Chris Chapman · Elea McDonnell Feit

R for Marketing Research and Analytics

Second Edition

Springer

Contents

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
  - 1.4.1 R Versus Python, Julia, and Others . . . 6
- 1.5 Which R? Base or Tidy? . . . 7
- 1.6 Using This Book . . . 8
  - 1.6.1 About the Text . . . 8
  - 1.6.2 About the Data . . . 9
  - 1.6.3 Online Material . . . 10
  - 1.6.4 When Things Go Wrong . . . 10
- 1.7 Key Points . . . 12

2 An Overview of the R Language . . . 13
- 2.1 Getting Started . . . 13
  - 2.1.1 Initial Steps . . . 13
  - 2.1.2 Starting R . . . 14
- 2.2 A Quick Tour of R's Capabilities . . . 15
- 2.3 Basics of Working with R Commands . . . 19
- 2.4 Basic Objects . . . 20
  - 2.4.1 Vectors . . . 20
  - 2.4.2 Help! A Brief Detour . . . 23
  - 2.4.3 More on Vectors and Indexing . . . 25
  - 2.4.4 aaRgh! A Digression for New Programmers . . . 27
  - 2.4.5 Missing and Interesting Values . . . 27
  - 2.4.6 Using R for Mathematical Computation . . . 29
  - 2.4.7 Lists . . . 29

2.5 Data Frames . . . 31
2.6 Loading and Saving Data . . . 34
2.6.1 Image Files . . . 35
2.6.2 CSV Files . . . 36
2.7 Writing Your Own Functions* . . . 37
2.7.1 Language Structures* . . . 39
2.7.2 Anonymous Functions* . . . 40
2.8 Clean Up! . . . 41
2.9 Key Points . . . 42
2.10 Learning More* . . . 43
2.11 Exercises . . . 44
2.11.1 Preliminary Note on Exercises . . . 44
2.11.2 Exercises . . . 44

Part II Fundamentals of Data Analysis

3 Describing Data . . . 49
3.1 Simulating Data . . . 49
3.1.1 Store Data: Setting the Structure . . . 50
3.1.2 Store Data: Simulating Data Points . . . 51
3.2 Functions to Summarize a Variable . . . 54
3.2.1 Discrete Variables . . . 54
3.2.2 Continuous Variables . . . 56
3.3 Summarizing Data Frames . . . 57
3.3.1 `summary()` . . . 58
3.3.2 `describe()` . . . 59
3.3.3 Recommended Approach to Inspecting Data . . . 60
3.3.4 `apply()*` . . . 60
3.4 Single Variable Visualization . . . 62
3.4.1 Histograms . . . 62
3.4.2 Boxplots . . . 66
3.4.3 QQ Plot to Check Normality* . . . 69
3.4.4 Cumulative Distribution* . . . 70
3.4.5 Language Brief: `by()` and `aggregate()` . . . 71
3.4.6 Maps . . . 73
3.5 Key Points . . . 75
3.6 Data Sources . . . 75
3.7 Learning More* . . . 76
3.8 Exercises . . . 76
3.8.1 E-Commerce Data for Exercises . . . 76
3.8.2 Exercises . . . 77

4 Relationships Between Continuous Variables 79
4.1 Retailer Data 79
4.1.1 Simulating the Data 80
4.1.2 Simulating Online and In-store Sales Data 81
4.1.3 Simulating Satisfaction Survey Responses 82
4.1.4 Simulating Non-response Data 83
4.2 Exploring Associations Between Variables with Scatterplots 84
4.2.1 Creating a Basic Scatterplot with `plot()` 85
4.2.2 Color-Coding Points on a Scatterplot 88
4.2.3 Adding a Legend to a Plot 89
4.2.4 Plotting on a Log Scale 90
4.3 Combining Plots in a Single Graphics Object 91
4.4 Scatterplot Matrices 93
4.4.1 `pairs()` 93
4.4.2 `scatterplotMatrix()` 95
4.5 Correlation Coefficients 96
4.5.1 Correlation Tests 97
4.5.2 Correlation Matrices 98
4.5.3 Transforming Variables Before Computing Correlations 99
4.5.4 Typical Marketing Data Transformations 101
4.5.5 Box-Cox Transformations* 102
4.6 Exploring Associations in Survey Responses 103
4.6.1 `jitter()` 104
4.6.2 `polychoric()*` 105
4.7 Key Points 106
4.8 Data Sources 107
4.9 Learning More* 107
4.10 Exercises 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: `for()` Loops 114
5.1.3 Language Brief: `if()` Blocks 115
5.1.4 Final Segment Data Generation 117
5.2 Finding Descriptives by Group 119
5.2.1 Language Brief: Basic Formula Syntax 122
5.2.2 Descriptives for Two-Way Groups 122
5.2.3 Visualization by Group: Frequencies and Proportions 124
5.2.4 Visualization by Group: Continuous Data 127

5.3 Key Points . . . 130
5.4 Data Sources . . . 131
5.5 Learning More* . . . 131
5.6 Exercises . . . 131

6 Comparing Groups: Statistical Tests . . . 133
6.1 Data for Comparing Groups . . . 133
6.2 Testing Group Frequencies: `chisq.test()` . . . 133
6.3 Testing Observed Proportions: `binom.test()` . . . 137
6.3.1 About Confidence Intervals . . . 137
6.3.2 More About `binom.test()` and Binomial Distributions . . . 138
6.4 Testing Group Means: `t.test()` . . . 139
6.5 Testing Multiple Group Means: Analysis of Variance (ANOVA) . . . 141
6.5.1 Model Comparison in ANOVA* . . . 143
6.5.2 Visualizing Group Confidence Intervals . . . 144
6.5.3 Variable Selection in ANOVA: Stepwise Modeling* . . . 145
6.6 Bayesian ANOVA: Getting Started* . . . 146
6.6.1 Why Bayes? . . . 147
6.6.2 Basics of Bayesian ANOVA* . . . 147
6.6.3 Inspecting the Posterior Draws* . . . 150
6.6.4 Plotting the Bayesian Credible Intervals* . . . 152
6.7 Key Points . . . 153
6.8 Learning More* . . . 154
6.9 Exercises . . . 154

7 Identifying Drivers of Outcomes: Linear Models . . . 157
7.1 Amusement Park Data . . . 158
7.1.1 Simulating the Amusement Park Data . . . 158
7.2 Fitting Linear Models with `lm()` . . . 160
7.2.1 Preliminary Data Inspection . . . 161
7.2.2 Recap: Bivariate Association . . . 163
7.2.3 Linear Model with a Single Predictor . . . 164
7.2.4 `lm` Objects . . . 164
7.2.5 Checking Model Fit . . . 167
7.3 Fitting Linear Models with Multiple Predictors . . . 170
7.3.1 Comparing Models . . . 172
7.3.2 Using a Model to Make Predictions . . . 174
7.3.3 Standardizing the Predictors . . . 174
7.4 Using Factors as Predictors . . . 176

7.5 Interaction Terms . . . 178
7.5.1 Language Brief: Advanced Formula Syntax* . . . 181
7.5.2 Caution! Overfitting . . . 182
7.5.3 Recommended Procedure for Linear Model Fitting . . . 182
7.5.4 Bayesian Linear Models with `MCMCregress()`* . . . 183
7.6 Key Points . . . 185
7.7 Data Sources . . . 186
7.8 Learning More* . . . 187
7.9 Exercises . . . 188
7.9.1 Simulated Hotel Satisfaction and Account Data . . . 188
7.9.2 Exercises . . . 188

Part III Advanced Marketing Applications

8 Reducing Data Complexity . . . 193
8.1 Consumer Brand Rating Data . . . 193
8.1.1 Rescaling the Data . . . 194
8.1.2 Aggregate Mean Ratings by Brand . . . 196
8.2 Principal Component Analysis and Perceptual Maps . . . 198
8.2.1 PCA Example . . . 198
8.2.2 Visualizing PCA . . . 200
8.2.3 PCA for Brand Ratings . . . 201
8.2.4 Perceptual Map of the Brands . . . 203
8.2.5 Cautions with Perceptual Maps . . . 205
8.3 Exploratory Factor Analysis . . . 206
8.3.1 Basic EFA Concepts . . . 207
8.3.2 Finding an EFA Solution . . . 208
8.3.3 EFA Rotations . . . 210
8.3.4 Using Factor Scores for Brands . . . 213
8.4 Multidimensional Scaling . . . 215
8.4.1 Non-metric MDS . . . 215
8.5 Key Points . . . 217
8.6 Data Sources . . . 218
8.7 Learning More* . . . 219
8.8 Exercises . . . 219
8.8.1 PRST Brand Data . . . 219
8.8.2 Exercises . . . 220

9 Additional Linear Modeling Topics . . . 223
9.1 Handling Highly Correlated Variables . . . 224
9.1.1 An Initial Linear Model of Online Spend . . . 224
9.1.2 Remediating Collinearity . . . 227

9.2 Linear Models for Binary Outcomes: Logistic Regression 229
9.2.1 Basics of the Logistic Regression Model 229
9.2.2 Data for Logistic Regression of Season Passes 230
9.2.3 Sales Table Data 231
9.2.4 Language Brief: Classes and Attributes of Objects* 232
9.2.5 Finalizing the Data 233
9.2.6 Fitting a Logistic Regression Model 234
9.2.7 Reconsidering the Model 236
9.2.8 Additional Discussion 238
9.3 Hierarchical Models 239
9.3.1 Some HLM Concepts 239
9.3.2 Ratings-Based Conjoint Analysis for the Amusement Park 240
9.3.3 Simulating Ratings-Based Conjoint Data 241
9.3.4 An Initial Linear Model 242
9.3.5 Initial Hierarchical Linear Model with `lme4` 244
9.3.6 Complete Hierarchical Linear Model 245
9.3.7 Conclusion for Classical HLM 247
9.4 Bayesian Hierarchical Linear Models* 247
9.4.1 Initial Linear Model with `MCMCregress()`* 248
9.4.2 Hierarchical Linear Model with `MCMChregress()`* 249
9.4.3 Inspecting Distribution of Preference* 252
9.5 A Quick Comparison of the Effects* 254
9.6 Key Points 258
9.7 Data Sources 259
9.8 Learning More* 260
9.9 Exercises 261
9.9.1 Online Visits and Sales Data for Exercises 261
9.9.2 Exercises for Collinearity and Logistic Regression . . . 262
9.9.3 Handbag Conjoint Analysis Data for Exercises 263
9.9.4 Exercises for Metric Conjoint and Hierarchical Linear Models 263

10 Confirmatory Factor Analysis and Structural Equation Modeling 265
10.1 The Motivation for Structural Models 266
10.1.1 Structural Models in This Chapter 267
10.2 Scale Assessment: Confirmatory Factor Analysis (CFA) 268
10.2.1 Simulating PIES CFA Data 270
10.2.2 Estimating the PIES CFA Model 273
10.2.3 Assessing the PIES CFA Model 276

10.3 General Models: Structural Equation Models . . . 280
10.3.1 The Repeat Purchase Model in R . . . 282
10.3.2 Assessing the Repeat Purchase Model . . . 283
10.4 The Partial Least Squares (PLS) Alternative . . . 285
10.4.1 PLS-SEM for Repeat Purchase . . . 286
10.4.2 Visualizing the Fitted PLS Model* . . . 288
10.4.3 Assessing the PLS-SEM Model . . . 289
10.4.4 PLS-SEM with the Larger Sample . . . 291
10.5 Key Points . . . 293
10.6 Learning More* . . . 294
10.7 Exercises . . . 295
10.7.1 Brand Data for Confirmatory Factor Analysis Exercises . . . 295
10.7.2 Exercises for Confirmatory Factor Analysis . . . 295
10.7.3 Purchase Intention Data for Structural Equation Model Exercises . . . 295
10.7.4 Exercises for Structural Equation Models and PLS SEM . . . 296

11 Segmentation: Clustering and Classification . . . 299
11.1 Segmentation Philosophy . . . 299
11.1.1 The Difficulty of Segmentation . . . 300
11.1.2 Segmentation as Clustering and Classification . . . 301
11.2 Segmentation Data . . . 302
11.3 Clustering . . . 302
11.3.1 The Steps of Clustering . . . 303
11.3.2 Hierarchical Clustering: `hclust()` Basics . . . 305
11.3.3 Hierarchical Clustering Continued: Groups from `hclust()` . . . 308
11.3.4 Mean-Based Clustering: `kmeans()` . . . 311
11.3.5 Model-Based Clustering: `Mclust()` . . . 314
11.3.6 Comparing Models with `BIC()` . . . 315
11.3.7 Latent Class Analysis: `poLCA()` . . . 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: `naiveBayes()` . . . 323
11.4.2 Random Forest Classification: `randomForest()` . . . 327
11.4.3 Random Forest Variable Importance . . . 330
11.5 Prediction: Identifying Potential Customers* . . . 332
11.6 Key Points . . . 336
11.7 Learning More* . . . 337

11.8 Exercises . . . 338
11.8.1 Music Subscription Data for Exercises . . . 338
11.8.2 Exercises . . . 339

12 Association Rules for Market Basket Analysis . . . 341
12.1 The Basics of Association Rules . . . 342
12.2 Retail Transaction Data: Market Baskets . . . 343
12.2.1 Example Data: `Groceries` . . . 344
12.2.2 Supermarket Data . . . 346
12.3 Finding and Visualizing Association Rules . . . 347
12.3.1 Finding and Plotting Subsets of Rules . . . 350
12.3.2 Using Profit Margin Data with Transactions: An Initial Start . . . 350
12.3.3 Language Brief: A Function for Margin Using an Object's `class`* . . . 352
12.4 Rules in Non-transactional Data: Exploring Segments Again . . . 356
12.4.1 Language Brief: Slicing Continuous Data with `cut()` . . . 357
12.4.2 Exploring Segment Associations . . . 358
12.5 Key Points . . . 360
12.6 Learning More* . . . 361
12.7 Exercises . . . 361
12.7.1 Retail Transactions Data for Exercises . . . 361
12.7.2 Exercises . . . 362

13 Choice Modeling . . . 363
13.1 Choice-Based Conjoint Analysis Surveys . . . 364
13.2 Simulating Choice Data* . . . 365
13.3 Fitting a Choice Model . . . 369
13.3.1 Inspecting Choice Data . . . 370
13.3.2 Fitting Choice Models with `mlogit()` . . . 371
13.3.3 Reporting Choice Model Findings . . . 374
13.3.4 Share Predictions for Identical Alternatives . . . 378
13.3.5 Planning the Sample Size for a Conjoint Study . . . 379
13.4 Adding Consumer Heterogeneity to Choice Models . . . 380
13.4.1 Estimating Mixed Logit Models with `mlogit()` . . . 381
13.4.2 Share Prediction for Heterogeneous Choice Models . . . 384
13.5 Hierarchical Bayes Choice Models . . . 385
13.5.1 Estimating Hierarchical Bayes Choice Models with `ChoiceModelR` . . . 385
13.5.2 Share Prediction for Hierarchical Bayes Choice Models . . . 391

13.6 Design of Choice-Based Conjoint Surveys* . . . 393
13.7 Key Points . . . 395
13.8 Data Sources . . . 396
13.9 Learning More* . . . 396
13.10 Excercises . . . 397

14 Behavior Sequences . . . 399
14.1 Web Log Data . . . 399
14.1.1 EPA Web Data . . . 400
14.1.2 Processing the Raw Data . . . 401
14.1.3 Cleaning the Data . . . 401
14.1.4 Handling Dates and Times . . . 402
14.1.5 Requests and Page Types . . . 403
14.1.6 Additional HTTP Data . . . 405
14.2 Basic Event Statistics . . . 405
14.2.1 Events . . . 405
14.2.2 Events by Time . . . 406
14.2.3 Errors . . . 407
14.2.4 Active Users . . . 408
14.3 Identifying Sequences (Sessions) . . . 409
14.3.1 Extracting Sessions . . . 409
14.3.2 Session Statistics . . . 412
14.4 Markov Chains for Behavior Transitions . . . 414
14.4.1 Key Concepts and Demonstration . . . 415
14.4.2 Formatting the EPA Data for `clickstream` Analysis . . . 416
14.4.3 Estimating the Markov Chain . . . 419
14.4.4 Visualizing the MC Results . . . 419
14.4.5 Higher Order Chains and Prediction . . . 420
14.5 Discussion and Questions . . . 423
14.6 Key Points . . . 424
14.7 Learning More* . . . 425
14.8 Exercises . . . 426

Conclusion . . . 429

Appendix A: R Versions and Related Software . . . 431

Appendix B: An Introduction to Reproducible Results with R Notebooks . . . 439

Appendix C: Scaling Up . . . 447

Appendix D: Packages Used . . . 459

Appendix E: Online Materials and Data Files 465

References 469

Index 479