
The R Book

The R Book

Second Edition

Michael J. Crawley

Imperial College London at Silwood Park, UK

<http://www.bio.ic.ac.uk/research/mjcraw/therbook/index.htm>



A John Wiley & Sons, Ltd., Publication

This edition first published 2013
© 2013 John Wiley & Sons, Ltd

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Crawley, Michael J.

The R book / Michael J. Crawley. – 2e.

pages cm

Includes bibliographical references and index.

ISBN 978-0-470-97392-9 (hardback)

1. R (Computer program language) 2. Mathematical statistics—Data processing. I. Title.

QA276.45.R3C73 2013

519.50285/5133—dc23

2012027339

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-97392-9

Set in 10/12pt Times by Aptara Inc., New Delhi, India.

Chapters

<i>Preface</i>	xxiii
1 Getting Started	1
2 Essentials of the R Language	12
3 Data Input	137
4 Dataframes	159
5 Graphics	189
6 Tables	244
7 Mathematics	258
8 Classical Tests	344
9 Statistical Modelling	388
10 Regression	449
11 Analysis of Variance	498
12 Analysis of Covariance	537
13 Generalized Linear Models	557
14 Count Data	579
15 Count Data in Tables	599
16 Proportion Data	628
17 Binary Response Variables	650
18 Generalized Additive Models	666
19 Mixed-Effects Models	681
20 Non-Linear Regression	715
21 Meta-Analysis	740
22 Bayesian Statistics	752

23	Tree Models	768
24	Time Series Analysis	785
25	Multivariate Statistics	809
26	Spatial Statistics	825
27	Survival Analysis	869
28	Simulation Models	893
29	Changing the Look of Graphics	907
	<i>References and Further Reading</i>	971
	<i>Index</i>	977

Detailed Contents

Preface

xxiii

1	Getting Started	1
1.1	How to use this book	1
1.1.1	Beginner in both computing and statistics	1
1.1.2	Student needing help with project work	2
1.1.3	Done some R and some statistics, but keen to learn more of both	2
1.1.4	Done regression and ANOVA, but want to learn more advanced statistical modelling	2
1.1.5	Experienced in statistics, but a beginner in R	2
1.1.6	Experienced in computing, but a beginner in R	2
1.1.7	Familiar with statistics and computing, but need a friendly reference manual	3
1.2	Installing R	3
1.3	Running R	3
1.4	The Comprehensive R Archive Network	4
1.4.1	Manuals	5
1.4.2	Frequently asked questions	5
1.4.3	Contributed documentation	5
1.5	Getting help in R	6
1.5.1	Worked examples of functions	6
1.5.2	Demonstrations of R functions	7
1.6	Packages in R	7
1.6.1	Contents of packages	8
1.6.2	Installing packages	8
1.7	Command line versus scripts	9
1.8	Data editor	9
1.9	Changing the look of the R screen	10
1.10	Good housekeeping	10
1.11	Linking to other computer languages	11
2	Essentials of the R Language	12
2.1	Calculations	13
2.1.1	Complex numbers in R	13
2.1.2	Rounding	14
2.1.3	Arithmetic	16
2.1.4	Modulo and integer quotients	17

2.1.5	Variable names and assignment	18
2.1.6	Operators	19
2.1.7	Integers	19
2.1.8	Factors	20
2.2	Logical operations	22
2.2.1	<code>TRUE</code> and <code>T</code> with <code>FALSE</code> and <code>F</code>	22
2.2.2	Testing for equality with real numbers	23
2.2.3	Equality of floating point numbers using <code>all.equal</code>	23
2.2.4	Summarizing differences between objects using <code>all.equal</code>	24
2.2.5	Evaluation of combinations of <code>TRUE</code> and <code>FALSE</code>	25
2.2.6	Logical arithmetic	25
2.3	Generating sequences	27
2.3.1	Generating repeats	28
2.3.2	Generating factor levels	29
2.4	Membership: Testing and coercing in R	30
2.5	Missing values, infinity and things that are not numbers	32
2.5.1	Missing values: <code>NA</code>	33
2.6	Vectors and subscripts	35
2.6.1	Extracting elements of a vector using subscripts	36
2.6.2	Classes of vector	38
2.6.3	Naming elements within vectors	38
2.6.4	Working with logical subscripts	39
2.7	Vector functions	41
2.7.1	Obtaining tables of means using <code>tapply</code>	42
2.7.2	The aggregate function for grouped summary statistics	44
2.7.3	Parallel minima and maxima: <code>pmin</code> and <code>pmax</code>	45
2.7.4	Summary information from vectors by groups	46
2.7.5	Addresses within vectors	46
2.7.6	Finding closest values	47
2.7.7	Sorting, ranking and ordering	47
2.7.8	Understanding the difference between <code>unique</code> and <code>duplicated</code>	49
2.7.9	Looking for runs of numbers within vectors	50
2.7.10	Sets: <code>union</code> , <code>intersect</code> and <code>setdiff</code>	52
2.8	Matrices and arrays	53
2.8.1	Matrices	54
2.8.2	Naming the rows and columns of matrices	55
2.8.3	Calculations on rows or columns of the matrix	56
2.8.4	Adding rows and columns to the matrix	58
2.8.5	The <code>sweep</code> function	59
2.8.6	Applying functions with <code>apply</code> , <code>sapply</code> and <code>lapply</code>	61
2.8.7	Using the <code>max.col</code> function	65
2.8.8	Restructuring a multi-dimensional array using <code>aperm</code>	67
2.9	Random numbers, sampling and shuffling	69
2.9.1	The <code>sample</code> function	70
2.10	Loops and repeats	71
2.10.1	Creating the binary representation of a number	73
2.10.2	Loop avoidance	74

2.10.3	The slowness of loops	75
2.10.4	Do not ‘grow’ data sets by concatenation or recursive function calls	76
2.10.5	Loops for producing time series	77
2.11	Lists	78
2.11.1	Lists and <code>lapply</code>	80
2.11.2	Manipulating and saving lists	82
2.12	Text, character strings and pattern matching	86
2.12.1	Pasting character strings together	87
2.12.2	Extracting parts of strings	88
2.12.3	Counting things within strings	89
2.12.4	Upper- and lower-case text	91
2.12.5	The <code>match</code> function and relational databases	91
2.12.6	Pattern matching	93
2.12.7	Dot <code>.</code> as the ‘anything’ character	95
2.12.8	Substituting text within character strings	96
2.12.9	Locations of a pattern within a vector using <code>regexpr</code>	97
2.12.10	Using <code>%in%</code> and <code>which</code>	98
2.12.11	More on pattern matching	98
2.12.12	Perl regular expressions	100
2.12.13	Stripping patterned text out of complex strings	100
2.13	Dates and times in R	101
2.13.1	Reading time data from files	102
2.13.2	The <code>strptime</code> function	103
2.13.3	The <code>difftime</code> function	104
2.13.4	Calculations with dates and times	105
2.13.5	The <code>difftime</code> and <code>as.difftime</code> functions	105
2.13.6	Generating sequences of dates	107
2.13.7	Calculating time differences between the rows of a dataframe	109
2.13.8	Regression using dates and times	111
2.13.9	Summary of dates and times in R	113
2.14	Environments	113
2.14.1	Using <code>with</code> rather than <code>attach</code>	113
2.14.2	Using <code>attach</code> in this book	114
2.15	Writing R functions	115
2.15.1	Arithmetic mean of a single sample	115
2.15.2	Median of a single sample	115
2.15.3	Geometric mean	116
2.15.4	Harmonic mean	118
2.15.5	Variance	119
2.15.6	Degrees of freedom	119
2.15.7	Variance ratio test	120
2.15.8	Using variance	121
2.15.9	Deparsing: A graphics function for error bars	123
2.15.10	The <code>switch</code> function	125
2.15.11	The evaluation environment of a function	126
2.15.12	Scope	126
2.15.13	Optional arguments	126

2.15.14	Variable numbers of arguments (. . .)	127
2.15.15	Returning values from a function	128
2.15.16	Anonymous functions	129
2.15.17	Flexible handling of arguments to functions	129
2.15.18	Structure of an object: <code>str</code>	130
2.16	Writing from R to file	133
2.16.1	Saving your work	133
2.16.2	Saving history	133
2.16.3	Saving graphics	134
2.16.4	Saving data produced within R to disc	134
2.16.5	Pasting into an Excel spreadsheet	135
2.16.6	Writing an Excel readable file from R	135
2.17	Programming tips	135
3	Data Input	137
3.1	Data input from the keyboard	137
3.2	Data input from files	138
3.2.1	The working directory	138
3.2.2	Data input using <code>read.table</code>	139
3.2.3	Common errors when using <code>read.table</code>	139
3.2.4	Separators and decimal points	140
3.2.5	Data input directly from the web	140
3.3	Input from files using <code>scan</code>	141
3.3.1	Reading a dataframe with <code>scan</code>	141
3.3.2	Input from more complex file structures using <code>scan</code>	143
3.4	Reading data from a file using <code>readLines</code>	145
3.4.1	Input a dataframe using <code>readLines</code>	145
3.4.2	Reading non-standard files using <code>readLines</code>	147
3.5	Warnings when you <code>attach</code> the dataframe	149
3.6	Masking	150
3.7	Input and output formats	150
3.8	Checking files from the command line	151
3.9	Reading dates and times from files	151
3.10	Built-in data files	152
3.11	File paths	152
3.12	Connections	153
3.13	Reading data from an external database	154
3.13.1	Creating the DSN for your computer	155
3.13.2	Setting up R to read from the database	155
4	Dataframes	159
4.1	Subscripts and indices	164
4.2	Selecting rows from the dataframe at random	165
4.3	Sorting dataframes	166
4.4	Using logical conditions to select rows from the dataframe	169
4.5	Omitting rows containing missing values, <code>NA</code>	172
4.5.1	Replacing <code>NAs</code> with zeros	174
4.6	Using <code>order</code> and <code>!duplicated</code> to eliminate pseudoreplication	174

4.7	Complex ordering with mixed directions	174
4.8	A dataframe with row names instead of row numbers	176
4.9	Creating a dataframe from another kind of object	177
4.10	Eliminating duplicate rows from a dataframe	180
4.11	Dates in dataframes	180
4.12	Using the <code>match</code> function in dataframes	182
4.13	Merging two dataframes	183
4.14	Adding margins to a dataframe	185
4.15	Summarizing the contents of dataframes	187
5	Graphics	189
5.1	Plots with two variables	189
5.2	Plotting with two continuous explanatory variables: Scatterplots	190
5.2.1	Plotting symbols: <code>pch</code>	195
5.2.2	Colour for symbols in plots	196
5.2.3	Adding text to scatterplots	197
5.2.4	Identifying individuals in scatterplots	198
5.2.5	Using a third variable to label a scatterplot	200
5.2.6	Joining the dots	201
5.2.7	Plotting stepped lines	202
5.3	Adding other shapes to a plot	203
5.3.1	Placing items on a plot with the cursor, using the <code>locator</code> function	204
5.3.2	Drawing more complex shapes with <code>polygon</code>	205
5.4	Drawing mathematical functions	206
5.4.1	Adding smooth parametric curves to a scatterplot	207
5.4.2	Fitting non-parametric curves through a scatterplot	209
5.5	Shape and size of the graphics window	211
5.6	Plotting with a categorical explanatory variable	212
5.6.1	Boxplots with notches to indicate significant differences	213
5.6.2	Barplots with error bars	214
5.6.3	Plots for multiple comparisons	217
5.6.4	Using colour palettes with categorical explanatory variables	219
5.7	Plots for single samples	220
5.7.1	Histograms and bar charts	220
5.7.2	Histograms	221
5.7.3	Histograms of integers	224
5.7.4	Overlaying histograms with smooth density functions	225
5.7.5	Density estimation for continuous variables	226
5.7.6	Index plots	227
5.7.7	Time series plots	228
5.7.8	Pie charts	230
5.7.9	The <code>stripchart</code> function	231
5.7.10	A plot to test for normality	232
5.8	Plots with multiple variables	234
5.8.1	The <code>pairs</code> function	234
5.8.2	The <code>coplot</code> function	236
5.8.3	Interaction plots	237

5.9	Special plots	238
5.9.1	Design plots	238
5.9.2	Bubble plots	239
5.9.3	Plots with many identical values	240
5.10	Saving graphics to file	242
5.11	Summary	242
6	Tables	244
6.1	Tables of counts	244
6.2	Summary tables	245
6.3	Expanding a table into a dataframe	250
6.4	Converting from a dataframe to a table	252
6.5	Calculating tables of proportions with <code>prop.table</code>	253
6.6	The <code>scale</code> function	254
6.7	The <code>expand.grid</code> function	254
6.8	The <code>model.matrix</code> function	255
6.9	Comparing <code>table</code> and <code>tabulate</code>	256
7	Mathematics	258
7.1	Mathematical functions	258
7.1.1	Logarithmic functions	259
7.1.2	Trigonometric functions	260
7.1.3	Power laws	261
7.1.4	Polynomial functions	262
7.1.5	Gamma function	264
7.1.6	Asymptotic functions	265
7.1.7	Parameter estimation in asymptotic functions	266
7.1.8	Sigmoid (S-shaped) functions	267
7.1.9	Biexponential model	269
7.1.10	Transformations of the response and explanatory variables	270
7.2	Probability functions	271
7.3	Continuous probability distributions	272
7.3.1	Normal distribution	274
7.3.2	The central limit theorem	278
7.3.3	Maximum likelihood with the normal distribution	282
7.3.4	Generating random numbers with exact mean and standard deviation	284
7.3.5	Comparing data with a normal distribution	285
7.3.6	Other distributions used in hypothesis testing	286
7.3.7	The chi-squared distribution	287
7.3.8	Fisher's F distribution	289
7.3.9	Student's t distribution	291
7.3.10	The gamma distribution	293
7.3.11	The exponential distribution	296
7.3.12	The beta distribution	296
7.3.13	The Cauchy distribution	298
7.3.14	The lognormal distribution	299
7.3.15	The logistic distribution	300
7.3.16	The log-logistic distribution	301

7.3.17	The Weibull distribution	301
7.3.18	Multivariate normal distribution	303
7.3.19	The uniform distribution	304
7.3.20	Plotting empirical cumulative distribution functions	306
7.4	Discrete probability distributions	307
7.4.1	The Bernoulli distribution	307
7.4.2	The binomial distribution	308
7.4.3	The geometric distribution	311
7.4.4	The hypergeometric distribution	312
7.4.5	The multinomial distribution	313
7.4.6	The Poisson distribution	314
7.4.7	The negative binomial distribution	315
7.4.8	The Wilcoxon rank-sum statistic	322
7.5	Matrix algebra	322
7.5.1	Matrix multiplication	323
7.5.2	Diagonals of matrices	324
7.5.3	Determinant	325
7.5.4	Inverse of a matrix	327
7.5.5	Eigenvalues and eigenvectors	328
7.5.6	Matrices in statistical models	331
7.5.7	Statistical models in matrix notation	334
7.6	Solving systems of linear equations using matrices	338
7.7	Calculus	339
7.7.1	Derivatives	339
7.7.2	Integrals	339
7.7.3	Differential equations	340
8	Classical Tests	344
8.1	Single samples	344
8.1.1	Data summary	345
8.1.2	Plots for testing normality	346
8.1.3	Testing for normality	347
8.1.4	An example of single-sample data	348
8.2	Bootstrap in hypothesis testing	349
8.3	Skew and kurtosis	350
8.3.1	Skew	350
8.3.2	Kurtosis	352
8.4	Two samples	353
8.4.1	Comparing two variances	354
8.4.2	Comparing two means	358
8.4.3	Student's t test	358
8.4.4	Wilcoxon rank-sum test	361
8.5	Tests on paired samples	362
8.6	The sign test	364
8.7	Binomial test to compare two proportions	365
8.8	Chi-squared contingency tables	365
8.8.1	Pearson's chi-squared	367
8.8.2	G test of contingency	369

8.8.3	Unequal probabilities in the null hypothesis	370
8.8.4	Chi-squared tests on table objects	370
8.8.5	Contingency tables with small expected frequencies: Fisher's exact test	371
8.9	Correlation and covariance	373
8.9.1	Data dredging	375
8.9.2	Partial correlation	375
8.9.3	Correlation and the variance of differences between variables	376
8.9.4	Scale-dependent correlations	377
8.10	Kolmogorov–Smirnov test	379
8.11	Power analysis	382
8.12	Bootstrap	385
9	Statistical Modelling	388
9.1	First things first	389
9.2	Maximum likelihood	390
9.3	The principle of parsimony (Occam's razor)	390
9.4	Types of statistical model	391
9.5	Steps involved in model simplification	393
9.5.1	Caveats	393
9.5.2	Order of deletion	394
9.6	Model formulae in R	395
9.6.1	Interactions between explanatory variables	396
9.6.2	Creating formula objects	397
9.7	Multiple error terms	398
9.8	The intercept as parameter 1	398
9.9	The <code>update</code> function in model simplification	399
9.10	Model formulae for regression	399
9.11	Box–Cox transformations	401
9.12	Model criticism	403
9.13	Model checking	404
9.13.1	Heteroscedasticity	404
9.13.2	Non-normality of errors	405
9.14	Influence	408
9.15	Summary of statistical models in R	411
9.16	Optional arguments in model-fitting functions	412
9.16.1	Subsets	413
9.16.2	Weights	413
9.16.3	Missing values	414
9.16.4	Offsets	415
9.16.5	Dataframes containing the same variable names	415
9.17	Akaike's information criterion	415
9.17.1	AIC as a measure of the fit of a model	416
9.18	Leverage	417
9.19	Misspecified model	418
9.20	Model checking in R	418
9.21	Extracting information from model objects	420
9.21.1	Extracting information by name	421
9.21.2	Extracting information by list subscripts	421

9.21.3	Extracting components of the model using <code>\$</code>	425
9.21.4	Using lists with models	425
9.22	The <code>summary</code> tables for continuous and categorical explanatory variables	426
9.23	Contrasts	430
9.23.1	Contrast coefficients	431
9.23.2	An example of contrasts in R	432
9.23.3	<i>A priori</i> contrasts	433
9.24	Model simplification by stepwise deletion	437
9.25	Comparison of the three kinds of contrasts	440
9.25.1	Treatment contrasts	440
9.25.2	Helmert contrasts	440
9.25.3	Sum contrasts	442
9.26	Aliasing	443
9.27	Orthogonal polynomial contrasts: <code>contr.poly</code>	443
9.28	Summary of statistical modelling	448
10	Regression	449
10.1	Linear regression	450
10.1.1	The famous five in R	453
10.1.2	Corrected sums of squares and sums of products	453
10.1.3	Degree of scatter	456
10.1.4	Analysis of variance in regression: $SSY = SSR + SSE$	458
10.1.5	Unreliability estimates for the parameters	460
10.1.6	Prediction using the fitted model	462
10.1.7	Model checking	463
10.2	Polynomial approximations to elementary functions	465
10.3	Polynomial regression	466
10.4	Fitting a mechanistic model to data	468
10.5	Linear regression after transformation	469
10.6	Prediction following regression	472
10.7	Testing for lack of fit in a regression	475
10.8	Bootstrap with regression	478
10.9	Jackknife with regression	481
10.10	Jackknife after bootstrap	483
10.11	Serial correlation in the residuals	484
10.12	Piecewise regression	485
10.13	Multiple regression	489
10.13.1	The multiple regression model	490
10.13.2	Common problems arising in multiple regression	497
11	Analysis of Variance	498
11.1	One-way ANOVA	498
11.1.1	Calculations in one-way ANOVA	502
11.1.2	Assumptions of ANOVA	503
11.1.3	A worked example of one-way ANOVA	503
11.1.4	Effect sizes	509
11.1.5	Plots for interpreting one-way ANOVA	511
11.2	Factorial experiments	516
11.3	Pseudoreplication: Nested designs and split plots	519

11.3.1	Split-plot experiments	519
11.3.2	Mixed-effects models	522
11.3.3	Fixed effect or random effect?	523
11.3.4	Removing the pseudoreplication	523
11.3.5	Derived variable analysis	524
11.4	Variance components analysis	524
11.5	Effect sizes in ANOVA: <code>ao</code> or <code>lm</code> ?	527
11.6	Multiple comparisons	531
11.7	Multivariate analysis of variance	535
12	Analysis of Covariance	537
12.1	Analysis of covariance in R	538
12.2	ANCOVA and experimental design	548
12.3	ANCOVA with two factors and one continuous covariate	548
12.4	Contrasts and the parameters of ANCOVA models	551
12.5	Order matters in <code>summary.ao</code>	554
13	Generalized Linear Models	557
13.1	Error structure	558
13.2	Linear predictor	559
13.3	Link function	559
13.3.1	Canonical link functions	560
13.4	Proportion data and binomial errors	560
13.5	Count data and Poisson errors	561
13.6	Deviance: Measuring the goodness of fit of a GLM	562
13.7	Quasi-likelihood	562
13.8	The <code>quasi</code> family of models	563
13.9	Generalized additive models	565
13.10	Offsets	566
13.11	Residuals	568
13.11.1	Misspecified error structure	569
13.11.2	Misspecified link function	569
13.12	Overdispersion	570
13.13	Bootstrapping a GLM	570
13.14	Binomial GLM with ordered categorical variables	574
14	Count Data	579
14.1	A regression with Poisson errors	579
14.2	Analysis of deviance with count data	581
14.3	Analysis of covariance with count data	586
14.4	Frequency distributions	588
14.5	Overdispersion in log-linear models	592
14.6	Negative binomial errors	595
15	Count Data in Tables	599
15.1	A two-class table of counts	599
15.2	Sample size for count data	600
15.3	A four-class table of counts	600
15.4	Two-by-two contingency tables	601
15.5	Using log-linear models for simple contingency tables	602

15.6	The danger of contingency tables	604
15.7	Quasi-Poisson and negative binomial models compared	606
15.8	A contingency table of intermediate complexity	608
15.9	Schoener's lizards: A complex contingency table	610
15.10	Plot methods for contingency tables	616
15.11	Graphics for count data: Spine plots and spinograms	621
16	Proportion Data	628
16.1	Analyses of data on one and two proportions	629
16.2	Count data on proportions	629
16.3	Odds	630
16.4	Overdispersion and hypothesis testing	631
16.5	Applications	632
16.5.1	Logistic regression with binomial errors	633
16.5.2	Estimating LD50 and LD90 from bioassay data	635
16.5.3	Proportion data with categorical explanatory variables	636
16.6	Averaging proportions	639
16.7	Summary of modelling with proportion count data	640
16.8	Analysis of covariance with binomial data	640
16.9	Converting complex contingency tables to proportions	643
16.9.1	Analysing Schoener's lizards as proportion data	645
17	Binary Response Variables	650
17.1	Incidence functions	652
17.2	Graphical tests of the fit of the logistic to data	653
17.3	ANCOVA with a binary response variable	655
17.4	Binary response with pseudoreplication	660
18	Generalized Additive Models	666
18.1	Non-parametric smoothers	667
18.2	Generalized additive models	669
18.2.1	Technical aspects	672
18.3	An example with strongly humped data	675
18.4	Generalized additive models with binary data	677
18.5	Three-dimensional graphic output from <code>gam</code>	679
19	Mixed-Effects Models	681
19.1	Replication and pseudoreplication	683
19.2	The <code>lme</code> and <code>lmer</code> functions	684
19.2.1	<code>lme</code>	684
19.2.2	<code>lmer</code>	685
19.3	Best linear unbiased predictors	685
19.4	Designed experiments with different spatial scales: Split plots	685
19.5	Hierarchical sampling and variance components analysis	691
19.6	Mixed-effects models with temporal pseudoreplication	695
19.7	Time series analysis in mixed-effects models	699
19.8	Random effects in designed experiments	703
19.9	Regression in mixed-effects models	704
19.10	Generalized linear mixed models	710
19.10.1	Hierarchically structured count data	710

20	Non-Linear Regression	715
20.1	Comparing Michaelis–Menten and asymptotic exponential	719
20.2	Generalized additive models	720
20.3	Grouped data for non-linear estimation	721
20.4	Non-linear time series models (temporal pseudoreplication)	726
20.5	Self-starting functions	728
20.5.1	Self-starting Michaelis–Menten model	729
20.5.2	Self-starting asymptotic exponential model	730
20.5.3	Self-starting logistic	730
20.5.4	Self-starting four-parameter logistic	731
20.5.5	Self-starting Weibull growth function	733
20.5.6	Self-starting first-order compartment function	734
20.6	Bootstrapping a family of non-linear regressions	735
21	Meta-Analysis	740
21.1	Effect size	741
21.2	Weights	741
21.3	Fixed versus random effects	741
21.3.1	Fixed-effect meta-analysis of scaled differences	742
21.3.2	Random effects with a scaled mean difference	746
21.4	Random-effects meta-analysis of binary data	748
22	Bayesian Statistics	752
22.1	Background	754
22.2	A continuous response variable	755
22.3	Normal prior and normal likelihood	755
22.4	Priors	756
22.4.1	Conjugate priors	757
22.5	Bayesian statistics for realistically complicated models	757
22.6	Practical considerations	758
22.7	Writing BUGS models	758
22.8	Packages in R for carrying out Bayesian analysis	758
22.9	Installing JAGS on your computer	759
22.10	Running JAGS in R	759
22.11	MCMC for a simple linear regression	760
22.12	MCMC for a model with temporal pseudoreplication	763
22.13	MCMC for a model with binomial errors	766
23	Tree Models	768
23.1	Background	769
23.2	Regression trees	771
23.3	Using <code>rpart</code> to fit tree models	772
23.4	Tree models as regressions	775
23.5	Model simplification	776
23.6	Classification trees with categorical explanatory variables	778
23.7	Classification trees for replicated data	780
23.8	Testing for the existence of humps	783
24	Time Series Analysis	785
24.1	Nicholson’s blowflies	785

24.2	Moving average	792
24.3	Seasonal data	793
24.3.1	Pattern in the monthly means	796
24.4	Built-in time series functions	797
24.5	Decompositions	797
24.6	Testing for a trend in the time series	798
24.7	Spectral analysis	800
24.8	Multiple time series	801
24.9	Simulated time series	803
24.10	Time series models	805
25	Multivariate Statistics	809
25.1	Principal components analysis	809
25.2	Factor analysis	813
25.3	Cluster analysis	816
25.3.1	Partitioning	816
25.3.2	Taxonomic use of <code>kmeans</code>	817
25.4	Hierarchical cluster analysis	819
25.5	Discriminant analysis	821
25.6	Neural networks	824
26	Spatial Statistics	825
26.1	Point processes	825
26.1.1	Random points in a circle	826
26.2	Nearest neighbours	829
26.2.1	Tessellation	833
26.3	Tests for spatial randomness	834
26.3.1	Ripley's K	834
26.3.2	Quadrat-based methods	838
26.3.3	Aggregated pattern and quadrat count data	839
26.3.4	Counting things on maps	842
26.4	Packages for spatial statistics	844
26.4.1	The <code>spatstat</code> package	845
26.4.2	The <code>spdep</code> package	849
26.4.3	Polygon lists	854
26.5	Geostatistical data	856
26.6	Regression models with spatially correlated errors: Generalized least squares	860
26.7	Creating a dot-distribution map from a relational database	867
27	Survival Analysis	869
27.1	A Monte Carlo experiment	869
27.2	Background	872
27.3	The survivor function	873
27.4	The density function	873
27.5	The hazard function	874
27.6	The exponential distribution	874
27.6.1	Density function	874
27.6.2	Survivor function	874
27.6.3	Hazard function	874

27.7	Kaplan–Meier survival distributions	875
27.8	Age-specific hazard models	876
27.9	Survival analysis in R	878
27.9.1	Parametric models	878
27.9.2	Cox proportional hazards model	878
27.9.3	Cox’s proportional hazard or a parametric model?	879
27.10	Parametric analysis	879
27.11	Cox’s proportional hazards	882
27.12	Models with censoring	883
27.12.1	Parametric models	884
27.12.2	Comparing <code>coxph</code> and <code>survreg</code> survival analysis	887
28	Simulation Models	893
28.1	Temporal dynamics: Chaotic dynamics in population size	893
28.1.1	Investigating the route to chaos	895
28.2	Temporal and spatial dynamics: A simulated random walk in two dimensions	896
28.3	Spatial simulation models	897
28.3.1	Metapopulation dynamics	898
28.3.2	Coexistence resulting from spatially explicit (local) density dependence	900
28.4	Pattern generation resulting from dynamic interactions	903
29	Changing the Look of Graphics	907
29.1	Graphs for publication	907
29.2	Colour	908
29.2.1	Palettes for groups of colours	910
29.2.2	The <code>RColorBrewer</code> package	913
29.2.3	Coloured plotting symbols with contrasting margins	914
29.2.4	Colour in legends	915
29.2.5	Background colours	916
29.2.6	Foreground colours	917
29.2.7	Different colours and font styles for different parts of the graph	917
29.2.8	Full control of colours in plots	918
29.3	Cross-hatching	920
29.4	Grey scale	921
29.5	Coloured convex hulls and other polygons	921
29.6	Logarithmic axes	922
29.7	Different font families for text	923
29.8	Mathematical and other symbols on plots	924
29.9	Phase planes	928
29.10	Fat arrows	929
29.11	Three-dimensional plots	930
29.12	Complex 3D plots with <code>wireframe</code>	933
29.13	An alphabetical tour of the graphics parameters	935
29.13.1	Text justification, <code>adj</code>	935
29.13.2	Annotation of graphs, <code>ann</code>	935
29.13.3	Delay moving on to the next in a series of plots, <code>ask</code>	935
29.13.4	Control over the axes, <code>axis</code>	938
29.13.5	Background colour for plots, <code>bg</code>	939

29.13.6	Boxes around plots, <code>bty</code>	939
29.13.7	Size of plotting symbols using the character expansion function, <code>cex</code>	940
29.13.8	Changing the shape of the plotting region, <code>plt</code>	941
29.13.9	Locating multiple graphs in non-standard layouts using <code>fig</code>	942
29.13.10	Two graphs with a common x scale but different y scales using <code>fig</code>	942
29.13.11	The <code>layout</code> function	943
29.13.12	Creating and controlling multiple screens on a single device	945
29.13.13	Orientation of numbers on the tick marks, <code>las</code>	947
29.13.14	Shapes for the ends and joins of lines, <code>lend</code> and <code>ljoin</code>	947
29.13.15	Line types, <code>lty</code>	948
29.13.16	Line widths, <code>lwd</code>	949
29.13.17	Several graphs on the same page, <code>mfrow</code> and <code>mfcop</code>	950
29.13.18	Margins around the plotting area, <code>mar</code>	950
29.13.19	Plotting more than one graph on the same axes, <code>new</code>	951
29.13.20	Two graphs on the same plot with different scales for their y axes	951
29.13.21	Outer margins, <code>oma</code>	952
29.13.22	Packing graphs closer together	954
29.13.23	Square plotting region, <code>pty</code>	955
29.13.24	Character rotation, <code>srt</code>	955
29.13.25	Rotating the axis labels	955
29.13.26	Tick marks on the axes	956
29.13.27	Axis styles	957
29.14	Trellis graphics	957
29.14.1	Panel box-and-whisker plots	959
29.14.2	Panel scatterplots	960
29.14.3	Panel barplots	965
29.14.4	Panels for conditioning plots	966
29.14.5	Panel histograms	967
29.14.6	Effect sizes	968
29.14.7	More panel functions	969
	<i>References and Further Reading</i>	971
	<i>Index</i>	977

Preface

R is a high-level language and an environment for data analysis and graphics. The design of R was heavily influenced by two existing languages: Becker, Chambers and Wilks' S and Sussman's Scheme. The resulting language is very similar in appearance to S, but the underlying implementation and semantics are derived from Scheme. This book is intended as an introduction to the riches of the R environment, aimed at beginners and intermediate users in disciplines ranging from science to economics and from medicine to engineering. I hope that the book can be read as a text as well as dipped into as a reference manual. The early chapters assume absolutely no background in statistics or computing, but the later chapters assume that the material in the earlier chapters has been studied. The book covers data handling, graphics, mathematical functions, and a wide range of statistical techniques all the way from elementary classical tests, through regression and analysis of variance and generalized linear modelling, up to more specialized topics such as Bayesian analysis, spatial statistics, multivariate methods, tree models, mixed-effects models and time series analysis. The idea is to introduce users to the assumptions that lie behind the tests, fostering a critical approach to statistical modelling, but involving little or no statistical theory and assuming no background in mathematics or statistics.

Why should you switch to using R when you have mastered a perfectly adequate statistical package already? At one level, there is no point in switching. If you only carry out a very limited range of statistical tests, and you do not intend to do more (or different) in the future, then fine. The main reason for switching to R is to take advantage of its unrivalled coverage and the availability of new, cutting-edge applications in fields such as generalized mixed-effects modelling and generalized additive models. The next reason for learning R is that you want to be able to understand the literature. More and more people are reporting their results in the context of R, and it is important to know what they are talking about. Third, look around your discipline to see who else is using R: many of the top people will have switched to R already. A large proportion of the world's leading statisticians use R, and this should tell you something (many, indeed, contribute to R, as you can see below). Another reason for changing to R is the quality of back-up and support available. There is a superb network of dedicated R wizards out there on the web, eager to answer your questions. If you intend to invest sufficient effort to become good at statistical computing, then the structure of R and the ease with which you can write your own functions are major attractions. Last, and certainly not least, the product is free. This is some of the finest integrated software in the world, and yet it is yours for absolutely nothing.

Although much of the text will equally apply to S-PLUS, there are some substantial differences, so in order not to confuse things I concentrate on describing R. I have made no attempt to show where S-PLUS is different from R, but if you have to work in S-PLUS, then try it and see if it works.

Acknowledgements

S is an elegant, widely accepted, and enduring software system with outstanding conceptual integrity, thanks to the insight, taste, and effort of John Chambers. In 1998, the Association for Computing Machinery (ACM) presented him with its Software System Award, for ‘the S system, which has forever altered the way people analyze, visualize, and manipulate data’. R was inspired by the S environment that was developed by John Chambers, and which had substantial input from Douglas Bates, Rick Becker, Bill Cleveland, Trevor Hastie, Daryl Pregibon and Allan Wilks.

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in New Zealand. Subsequently, a large group of individuals contributed to R by sending code and bug reports. John Chambers graciously contributed advice and encouragement in the early days of R, and later became a member of the core team. The current R is the result of a collaborative effort with contributions from all over the world.

Since mid-1997 there has been a core group with write access to the R source, currently consisting of Douglas Bates, John Chambers, Peter Dalgaard, Seth Falcon, Robert Gentleman, Kurt Hornik, Stefano Iacus, Ross Ihaka, Friedrich Leisch, Uwe Ligges, Thomas Lumley, Martin Maechler, Guido Masarotto (up to June 2003), Duncan Murdoch, Paul Murrell, Martyn Plummer, Brian Ripley, Deepayan Sarkar, Heiner Schwarte (up to October 1999), Duncan Temple Lang, Luke Tierney and Simon Urbanek.

R would not be what it is today without the invaluable help of the following people, who contributed by donating code, bug fixes and documentation: Valerio Aimale, Thomas Baier, Roger Bivand, Ben Bolker, David Brahm, Göran Broström, Patrick Burns, Vince Carey, Saikat DebRoy, Brian D’Urso, Lyndon Drake, Dirk Eddebuettel, John Fox, Paul Gilbert, Torsten Hothorn, Robert King, Kjetil Kjærnsmo, Philippe Lambert, Jan de Leeuw, Jim Lindsey, Patrick Lindsey, Catherine Loader, Gordon Maclean, John Maindonald, David Meyer, Jens Oehlschlägel, Steve Oncley, Richard O’Keefe, Hubert Palme, José C. Pinheiro, Anthony Rossini, Jonathan Rougier, Günther Sawitzki, Bill Simpson, Gordon Smyth, Adrian Trapletti, Terry Therneau, Bill Venables, Gregory R. Warnes, Andreas Weingessel, Morten Welinder, Simon Wood, and Achim Zeileis.

If you use R you should cite it in your written work. To cite the base package, put:

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

You can see the most up-to-date citation by typing `citation()` at the prompt. To cite individual contributed packages, you may find the appropriate citation in the description of the package, but failing that you will need to construct the citation from the author’s name, date, and title of the package from the reference manual for the package that is available on CRAN (see p. 3).

Special thanks are due to the generations of graduate students on the annual GLIM course at Silwood. It was their feedback that enabled me to understand those aspects of R that are most difficult for beginners, and highlighted the concepts that require the most detailed explanation. Please tell me about the errors and omissions you find, and send suggestions for changes and additions to m.crawley@imperial.ac.uk.

The data files used in this book can be downloaded from <http://www.bio.ic.ac.uk/research/mjcraw/therbook/index.htm>.

M.J. Crawley
Ascot
September 2012