
Contents

Preface	xiii
Content – How the Chapters Fit Together	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	7
1.2 <i>Data Input, Packages and the Search List</i>	8
1.2.1 Input of data from a file	8
1.2.2 R Packages	8
1.3 <i>Vectors in R</i>	9
1.3.1 Vectors	9
1.3.2 Concatenation – joining vector objects	10
1.3.3 Subsets of vectors	10
1.3.4 Patterned data	11
1.3.5 Missing values	11
1.3.6 Factors	12
1.3.7 Time series	13
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	16
1.4.3 * Data frames and matrices	17
1.5 <i>Functions, Operators and Loops</i>	18
1.5.1 Common useful 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	22
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 <i>plot</i> () and allied functions	24
1.6.2	The use of color	26
1.6.3	The importance of aspect ratio	27
1.6.4	Dimensions and other settings for graphics devices	27
1.6.5	The plotting of expressions and mathematical symbols	28
1.6.6	Identification and location on the figure region	28
1.6.7	Plot methods for objects other than vectors	29
1.6.8	Lattice (trellis) Graphics	29
1.6.9	Good and bad graphs	32
1.6.10	Further information on graphics	32
1.7	<i>Additional Points on the Use of R</i>	33
1.8	<i>Recap</i>	35
1.9	<i>Further Reading</i>	36
1.10	<i>Exercises</i>	37
2	Styles of Data Analysis	42
2.1	<i>Revealing Views of the Data</i>	42
2.1.1	Views of a single sample	43
2.1.2	Patterns in univariate time series	47
2.1.3	Patterns in bivariate data	48
2.1.4	Patterns in grouped data	51
2.1.5	* Multiple variables and times	52
2.1.6	Scatterplots, broken down by multiple factors	55
2.1.7	What to look for in plots	57
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?	69
2.3.2	Helpful and unhelpful questions	70
2.3.3	How will results be used?	70
2.3.4	Formal and informal assessments	71
2.3.5	Statistical Analysis Strategies	72
2.3.6	Planning the formal analysis	72
2.3.7	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
3	Statistical Models	77
3.1	<i>Statistical Models</i>	77
3.1.1	Incorporation of a error or noise component	78

3.1.2	Fitting models – the model formula	80
3.2	<i>Distributions: Models for the Random Component</i>	81
3.2.1	Discrete distributions – models for counts	81
3.2.2	Continuous distributions	84
3.3	<i>Simulation of Random Numbers and Random Samples</i>	86
3.3.1	Simulation of the sampling distribution of the mean	88
3.3.2	Sampling from finite populations	89
3.4	<i>Model Assumptions</i>	90
3.4.1	Random sampling assumptions – independence	91
3.4.2	Checks for normality	91
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
4	A Review of Inference Concepts	102
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	117
4.3.2	Additional notes	119
4.4	<i>One-Way Unstructured Comparisons</i>	120
4.4.1	Multiple comparisons	122
4.4.2	Data with a two-way structure, i.e. two factors	124
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	128
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
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>	133
4.8.1	Maximum likelihood estimation	134
4.8.2	Bayesian estimation	135
4.8.3	If there is strong prior information, use it!	135
4.9	<i>Recap</i>	136
4.10	<i>Further Reading</i>	136
4.11	<i>Exercises</i>	138
5	Regression with a Single Predictor	143
5.1	<i>Fitting a Line to Data</i>	143
5.1.1	Summary information – lawn roller example	144
5.1.2	Residual plots	144
5.1.3	Iron slag example: is there a pattern in the residuals?	146
5.1.4	The analysis of variance table	147
5.2	<i>Outliers, Influence and Robust Regression</i>	149
5.3	<i>Standard Errors and Confidence Intervals</i>	151
5.3.1	Confidence intervals and tests for the slope	151
5.3.2	SEs and confidence intervals for predicted values	151
5.3.3	* Implications for design	152
5.4	<i>Regression versus Qualitative anova Comparisons – Issues of Power</i> . . .	154
5.5	<i>Assessing Predictive Accuracy</i>	155
5.5.1	Training/test sets, and cross-validation	155
5.5.2	Cross-validation – an example	156
5.5.3	* Bootstrapping	158
5.6	* <i>A Note on Power Transformations</i>	161
5.7	<i>Size and Shape Data</i>	162
5.7.1	Allometric growth	163
5.7.2	There are two regression lines!	164
5.8	<i>The Model Matrix in Regression</i>	165
5.9	<i>Recap</i>	166
5.10	<i>Methodological References</i>	167
5.11	<i>Exercises</i>	167
6	Multiple Linear Regression	170
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
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?	188
6.4	<i>A Strategy for Fitting Multiple Regression Models</i>	190
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 (VIF)	201
6.6.2	Remedies for multicollinearity	203
6.7	<i>Multiple Regression Models – Additional Points</i>	204
6.7.1	Errors in x	204
6.7.2	Implications for variable selection	207
6.7.3	Confusion between explanatory and response variables	208
6.7.4	Missing explanatory variables	208
6.7.5	* The use of transformations	210
6.7.6	* Non-linear methods – an alternative to transformation?	210
6.8	<i>Recap</i>	212
6.9	<i>Further Reading</i>	212
6.10	<i>Exercises</i>	214
7	Exploiting the Linear Model Framework	218
7.1	<i>Levels of a Factor – Using Indicator Variables</i>	218
7.1.1	Example – sugar weight	218
7.1.2	Different choices for the model matrix when there are factors	221
7.2	<i>Block Designs and Balanced Incomplete Block Designs</i>	223
7.2.1	Analysis of the rice data, allowing for block effects	223
7.2.2	A balanced incomplete block design	224
7.3	<i>Fitting Multiple Lines</i>	226
7.4	<i>Polynomial Regression</i>	230
7.4.1	Issues in the choice of model	232
7.5	<i>Methods for Passing Smooth Curves through Data</i>	233
7.5.1	Scatterplot smoothing – regression splines	234
7.5.2	*Penalized splines and generalized additive models	237
7.5.3	Other smoothing methods	238
7.6	<i>Smoothing Terms in Additive Models</i>	239
7.6.1	*The fitting of penalized spline terms	242
7.7	<i>Further Reading</i>	242
7.8	<i>Exercises</i>	242
8	Generalized Linear Models and Survival Analysis	245
8.1	<i>Generalized Linear Models</i>	245
8.1.1	Transformation of the expected value on the left	245
8.1.2	Noise terms need not be normal	246

8.1.3	Log odds in contingency tables	246
8.1.4	Logistic regression with a continuous explanatory variable	247
8.2	<i>Logistic Multiple Regression</i>	250
8.2.1	Selection of model terms, and fitting the model	253
8.2.2	Fitted values	255
8.2.3	A plot of contributions of explanatory variables	256
8.2.4	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	262
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	275
8.7	<i>Survival analysis</i>	275
8.7.1	Analysis of the <code>Aids2</code> data	276
8.7.2	Right censoring prior to the termination of the study	278
8.7.3	The survival curve for male homosexuals	278
8.7.4	Hazard rates	279
8.7.5	The Cox proportional hazards model	279
8.8	<i>Transformations for Count Data</i>	281
8.9	<i>Further Reading</i>	282
8.10	<i>Exercises</i>	283
9	Time Series Models	285
9.1	<i>Time Series – Some Basic Ideas</i>	285
9.1.1	Preliminary graphical explorations	285
9.1.2	The autocorrelation and partial autocorrelation function	286
9.1.3	Autoregressive (AR) models	288
9.1.4	* Autoregressive moving average (ARMA) models – theory	290
9.1.5	Automatic model selection?	290
9.1.6	A time series forecast	292
9.2	<i>Regression modeling with moving average errors</i>	294
9.3	* <i>Nonlinear time series</i>	301
9.4	<i>Other time series packages</i>	303
9.5	<i>Further Reading</i>	303
9.6	<i>Exercises</i>	304
10	Multi-level Models, and Repeated Measures	306
10.1	<i>A One-Way Random Effects Model</i>	307
10.1.1	Analysis with <code>aov()</code>	308

10.1.2	A More Formal Approach	311
10.1.3	Analysis using <code>lmer()</code>	313
10.2	<i>Survey Data, with Clustering</i>	316
10.2.1	Alternative models	316
10.2.2	Instructive, though faulty, analyses	321
10.2.3	Predictive accuracy	322
10.3	<i>A Multi-level Experimental Design</i>	322
10.3.1	The anova table	324
10.3.2	Expected values of mean squares	325
10.3.3	* The sums of squares breakdown	326
10.3.4	The variance components	328
10.3.5	The mixed model analysis	329
10.3.6	Predictive accuracy	332
10.3.7	Different sources of variance – complication or focus of interest?	332
10.4	<i>Within and Between Subject Effects</i>	333
10.4.1	Model selection	333
10.4.2	Estimates of model parameters	335
10.5	<i>A Generalized Linear Mixed Model</i>	336
10.6	<i>Repeated Measures in Time</i>	338
10.6.1	Example – random variation between profiles	340
10.6.2	Orthodontic measurements on children	345
10.7	<i>Error structure considerations</i>	349
10.7.1	Predictions from models with a complex error structure	349
10.7.2	Error structure in explanatory variables	350
10.8	<i>Further Notes on Multi-level and Other Models with Correlated Errors</i>	350
10.8.1	An historical perspective on multi-level models	350
10.8.2	Meta-analysis	351
10.8.3	Functional data analysis	352
10.9	<i>Recap</i>	352
10.10	<i>Further Reading</i>	352
10.11	<i>Exercises</i>	354
11	Tree-based Classification and Regression	356
11.1	<i>The Uses of Tree-based Methods</i>	357
11.1.1	Problems for which tree-based regression may be used	357
11.2	<i>Detecting Email Spam – an Example</i>	358
11.2.1	Choosing the number of splits	361
11.3	<i>Terminology and Methodology</i>	361
11.3.1	Choosing the split – regression trees	361
11.3.2	Within and between sums of squares	362
11.3.3	Choosing the split – classification trees	364
11.3.4	Tree-based regression versus loess regression smoothing	364
11.4	<i>Predictive Accuracy, and the Cost-complexity Tradeoff</i>	367
11.4.1	Cross-validation	367
11.4.2	The cost-complexity parameter	368

11.4.3	Prediction error versus tree size	368
11.5	<i>Data for female heart attack patients</i>	369
11.5.1	The one-standard-deviation rule	371
11.5.2	Printed Information on Each Split	371
11.6	<i>Detecting Email Spam – the Optimal Tree</i>	372
11.7	<i>The randomForest Package</i>	374
11.8	<i>Additional Notes on Tree-Based Methods</i>	378
11.9	<i>Further Reading and Extensions</i>	379
11.10	<i>Exercises</i>	380
12	Multivariate Data Exploration and Discrimination	382
12.1	<i>Multivariate Exploratory Data Analysis</i>	383
12.1.1	Scatterplot matrices	383
12.1.2	Principal components analysis	384
12.1.3	Multi-dimensional scaling	390
12.2	<i>Discriminant Analysis</i>	391
12.2.1	Example – plant architecture	392
12.2.2	Logistic discriminant analysis	393
12.2.3	Linear discriminant analysis	394
12.2.4	An example with more than two groups	396
12.3	<i>*High-dimensional data, classification, and plots</i>	397
12.3.1	Classifications and associated graphs	400
12.3.2	Flawed graphs	400
12.3.3	Accuracies and Scores for test data	404
12.3.4	Graphs derived from the cross-validation process	410
12.4	<i>Further Reading</i>	412
12.5	<i>Exercises</i>	413
13	Regression on Principal Component or Discriminant Scores	416
13.1	<i>Principal Component Scores in Regression</i>	416
13.2*	<i>Propensity Scores in Regression Comparisons – Labor Training Data</i>	420
13.2.1	Regression comparisons	423
13.2.2	A strategy that uses propensity scores	425
13.3	<i>Further Reading</i>	432
13.4	<i>Exercises</i>	432
14	The R System – Additional Topics	434
14.1	<i>Graphical User Interfaces to R</i>	434
14.1.1	The R Commander Graphical User Interface	434
14.1.2	Basics of the R Commander’s interface	435
14.1.3	The <i>rattle</i> GUI	436
14.1.4	Creating GUIs	436
14.2	<i>Working Directories, Workspaces and the Search List</i>	437
14.2.1	*The search path	437
14.2.2	Workspace management	438

14.2.3	Utility functions	439
14.3	<i>Data Input and Output</i>	439
14.3.1	Input of data	440
14.3.2	Data output	443
14.4	<i>Functions and operators – Some Further Details</i>	444
14.4.1	Function arguments	445
14.4.2	Character string and vector functions	446
14.4.3	Anonymous functions	447
14.4.4	Functions for working with dates (and times)	447
14.4.5	Creating Groups	449
14.4.6	Logical operators	449
14.5	<i>Factors</i>	449
14.6	<i>Missing Values</i>	452
14.7	<i>Matrices and Arrays</i>	454
14.7.1	Matrix arithmetic	456
14.7.2	Outer products	456
14.7.3	Arrays	457
14.8	<i>Manipulations with Lists, Data Frames and Matrices</i>	458
14.8.1	Lists – an extension of the notion of “vector”	458
14.8.2	Changing the shape of data frames (or matrices)	460
14.8.3	* Merging data frames – <i>merge()</i>	461
14.8.4	Joining data frames, matrices and vectors – <i>cbind()</i>	461
14.8.5	The <i>apply</i> family of functions	461
14.8.6	Splitting vectors and data frames into lists – <i>split()</i>	463
14.8.7	Multivariate time series	463
14.9	<i>Classes and Methods</i>	464
14.9.1	Printing and summarizing model objects	465
14.9.2	Extracting information from model objects	465
14.9.3	S4 classes and methods	466
14.10	<i>Manipulation of Language Constructs</i>	466
14.10.1	Model and graphics formulae	467
14.10.2	The use of a list to pass arguments	468
14.10.3	Expressions	469
14.10.4	Environments	469
14.10.5	Function environments, and lazy evaluation	470
14.11	<i>Creation of R Packages</i>	471
14.12	<i>Document Preparation – Sweave() and xtable()</i>	473
14.13	<i>Graphs in R</i>	473
14.13.1	Hardcopy graphics devices	473
14.13.2	Multiple graphs on a single base graphics page	474
14.13.3	Plotting characters, symbols, line types and colors	474
14.13.4	Formatting and plotting of text and equations	477
14.14	<i>Lattice Graphics, and the grid Package</i>	479
14.14.1	Lattice Graphics vs Base Graphics	479

14.14.2 Groups within data, and/or columns in parallel	479
14.14.3 Lattice Parameters and Graphics Features	481
14.14.4 A further example	483
14.14.5 Keys – <code>auto.key</code> , <code>key</code> & <code>legend</code>	484
14.14.6 Panel Functions and Interaction with Plots	485
14.14.7 Interaction with lattice plots – <code>focus</code> , <code>interact</code> , <code>unfocus</code>	486
14.14.8 Multiple lattice graphs on a graphics page	488
<i>14.15 An Implementation of Wilkinson's Grammar of Graphics</i>	<i>489</i>
14.15.1 Australian rain data	489
<i>14.16 Dynamic Graphics – the <code>rgl</code> and <code>rggobi</code> packages</i>	<i>492</i>
<i>14.17 Further Reading</i>	<i>493</i>
<i>14.18 Exercises</i>	<i>494</i>