

## BRIEF CONTENTS

<i>Why Use This Book</i>	page xxi
<i>Simplified Notation</i>	xxiv
<i>Acknowledgments</i>	xxv
<b>I DATA EXPLORATION</b>	<b>1</b>
1 Origins of Data	3
2 Preparing Data for Analysis	30
3 Exploratory Data Analysis	58
4 Comparison and Correlation	96
5 Generalizing from Data	118
6 Testing Hypotheses	143
<b>II REGRESSION ANALYSIS</b>	<b>169</b>
7 Simple Regression	171
8 Complicated Patterns and Messy Data	200
9 Generalizing Results of a Regression	236
10 Multiple Linear Regression	266
11 Modeling Probabilities	297
12 Regression with Time Series Data	329
<b>III PREDICTION</b>	<b>363</b>
13 A Framework for Prediction	365
14 Model Building for Prediction	391
15 Regression Trees	417
16 Random Forest and Boosting	438

---

<b>vi</b>	<b>Brief Contents</b>	
	<hr/>	
<b>17</b>	<b>Probability Prediction and Classification</b>	<b>457</b>
<b>18</b>	<b>Forecasting from Time Series Data</b>	<b>487</b>
<b>IV</b>	<b>CAUSAL ANALYSIS</b>	<b>517</b>
	<hr/>	
<b>19</b>	<b>A Framework for Causal Analysis</b>	<b>519</b>
<b>20</b>	<b>Designing and Analyzing Experiments</b>	<b>555</b>
<b>21</b>	<b>Regression and Matching with Observational Data</b>	<b>588</b>
<b>22</b>	<b>Difference-in-Differences</b>	<b>620</b>
<b>23</b>	<b>Methods for Panel Data</b>	<b>649</b>
<b>24</b>	<b>Appropriate Control Groups for Panel Data</b>	<b>681</b>
	<i>References</i>	704
	<i>Index</i>	709

## CONTENTS

<i>Why Use This Book</i>	<i>page</i> xxi
<i>Simplified Notation</i>	xxiv
<i>Acknowledgments</i>	xxv
<b>I DATA EXPLORATION</b>	<b>1</b>
<b>1 Origins of Data</b>	<b>3</b>
1.1 What Is Data?	4
1.2 Data Structures	5
1.A1 CASE STUDY – Finding a Good Deal among Hotels: Data Collection	6
1.3 Data Quality	7
1.B1 CASE STUDY – Comparing Online and Offline Prices: Data Collection	9
1.C1 CASE STUDY – Management Quality and Firm Performance: Data Collection	10
1.4 How Data Is Born: The Big Picture	11
1.5 Collecting Data from Existing Sources	12
1.A2 CASE STUDY – Finding a Good Deal among Hotels: Data Collection	14
1.B2 CASE STUDY – Comparing Online and Offline Prices: Data Collection	15
1.6 Surveys	16
1.C2 CASE STUDY – Management Quality and Firm Size: Data Collection	18
1.7 Sampling	18
1.8 Random Sampling	19
1.B3 CASE STUDY – Comparing Online and Offline Prices: Data Collection	21
1.C3 CASE STUDY – Management Quality and Firm Size: Data Collection	21
1.9 Big Data	22
1.10 Good Practices in Data Collection	24
1.11 Ethical and Legal Issues of Data Collection	26
1.12 Main Takeaways	27
Practice Questions	27
Data Exercises	28
References and Further Reading	28
<b>2 Preparing Data for Analysis</b>	<b>30</b>
2.1 Types of Variables	31
2.2 Stock Variables, Flow Variables	33
2.3 Types of Observations	33
2.4 Tidy Data	35
2.A1 CASE STUDY – Finding a Good Deal among Hotels: Data Preparation	36
2.5 Tidy Approach for Multi-dimensional Data	37
2.B1 CASE STUDY – Displaying Immunization Rates across Countries	37
2.6 Relational Data and Linking Data Tables	38

2.C1	CASE STUDY – Identifying Successful Football Managers	40
2.7	Entity Resolution: Duplicates, Ambiguous Identification, and Non-entity Rows	42
2.C2	CASE STUDY – Identifying Successful Football Managers	43
2.8	Discovering Missing Values	44
2.9	Managing Missing Values	46
2.A2	CASE STUDY – Finding a Good Deal among Hotels: Data Preparation	47
2.10	The Process of Cleaning Data	48
2.11	Reproducible Workflow: Write Code and Document Your Steps	49
2.12	Organizing Data Tables for a Project	50
2.C3	CASE STUDY – Identifying Successful Football Managers	52
2.C4	CASE STUDY – Identifying Successful Football Managers	53
2.13	Main Takeaways	54
	Practice Questions	54
	Data Exercises	55
	References and Further Reading	56
2.U1	Under the Hood: Naming Files	56
<b>3</b>	<b>Exploratory Data Analysis</b>	<b>58</b>
3.1	Why Do Exploratory Data Analysis?	59
3.2	Frequencies and Probabilities	60
3.3	Visualizing Distributions	61
3.A1	CASE STUDY – Finding a Good Deal among Hotels: Data Exploration	62
3.4	Extreme Values	65
3.A2	CASE STUDY – Finding a Good Deal among Hotels: Data Exploration	66
3.5	Good Graphs: Guidelines for Data Visualization	68
3.A3	CASE STUDY – Finding a Good Deal among Hotels: Data Exploration	71
3.6	Summary Statistics for Quantitative Variables	72
3.B1	CASE STUDY – Comparing Hotel Prices in Europe: Vienna vs. London	74
3.7	Visualizing Summary Statistics	77
3.C1	CASE STUDY – Measuring Home Team Advantage in Football	78
3.8	Good Tables	80
3.C2	CASE STUDY – Measuring Home Team Advantage in Football	82
3.9	Theoretical Distributions	83
3.D1	CASE STUDY – Distributions of Body Height and Income	85
3.10	Steps of Exploratory Data Analysis	87
3.11	Main Takeaways	88
	Practice Questions	88
	Data Exercises	89
	References and Further Reading	90
3.U1	Under the Hood: More on Theoretical Distributions	90
	Bernoulli Distribution	91
	Binomial Distribution	91
	Uniform Distribution	92
	Power-Law Distribution	92

Contents	ix
<b>4 Comparison and Correlation</b>	<b>96</b>
4.1 The $y$ and the $x$	97
4.A1 CASE STUDY – Management Quality and Firm Size: Describing Patterns of Association	98
4.2 Conditioning	100
4.3 Conditional Probabilities	101
4.A2 CASE STUDY – Management Quality and Firm Size: Describing Patterns of Association	102
4.4 Conditional Distribution, Conditional Expectation	103
4.5 Conditional Distribution, Conditional Expectation with Quantitative $x$	104
4.A3 CASE STUDY – Management Quality and Firm Size: Describing Patterns of Association	105
4.6 Dependence, Covariance, Correlation	108
4.7 From Latent Variables to Observed Variables	110
4.A4 CASE STUDY – Management Quality and Firm Size: Describing Patterns of Association	111
4.8 Sources of Variation in $x$	113
4.9 Main Takeaways	114
Practice Questions	115
Data Exercises	115
References and Further Reading	116
4.U1 Under the Hood: Inverse Conditional Probabilities, Bayes' Rule	116
<b>5 Generalizing from Data</b>	<b>118</b>
5.1 When to Generalize and to What?	119
5.A1 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	121
5.2 Repeated Samples, Sampling Distribution, Standard Error	122
5.A2 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	123
5.3 Properties of the Sampling Distribution	125
5.A3 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	127
5.4 The confidence interval	128
5.A4 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	129
5.5 Discussion of the CI: Confidence or Probability?	129
5.6 Estimating the Standard Error with the Bootstrap Method	130
5.A5 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	132
5.7 The Standard Error Formula	133
5.A6 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	134
5.8 External Validity	135
5.A7 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	136
5.9 Big Data, Statistical Inference, External Validity	137
5.10 Main Takeaways	138
Practice Questions	138
Data Exercises	139
References and Further Reading	139
5.U1 Under the Hood: The Law of Large Numbers and the Central Limit Theorem	140

x	Contents	
	<b>6 Testing Hypotheses</b>	<b>143</b>
	6.1 The Logic of Testing Hypotheses	144
	6.A1 CASE STUDY – Comparing Online and Offline Prices: Testing the Difference	145
	6.2 Null Hypothesis, Alternative Hypothesis	148
	6.3 The t-Test	149
	6.4 Making a Decision; False Negatives, False Positives	150
	6.5 The p-Value	154
	6.A2 CASE STUDY – Comparing Online and Offline Prices: Testing the Difference	155
	6.6 Steps of Hypothesis Testing	157
	6.7 One-Sided Alternatives	158
	6.B1 CASE STUDY – Testing the Likelihood of Loss on a Stock Portfolio	159
	6.8 Testing Multiple Hypotheses	160
	6.A3 CASE STUDY – Comparing Online and Offline Prices: Testing the Difference	161
	6.9 p-Hacking	162
	6.10 Testing Hypotheses with Big Data	164
	6.11 Main Takeaways	165
	Practice Questions	165
	Data Exercises	166
	References and Further Reading	167
	<b>II REGRESSION ANALYSIS</b>	<b>169</b>
	<b>7 Simple Regression</b>	<b>171</b>
	7.1 When and Why Do Simple Regression Analysis?	172
	7.2 Regression: Definition	172
	7.3 Non-parametric Regression	174
	7.A1 CASE STUDY – Finding a Good Deal among Hotels with Simple Regression	175
	7.4 Linear Regression: Introduction	178
	7.5 Linear Regression: Coefficient Interpretation	179
	7.6 Linear Regression with a Binary Explanatory Variable	180
	7.7 Coefficient Formula	181
	7.A2 CASE STUDY – Finding a Good Deal among Hotels with Simple Regression	183
	7.8 Predicted Dependent Variable and Regression Residual	184
	7.A3 CASE STUDY – Finding a Good Deal among Hotels with Simple Regression	185
	7.9 Goodness of Fit, R-Squared	188
	7.10 Correlation and Linear Regression	189
	7.11 Regression Analysis, Regression toward the Mean, Mean Reversion	190
	7.12 Regression and Causation	190
	7.A4 CASE STUDY – Finding a Good Deal among Hotels with Simple Regression	192
	7.13 Main Takeaways	192
	Practice Questions	193
	Data Exercises	193
	References and Further Reading	194

Contents	xi
7.U1 Under the Hood: Derivation of the OLS Formulae for the Intercept and Slope Coefficients	194
7.U2 Under the Hood: More on Residuals and Predicted Values with OLS	197
<b>8 Complicated Patterns and Messy Data</b>	<b>200</b>
8.1 When and Why Care about the Shape of the Association between $y$ and $x$ ?	201
8.2 Taking Relative Differences or Log	202
8.3 Log Transformation and Non-positive Values	204
8.4 Interpreting Log Values in a Regression	206
8.A1 CASE STUDY – Finding a Good Deal among Hotels with Nonlinear Function	207
8.5 Other Transformations of Variables	210
8.B1 CASE STUDY – How is Life Expectancy Related to the Average Income of a Country?	210
8.6 Regression with a Piecewise Linear Spline	215
8.7 Regression with Polynomial	216
8.8 Choosing a Functional Form in a Regression	218
8.B2 CASE STUDY – How is Life Expectancy Related to the Average Income of a Country?	219
8.9 Extreme Values and Influential Observations	221
8.10 Measurement Error in Variables	222
8.11 Classical Measurement Error	223
8.C1 CASE STUDY – Hotel Ratings and Measurement Error	225
8.12 Non-classical Measurement Error and General Advice	227
8.13 Using Weights in Regression Analysis	228
8.B3 CASE STUDY – How is Life Expectancy Related to the Average Income of a Country?	229
8.14 Main Takeaways	230
Practice Questions	231
Data Exercises	232
References and Further Reading	232
8.U1 Under the Hood: Details of the Log Approximation	233
8.U2 Under the Hood: Deriving the Consequences of Classical Measurement Error	234
<b>9 Generalizing Results of a Regression</b>	<b>236</b>
9.1 Generalizing Linear Regression Coefficients	237
9.2 Statistical Inference: CI and SE of Regression Coefficients	238
9.A1 CASE STUDY – Estimating Gender and Age Differences in Earnings	240
9.3 Intervals for Predicted Values	243
9.A2 CASE STUDY – Estimating Gender and Age Differences in Earnings	245
9.4 Testing Hypotheses about Regression Coefficients	249
9.5 Testing More Complex Hypotheses	251
9.A3 CASE STUDY – Estimating Gender and Age Differences in Earnings	252
9.6 Presenting Regression Results	253
9.A4 CASE STUDY – Estimating Gender and Age Differences in Earnings	254
9.7 Data Analysis to Help Assess External Validity	256

9.B1	CASE STUDY – How Stable is the Hotel Price–Distance to Center Relationship?	256
9.8	Main Takeaways	260
	Practice Questions	261
	Data Exercises	261
	References and Further Reading	262
9.U1	Under the Hood: The Simple SE Formula for Regression Intercept	262
9.U2	Under the Hood: The Law of Large Numbers for $\hat{\beta}$	263
9.U3	Under the Hood: Deriving $SE(\hat{\beta})$ with the Central Limit Theorem	264
9.U4	Under the Hood: Degrees of Freedom Adjustment for the SE Formula	265
<b>10</b>	<b>Multiple Linear Regression</b>	<b>266</b>
10.1	Multiple Regression: Why and When?	267
10.2	Multiple Linear Regression with Two Explanatory Variables	267
10.3	Multiple Regression and Simple Regression: Omitted Variable Bias	268
10.A1	CASE STUDY – Understanding the Gender Difference in Earnings	270
10.4	Multiple Linear Regression Terminology	272
10.5	Standard Errors and Confidence Intervals in Multiple Linear Regression	273
10.6	Hypothesis Testing in Multiple Linear Regression	275
10.A2	CASE STUDY – Understanding the Gender Difference in Earnings	275
10.7	Multiple Linear Regression with Three or More Explanatory Variables	276
10.8	Nonlinear Patterns and Multiple Linear Regression	277
10.A3	CASE STUDY – Understanding the Gender Difference in Earnings	278
10.9	Qualitative Right-Hand-Side Variables	279
10.A4	CASE STUDY – Understanding the Gender Difference in Earnings	280
10.10	Interactions: Uncovering Different Slopes across Groups	282
10.A5	CASE STUDY – Understanding the Gender Difference in Earnings	284
10.11	Multiple Regression and Causal Analysis	286
10.A6	CASE STUDY – Understanding the Gender Difference in Earnings	287
10.12	Multiple Regression and Prediction	290
10.B1	CASE STUDY – Finding a Good Deal among Hotels with Multiple Regression	292
10.13	Main Takeaways	294
	Practice Questions	294
	Data Exercises	295
	References and Further Reading	296
10.U1	Under the Hood: A Two-Step Procedure to Get the Multiple Regression Coefficient	296
<b>11</b>	<b>Modeling Probabilities</b>	<b>297</b>
11.1	The Linear Probability Model	298
11.2	Predicted Probabilities in the Linear Probability Model	299
11.A1	CASE STUDY – Does Smoking Pose a Health Risk?	301
11.3	Logit and Probit	307
11.A2	CASE STUDY – Does Smoking Pose a Health Risk?	308
11.4	Marginal Differences	309
11.A3	CASE STUDY – Does Smoking Pose a Health Risk?	311

11.5	Goodness of Fit: R-Squared and Alternatives	312
11.6	The Distribution of Predicted Probabilities	314
11.7	Bias and Calibration	314
11.B1	CASE STUDY – Are Australian Weather Forecasts Well Calibrated?	315
11.8	Refinement	317
11.A4	CASE STUDY – Does Smoking Pose a Health risk?	318
11.9	Using Probability Models for Other Kinds of $y$ Variables	321
11.10	Main Takeaways	323
	Practice Questions	323
	Data Exercises	324
	References and Further Reading	325
11.U1	Under the Hood: Saturated Models	325
11.U2	Under the Hood: Maximum Likelihood Estimation and Search Algorithms	326
11.U3	Under the Hood: From Logit and Probit Coefficients to Marginal Differences	327
<b>12</b>	<b>Regression with Time Series Data</b>	<b>329</b>
12.1	Preparation of Time Series Data	330
12.2	Trend and Seasonality	332
12.3	Stationarity, Non-stationarity, Random Walk	333
12.A1	CASE STUDY – Returns on a Company Stock and Market Returns	335
12.4	Time Series Regression	338
12.A2	CASE STUDY – Returns on a Company Stock and Market Returns	339
12.5	Trends, Seasonality, Random Walks in a Regression	343
12.B1	CASE STUDY – Electricity Consumption and Temperature	346
12.6	Serial Correlation	349
12.7	Dealing with Serial Correlation in Time Series Regressions	350
12.B2	CASE STUDY – Electricity Consumption and Temperature	352
12.8	Lags of $x$ in a Time Series Regression	355
12.B3	CASE STUDY – Electricity Consumption and Temperature	357
12.9	The Process of Time Series Regression Analysis	359
12.10	Main Takeaways	360
	Practice Questions	360
	Data Exercises	361
	References and Further Reading	362
12.U1	Under the Hood: Testing for Unit Root	362
<b>III</b>	<b>PREDICTION</b>	<b>363</b>
<b>13</b>	<b>A Framework for Prediction</b>	<b>365</b>
13.1	Prediction Basics	366
13.2	Various Kinds of Prediction	367
13.A1	CASE STUDY – Predicting Used Car Value with Linear Regressions	369
13.3	The Prediction Error and Its Components	369
13.A2	CASE STUDY – Predicting Used Car Value with Linear Regressions	371
13.4	The Loss Function	373

13.5	Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)	375
13.6	Bias and Variance of Predictions	376
13.7	The Task of Finding the Best Model	377
13.8	Finding the Best Model by Best Fit and Penalty: The BIC	379
13.9	Finding the Best Model by Training and Test Samples	380
13.10	Finding the Best Model by Cross-Validation	382
13.A3	CASE STUDY – Predicting Used Car Value with Linear Regressions	383
13.11	External Validity and Stable Patterns	384
13.A4	CASE STUDY – Predicting Used Car Value with Linear Regressions	386
13.12	Machine Learning and the Role of Algorithms	387
13.13	Main Takeaways	389
	Practice Questions	389
	Data Exercises	390
	References and Further Reading	390
<b>14</b>	<b>Model Building for Prediction</b>	<b>391</b>
14.1	Steps of Prediction	392
14.2	Sample Design	393
14.3	Label Engineering and Predicting Log $y$	394
14.A1	CASE STUDY – Predicting Used Car Value: Log Prices	395
14.4	Feature Engineering: Dealing with Missing Values	397
14.5	Feature Engineering: What $x$ Variables to Have and in What Functional Form	398
14.B1	CASE STUDY – Predicting Airbnb Apartment Prices: Selecting a Regression Model	399
14.6	We Can't Try Out All Possible Models	402
14.7	Evaluating the Prediction Using a Holdout Set	403
14.B2	CASE STUDY – Predicting Airbnb Apartment Prices: Selecting a Regression Model	404
14.8	Selecting Variables in Regressions by LASSO	407
14.B3	CASE STUDY – Predicting Airbnb Apartment Prices: Selecting a Regression Model	409
14.9	Diagnostics	410
14.B4	CASE STUDY – Predicting Airbnb Apartment Prices: Selecting a Regression Model	411
14.10	Prediction with Big Data	412
14.11	Main Takeaways	414
	Practice Questions	414
	Data Exercises	415
	References and Further Reading	415
14.U1	Under the Hood: Text Parsing	415
14.U2	Under the Hood: Log Correction	416
<b>15</b>	<b>Regression Trees</b>	<b>417</b>
15.1	The Case for Regression Trees	418
15.2	Regression Tree Basics	419

15.3	Measuring Fit and Stopping Rules	420
15.A1	CASE STUDY – Predicting Used Car Value with a Regression Tree	421
15.4	Regression Tree with Multiple Predictor Variables	425
15.5	Pruning a Regression Tree	426
15.6	A Regression Tree is a Non-parametric Regression	426
15.A2	CASE STUDY – Predicting Used Car Value with a Regression Tree	427
15.7	Variable Importance	430
15.8	Pros and Cons of Using a Regression Tree for Prediction	431
15.A3	CASE STUDY – Predicting Used Car Value with a Regression Tree	433
15.9	Main Takeaways	435
	Practice Questions	435
	Data Exercises	436
	References and Further Reading	437
<b>16</b>	<b>Random Forest and Boosting</b>	<b>438</b>
16.1	From a Tree to a Forest: Ensemble Methods	439
16.2	Random Forest	440
16.3	The Practice of Prediction with Random Forest	442
16.A1	CASE STUDY – Predicting Airbnb Apartment Prices with Random Forest	443
16.4	Diagnostics: The Variable Importance Plot	444
16.5	Diagnostics: The Partial Dependence Plot	445
16.6	Diagnostics: Fit in Various Subsets	446
16.A2	CASE STUDY – Predicting Airbnb Apartment Prices with Random Forest	446
16.7	An Introduction to Boosting and the GBM Model	449
16.A3	CASE STUDY – Predicting Airbnb Apartment Prices with Random Forest	450
16.8	A Review of Different Approaches to Predict a Quantitative $y$	452
16.9	Main Takeaways	454
	Practice Questions	454
	Data Exercises	455
	References and Further Reading	456
<b>17</b>	<b>Probability Prediction and Classification</b>	<b>457</b>
17.1	Predicting a Binary $y$ : Probability Prediction and Classification	458
17.A1	CASE STUDY – Predicting Firm Exit: Probability and Classification	459
17.2	The Practice of Predicting Probabilities	462
17.A2	CASE STUDY – Predicting Firm Exit: Probability and Classification	463
17.3	Classification and the Confusion Table	466
17.4	Illustrating the Trade-Off between Different Classification Thresholds: The ROC Curve	468
17.A3	CASE STUDY – Predicting Firm Exit: Probability and Classification	469
17.5	Loss Function and Finding the Optimal Classification Threshold	471
17.A4	CASE STUDY – Predicting Firm Exit: Probability and Classification	473
17.6	Probability Prediction and Classification with Random Forest	475
17.A5	CASE STUDY – Predicting Firm Exit: Probability and Classification	477
17.7	Class Imbalance	480
17.8	The Process of Prediction with a Binary Target Variable	481

17.9	Main Takeaways	482
	Practice Questions	482
	Data Exercises	483
	References and Further Reading	484
17.U1	Under the Hood: The Gini Node Impurity Measure and MSE	484
17.U2	Under the Hood: On the Method of Finding an Optimal Threshold	485
<b>18</b>	<b>Forecasting from Time Series Data</b>	<b>487</b>
18.1	Forecasting: Prediction Using Time Series Data	488
18.2	Holdout, Training, and Test Samples in Time Series Data	489
18.3	Long-Horizon Forecasting: Seasonality and Predictable Events	491
18.4	Long-Horizon Forecasting: Trends	492
18.A1	CASE STUDY – Forecasting Daily Ticket Volumes for a Swimming Pool	494
18.5	Forecasting for a Short Horizon Using the Patterns of Serial Correlation	500
18.6	Modeling Serial Correlation: AR(1)	500
18.7	Modeling Serial Correlation: ARIMA	501
18.B1	CASE STUDY – Forecasting a Home Price Index	503
18.8	VAR: Vector Autoregressions	505
18.B2	CASE STUDY – Forecasting a Home Price Index	507
18.9	External Validity of Forecasts	509
18.B3	CASE STUDY – Forecasting a Home Price Index	510
18.10	Main Takeaways	512
	Practice Questions	512
	Data Exercises	513
	References and Further Reading	514
18.U1	Under the Hood: Details of the ARIMA Model	514
18.U2	Under the Hood: Auto-Arima	516
<b>IV</b>	<b>CAUSAL ANALYSIS</b>	<b>517</b>
<b>19</b>	<b>A Framework for Causal Analysis</b>	<b>519</b>
19.1	Intervention, Treatment, Subjects, Outcomes	520
19.2	Potential Outcomes	522
19.3	The Individual Treatment Effect	523
19.4	Heterogeneous Treatment Effects	524
19.5	ATE: The Average Treatment Effect	525
19.6	Average Effects in Subgroups and ATET	527
19.7	Quantitative Causal Variables	527
19.A1	CASE STUDY – Food and Health	528
19.8	Ceteris Paribus: Other Things Being the Same	530
19.9	Causal Maps	531
19.10	Comparing Different Observations to Uncover Average Effects	533
19.11	Random Assignment	535
19.12	Sources of Variation in the Causal Variable	536
19.A2	CASE STUDY – Food and Health	537

Contents	xvii
19.13 Experimenting versus Conditioning	539
19.14 Confounders in Observational Data	541
19.15 From Latent Variables to Measured Variables	543
19.16 Bad Conditioners: Variables Not to Condition On	544
19.A3 CASE STUDY – Food and Health	545
19.17 External Validity, Internal Validity	549
19.18 Constructive Skepticism	551
19.19 Main Takeaways	552
Practice Questions	552
Data Exercises	553
References and Further Reading	554
<b>20 Designing and Analyzing Experiments</b>	<b>555</b>
20.1 Randomized Experiments and Potential Outcomes	556
20.2 Field Experiments, A/B Testing, Survey Experiments	557
20.A1 CASE STUDY – Working from Home and Employee Performance	558
20.B1 CASE STUDY – Fine Tuning Social Media Advertising	559
20.3 The Experimental Setup: Definitions	560
20.4 Random Assignment in Practice	560
20.5 Number of Subjects and Proportion Treated	562
20.6 Random Assignment and Covariate Balance	563
20.A2 CASE STUDY – Working from Home and Employee Performance	565
20.7 Imperfect Compliance and Intent-to-Treat	567
20.A3 CASE STUDY – Working from Home and Employee Performance	569
20.8 Estimation and Statistical Inference	570
20.B2 CASE STUDY – Fine Tuning Social Media Advertising	571
20.9 Including Covariates in a Regression	572
20.A4 CASE STUDY – Working from Home and Employee Performance	573
20.10 Spillovers	576
20.11 Additional Threats to Internal Validity	577
20.A5 CASE STUDY – Working from Home and Employee Performance	579
20.12 External Validity, and How to Use the Results in Decision Making	581
20.A6 CASE STUDY – Working from Home and Employee Performance	582
20.13 Main Takeaways	583
Practice Questions	584
Data Exercises	585
References and Further Reading	585
20.U1 Under the Hood: LATE: The Local Average Treatment Effect	586
20.U2 Under the Hood: The Formula for Sample Size Calculation	586
<b>21 Regression and Matching with Observational Data</b>	<b>588</b>
21.1 Thought Experiments	589
21.A1 CASE STUDY – Founder/Family Ownership and Quality of Management	590
21.2 Variables to Condition on, Variables Not to Condition On	591
21.A2 CASE STUDY – Founder/Family Ownership and Quality of Management	592

21.3	Conditioning on Confounders by Regression	595
21.4	Selection of Variables and Functional Form in a Regression for Causal Analysis	597
21.A3	CASE STUDY – Founder/Family Ownership and Quality of Management	598
21.5	Matching	601
21.6	Common Support	603
21.7	Matching on the Propensity Score	604
21.A4	CASE STUDY – Founder/Family Ownership and Quality of Management	605
21.8	Comparing Linear Regression and Matching	607
21.A5	CASE STUDY – Founder/Family Ownership and Quality of Management	609
21.9	Instrumental Variables	610
21.10	Regression-Discontinuity	613
21.11	Main Takeaways	614
	Practice Questions	614
	Data Exercises	615
	References and Further Reading	616
21.U1	Under the Hood: Unobserved Heterogeneity and Endogenous $x$ in a Regression	616
21.U2	Under the hood: LATE is IV	618
<b>22</b>	<b>Difference-in-Differences</b>	<b>620</b>
22.1	Conditioning on Pre-intervention Outcomes	621
22.2	Basic Difference-in-Differences Analysis: Comparing Average Changes	622
22.A1	CASE STUDY – How Does a Merger between Airlines Affect Prices?	625
22.3	The Parallel Trends Assumption	629
22.A2	CASE STUDY – How Does a Merger between Airlines Affect Prices?	631
22.4	Conditioning on Additional Confounders in Diff-in-Diffs Regressions	633
22.A3	CASE STUDY – How Does a Merger between Airlines Affect Prices?	635
22.5	Quantitative Causal Variable	637
22.A4	CASE STUDY – How Does a Merger between Airlines Affect Prices?	638
22.6	Difference-in-Differences with Pooled Cross-Sections	640
22.A5	CASE STUDY – How Does a Merger between Airlines Affect Prices?	643
22.7	Main Takeaways	645
	Practice Questions	646
	Data Exercises	647
	References and Further Reading	648
<b>23</b>	<b>Methods for Panel Data</b>	<b>649</b>
23.1	Multiple Time Periods Can Be Helpful	650
23.2	Estimating Effects Using Observational Time Series	651
23.3	Lags to Estimate the Time Path of Effects	653
23.4	Leads to Examine Pre-trends and Reverse Effects	653
23.5	Pooled Time Series to Estimate the Effect for One Unit	654
23.A1	CASE STUDY – Import Demand and Industrial Production	656
23.6	Panel Regression with Fixed Effects	659
23.7	Aggregate Trend	661

23.B1	CASE STUDY – Immunization against Measles and Saving Children	662
23.8	Clustered Standard Errors	665
23.9	Panel Regression in First Differences	666
23.10	Lags and Leads in FD Panel Regressions	667
23.B2	CASE STUDY – Immunization against Measles and Saving Children	669
23.11	Aggregate Trend and Individual Trends in FD Models	671
23.B3	CASE STUDY – Immunization against Measles and Saving Children	672
23.12	Panel Regressions and Causality	674
23.13	First Differences or Fixed Effects?	675
23.14	Dealing with Unbalanced Panels	677
23.15	Main Takeaways	678
	Practice Questions	678
	Data Exercises	680
	References and Further Reading	680
<b>24</b>	<b>Appropriate Control Groups for Panel Data</b>	<b>681</b>
24.1	When and Why to Select a Control Group in xt Panel Data	682
24.2	Comparative Case Studies	682
24.3	The Synthetic Control Method	683
24.A1	CASE STUDY – Estimating the Effect of the 2010 Haiti Earthquake on GDP	684
24.4	Event Studies	687
24.B1	CASE STUDY – Estimating the Impact of Replacing Football Team Managers	690
24.5	Selecting a Control Group in Event Studies	694
24.B2	CASE STUDY – Estimating the Impact of Replacing Football Team Managers	696
24.6	Main Takeaways	701
	Practice Questions	701
	Data Exercises	702
	References and Further Reading	703
	<i>References</i>	704
	<i>Index</i>	709