
Contents

Preface

page **xix**

1	A brief introduction to R	1
1.1	<i>An overview of R</i>	1
1.1.1	A short R session	1
1.1.2	The uses of R	5
1.1.3	Online help	6
1.1.4	Further steps in learning R	8
1.2	<i>Data input, packages and the search list</i>	8
1.2.1	Reading data from a file	8
1.2.2	R packages	9
1.3	<i>Vectors, factors and univariate time series</i>	10
1.3.1	Vectors in R	10
1.3.2	Concatenation – joining vector objects	10
1.3.3	Subsets of vectors	11
1.3.4	Patterned data	11
1.3.5	Missing values	12
1.3.6	Factors	13
1.3.7	Time series	14
1.4	<i>Data frames and matrices</i>	14
1.4.1	The attaching of data frames	16
1.4.2	Aggregation, stacking and unstacking	17
1.4.3*	Data frames and matrices	17
1.5	<i>Functions, operators and loops</i>	18
1.5.1	Built-in functions	18
1.5.2	Generic functions and the class of an object	20
1.5.3	User-written functions	21
1.5.4	Relational and logical operators and operations	22
1.5.5	Selection and matching	23
1.5.6	Functions for working with missing values	23
1.5.7*	Looping	24
1.6	<i>Graphics in R</i>	24
1.6.1	The function <code>plot()</code> and allied functions	25
1.6.2	The use of color	27

1.6.3	The importance of aspect ratio	27
1.6.4	Dimensions and other settings for graphics devices	28
1.6.5	The plotting of expressions and mathematical symbols	28
1.6.6	Identification and location on the figure region	29
1.6.7	Plot methods for objects other than vectors	29
1.6.8	Lattice graphics versus base graphics – <code>xyplot()</code> versus <code>plot()</code>	30
1.6.9	Further information on graphics	30
1.6.10	Good and bad graphs	30
1.7	<i>Lattice (trellis) graphics</i>	31
1.8	<i>Additional points on the use of R</i>	33
1.9	<i>Recap</i>	36
1.10	<i>Further reading</i>	36
1.10.1	References for further reading	37
1.11	<i>Exercises</i>	37
2	Styles of data analysis	43
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	48
2.1.3	Patterns in bivariate data	50
2.1.4	Patterns in grouped data	52
2.1.5*	Multiple variables and times	54
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	60
2.2.2	Summaries of information from data frames	63
2.2.3	Standard deviation and inter-quartile range	66
2.2.4	Correlation	68
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	Helpful and unhelpful questions	70
2.3.3	How will results be used?	71
2.3.4	Formal and informal assessments	72
2.3.5	Statistical analysis strategies	73
2.3.6	Planning the formal analysis	73
2.3.7	Changes to the intended plan of analysis	74
2.4	<i>Recap</i>	74
2.5	<i>Further reading</i>	75
2.5.1	References for further reading	75
2.6	<i>Exercises</i>	75
3	Statistical models	78
3.1	<i>Regularities</i>	79
3.1.1	Deterministic models	79

3.1.2	Models that include a random component	79
3.1.3	Fitting models – the model formula	82
3.2	<i>Distributions: models for the random component</i>	83
3.2.1	Discrete distributions	84
3.2.2	Continuous distributions	86
3.3	<i>The uses of random numbers</i>	88
3.3.1	Simulation	88
3.3.2	Sampling from populations	89
3.4	<i>Model assumptions</i>	90
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	95
3.5	<i>Recap</i>	96
3.6	<i>Further reading</i>	97
3.6.1	References for further reading	97
3.7	<i>Exercises</i>	97
4	An introduction to formal inference	101
4.1	<i>Basic concepts of estimation</i>	101
4.1.1	Population parameters and sample statistics	101
4.1.2	Sampling distributions	102
4.1.3	Assessing accuracy – the standard error	102
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	104
4.2	<i>Confidence intervals and hypothesis tests</i>	107
4.2.1	One- and two-sample intervals and tests for means	107
4.2.2	Confidence intervals and tests for proportions	113
4.2.3	Confidence intervals for the correlation	113
4.2.4	Confidence intervals versus hypothesis tests	114
4.3	<i>Contingency tables</i>	115
4.3.1	Rare and endangered plant species	117
4.3.2	Additional notes	119
4.4	<i>One-way unstructured comparisons</i>	120
4.4.1	Displaying means for the one-way layout	123
4.4.2	Multiple comparisons	124
4.4.3	Data with a two-way structure, that is, two factors	125
4.4.4	Presentation issues	126
4.5	<i>Response curves</i>	126
4.6	<i>Data with a nested variation structure</i>	127
4.6.1	Degrees of freedom considerations	128
4.6.2	General multi-way analysis of variance designs	129

4.7	<i>Resampling methods for standard errors, tests and confidence intervals</i>	129
4.7.1	The one-sample permutation test	129
4.7.2	The two-sample permutation test	130
4.7.3*	Estimating the standard error of the median: bootstrapping	131
4.7.4	Bootstrap estimates of confidence intervals	133
4.8*	<i>Theories of inference</i>	134
4.8.1	Maximum likelihood estimation	135
4.8.2	Bayesian estimation	136
4.8.3	If there is strong prior information, use it!	136
4.9	<i>Recap</i>	137
4.10	<i>Further reading</i>	138
4.10.1	References for further reading	138
4.11	<i>Exercises</i>	139
5	Regression with a single predictor	144
5.1	<i>Fitting a line to data</i>	144
5.1.1	Lawn roller example	145
5.1.2	Calculating fitted values and residuals	146
5.1.3	Residual plots	147
5.1.4	Iron slag example: is there a pattern in the residuals?	148
5.1.5	The analysis of variance table	150
5.2	<i>Outliers, influence and robust regression</i>	151
5.3	<i>Standard errors and confidence intervals</i>	153
5.3.1	Confidence intervals and tests for the slope	153
5.3.2	SEs and confidence intervals for predicted values	154
5.3.3*	Implications for design	155
5.4	<i>Regression versus qualitative anova comparisons</i>	157
5.4.1	Issues of power	157
5.4.2	The pattern of change	158
5.5	<i>Assessing predictive accuracy</i>	158
5.5.1	Training/test sets and cross-validation	158
5.5.2	Cross-validation – an example	159
5.5.3*	Bootstrapping	161
5.6*	<i>A note on power transformations</i>	164
5.6.1*	General power transformations	164
5.7	<i>Size and shape data</i>	165
5.7.1	Allometric growth	166
5.7.2	There are two regression lines!	167
5.8	<i>The model matrix in regression</i>	168
5.9	<i>Recap</i>	169
5.10	<i>Methodological references</i>	170
5.11	<i>Exercises</i>	170

6	Multiple linear regression	173
6.1	<i>Basic ideas: book weight and brain weight examples</i>	173
6.1.1	Omission of the intercept term	176
6.1.2	Diagnostic plots	176
6.1.3	Example: brain weight	178
6.1.4	Plots that show the contribution of individual terms	180
6.2	<i>Multiple regression assumptions and diagnostics</i>	182
6.2.1	Influential outliers and Cook's distance	183
6.2.2	Influence on the regression coefficients	184
6.2.3*	Additional diagnostic plots	185
6.2.4	Robust and resistant methods	185
6.2.5	The uses of model diagnostics	185
6.3	<i>A strategy for fitting multiple regression models</i>	186
6.3.1	Preliminaries	186
6.3.2	Model fitting	187
6.3.3	An example – the Scottish hill race data	187
6.4	<i>Measures for the assessment and comparison of regression models</i>	193
6.4.1	R^2 and adjusted R^2	193
6.4.2	AIC and related statistics	194
6.4.3	How accurately does the equation predict?	194
6.5	<i>Interpreting regression coefficients</i>	196
6.5.1	Book dimensions and book weight	196
6.6	<i>Problems with many explanatory variables</i>	199
6.6.1	Variable selection issues	200
6.7	<i>Multicollinearity</i>	202
6.7.1	A contrived example	202
6.7.2	The variance inflation factor	206
6.7.3	Remedies for multicollinearity	206
6.8	<i>Multiple regression models – additional points</i>	207
6.8.1	Errors in x	207
6.8.2	Confusion between explanatory and response variables	210
6.8.3	Missing explanatory variables	210
6.8.4*	The use of transformations	212
6.8.5*	Non-linear methods – an alternative to transformation?	212
6.9	<i>Recap</i>	214
6.10	<i>Further reading</i>	214
6.10.1	References for further reading	215
6.11	<i>Exercises</i>	216
7	Exploiting the linear model framework	219
7.1	<i>Levels of a factor – using indicator variables</i>	220
7.1.1	Example – sugar weight	220
7.1.2	Different choices for the model matrix when there are factors	223

7.2	<i>Block designs and balanced incomplete block designs</i>	224
7.2.1	Analysis of the rice data, allowing for block effects	224
7.2.2	A balanced incomplete block design	226
7.3	<i>Fitting multiple lines</i>	227
7.4	<i>Polynomial regression</i>	231
7.4.1	Issues in the choice of model	233
7.5*	<i>Methods for passing smooth curves through data</i>	234
7.5.1	Scatterplot smoothing – regression splines	235
7.5.2*	Penalized splines and generalized additive models	239
7.5.3	Other smoothing methods	239
7.6	<i>Smoothing terms in additive models</i>	241
7.6.1*	The fitting of penalized spline terms	243
7.7	<i>Further reading</i>	243
7.7.1	References for further reading	243
7.8	<i>Exercises</i>	243
8	Generalized linear models and survival analysis	246
8.1	<i>Generalized linear models</i>	246
8.1.1	Transformation of the expected value on the left	246
8.1.2	Noise terms need not be normal	247
8.1.3	Log odds in contingency tables	247
8.1.4	Logistic regression with a continuous explanatory variable	248
8.2	<i>Logistic multiple regression</i>	251
8.2.1	Selection of model terms and fitting the model	253
8.2.2	A plot of contributions of explanatory variables	256
8.2.3	Cross-validation estimates of predictive accuracy	257
8.3	<i>Logistic models for categorical data – an example</i>	258
8.4	<i>Poisson and quasi-Poisson regression</i>	260
8.4.1	Data on aberrant crypt foci	260
8.4.2	Moth habitat example	263
8.5	<i>Additional notes on generalized linear models</i>	269
8.5.1*	Residuals, and estimating the dispersion	269
8.5.2	Standard errors and z- or t-statistics for binomial models	270
8.5.3	Leverage for binomial models	270
8.6	<i>Models with an ordered categorical or categorical response</i>	271
8.6.1	Ordinal regression models	271
8.6.2*	Loglinear models	274
8.7	<i>Survival analysis</i>	275
8.7.1	Analysis of the Aids2 data	276
8.7.2	Right censoring prior to the termination of the study	278
8.7.3	The survival curve for male homosexuals	279
8.7.4	Hazard rates	279
8.7.5	The Cox proportional hazards model	280
8.8	<i>Transformations for count data</i>	282
8.9	<i>Further reading</i>	283

8.9.1	References for further reading	283
8.10	<i>Exercises</i>	284
9	Time series models	286
9.1	<i>Time series – some basic ideas</i>	286
9.1.1	Preliminary graphical explorations	286
9.1.2	The autocorrelation function	287
9.1.3	Autoregressive models	288
9.1.4*	Autoregressive moving average models – theory	290
9.2*	<i>Regression modeling with moving average errors</i>	291
9.3*	<i>Non-linear time series</i>	297
9.4	<i>Other time series packages</i>	298
9.5	<i>Further reading</i>	298
9.5.1	Spatial statistics	299
9.5.2	References for further reading	299
9.6	<i>Exercises</i>	299
10	Multi-level models and repeated measures	301
10.1	<i>A one-way random effects model</i>	302
10.1.1	Analysis with <code>aoV()</code>	303
10.1.2	A more formal approach	306
10.1.3	Analysis using <code>lmer()</code>	308
10.2	<i>Survey data, with clustering</i>	311
10.2.1	Alternative models	311
10.2.2	Instructive, though faulty, analyses	316
10.2.3	Predictive accuracy	317
10.3	<i>A multi-level experimental design</i>	317
10.3.1	The anova table	319
10.3.2	Expected values of mean squares	320
10.3.3*	The sums of squares breakdown	321
10.3.4	The variance components	324
10.3.5	The mixed model analysis	325
10.3.6	Predictive accuracy	327
10.3.7	Different sources of variance – complication or focus of interest?	327
10.4	<i>Within- and between-subject effects</i>	328
10.4.1	Model selection	329
10.4.2	Estimates of model parameters	330
10.5	<i>Repeated measures in time</i>	332
10.5.1	Example – random variation between profiles	334
10.5.2	Orthodontic measurements on children	339
10.6	<i>Error structure considerations</i>	343
10.6.1	Predictions from models with a complex error structure	343
10.6.2	Error structure in explanatory variables	344

10.7	<i>Further notes on multi-level and other models with correlated errors</i>	344
10.7.1	An historical perspective on multi-level models	344
10.7.2	Meta-analysis	346
10.7.3	Functional data analysis	346
10.8	<i>Recap</i>	346
10.9	<i>Further reading</i>	347
10.9.1	References for further reading	347
10.10	<i>Exercises</i>	348
11	Tree-based classification and regression	350
11.1	<i>The uses of tree-based methods</i>	351
11.1.1	Problems for which tree-based regression may be used	351
11.2	<i>Detecting email spam – an example</i>	352
11.2.1	Choosing the number of splits	355
11.3	<i>Terminology and methodology</i>	355
11.3.1	Choosing the split – regression trees	355
11.3.2	Within and between sums of squares	356
11.3.3	Choosing the split – classification trees	357
11.3.4	Tree-based regression versus loess regression smoothing	358
11.4	<i>Predictive accuracy and the cost–complexity tradeoff</i>	360
11.4.1	Cross-validation	361
11.4.2	The cost–complexity parameter	361
11.4.3	Prediction error versus tree size	362
11.5	<i>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	365
11.6	<i>Detecting email spam – the optimal tree</i>	366
11.7	<i>The randomForest package</i>	368
11.8	<i>Additional notes on tree-based methods</i>	371
11.8.1	The combining of tree-based methods with other approaches	371
11.8.2	Models with a complex error structure	372
11.8.3	Pruning as variable selection	372
11.8.4	Other types of tree	372
11.8.5	Factors as predictors	372
11.8.6	Summary of pluses and minuses of tree-based methods	372
11.9	<i>Further reading</i>	373
11.9.1	References for further reading	373
11.10	<i>Exercises</i>	374
12	Multivariate data exploration and discrimination	375
12.1	<i>Multivariate exploratory data analysis</i>	376
12.1.1	Scatterplot matrices	376
12.1.2	Principal components analysis	377
12.1.3	Multi-dimensional scaling	383

12.2	<i>Discriminant analysis</i>	384
12.2.1	Example – plant architecture	384
12.2.2	Logistic discriminant analysis	386
12.2.3	Linear discriminant analysis	387
12.2.4	An example with more than two groups	388
12.3*	<i>High-dimensional data, classification and plots</i>	390
12.3.1	Classifications and associated graphs	392
12.3.2	Flawed graphs	393
12.3.3	Accuracies and scores for test data	397
12.3.4	Graphs derived from the cross-validation process	403
12.4	<i>Further reading</i>	405
12.4.1	References for further reading	406
12.5	<i>Exercises</i>	406
13	Regression on principal component or discriminant scores	408
13.1	<i>Principal component scores in regression</i>	408
13.2*	<i>Propensity scores in regression comparisons – labor training data</i>	412
13.2.1	Regression analysis, using all covariates	415
13.2.2	The use of propensity scores	417
13.3	<i>Further reading</i>	419
13.3.1	References for further reading	419
13.4	<i>Exercises</i>	420
14	The R system – additional topics	421
14.1	<i>Working directories, workspaces and the search list</i>	421
14.1.1*	The search path	421
14.1.2	Workspace management	421
14.1.3	Utility functions	423
14.2	<i>Data input and output</i>	423
14.2.1	Input of data	424
14.2.2	Data output	428
14.3	<i>Functions and operators – some further details</i>	429
14.3.1	Function arguments	430
14.3.2	Character string and vector functions	431
14.3.3	Anonymous functions	431
14.3.4	Functions for working with dates (and times)	432
14.3.5	Creating groups	433
14.3.6	Logical operators	434
14.4	<i>Factors</i>	434
14.5	<i>Missing values</i>	437
14.6*	<i>Matrices and arrays</i>	439
14.6.1	Matrix arithmetic	440
14.6.2	Outer products	441
14.6.3	Arrays	442

14.7	<i>Manipulations with lists, data frames and matrices</i>	443
14.7.1	Lists – an extension of the notion of “vector”	443
14.7.2	Changing the shape of data frames	445
14.7.3*	Merging data frames – <code>merge()</code>	445
14.7.4	Joining data frames, matrices and vectors – <code>cbind()</code>	446
14.7.5	The <code>apply</code> family of functions	446
14.7.6	Splitting vectors and data frames into lists – <code>split()</code>	448
14.7.7	Multivariate time series	448
14.8	<i>Classes and methods</i>	449
14.8.1	Printing and summarizing model objects	449
14.8.2	Extracting information from model objects	450
14.8.3	S4 classes and methods	450
14.9	<i>Manipulation of language constructs</i>	451
14.9.1	Model and graphics formulae	451
14.9.2	The use of a list to pass parameter values	452
14.9.3	Expressions	453
14.9.4	Environments	453
14.9.5	Function environments and lazy evaluation	455
14.10	<i>Document preparation — Sweave()</i>	456
14.11	<i>Graphs in R</i>	457
14.11.1	Hardcopy graphics devices	457
14.11.2	Multiple graphs on a single graphics page	457
14.11.3	Plotting characters, symbols, line types and colors	457
14.12	<i>Lattice graphics and the grid package</i>	462
14.12.1	Interaction with plots	464
14.12.2*	Use of <code>grid.text()</code> to label points	464
14.12.3*	Multiple lattice graphs on a graphics page	465
14.13	<i>Further reading</i>	466
14.13.1	Vignettes	466
14.13.2	References for further reading	466
14.14	<i>Exercises</i>	467
Epilogue – models		470
References		474
Index of R Symbols and Functions		485
Index of Terms		491
Index of Authors		501
Color Plates after Page 502		