CONTENTS

Preface xi

1 Introduction 1
- 1.1 Book Overview . . . 3
- 1.2 Chapter Summaries . . . 4
- 1.3 How to Use This Book . . . 5
- 1.4 Why Learn to Analyze Data? . . . 6
  - 1.4.1 Learning to Code . . . 6
- 1.5 Getting Ready . . . 7
- 1.6 Introduction to R . . . 8
  - 1.6.1 Doing Calculations in R . . . 9
  - 1.6.2 Creating Objects in R . . . 10
  - 1.6.3 Using Functions in R . . . 12
- 1.7 Loading and Making Sense of Data . . . 14
  - 1.7.1 Setting the Working Directory . . . 15
  - 1.7.2 Loading the Dataset . . . 15
  - 1.7.3 Understanding the Data . . . 16
  - 1.7.4 Identifying the Types of Variables Included 19
  - 1.7.5 Identifying the Number of Observations . . 20
- 1.8 Computing and Interpreting Means . . . 21
  - 1.8.1 Accessing Variables inside Dataframes . . . 21
  - 1.8.2 Means . . . 22
- 1.9 Summary . . . 24
- 1.10 Cheatsheets . . . 25
  - 1.10.1 Concepts and Notation . . . 25
  - 1.10.2 R Symbols and Operators . . . 26
  - 1.10.3 R Functions . . . 26

2 Estimating Causal Effects with Randomized Experiments 27
- 2.1 Project STAR . . . 27
- 2.2 Treatment and Outcome Variables . . . 28
  - 2.2.1 Treatment Variables . . . 29
  - 2.2.2 Outcome Variables . . . 29
- 2.3 Individual Causal Effects . . . 29
- 2.4 Average Causal Effects . . . 33
  - 2.4.1 Randomized Experiments and the Difference-in-Means Estimator . . . 35
- 2.5 Do Small Classes Improve Student Performance? . 39

2.5.1 Relational Operators in R 39
2.5.2 Creating New Variables 40
2.5.3 Subsetting Variables 42
2.6 Summary 46
2.7 Cheatsheets 47
2.7.1 Concepts and Notation 47
2.7.2 R Symbols and Operators 50
2.7.3 R Functions 50

3 Inferring Population Characteristics via Survey Research 51
3.1 The EU Referendum in the UK 51
3.2 Survey Research 52
3.2.1 Random Sampling 53
3.2.2 Potential Challenges 54
3.3 Measuring Support for Brexit 55
3.3.1 Predicting the Referendum Outcome 56
3.3.2 Frequency Tables 57
3.3.3 Tables of Proportions 57
3.4 Who Supported Brexit? 58
3.4.1 Handling Missing Data 59
3.4.2 Two-Way Frequency Tables 62
3.4.3 Two-Way Tables of Proportions 64
3.4.4 Histograms 66
3.4.5 Density Histograms 68
3.4.6 Descriptive Statistics 71
3.5 Relationship between Education and the Leave Vote in the Entire UK 76
3.5.1 Scatter Plots 78
3.5.2 Correlation 82
3.6 Summary 88
3.7 Cheatsheets 90
3.7.1 Concepts and Notation 90
3.7.2 R Symbols and Operators 96
3.7.3 R Functions 96

4 Predicting Outcomes Using Linear Regression 98
4.1 GDP and Night-Time Light Emissions 98
4.2 Predictors, Observed vs. Predicted Outcomes, and Prediction Errors 99
4.3 Summarizing the Relationship between Two Variables with a Line 100
4.3.1 The Linear Regression Model 101
4.3.2 The Intercept Coefficient 103
4.3.3 The Slope Coefficient 104
4.3.4 The Least Squares Method 106
4.4 Predicting GDP Using Prior GDP 107
4.4.1 Relationship between GDP and Prior GDP 109
4.4.2 With Natural Logarithm Transformations . . 113
4.5 Predicting GDP Growth Using Night-Time Light Emissions 116

4.6 Measuring How Well the Model Fits the Data with the Coefficient of Determination, R^2 120
4.6.1 How Well Do the Three Predictive Models in This Chapter Fit the Data? 122
4.7 Summary 123
4.8 Appendix: Interpretation of the Slope in the Log-Log Linear Model 124
4.9 Cheatsheets 126
4.9.1 Concepts and Notation 126
4.9.2 R Functions 128

5 Estimating Causal Effects with Observational Data 129
5.1 Russian State-Controlled TV Coverage of 2014 Ukrainian Affairs 129
5.2 Challenges of Estimating Causal Effects with Observational Data 130
5.2.1 Confounding Variables 130
5.2.2 Why Are Confounders a Problem? 131
5.2.3 Confounders in Randomized Experiments . 133
5.3 The Effect of Russian TV on Ukrainians' Voting Behavior 135
5.3.1 Using the Simple Linear Model to Compute the Difference-in-Means Estimator 136
5.3.2 Controlling for Confounders Using a Multiple Linear Regression Model 142
5.4 The Effect of Russian TV on Ukrainian Electoral Outcomes 147
5.4.1 Using the Simple Linear Model to Compute the Difference-in-Means Estimator 149
5.4.2 Controlling for Confounders Using a Multiple Linear Regression Model 151
5.5 Internal and External Validity 153
5.5.1 Randomized Experiments vs. Observational Studies 153
5.5.2 The Role of Randomization 154
5.5.3 How Good Are the Two Causal Analyses in This Chapter? 155
5.5.4 How Good Was the Causal Analysis in Chapter 2? 156
5.5.5 The Coefficient of Determination, R^2 157
5.6 Summary 157
5.7 Cheatsheets 159
5.7.1 Concepts and Notation 159
5.7.2 R Functions 161

6 Probability 162
6.1 What Is Probability? 162
6.2 Axioms of Probability 163
6.3 Events, Random Variables, and Probability Distributions 165

6.4 Probability Distributions 166
6.4.1 The Bernoulli Distribution 166
6.4.2 The Normal Distribution 169
6.4.3 The Standard Normal Distribution 173
6.4.4 Recap 179
6.5 Population Parameters vs. Sample Statistics 179
6.5.1 The Law of Large Numbers 180
6.5.2 The Central Limit Theorem 183
6.5.3 Sampling Distribution of the Sample Mean 188
6.6 Summary 189
6.7 Appendix: For Loops 190
6.8 Cheatsheets 192
6.8.1 Concepts and Notation 192
6.8.2 R Symbols and Operators 194
6.8.3 R Functions 195

7 Quantifying Uncertainty 196
7.1 Estimators and Their Sampling Distributions . . . 196
7.2 Confidence Intervals 202
7.2.1 For the Sample Mean 203
7.2.2 For the Difference-in-Means Estimator . . 206
7.2.3 For Predicted Outcomes 209
7.3 Hypothesis Testing 211
7.3.1 With the Difference-in-Means Estimator . . 218
7.3.2 With Estimated Regression Coefficients . . 220
7.4 Statistical vs. Scientific Significance 224
7.5 Summary 225
7.6 Cheatsheets 226
7.6.1 Concepts and Notation 226
7.6.2 R Symbols and Operators 229
7.6.3 R Functions 229

Index of Concepts 231

Index of Mathematical Notation 235

Index of R and RStudio 237