
        8.4.3 Contour plots ............................................................................ 130
        8.4.4 3-D plots .................................................................................. 130
CONTENTS

8.5 Special-purpose plots .................................................. 130
  8.5.1 Choropleth maps .................................................. 130
  8.5.2 Interaction plots .................................................. 130
  8.5.3 Plots for categorical data ........................................ 131
  8.5.4 Circular plot ..................................................... 131
  8.5.5 Plot an arbitrary function ....................................... 131
  8.5.6 Normal quantile–quantile plot .................................. 131
  8.5.7 Receiver operating characteristic (ROC) curve ............... 132
  8.5.8 Plot confidence intervals for the mean ......................... 132
  8.5.9 Plot prediction limits from a simple linear regression ...... 132
  8.5.10 Plot predicted lines for each value of a variable ........... 133
  8.5.11 Kaplan–Meier plot ............................................. 133
  8.5.12 Hazard function plotting ....................................... 133
  8.5.13 Mean–diference plots ........................................... 133

8.6 Further resources ...................................................... 134

8.7 Examples .................................................................. 134
  8.7.1 Scatterplot with multiple axes .................................... 134
  8.7.2 Conditioning plot .................................................. 135
  8.7.3 Scatterplot with marginal histograms ......................... 135
  8.7.4 Kaplan–Meier plot ................................................. 137
  8.7.5 ROC curve .......................................................... 138
  8.7.6 Pairs plot ........................................................... 138
  8.7.7 Visualize correlation matrix ...................................... 141

9 Graphical options and configuration .................................... 145
  9.1 Adding elements ...................................................... 145
    9.1.1 Arbitrary straight line .......................................... 145
    9.1.2 Plot symbols ..................................................... 145
    9.1.3 Add points to an existing graphic ............................. 146
    9.1.4 Jitter points ....................................................... 146
    9.1.5 Regression line fit to points .................................... 146
    9.1.6 Smoothed line .................................................... 146
    9.1.7 Normal density ................................................... 147
    9.1.8 Marginal rug plot ............................................... 147
    9.1.9 Titles ............................................................... 147
    9.1.10 Footnotes .......................................................... 147
    9.1.11 Text ............................................................... 147
    9.1.12 Mathematical symbols .......................................... 148
    9.1.13 Arrows and shapes ............................................. 148
    9.1.14 Add grid ......................................................... 148
    9.1.15 Legend ........................................................... 148
    9.1.16 Identifying and locating points ............................... 148

9.2 Options and parameters .............................................. 149
  9.2.1 Graph size ......................................................... 149
  9.2.2 Grid of plots per page ............................................ 149
  9.2.3 More general page layouts ....................................... 149
  9.2.4 Fonts ............................................................... 150
  9.2.5 Point and text size ............................................... 150
  9.2.6 Box around plots ................................................ 150
  9.2.7 Size of margins ................................................... 150
  9.2.8 Graphical settings ............................................... 150
CONTENTS

9.2.9 Axis range and style ........................................ 151
9.2.10 Axis labels, values, and tick marks ...................... 151
9.2.11 Line styles .............................................. 151
9.2.12 Line widths .............................................. 151
9.2.13 Colors .................................................. 151
9.2.14 Log scale ............................................... 152
9.2.15 Omit axes ............................................... 152
9.3 Saving graphs .............................................. 152
  9.3.1 PDF ................................................... 152
  9.3.2 Postscript .............................................. 152
  9.3.3 RTF ................................................... 152
  9.3.4 JPEG .................................................. 153
  9.3.5 Windows Metafile ....................................... 153
  9.3.6 Bitmap image file (BMP) ................................ 153
  9.3.7 Tagged Image File Format ............................... 153
  9.3.8 PNG .................................................. 153
  9.3.9 Closing a graphic device ............................... 153

10 Simulation ...................................................... 155
  10.1 Generating data .......................................... 155
      10.1.1 Generate categorical data ........................... 155
      10.1.2 Generate data from a logistic regression .......... 156
      10.1.3 Generate data from a generalized linear mixed model 156
      10.1.4 Generate correlated binary data ..................... 157
      10.1.5 Generate data from a Cox model .................... 158
      10.1.6 Sampling from a challenging distribution .......... 159
  10.2 Simulation applications ................................... 161
      10.2.1 Simulation study of Student’s $t$-test ............. 161
      10.2.2 Diploma (or hat-check) problem ...................... 162
      10.2.3 Monty Hall problem ................................ 163
      10.2.4 Censored survival .................................. 165
  10.3 Further resources ......................................... 165

11 Special topics .................................................. 167
  11.1 Processing by group ...................................... 167
      11.1.1 Means by group .................................... 167
      11.1.2 Linear models stratified by each value of a grouping variable 168
  11.2 Simulation-based power calculations ...................... 169
  11.3 Reproducible analysis and output ........................ 171
  11.4 Advanced statistical methods ............................. 173
      11.4.1 Bayesian methods .................................. 173
      11.4.2 Propensity scores .................................. 177
      11.4.3 Bootstrapping ..................................... 181
      11.4.4 Missing data ...................................... 182
      11.4.5 Finite mixture models with concomitant variables .... 185
  11.5 Further resources ......................................... 186
12 Case studies
12.1 Data management and related tasks ........................................ 187
  12.1.1 Finding two closest values in a vector .................................. 187
  12.1.2 Tabulate binomial probabilities ........................................... 188
  12.1.3 Calculate and plot a running average .................................... 188
  12.1.4 Create a Fibonacci sequence .............................................. 189
12.2 Read variable format files ....................................................... 190
12.3 Plotting maps ................................................................. 192
  12.3.1 Massachusetts counties, continued ....................................... 192
  12.3.2 Bike ride plot ............................................................. 193
  12.3.3 Choropleth maps .......................................................... 193
12.4 Data scraping ................................................................. 195
  12.4.1 Scraping data from HTML files ......................................... 195
  12.4.2 Reading data with two lines per observation ............................ 196
  12.4.3 Plotting time series data .................................................. 197
  12.4.4 Reading tables from HTML ................................................. 198
  12.4.5 URL APIs and truly random numbers .................................... 199
  12.4.6 Reading from a web API ................................................... 200
12.5 Text mining ................................................................. 202
  12.5.1 Retrieving data from arXiv.org ......................................... 202
  12.5.2 Exploratory text mining .................................................... 202
12.6 Interactive visualization ...................................................... 203
  12.6.1 Visualization using the grammar of graphics (ggvis) ................. 203
  12.6.2 Shiny in Markdown ........................................................ 205
  12.6.3 Creating a standalone Shiny app ......................................... 206
12.7 Manipulating bigger datasets .................................................. 207
12.8 Constrained optimization: the knapsack problem ......................... 208

A Introduction to R and RStudio ................................................. 211
A.1 Installation ............................................................................. 212
  A.1.1 Installation under Windows .................................................. 212
  A.1.2 Installation under Mac OS X ................................................. 213
  A.1.3 RStudio ............................................................................. 213
  A.1.4 Other graphical interfaces .................................................... 213
A.2 Running R and sample session .................................................. 214
  A.2.1 Replicating examples from the book and sourcing commands ........ 215
  A.2.2 Batch mode ...................................................................... 216
A.3 Learning R ............................................................................. 216
  A.3.1 Getting help ........................................................................ 216
  A.3.2 swirl ................................................................................. 217
A.4 Fundamental structures and objects ............................................ 220
  A.4.1 Objects and vectors ........................................................... 221
  A.4.2 Indexing .......................................................................... 221
  A.4.3 Operators ......................................................................... 222
  A.4.4 Lists ................................................................................. 222
  A.4.5 Matrices .......................................................................... 223
  A.4.6 Dataframes .................................................................... 223
  A.4.7 Attributes and classes ....................................................... 226
  A.4.8 Options .......................................................................... 226
A.5 Functions ............................................................................... 226
  A.5.1 Calling functions .............................................................. 226
CONTENTS

A.5.2 The apply family of functions ........................................... 227
A.5.3 Pipes and connections between functions ............................. 228
A.6 Add-ons: packages .............................................................. 229
   A.6.1 Introduction to packages .............................................. 229
   A.6.2 Packages and name conflicts ........................................ 230
   A.6.3 Maintaining packages ................................................. 231
   A.6.4 CRAN task views ...................................................... 231
   A.6.5 Installed libraries and packages ..................................... 231
   A.6.6 Packages referenced in this book .................................... 233
   A.6.7 Datasets available with R ............................................ 236
A.7 Support and bugs ............................................................... 236

B The HELP study dataset ....................................................... 237
   B.1 Background on the HELP study ........................................ 237
   B.2 Roadmap to analyses of the HELP dataset ........................... 237
   B.3 Detailed description of the dataset ................................... 239

C References .............................................................................. 243

D Indices ................................................................................... 255
   D.1 Subject index .................................................................... 255
   D.2 R index ........................................................................... 276