

---

## Contents

<b>Preface</b>	<i>page</i> <b>xix</b>
<b>Content – how the chapters fit together</b>	<b>xxv</b>
<b>1 A brief introduction to R</b>	<b>1</b>
1.1 <i>An overview of R</i>	1
1.1.1 A short R session	1
1.1.2 The uses of R	6
1.1.3 Online help	7
1.1.4 Input of data from a file	8
1.1.5 R packages	9
1.1.6 Further steps in learning R	9
1.2 <i>Vectors, factors, and univariate time series</i>	10
1.2.1 Vectors	10
1.2.2 Concatenation – joining vector objects	10
1.2.3 The use of relational operators to compare vector elements	11
1.2.4 The use of square brackets to extract subsets of vectors	11
1.2.5 Patterned data	11
1.2.6 Missing values	12
1.2.7 Factors	13
1.2.8 Time series	14
1.3 <i>Data frames and matrices</i>	14
1.3.1 Accessing the columns of data frames – <code>with()</code> and <code>attach()</code>	17
1.3.2 Aggregation, stacking, and unstacking	17
1.3.3* Data frames and matrices	18
1.4 <i>Functions, operators, and loops</i>	19
1.4.1 Common useful built-in functions	19
1.4.2 Generic functions, and the class of an object	21
1.4.3 User-written functions	22
1.4.4 <code>if</code> Statements	23
1.4.5 Selection and matching	23
1.4.6 Functions for working with missing values	24
1.4.7* Looping	24

1.5	<i>Graphics in R</i>	25
1.5.1	The function <code>plot( )</code> and allied functions	25
1.5.2	The use of color	27
1.5.3	The importance of aspect ratio	28
1.5.4	Dimensions and other settings for graphics devices	28
1.5.5	The plotting of expressions and mathematical symbols	29
1.5.6	Identification and location on the figure region	29
1.5.7	Plot methods for objects other than vectors	30
1.5.8	Lattice (trellis) graphics	30
1.5.9	Good and bad graphs	32
1.5.10	Further information on graphics	33
1.6	<i>Additional points on the use of R</i>	33
1.7	<i>Recap</i>	35
1.8	<i>Further reading</i>	36
1.9	<i>Exercises</i>	37
<b>2</b>	<b>Styles of data analysis</b>	<b>43</b>
2.1	<i>Revealing views of the data</i>	43
2.1.1	Views of a single sample	44
2.1.2	Patterns in univariate time series	47
2.1.3	Patterns in bivariate data	49
2.1.4	Patterns in grouped data – lengths of cuckoo eggs	52
2.1.5*	Multiple variables and times	53
2.1.6	Scatterplots, broken down by multiple factors	56
2.1.7	What to look for in plots	58
2.2	<i>Data summary</i>	59
2.2.1	Counts	59
2.2.2	Summaries of information from data frames	63
2.2.3	Standard deviation and inter-quartile range	65
2.2.4	Correlation	67
2.3	<i>Statistical analysis questions, aims, and strategies</i>	69
2.3.1	How relevant and how reliable are the data?	70
2.3.2	How will results be used?	70
2.3.3	Formal and informal assessments	71
2.3.4	Statistical analysis strategies	72
2.3.5	Planning the formal analysis	72
2.3.6	Changes to the intended plan of analysis	73
2.4	<i>Recap</i>	73
2.5	<i>Further reading</i>	74
2.6	<i>Exercises</i>	74
<b>3</b>	<b>Statistical models</b>	<b>77</b>
3.1	<i>Statistical models</i>	77
3.1.1	Incorporation of an error or noise component	78
3.1.2	Fitting models – the model formula	80

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Table of Contents

[More information](#)

<i>Contents</i>		xi
3.2	<i>Distributions: models for the random component</i>	81
3.2.1	Discrete distributions – models for counts	82
3.2.2	Continuous distributions	84
3.3	<i>Simulation of random numbers and random samples</i>	86
3.3.1	Sampling from the normal and other continuous distributions	87
3.3.2	Simulation of regression data	88
3.3.3	Simulation of the sampling distribution of the mean	88
3.3.4	Sampling from finite populations	90
3.4	<i>Model assumptions</i>	91
3.4.1	Random sampling assumptions – independence	91
3.4.2	Checks for normality	92
3.4.3	Checking other model assumptions	95
3.4.4	Are non-parametric methods the answer?	95
3.4.5	Why models matter – adding across contingency tables	96
3.5	<i>Recap</i>	97
3.6	<i>Further reading</i>	98
3.7	<i>Exercises</i>	98
<b>4</b>	<b>A review of inference concepts</b>	<b>102</b>
4.1	<i>Basic concepts of estimation</i>	102
4.1.1	Population parameters and sample statistics	102
4.1.2	Sampling distributions	102
4.1.3	Assessing accuracy – the standard error	103
4.1.4	The standard error for the difference of means	103
4.1.5*	The standard error of the median	104
4.1.6	The sampling distribution of the <i>t</i> -statistic	105
4.2	<i>Confidence intervals and tests of hypotheses</i>	106
4.2.1	A summary of one- and two-sample calculations	109
4.2.2	Confidence intervals and tests for proportions	112
4.2.3	Confidence intervals for the correlation	113
4.2.4	Confidence intervals versus hypothesis tests	113
4.3	<i>Contingency tables</i>	114
4.3.1	Rare and endangered plant species	116
4.3.2	Additional notes	119
4.4	<i>One-way unstructured comparisons</i>	119
4.4.1	Multiple comparisons	122
4.4.2	Data with a two-way structure, i.e., two factors	123
4.4.3	Presentation issues	124
4.5	<i>Response curves</i>	125
4.6	<i>Data with a nested variation structure</i>	126
4.6.1	Degrees of freedom considerations	127
4.6.2	General multi-way analysis of variance designs	127
4.7	<i>Resampling methods for standard errors, tests, and confidence intervals</i>	128
4.7.1	The one-sample permutation test	128
4.7.2	The two-sample permutation test	129

xii	<i>Contents</i>	
	4.7.3*	Estimating the standard error of the median: bootstrapping 130
	4.7.4	Bootstrap estimates of confidence intervals 131
	4.8*	<i>Theories of inference</i> 132
	4.8.1	Maximum likelihood estimation 133
	4.8.2	Bayesian estimation 133
	4.8.3	If there is strong prior information, use it! 135
	4.9	<i>Recap</i> 135
	4.10	<i>Further reading</i> 136
	4.11	<i>Exercises</i> 137
<b>5</b>	<b>Regression with a single predictor</b>	<b>142</b>
	5.1	<i>Fitting a line to data</i> 142
	5.1.1	Summary information – lawn roller example 143
	5.1.2	Residual plots 143
	5.1.3	Iron slag example: is there a pattern in the residuals? 145
	5.1.4	The analysis of variance table 147
	5.2	<i>Outliers, influence, and robust regression</i> 147
	5.3	<i>Standard errors and confidence intervals</i> 149
	5.3.1	Confidence intervals and tests for the slope 150
	5.3.2	SEs and confidence intervals for predicted values 150
	5.3.3*	Implications for design 151
	5.4	<i>Assessing predictive accuracy</i> 152
	5.4.1	Training/test sets and cross-validation 153
	5.4.2	Cross-validation – an example 153
	5.4.3*	Bootstrapping 155
	5.5	<i>Regression versus qualitative anova comparisons – issues of power</i> 158
	5.6	<i>Logarithmic and other transformations</i> 160
	5.6.1*	A note on power transformations 160
	5.6.2	Size and shape data – allometric growth 161
	5.7	<i>There are two regression lines!</i> 162
	5.8	<i>The model matrix in regression</i> 163
	5.9*	<i>Bayesian regression estimation using the MCMCpack package</i> 165
	5.10	<i>Recap</i> 166
	5.11	<i>Methodological references</i> 167
	5.12	<i>Exercises</i> 167
<b>6</b>	<b>Multiple linear regression</b>	<b>170</b>
	6.1	<i>Basic ideas: a book weight example</i> 170
	6.1.1	Omission of the intercept term 172
	6.1.2	Diagnostic plots 173
	6.2	<i>The interpretation of model coefficients</i> 174
	6.2.1	Times for Northern Irish hill races 174
	6.2.2	Plots that show the contribution of individual terms 177
	6.2.3	Mouse brain weight example 179
	6.2.4	Book dimensions, density, and book weight 181

<i>Contents</i>	xiii
6.3 <i>Multiple regression assumptions, diagnostics, and efficacy measures</i>	183
6.3.1 Outliers, leverage, influence, and Cook's distance	183
6.3.2 Assessment and comparison of regression models	186
6.3.3 How accurately does the equation predict?	187
6.4 <i>A strategy for fitting multiple regression models</i>	189
6.4.1 Suggested steps	190
6.4.2 Diagnostic checks	191
6.4.3 An example – Scottish hill race data	191
6.5 <i>Problems with many explanatory variables</i>	196
6.5.1 Variable selection issues	197
6.6 <i>Multicollinearity</i>	199
6.6.1 The variance inflation factor	201
6.6.2 Remedies for multicollinearity	203
6.7 <i>Errors in x</i>	203
6.8 <i>Multiple regression models – additional points</i>	208
6.8.1 Confusion between explanatory and response variables	208
6.8.2 Missing explanatory variables	208
6.8.3* The use of transformations	210
6.8.4* Non-linear methods – an alternative to transformation?	210
6.9 <i>Recap</i>	212
6.10 <i>Further reading</i>	212
6.11 <i>Exercises</i>	214
<b>7 Exploiting the linear model framework</b>	<b>217</b>
7.1 <i>Levels of a factor – using indicator variables</i>	217
7.1.1 Example – sugar weight	217
7.1.2 Different choices for the model matrix when there are factors	220
7.2 <i>Block designs and balanced incomplete block designs</i>	222
7.2.1 Analysis of the rice data, allowing for block effects	222
7.2.2 A balanced incomplete block design	223
7.3 <i>Fitting multiple lines</i>	224
7.4 <i>Polynomial regression</i>	228
7.4.1 Issues in the choice of model	229
7.5* <i>Methods for passing smooth curves through data</i>	231
7.5.1 Scatterplot smoothing – regression splines	232
7.5.2* Roughness penalty methods and generalized additive models	235
7.5.3 Distributional assumptions for automatic choice of roughness penalty	236
7.5.4 Other smoothing methods	236
7.6 <i>Smoothing with multiple explanatory variables</i>	238
7.6.1 An additive model with two smooth terms	238
7.6.2* A smooth surface	240
7.7 <i>Further reading</i>	240
7.8 <i>Exercises</i>	240

xiv	<i>Contents</i>	
<b>8</b>	<b>Generalized linear models and survival analysis</b>	<b>244</b>
8.1	<i>Generalized linear models</i>	244
8.1.1	Transformation of the expected value on the left	244
8.1.2	Noise terms need not be normal	245
8.1.3	Log odds in contingency tables	245
8.1.4	Logistic regression with a continuous explanatory variable	246
8.2	<i>Logistic multiple regression</i>	249
8.2.1	Selection of model terms, and fitting the model	252
8.2.2	Fitted values	254
8.2.3	A plot of contributions of explanatory variables	255
8.2.4	Cross-validation estimates of predictive accuracy	255
8.3	<i>Logistic models for categorical data – an example</i>	256
8.4	<i>Poisson and quasi-Poisson regression</i>	258
8.4.1	Data on aberrant crypt foci	258
8.4.2	Moth habitat example	261
8.5	<i>Additional notes on generalized linear models</i>	266
8.5.1*	Residuals, and estimating the dispersion	266
8.5.2	Standard errors and $z$ - or $t$ -statistics for binomial models	267
8.5.3	Leverage for binomial models	268
8.6	<i>Models with an ordered categorical or categorical response</i>	268
8.6.1	Ordinal regression models	269
8.6.2*	Loglinear models	272
8.7	<i>Survival analysis</i>	272
8.7.1	Analysis of the <code>Aids2</code> data	273
8.7.2	Right-censoring prior to the termination of the study	275
8.7.3	The survival curve for male homosexuals	276
8.7.4	Hazard rates	276
8.7.5	The Cox proportional hazards model	277
8.8	<i>Transformations for count data</i>	279
8.9	<i>Further reading</i>	280
8.10	<i>Exercises</i>	281
<b>9</b>	<b>Time series models</b>	<b>283</b>
9.1	<i>Time series – some basic ideas</i>	283
9.1.1	Preliminary graphical explorations	283
9.1.2	The autocorrelation and partial autocorrelation function	284
9.1.3	Autoregressive models	285
9.1.4*	Autoregressive moving average models – theory	287
9.1.5	Automatic model selection?	288
9.1.6	A time series forecast	289
9.2*	<i>Regression modeling with ARIMA errors</i>	291
9.3*	<i>Non-linear time series</i>	298
9.4	<i>Further reading</i>	300
9.5	<i>Exercises</i>	301

<b>10 Multi-level models and repeated measures</b>	<b>303</b>
10.1 <i>A one-way random effects model</i>	304
10.1.1 Analysis with <code>aov()</code>	305
10.1.2 A more formal approach	308
10.1.3 Analysis using <code>lmer()</code>	310
10.2 <i>Survey data, with clustering</i>	313
10.2.1 Alternative models	313
10.2.2 Instructive, though faulty, analyses	318
10.2.3 Predictive accuracy	319
10.3 <i>A multi-level experimental design</i>	319
10.3.1 The anova table	321
10.3.2 Expected values of mean squares	322
10.3.3* The analysis of variance sums of squares breakdown	323
10.3.4 The variance components	325
10.3.5 The mixed model analysis	326
10.3.6 Predictive accuracy	328
10.4 <i>Within- and between-subject effects</i>	329
10.4.1 Model selection	329
10.4.2 Estimates of model parameters	331
10.5 <i>A generalized linear mixed model</i>	332
10.6 <i>Repeated measures in time</i>	334
10.6.1 Example – random variation between profiles	336
10.6.2 Orthodontic measurements on children	340
10.7 <i>Further notes on multi-level and other models with correlated errors</i>	344
10.7.1 Different sources of variance – complication or focus of interest?	344
10.7.2 Predictions from models with a complex error structure	345
10.7.3 An historical perspective on multi-level models	345
10.7.4 Meta-analysis	347
10.7.5 Functional data analysis	347
10.7.6 Error structure in explanatory variables	347
10.8 <i>Recap</i>	347
10.9 <i>Further reading</i>	348
10.10 <i>Exercises</i>	349
<b>11 Tree-based classification and regression</b>	<b>351</b>
11.1 <i>The uses of tree-based methods</i>	352
11.1.1 Problems for which tree-based regression may be used	352
11.2 <i>Detecting email spam – an example</i>	353
11.2.1 Choosing the number of splits	356
11.3 <i>Terminology and methodology</i>	356
11.3.1 Choosing the split – regression trees	357
11.3.2 Within and between sums of squares	357
11.3.3 Choosing the split – classification trees	358
11.3.4 Tree-based regression versus loess regression smoothing	359

xvi	<i>Contents</i>	
	<i>11.4 Predictive accuracy and the cost–complexity trade-off</i>	361
	11.4.1 Cross-validation	361
	11.4.2 The cost–complexity parameter	362
	11.4.3 Prediction error versus tree size	363
	<i>11.5 Data for female heart attack patients</i>	363
	11.5.1 The one-standard-deviation rule	365
	11.5.2 Printed information on each split	366
	<i>11.6 Detecting email spam – the optimal tree</i>	366
	<i>11.7 The randomForest package</i>	369
	<i>11.8 Additional notes on tree-based methods</i>	372
	<i>11.9 Further reading and extensions</i>	373
	<i>11.10 Exercises</i>	374
<b>12</b>	<b>Multivariate data exploration and discrimination</b>	<b>377</b>
	<i>12.1 Multivariate exploratory data analysis</i>	378
	12.1.1 Scatterplot matrices	378
	12.1.2 Principal components analysis	379
	12.1.3 Multi-dimensional scaling	383
	<i>12.2 Discriminant analysis</i>	385
	12.2.1 Example – plant architecture	386
	12.2.2 Logistic discriminant analysis	387
	12.2.3 Linear discriminant analysis	388
	12.2.4 An example with more than two groups	390
	<i>12.3* High-dimensional data, classification, and plots</i>	392
	12.3.1 Classifications and associated graphs	394
	12.3.2 Flawed graphs	394
	12.3.3 Accuracies and scores for test data	398
	12.3.4 Graphs derived from the cross-validation process	404
	<i>12.4 Further reading</i>	406
	<i>12.5 Exercises</i>	407
<b>13</b>	<b>Regression on principal component or discriminant scores</b>	<b>410</b>
	<i>13.1 Principal component scores in regression</i>	410
	<i>13.2* Propensity scores in regression comparisons – labor training data</i>	414
	13.2.1 Regression comparisons	417
	13.2.2 A strategy that uses propensity scores	419
	<i>13.3 Further reading</i>	426
	<i>13.4 Exercises</i>	426
<b>14</b>	<b>The R system – additional topics</b>	<b>427</b>
	<i>14.1 Graphical user interfaces to R</i>	427
	14.1.1 The R Commander’s interface – a guide to getting started	428
	14.1.2 The <i>rattle</i> GUI	429
	14.1.3 The creation of simple GUIs – the <i>fgui</i> package	429
	<i>14.2 Working directories, workspaces, and the search list</i>	430



	<i>Contents</i>	xvii
14.2.1*	The search path	430
14.2.2	Workspace management	430
14.2.3	Utility functions	431
14.3	<i>R system configuration</i>	432
14.3.1	The R Windows installation directory tree	432
14.3.2	The library directories	433
14.3.3	The startup mechanism	433
14.4	<i>Data input and output</i>	433
14.4.1	Input of data	434
14.4.2	Data output	437
14.4.3	Database connections	438
14.5	<i>Functions and operators – some further details</i>	438
14.5.1	Function arguments	439
14.5.2	Character string and vector functions	440
14.5.3	Anonymous functions	441
14.5.4	Functions for working with dates (and times)	441
14.5.5	Creating groups	443
14.5.6	Logical operators	443
14.6	<i>Factors</i>	444
14.7	<i>Missing values</i>	446
14.8*	<i>Matrices and arrays</i>	448
14.8.1	Matrix arithmetic	450
14.8.2	Outer products	451
14.8.3	Arrays	451
14.9	<i>Manipulations with lists, data frames, matrices, and time series</i>	452
14.9.1	Lists – an extension of the notion of “vector”	452
14.9.2	Changing the shape of data frames (or matrices)	454
14.9.3*	Merging data frames – <code>merge()</code>	455
14.9.4	Joining data frames, matrices, and vectors – <code>cbind()</code>	455
14.9.5	The <code>apply</code> family of functions	456
14.9.6	Splitting vectors and data frames into lists – <code>split()</code>	457
14.9.7	Multivariate time series	458
14.10	<i>Classes and methods</i>	458
14.10.1	Printing and summarizing model objects	459
14.10.2	Extracting information from model objects	460
14.10.3	S4 classes and methods	460
14.11	<i>Manipulation of language constructs</i>	461
14.11.1	Model and graphics formulae	461
14.11.2	The use of a list to pass arguments	462
14.11.3	Expressions	463
14.11.4	Environments	463
14.11.5	Function environments and lazy evaluation	464
14.12*	<i>Creation of R packages</i>	465
14.13	<i>Document preparation – <code>Sweave()</code> and <code>xtable()</code></i>	467
14.14	<i>Further reading</i>	468
14.15	<i>Exercises</i>	469

Cambridge University Press

978-0-521-76293-9 - Data Analysis and Graphics Using R – an Example-Based Approach: Third Edition

John Maindonald and W. John Braun

Table of Contents

[More information](#)

xviii

Contents

<b>15</b>	<b>Graphs in R</b>	<b>472</b>
15.1	<i>Hardcopy graphics devices</i>	472
15.2	<i>Plotting characters, symbols, line types, and colors</i>	472
15.3	<i>Formatting and plotting of text and equations</i>	474
	15.3.1 Symbolic substitution of symbols in an expression	475
	15.3.2 Plotting expressions in parallel	475
15.4	<i>Multiple graphs on a single graphics page</i>	476
15.5	<i>Lattice graphics and the grid package</i>	477
	15.5.1 Groups within data, and/or columns in parallel	478
	15.5.2 Lattice parameter settings	480
	15.5.3 Panel functions, strip functions, strip labels, and other annotation	483
	15.5.4 Interaction with lattice (and other) plots – the <i>playwith</i> package	485
	15.5.5 Interaction with lattice plots – focus, interact, unfocus	485
	15.5.6 Overlaid plots with different scales	486
15.6	<i>An implementation of Wilkinson’s Grammar of Graphics</i>	487
15.7	<i>Dynamic graphics – the <code>rgl</code> and <code>rggobi</code> packages</i>	491
15.8	<i>Further reading</i>	492
	<b>Epilogue</b>	<b>493</b>
	<b>References</b>	<b>495</b>
	<b>Index of R symbols and functions</b>	<b>507</b>
	<b>Index of terms</b>	<b>514</b>
	<b>Index of authors</b>	<b>523</b>

The color plates will be found between pages 328 and 329.