
Contents

Preface	vii
---------------	-----

Part I Basics of R

1	Welcome to R	3
1.1	What Is R?	3
1.2	Why R?	4
1.3	Why Not R?	5
1.4	When R?	6
1.5	Using This Book	6
1.5.1	About the Text	6
1.5.2	About the Data	7
1.5.3	Online Material	8
1.5.4	When Things Go Wrong	9
1.6	Key Points	10
2	An Overview of the R Language	11
2.1	Getting Started	11
2.1.1	Initial Steps	11
2.1.2	Starting R	12
2.2	A Quick Tour of R's Capabilities	13
2.3	Basics of Working with R Commands	17
2.4	Basic Objects	18
2.4.1	Vectors	19
2.4.2	Help! A Brief Detour	21
2.4.3	More on Vectors and Indexing	24
2.4.4	aaRgh! A Digression for New Programmers	26
2.4.5	Missing and Interesting Values	26
2.4.6	Using R for Mathematical Computation	28
2.4.7	Lists	28

2.5	Data Frames	30
2.6	Loading and Saving Data	34
2.6.1	Image Files	36
2.6.2	CSV Files	36
2.7	Writing Your Own Functions*	38
2.7.1	Language Structures*	40
2.7.2	Anonymous Functions*	41
2.8	Clean Up!	42
2.9	Learning More*	43
2.10	Key Points	44

Part II Fundamentals of Data Analysis

3	Describing Data	47
3.1	Simulating Data	47
3.1.1	Store Data: Setting the Structure	48
3.1.2	Store Data: Simulating Data Points	50
3.2	Functions to Summarize a Variable	52
3.2.1	Discrete Variables	52
3.2.2	Continuous Variables	54
3.3	Summarizing Data Frames	56
3.3.1	<code>summary()</code>	57
3.3.2	<code>describe()</code>	58
3.3.3	Recommended Approach to Inspecting Data	59
3.3.4	<code>apply()</code> *	59
3.4	Single Variable Visualization	61
3.4.1	Histograms	61
3.4.2	Boxplots	66
3.4.3	QQ Plot to Check Normality*	68
3.4.4	Cumulative Distribution*	69
3.4.5	Language Brief: <code>by()</code> and <code>aggregate()</code>	70
3.4.6	Maps	72
3.5	Learning More*	74
3.6	Key Points	75
4	Relationships Between Continuous Variables	77
4.1	Retailer Data	77
4.1.1	Simulating Customer Data	78
4.1.2	Simulating Online and In-Store Sales Data	79
4.1.3	Simulating Satisfaction Survey Responses	80
4.1.4	Simulating Non-Response Data	82
4.2	Exploring Associations Between Variables with Scatterplots	83
4.2.1	Creating a Basic Scatterplot with <code>plot()</code>	83
4.2.2	Color-Coding Points on a Scatterplot	86

4.2.3	Adding a Legend to a Plot	88
4.2.4	Plotting on a Log Scale	89
4.3	Combining Plots in a Single Graphics Object	90
4.4	Scatterplot Matrices	92
4.4.1	<code>pairs()</code>	92
4.4.2	<code>scatterplotMatrix()</code>	93
4.5	Correlation Coefficients	95
4.5.1	Correlation Tests	97
4.5.2	Correlation Matrices	98
4.5.3	Transforming Variables before Computing Correlations	100
4.5.4	Typical Marketing Data Transformations	102
4.5.5	Box–Cox Transformations*	102
4.6	Exploring Associations in Survey Responses*	104
4.6.1	<code>jitter()</code> *	105
4.6.2	<code>polychoric()</code> *	106
4.7	Learning More*	107
4.8	Key Points	108
5	Comparing Groups: Tables and Visualizations	111
5.1	Simulating Consumer Segment Data	111
5.1.1	Segment Data Definition	112
5.1.2	Language Brief: <code>for()</code> Loops	114
5.1.3	Language Brief: <code>if()</code> Blocks	116
5.1.4	Final Segment Data Generation	118
5.2	Finding Descriptives by Group	120
5.2.1	Language Brief: Basic Formula Syntax	123
5.2.2	Descriptives for Two-Way Groups	124
5.2.3	Visualization by Group: Frequencies and Proportions	126
5.2.4	Visualization by Group: Continuous Data	129
5.3	Learning More*	132
5.4	Key Points	133
6	Comparing Groups: Statistical Tests	135
6.1	Data for Comparing Groups	135
6.2	Testing Group Frequencies: <code>chisq.test()</code>	136
6.3	Testing Observed Proportions: <code>binom.test()</code>	139
6.3.1	About Confidence Intervals	140
6.3.2	More About <code>binom.test()</code> and Binomial Distributions .	141
6.4	Testing Group Means: <code>t.test()</code>	142
6.5	Testing Multiple Group Means: ANOVA	144
6.5.1	Model Comparison in ANOVA*	146
6.5.2	Visualizing Group Confidence Intervals	147
6.5.3	Variable Selection in ANOVA: Stepwise Modeling*	148
6.6	Bayesian ANOVA: Getting Started*	149
6.6.1	Why Bayes?	150

6.6.2	Basics of Bayesian ANOVA*	150
6.6.3	Inspecting the Posterior Draws*	152
6.6.4	Plotting the Bayesian Credible Intervals*	155
6.7	Learning More*	156
6.8	Key Points	157
7	Identifying Drivers of Outcomes: Linear Models	159
7.1	Amusement Park Data	160
7.1.1	Simulating the Amusement Park Data	160
7.2	Fitting Linear Models with <code>lm()</code>	162
7.2.1	Preliminary Data Inspection	163
7.2.2	Recap: Bivariate Association	165
7.2.3	Linear Model with a Single Predictor	165
7.2.4	<code>lm</code> Objects	166
7.2.5	Checking Model Fit	169
7.3	Fitting Linear Models with Multiple Predictors	173
7.3.1	Comparing Models	175
7.3.2	Using a Model to Make Predictions	176
7.3.3	Standardizing the Predictors	177
7.4	Using Factors as Predictors	179
7.5	Interaction Terms	182
7.5.1	Language Brief: Advanced Formula Syntax*	183
7.6	Caution! Overfitting	185
7.7	Recommended Procedure for Linear Model Fitting	186
7.8	Bayesian Linear Models with <code>MCMCregress()</code> *	186
7.9	Learning More*	188
7.10	Key Points	190

Part III Advanced Marketing Applications

8	Reducing Data Complexity	195
8.1	Consumer Brand Rating Data	195
8.1.1	Rescaling the Data	197
8.1.2	Aggregate Mean Ratings by Brand	198
8.2	Principal Component Analysis and Perceptual Maps	200
8.2.1	PCA Example	200
8.2.2	Visualizing PCA	203
8.2.3	PCA for Brand Ratings	204
8.2.4	Perceptual Map of the Brands	206
8.2.5	Cautions with Perceptual Maps	208
8.3	Exploratory Factor Analysis	209
8.3.1	Basic EFA Concepts	210
8.3.2	Finding an EFA Solution	211

8.3.3	EFA Rotations	213
8.3.4	Using Factor Scores for Brands	216
8.4	Multidimensional Scaling	218
8.4.1	Non-metric MDS	219
8.5	Learning More*	221
8.5.1	Principal Component Analysis	221
8.5.2	Factor Analysis	221
8.5.3	Multidimensional Scaling	222
8.6	Key Points	222
8.6.1	Principal Component Analysis	222
8.6.2	Exploratory Factor Analysis	222
8.6.3	Multidimensional Scaling	223
9	Additional Linear Modeling Topics	225
9.1	Handling Highly Correlated Variables	226
9.1.1	An Initial Linear Model of Online Spend	226
9.1.2	Remediating Collinearity	229
9.2	Linear Models for Binary Outcomes: Logistic Regression	231
9.2.1	Basics of the Logistic Regression Model	231
9.2.2	Data for Logistic Regression of Season Passes	232
9.2.3	Sales Table Data	233
9.2.4	Language Brief: Classes and Attributes of Objects*	234
9.2.5	Finalizing the Data	236
9.2.6	Fitting a Logistic Regression Model	237
9.2.7	Reconsidering the Model	239
9.2.8	Additional Discussion	242
9.3	Hierarchical Linear Models	242
9.3.1	Some HLM Concepts	243
9.3.2	Ratings-Based Conjoint Analysis for the Amusement Park	244
9.3.3	Simulating Ratings-Based Conjoint Data	245
9.3.4	An Initial Linear Model	246
9.3.5	Hierarchical Linear Model with lme4	248
9.3.6	The Complete Hierarchical Linear Model	249
9.3.7	Summary of HLM with lme4	251
9.4	Bayesian Hierarchical Linear Models*	252
9.4.1	Initial Linear Model with MCMCregress()	253
9.4.2	Hierarchical Linear Model with MCMChregress()	253
9.4.3	Inspecting Distribution of Preference*	256
9.5	A Quick Comparison of Frequentist & Bayesian HLMs*	259
9.6	Learning More*	263
9.6.1	Collinearity	263
9.6.2	Logistic Regression	263
9.6.3	Hierarchical Models	263
9.6.4	Bayesian Hierarchical Models	263
9.7	Key Points	264
9.7.1	Collinearity	264

9.7.2	Logistic Regression	264
9.7.3	Hierarchical Linear Models	265
9.7.4	Bayesian Methods for Hierarchical Linear Models	266
10	Confirmatory Factor Analysis and Structural Equation Modeling ...	267
10.1	The Motivation for Structural Models	268
10.1.1	Structural Models in This Chapter	269
10.2	Scale Assessment: CFA	270
10.2.1	Simulating PIES CFA Data	272
10.2.2	Estimating the PIES CFA Model	277
10.2.3	Assessing the PIES CFA Model	278
10.3	General Models: Structural Equation Models	283
10.3.1	The Repeat Purchase Model in R	284
10.3.2	Assessing the Repeat Purchase Model	286
10.4	The Partial Least Squares (PLS) Alternative	288
10.4.1	PLS-SEM for Repeat Purchase	289
10.4.2	Visualizing the Fitted PLS Model*	292
10.4.3	Assessing the PLS-SEM Model	293
10.4.4	PLS-SEM with the Larger Sample	295
10.5	Learning More*	297
10.6	Key Points	297
11	Segmentation: Clustering and Classification	299
11.1	Segmentation Philosophy	299
11.1.1	The Difficulty of Segmentation	299
11.1.2	Segmentation as Clustering and Classification	300
11.2	Segmentation Data	302
11.3	Clustering	302
11.3.1	The Steps of Clustering	303
11.3.2	Hierarchical Clustering: <code>hclust()</code> Basics	305
11.3.3	Hierarchical Clustering Continued: Groups from <code>hclust()</code>	309
11.3.4	Mean-Based Clustering: <code>kmeans()</code>	311
11.3.5	Model-Based Clustering: <code>Mclust()</code>	314
11.3.6	Comparing Models with <code>BIC()</code>	315
11.3.7	Latent Class Analysis: <code>poLCA()</code>	317
11.3.8	Comparing Cluster Solutions	320
11.3.9	Recap of Clustering	322
11.4	Classification	322
11.4.1	Naive Bayes Classification: <code>naiveBayes()</code>	323
11.4.2	Random Forest Classification: <code>randomForest()</code>	327
11.4.3	Random Forest Variable Importance	330
11.5	Prediction: Identifying Potential Customers*	333
11.6	Learning More*	336
11.7	Key Points	337

12 Association Rules for Market Basket Analysis	339
12.1 The Basics of Association Rules	340
12.1.1 Metrics	340
12.2 Retail Transaction Data: Market Baskets	341
12.2.1 Example Data: Groceries	342
12.2.2 Supermarket Data	344
12.3 Finding and Visualizing Association Rules	346
12.3.1 Finding and Plotting Subsets of Rules	348
12.3.2 Using Profit Margin Data with Transactions: An Initial Start	349
12.3.3 Language Brief: A Function for Margin Using an Object's <code>class</code> *	351
12.4 Rules in Non-Transactional Data: Exploring Segments Again	356
12.4.1 Language Brief: Slicing Continuous Data with <code>cut()</code>	356
12.4.2 Exploring Segment Associations	357
12.5 Learning More*	360
12.6 Key Points	360
13 Choice Modeling	363
13.1 Choice-Based Conjoint Analysis Surveys	364
13.2 Simulating Choice Data*	365
13.3 Fitting a Choice Model	370
13.3.1 Inspecting Choice Data	371
13.3.2 Fitting Choice Models with <code>mlogit()</code>	372
13.3.3 Reporting Choice Model Findings	375
13.3.4 Share Predictions for Identical Alternatives	380
13.3.5 Planning the Sample Size for a Conjoint Study	381
13.4 Adding Consumer Heterogeneity to Choice Models	383
13.4.1 Estimating Mixed Logit Models with <code>mlogit()</code>	383
13.4.2 Share Prediction for Heterogeneous Choice Models	386
13.5 Hierarchical Bayes Choice Models	388
13.5.1 Estimating Hierarchical Bayes Choice Models with <code>ChoiceModelR</code>	388
13.5.2 Share Prediction for Hierarchical Bayes Choice Models	395
13.6 Design of Choice-Based Conjoint Surveys*	397
13.7 Learning More*	398
13.8 Key Points	399
Conclusion	401
A Appendix: R Versions and Related Software	403
A.1 R Base	403
A.2 RStudio	404
A.3 Emacs Speaks Statistics	405
A.4 Eclipse + StatET	406
A.5 Revolution R	407

A.6	Other Options	408
A.6.1	Text Editors	408
A.6.2	R Commander	408
A.6.3	Rattle	409
A.6.4	Deducer	409
A.6.5	TIBCO Enterprise Runtime for R	409
B	Appendix: Scaling Up	411
B.1	Handling Data	411
B.1.1	Data Wrangling	411
B.1.2	Microsoft Excel: <code>gdata</code>	412
B.1.3	SAS, SPSS, and Other Statistics Packages: <code>foreign</code>	412
B.1.4	SQL: <code>RSQLite</code> , <code>sqldf</code> and <code>RODBC</code>	413
B.2	Handling Large Data Sets	415
B.3	Speeding Up Computation	416
B.3.1	Efficient Coding and Data Storage	416
B.3.2	Enhancing the R Engine	417
B.4	Time Series Analysis, Repeated Measures, and Longitudinal Analysis	418
B.5	Automated and Interactive Reporting	419
C	Appendix: Packages Used	423
C.1	Core and Frequentist Statistics	424
C.2	Graphics	424
C.3	Bayesian Methods	425
C.4	Advanced Statistics	426
C.5	Machine Learning	426
C.6	Data Handling	427
C.7	Other Packages	428
D	Appendix: Online Materials and Data Files	431
D.1	Data File Structure	431
D.2	Data File URL Cross-Reference	432
D.2.1	Update on Data Locations	432
	References	435
	Index	447