

contents

preface	xiii
acknowledgments	xv
about this book	xvii
about the cover illustration	xxii

PART I GETTING STARTED.....1

1 Introduction to R 3

1.1	Why use R?	5
1.2	Obtaining and installing R	7
1.3	Working with R	7
	Getting started	8
	Getting help	11
	The workspace	11
	Input and output	13
1.4	Packages	14
	What are packages?	15
	Installing a package	16
	Loading a package	16
	Learning about a package	16
1.5	Batch processing	17
1.6	Using output as input—reusing results	18
1.7	Working with large datasets	18

1.8 Working through an example 18

1.9 Summary 20

2 **Creating a dataset** 21

2.1 Understanding datasets 22

2.2 Data structures 23

Vectors 24 ■ *Matrices* 24 ■ *Arrays* 26 ■ *Data frames* 27

Factors 30 ■ *Lists* 32

2.3 Data input 33

Entering data from the keyboard 34 ■ *Importing data from a delimited text*

file 35 ■ *Importing data from Excel* 36 ■ *Importing data from XML* 37

Webscraping 37 ■ *Importing data from SPSS* 38 ■ *Importing data from SAS* 38

Importing data from Stata 38 ■ *Importing data from netCDF* 39

Importing data from HDF5 39 ■ *Accessing database management systems*

(DBMSs) 39 ■ *Importing data via Stat/Transfer* 41

2.4 Annotating datasets 42

Variable labels 42 ■ *Value labels* 42

2.5 Useful functions for working with data objects 42

2.6 Summary 43

3 **Getting started with graphs** 45

3.1 Working with graphs 46

3.2 A simple example 48

3.3 Graphical parameters 49

Symbols and lines 50 ■ *Colors* 52 ■ *Text characteristics* 53

Graph and margin dimensions 54

3.4 Adding text, customized axes, and legends 56

Titles 57 ■ *Axes* 57 ■ *Reference lines* 60 ■ *Legend* 60

Text annotations 62

3.5 Combining graphs 65

Creating a figure arrangement with fine control 69

3.6 Summary 71

4 **Basic data management** 73

4.1 A working example 73

4.2 Creating new variables 75

4.3 Recoding variables 76

- 4.4 Renaming variables 78
- 4.5 Missing values 79
 - Recoding values to missing 80 ■ Excluding missing values from analyses 80*
- 4.6 Date values 81
 - Converting dates to character variables 83 ■ Going further 83*
- 4.7 Type conversions 83
- 4.8 Sorting data 84
- 4.9 Merging datasets 85
 - Adding columns 85 ■ Adding rows 85*
- 4.10 Subsetting datasets 86
 - Selecting (keeping) variables 86 ■ Excluding (dropping) variables 86*
 - Selecting observations 87 ■ The subset() function 88 ■ Random samples 89*
- 4.11 Using SQL statements to manipulate data frames 89
- 4.12 Summary 90

5 **Advanced data management 91**

- 5.1 A data management challenge 92
- 5.2 Numerical and character functions 93
 - Mathematical functions 93 ■ Statistical functions 94 ■ Probability functions 96*
 - Character functions 99 ■ Other useful functions 101 ■ Applying functions to matrices and data frames 102*
- 5.3 A solution for our data management challenge 103
- 5.4 Control flow 107
 - Repetition and looping 107 ■ Conditional execution 108*
- 5.5 User-written functions 109
- 5.6 Aggregation and restructuring 112
 - Transpose 112 ■ Aggregating data 112 ■ The reshape package 113*
- 5.7 Summary 116

PART II BASIC METHODS 117

6 **Basic graphs 119**

- 6.1 Bar plots 120
 - Simple bar plots 120 ■ Stacked and grouped bar plots 121 ■ Mean bar plots 122*
 - Tweaking bar plots 123 ■ Spinograms 124*
- 6.2 Pie charts 125
- 6.3 Histograms 128

- 6.4 Kernel density plots 130
- 6.5 Box plots 133
 - Using parallel box plots to compare groups* 134 ■ *Violin plots* 137
- 6.6 Dot plots 138
- 6.7 Summary 140

7 Basic statistics 141

- 7.1 Descriptive statistics 142
 - A menagerie of methods* 142 ■ *Descriptive statistics by group* 146
 - Visualizing results* 149
- 7.2 Frequency and contingency tables 149
 - Generating frequency tables* 150 ■ *Tests of independence* 156
 - Measures of association* 157 ■ *Visualizing results* 158
 - Converting tables to flat files* 158
- 7.3 Correlations 159
 - Types of correlations* 160 ■ *Testing correlations for significance* 162
 - Visualizing correlations* 164
- 7.4 t-tests 164
 - Independent t-test* 164 ■ *Dependent t-test* 165 ■ *When there are more than two groups* 166
- 7.5 Nonparametric tests of group differences 166
 - Comparing two groups* 166 ■ *Comparing more than two groups* 168
- 7.6 Visualizing group differences 170
- 7.7 Summary 170

PART III INTERMEDIATE METHODS 171

8 Regression 173

- 8.1 The many faces of regression 174
 - Scenarios for using OLS regression* 175 ■ *What you need to know* 176
- 8.2 OLS regression 177
 - Fitting regression models with `lm()`* 178 ■ *Simple linear regression* 179
 - Polynomial regression* 181 ■ *Multiple linear regression* 184
 - Multiple linear regression with interactions* 186
- 8.3 Regression diagnostics 188
 - A typical approach* 189 ■ *An enhanced approach* 192 ■ *Global validation of linear model assumption* 199 ■ *Multicollinearity* 199
- 8.4 Unusual observations 200
 - Outliers* 200 ■ *High leverage points* 201 ■ *Influential observations* 202

- 8.5 Corrective measures 205
 - Deleting observations* 205 ■ *Transforming variables* 205 ■ *Adding or deleting variables* 207 ■ *Trying a different approach* 207
- 8.6 Selecting the “best” regression model 207
 - Comparing models* 208 ■ *Variable selection* 209
- 8.7 Taking the analysis further 213
 - Cross-validation* 213 ■ *Relative importance* 215
- 8.8 Summary 218

9 Analysis of variance 219

- 9.1 A crash course on terminology 220
- 9.2 Fitting ANOVA models 222
 - The aov() function* 222 ■ *The order of formula terms* 223
- 9.3 One-way ANOVA 225
 - Multiple comparisons* 227 ■ *Assessing test assumptions* 229
- 9.4 One-way ANCOVA 230
 - Assessing test assumptions* 232 ■ *Visualizing the results* 232
- 9.5 Two-way factorial ANOVA 234
- 9.6 Repeated measures ANOVA 237
- 9.7 Multivariate analysis of variance (MANOVA) 239
 - Assessing test assumptions* 241 ■ *Robust MANOVA* 242
- 9.8 ANOVA as regression 243
- 9.9 Summary 245

10 Power analysis 246

- 10.1 A quick review of hypothesis testing 247
- 10.2 Implementing power analysis with the pwr package 249
 - t-tests* 250 ■ *ANOVA* 252 ■ *Correlations* 253 ■ *Linear models* 253
 - Tests of proportions* 254 ■ *Chi-square tests* 255 ■ *Choosing an appropriate effect size in novel situations* 257
- 10.3 Creating power analysis plots 258
- 10.4 Other packages 260
- 10.5 Summary 261

11 Intermediate graphs 263

- 11.1 Scatter plots 264
 - Scatter plot matrices* 267 ■ *High-density scatter plots* 271 ■ *3D scatter plots* 274
 - Bubble plots* 278

- 11.2 Line charts 280
- 11.3 Correlograms 283
- 11.4 Mosaic plots 288
- 11.5 Summary 290

12 **Resampling statistics and bootstrapping 291**

- 12.1 Permutation tests 292
- 12.2 Permutation test with the coin package 294
 - Independent two-sample and k-sample tests 295* ■ *Independence in contingency tables 296* ■ *Independence between numeric variables 297*
 - Dependent two-sample and k-sample tests 297* ■ *Going further 298*
- 12.3 Permutation tests with the lmPerm package 298
 - Simple and polynomial regression 299* ■ *Multiple regression 300*
 - One-way ANOVA and ANCOVA 301* ■ *Two-way ANOVA 302*
- 12.4 Additional comments on permutation tests 302
- 12.5 Bootstrapping 303
- 12.6 Bootstrapping with the boot package 304
 - Bootstrapping a single statistic 305* ■ *Bootstrapping several statistics 307*
- 12.7 Summary 309

PART IV **ADVANCED METHODS311**

13 **Generalized linear models 313**

- 13.1 Generalized linear models and the glm() function 314
 - The glm() function 315* ■ *Supporting functions 316* ■ *Model fit and regression diagnostics 317*
- 13.2 Logistic regression 317
 - Interpreting the model parameters 320* ■ *Assessing the impact of predictors on the probability of an outcome 321* ■ *Overdispersion 322* ■ *Extensions 323*
- 13.3 Poisson regression 324
 - Interpreting the model parameters 326* ■ *Overdispersion 327* ■ *Extensions 328*
- 13.4 Summary 330

14 **Principal components and factor analysis 331**

- 14.1 Principal components and factor analysis in R 333
- 14.2 Principal components 334
 - Selecting the number of components to extract 335*

- Extracting principal components* 336 ■ *Rotating principal components* 339
- Obtaining principal components scores* 341
- 14.3 Exploratory factor analysis 342
 - Deciding how many common factors to extract* 343 ■ *Extracting common factors* 344 ■ *Rotating factors* 345 ■ *Factor scores* 349 ■ *Other EFA-related packages* 349
- 14.4 Other latent variable models 349
- 14.5 Summary 350

15 **Advanced methods for missing data** 352

- 15.1 Steps in dealing with missing data 353
- 15.2 Identifying missing values 355
- 15.3 Exploring missing values patterns 356
 - Tabulating missing values* 357 ■ *Exploring missing data visually* 357 ■ *Using correlations to explore missing values* 360
- 15.4 Understanding the sources and impact of missing data 362
- 15.5 Rational approaches for dealing with incomplete data 363
- 15.6 Complete-case analysis (listwise deletion) 364
- 15.7 Multiple imputation 365
- 15.8 Other approaches to missing data 370
 - Pairwise deletion* 370 ■ *Simple (nonstochastic) imputation* 371
- 15.9 Summary 371

16 **Advanced graphics** 373

- 16.1 The four graphic systems in R 374
- 16.2 The lattice package 375
 - Conditioning variables* 379 ■ *Panel functions* 381 ■ *Grouping variables* 383
 - Graphic parameters* 387 ■ *Page arrangement* 388
- 16.3 The ggplot2 package 390
- 16.4 Interactive graphs 394
 - Interacting with graphs: identifying points* 394 ■ *playwith* 394
 - lattice* 396 ■ *Interactive graphics with the iplots package* 397 ■ *rggobi* 399
- 16.5 Summary 399

afterword *Into the rabbit hole* 400

<i>appendix A</i>	<i>Graphic user interfaces</i>	403
<i>appendix B</i>	<i>Customizing the startup environment</i>	406
<i>appendix C</i>	<i>Exporting data from R</i>	408
<i>appendix D</i>	<i>Creating publication-quality output</i>	410
<i>appendix E</i>	<i>Matrix Algebra in R</i>	419
<i>appendix F</i>	<i>Packages used in this book</i>	421
<i>appendix G</i>	<i>Working with large datasets</i>	429
<i>appendix H</i>	<i>Updating an R installation</i>	432
	<i>references</i>	434
	<i>index</i>	438