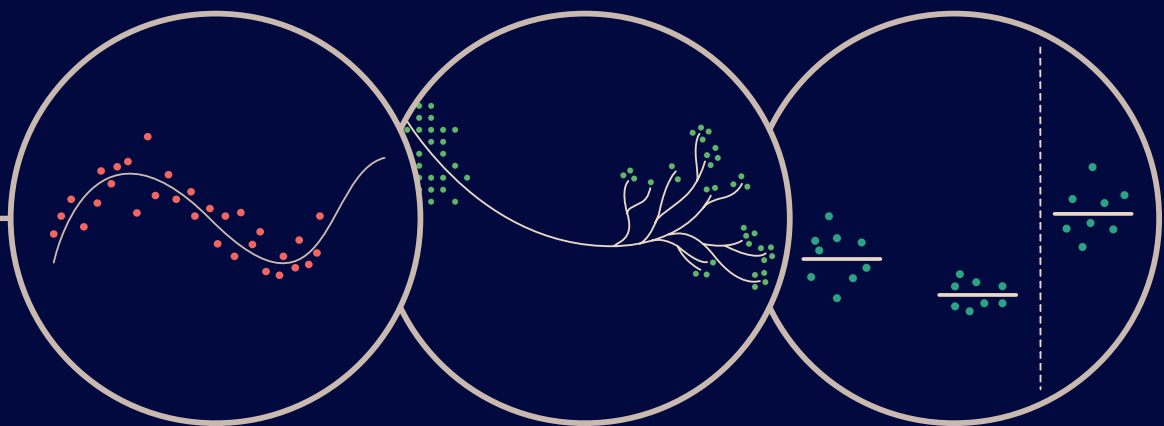


DATA ANALYSIS

FOR BUSINESS, ECONOMICS, AND POLICY

Gábor Békés | Gábor Kézdi



DATA ANALYSIS FOR BUSINESS, ECONOMICS, AND POLICY

This textbook provides future data analysts with the tools, methods, and skills needed to answer data-focused, real-life questions; to carry out data analysis; and to visualize and interpret results to support better decisions in business, economics, and public policy.

Data wrangling and exploration, regression analysis, machine learning, and causal analysis are comprehensively covered, as well as when, why, and how the methods work, and how they relate to each other.

As the most effective way to communicate data analysis, running case studies play a central role in this textbook. Each case starts with an industry-relevant question and answers it by using real-world data and applying the tools and methods covered in the textbook. Learning is then consolidated by over 420 practice questions and 117 data exercises.

Extensive online resources, including raw and cleaned data and codes for all analysis in Stata, R, and Python, can be found at www.gabors-data-analysis.com.

Gábor Békés is an Assistant Professor at the Department of Economics and Business of the Central European University, and director of the Business Analytics Program. He is a research affiliate at the Center for Economic Policy Research (CEPR). He has published in top economics journals on multinational firm activities and productivity, business clusters, and innovation spillovers. He managed international data collection projects on firm performance and supply chains. He has done both policy advising (the European Commission, ECB) as well as private-sector consultancy (in finance, business intelligence, and real estate). He has taught graduate-level data analysis and economic geography courses since 2012.

Gábor Kézdi is a research associate professor at the University of Michigan's Institute for Social Research. He has published in top journals in economics, statistics, and political science on topics including household finances, health, education, demography, and ethnic disadvantages and prejudice. He has managed several data collection projects in Europe; currently, he is co-investigator of the Health and Retirement Study in the US. He has consulted for various governmental and non-governmental institutions on the disadvantage of the Roma minority and the evaluation of social interventions. He has taught data analysis, econometrics, and labor economics from undergraduate to Ph.D. levels since 2002, and supervised a number of MA and Ph.D. students.

PROOF

DATA ANALYSIS FOR BUSINESS, ECONOMICS, AND POLICY

Gábor Békés

Central European University, Budapest

Gábor Kézdi

University of Michigan, Ann Arbor



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108483018

DOI: 10.1017/9781108591102

© Gábor Békés and Gábor Kézdi 2021

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2021

Printed in Singapore by Markono Print Media Pte Ltd 2021

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

ISBN 978-1-108-48301-8 Hardback

ISBN 978-1-108-71620-8 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

BRIEF CONTENTS

<i>Why Read This Book</i>	<i>page</i> <i>xxi</i>
<i>Simplified notation</i>	<i>xxiv</i>
<i>Acknowledgments</i>	<i>xxv</i>
I DATA EXPLORATION	1
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
II REGRESSION ANALYSIS	169
7 Simple Regression	171
8 Complicated Patterns and Messy Data	198
9 Generalizing Results of a Regression	234
10 Multiple Linear Regression	265
11 Modeling Probabilities	297
12 Regression with Time Series Data	330
III PREDICTION	365
13 A Framework for Prediction	367
14 Model Building for Prediction	393
15 Regression Trees	420
16 Random Forest and Boosting	441

17	Probability Prediction and Classification	459
18	Forecasting from Time Series Data	489
IV	CAUSAL ANALYSIS	519
19	A Framework for Causal Analysis	521
20	Designing and Analyzing Experiments	557
21	Regression and Matching with Observational Data	590
22	Difference-in-Differences	622
23	Methods for Panel Data	651
24	Appropriate Control Groups for Panel Data	683
	<i>Bibliography</i>	705
	<i>Index</i>	710

CONTENTS

<i>Why Read This Book</i>	page xxi
<i>Simplified notation</i>	xxiv
<i>Acknowledgments</i>	xxv

I DATA EXPLORATION 1

1 Origins of Data	3
1.1 What Is Data?	4
1.2 Data Structures	5
1.3 A1 CASE STUDY – Finding a Good Deal among Hotels: Data Collection	6
1.4 Data Quality	8
1.5 B1 CASE STUDY – Comparing Online and Offline Prices: Data Collection	9
1.6 C1 CASE STUDY – Management Quality and Firm Performance: Data Collection	10
1.7 How Data Is Born: The Big Picture	11
1.8 Collecting Data from Existing Sources	12
1.9 A2 CASE STUDY – Finding a Good Deal among Hotels: Data Collection	14
1.10 B2 CASE STUDY – Comparing Online and Offline Prices: Data Collection	15
1.11 Surveys	16
1.12 C2 CASE STUDY – Management Quality and Firm Size: Data Collection	18
1.13 Sampling	18
1.14 Random Sampling	19
1.15 B3 CASE STUDY – Comparing Online and Offline Prices: Data Collection	21
1.16 C3 CASE STUDY – Management Quality and Firm Size: Data Collection	22
1.17 Big Data	22
1.18 Good Practices in Data Collection	25
1.19 Ethical and Legal Issues of Data Collection	26
1.20 Summary and Practice	27
1.20.1 Main Takeaways	27
1.20.2 Practice Questions	27
1.20.3 Data Exercises	28
1.21 References and Further Reading	29
2 Preparing Data for Analysis	30
2.1 Types of Variables	31
2.2 Stock Variables, Flow Variables	33
2.3 Types of Observations	33
2.4 Tidy Data	35
2.5 A1 CASE STUDY – Finding a Good Deal among Hotels: Data Preparation	36
2.6 Tidy Approach for Multi-dimensional Data	37
2.7 B1 CASE STUDY – Displaying Immunization Rates across Countries	38
2.8 Relational Data and Linking Data Tables	39

2.9	C1 CASE STUDY – Identifying Successful Football Managers	40
2.10	Entity Resolution: Duplicates, Ambiguous Identification, and Non-entity Rows	43
2.11	C2 CASE STUDY – Identifying Successful Football Managers	44
2.12	Discovering Missing Values	45
2.13	Managing Missing Values	46
2.14	A2 CASE STUDY – Finding a Good Deal among Hotels: Data Preparation	47
2.15	The Process of Cleaning Data	48
2.16	Reproducible Workflow: Write Code and Document Your Steps	49
2.17	Organizing Data Tables for a Project	51
2.18	B3 CASE STUDY – Identifying Successful Football Managers	52
2.19	B4 CASE STUDY – Identifying Successful Football Managers	53
2.20	Summary and Practice	54
2.20.1	Main Takeaways	54
2.20.2	Practice Questions	54
2.20.3	Data Exercises	55
2.21	Under the Hood: Naming Files	56
2.22	References and Further Reading	57
3	Exploratory Data Analysis	58
3.1	Why Do Exploratory Data Analysis?	59
3.2	Frequencies and Probabilities	60
3.3	Visualizing Distributions	61
3.4	A1: CASE STUDY – Finding a Good Deal among Hotels: Data Exploration	62
3.5	Extreme Values	65
3.6	A2: CASE STUDY – Finding a Good Deal among Hotels: Data Exploration	66
3.7	Good Graphs: Guidelines for Data Visualization	68
3.8	A3 CASE STUDY – Finding a Good Deal among Hotels: Data Exploration	71
3.9	Summary Statistics for Quantitative Variables	72
3.10	B1 CASE STUDY – Comparing Hotel Prices in Europe: Vienna vs. London	74
3.11	Visualizing Summary Statistics	77
3.12	C1: CASE STUDY – Measuring Home Team Advantage in Football	78
3.13	Good Tables	80
3.14	C2 CASE STUDY – Measuring Home Team Advantage in Football	82
3.15	Theoretical Distributions	83
3.16	D1 CASE STUDY – Distributions of Body Height and Income	85
3.17	Steps of Exploratory Data Analysis	87
3.18	Summary and Practice	88
3.18.1	Main Takeaways	88
3.18.2	Practice Questions	88
3.18.3	Data Exercises	89
3.19	Under the Hood: More on Theoretical Distributions	90
3.19.1	Bernoulli Distribution	90
3.19.2	Binomial Distribution	91
3.19.3	Uniform Distribution	91
3.19.4	Power-Law Distribution	92

3.20	References and Further Reading	95
4	Comparison and Correlation	96
4.1	The y and the x	97
4.2	A1 CASE STUDY – Management Quality and Firm Size: Describing Patterns of Association	98
4.3	Conditioning	100
4.4	Conditional Probabilities	100
4.5	A2 CASE STUDY – Management Quality and Firm Size: Describing Patterns of Association	102
4.6	Conditional Distribution, Conditional Expectation	103
4.7	Conditional Distribution, Conditional Expectation with Quantitative x	104
4.8	A3 CASE STUDY – Management Quality and Firm Size: Describing Patterns of Association	105
4.9	Dependence, Covariance, Correlation	108
4.10	From Latent Variables to Observed Variables	109
4.11	A4 CASE STUDY – Management Quality and Firm Size: Describing Patterns of Association	111
4.12	Sources of Variation in x	113
4.13	Summary and Practice	114
	4.13.1 Main Takeaways	114
	4.13.2 Practice Questions	114
	4.13.3 Data Exercises	115
4.14	Under the Hood: Inverse Conditional Probabilities, Bayes' Rule	116
4.15	References and Further Reading	117
5	Generalizing from Data	118
5.1	When to Generalize and to What?	119
5.2	A1 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	121
5.3	Repeated Samples, Sampling Distribution, Standard Error	122
5.4	A2 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	123
5.5	Properties of the Sampling Distribution	125
5.6	A3 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	127
5.7	The confidence interval	128
5.8	A4 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	129
5.9	Discussion of the CI: Confidence or Probability?	129
5.10	Estimating the Standard Error with the Bootstrap Method	130
5.11	A5 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	132
5.12	The Standard Error Formula	134
5.13	A6 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	135
5.14	External Validity	135
5.15	A7 CASE STUDY – What Likelihood of Loss to Expect on a Stock Portfolio?	136
5.16	Big Data, Statistical Inference, External Validity	137
5.17	Summary and Practice	138
	5.17.1 Main Takeaways	138
	5.17.2 Practice Questions	138

5.17.3 Data exercises	139
5.18 Under the Hood: The Law of Large Numbers and the Central Limit Theorem	140
5.19 References and Further Reading	142
6 Testing Hypotheses	143
6.1 The Logic of Testing Hypotheses	144
6.2 A1 CASE STUDY – Comparing Online and Offline Prices: Testing the Difference	145
6.3 Null Hypothesis, Alternative Hypothesis	148
6.4 t-Test	149
6.5 Making a Decision; False Negatives, False Positives	150
6.6 p-Value	154
6.7 A2 CASE STUDY – Comparing Online and Offline Prices: Testing the Difference	155
6.8 Steps of Hypothesis Testing	157
6.9 One-Sided Alternatives	158
6.10 B1 CASE STUDY – Testing the Likelihood of Loss on a Stock Portfolio	159
6.11 Testing Multiple Hypotheses	160
6.12 A3 CASE STUDY – Comparing Online and Offline Prices: Testing the Difference	161
6.13 p-Hacking	163
6.14 Testing Hypotheses with Big Data	164
6.15 Summary and Practice	165
6.15.1 Main Takeaway	165
6.15.2 Practice Questions	165
6.15.3 Data Exercises	166
6.16 References and Further Reading	167
II REGRESSION ANALYSIS	169
7 Simple Regression	171
7.1 When and Why Do Simple Regression Analysis?	172
7.2 Regression: Definition	172
7.3 Non-parametric Regression	174
7.4 A1 Case Study – Finding a Good Deal among Hotels with Simple Regression	175
7.5 Linear Regression: Introduction	178
7.6 Linear Regression: Coefficient Interpretation	179
7.7 Linear Regression with a Binary Explanatory Variable	180
7.8 Coefficient Formula	181
7.9 A2 Case Study – Finding a Good Deal among Hotels with Simple regression	183
7.10 Predicted Dependent Variable and Regression Residual	184
7.11 A3 Case Study – Finding a Good Deal among Hotels with Simple regression	185
7.12 Goodness of Fit, R-Squared	188
7.13 Correlation and Linear Regression	189
7.14 Regression Analysis, Regression toward the Mean, Mean Reversion	190

Contents

xi

7.15	Regression and Causation	190
7.16	A4 Case Study – Finding a Good Deal among Hotels with Simple Regression	192
7.17	Summary and Practice	193
7.17.1	Main Takeaway	193
7.17.2	Practice Questions	193
7.17.3	Data Exercises	194
7.18	Under the Hood: Derivation of the OLS Formulae of the Intercept and Slope Coefficients	195
7.19	Under the Hood: More on Residuals and Predicted Values with OLS	196
7.20	References and Further Reading	197
8	Complicated Patterns and Messy Data	198
8.1	When and Why Care about the Shape of the Association between y and x ?	199
8.2	Taking Relative Differences or Log	200
8.3	Log Transformation and Non-positive Values	202
8.4	Interpreting Log Values in a Regression	204
8.5	A1 Case Study – Finding a Good Deal among Hotels with Nonlinear Function	205
8.6	Other Transformations of Variables	208
8.7	B1 Case Study – How is Life Expectancy Related to the Average Income of a Country?	208
8.8	Regression with a Piecewise Linear Spline	213
8.9	Regression with Polynomial	214
8.10	Choosing a Functional Form in a Regression	216
8.11	B2 Case Study – How is Life Expectancy Related to the Average Income of a Country?	217
8.12	Extreme Values and Influential Observations	219
8.13	Measurement Error in Variables	220
8.14	Classical Measurement Error	221
8.15	C1 Case Study – Hotel Ratings and Measurement Error	223
8.16	Non-classical Measurement Error and General Advice	225
8.17	Using Weights in Regression Analysis	226
8.18	B3 Case Study – How is Life Expectancy Related to the Average Income of a Country?	227
8.19	Summary and Practice	228
8.19.1	Main Takeaway	228
8.19.2	Practice Questions	229
8.19.3	Data Exercises	230
8.20	Under the Hood: Details of the Log Approximation	231
8.21	Under the Hood: Deriving the Consequences of Classical Measurement error	232
8.22	References and Further Reading	233
9	Generalizing Results of a Regression	234
9.1	Generalizing Linear Regression Coefficients	235
9.2	Statistical Inference: CI and SE of Regression Coefficients	236

9.3	A1 CASE STUDY – Estimating Gender and Age Differences in Earnings	238
9.4	Intervals for Predicted Values	241
9.5	A2 CASE STUDY – Estimating Gender and Age Differences in Earnings	243
9.6	Testing Hypotheses about Regression Coefficients	247
9.7	Testing More Complex Hypotheses	249
9.8	A3 CASE STUDY – Estimating Gender and Age Differences in Earnings	250
9.9	Presenting Regression Results	251
9.10	A4 CASE STUDY – Estimating Gender and Age Differences in Earnings	252
9.11	Data Analysis to Help Assess External Validity	254
9.12	B1 CASE STUDY – How Stable is the Hotel Price–Distance to Center Relationship?	255
9.13	Summary and Practice	259
9.13.1	Main Takeaway	259
9.13.2	Practice Questions	259
9.13.3	Data Exercises	260
9.14	Under the Hood: The Simple SE Formula for Regression Intercept	260
9.15	Under the Hood: The Law of Large Numbers for $\hat{\beta}$	261
9.16	Under the Hood: Deriving $SE(\hat{\beta})$ with the Central Limit Theorem	262
9.17	Under the Hood: Degrees of Freedom Adjustment for the SE Formula	263
9.18	References and Further Reading	264
10	Multiple Linear Regression	265
10.1	Multiple Regression: Why and When?	266
10.2	Multiple Linear Regression with Two Explanatory Variables	266
10.3	Multiple Regression and Simple Regression: Omitted Variable Bias	267
10.4	A1 CASE STUDY – Understanding the Gender Difference in Earnings	269
10.5	Multiple Linear Regression Terminology	271
10.6	Standard Errors and Confidence Intervals in Multiple Linear Regression	272
10.7	Hypothesis Testing in Multiple Linear Regression	274
10.8	A2 CASE STUDY – Understanding the Gender Difference in Earnings	274
10.9	Multiple Linear Regression with Three or More Explanatory Variables	275
10.10	Nonlinear Patterns and Multiple Linear Regression	276
10.11	A3 CASE STUDY – Understanding the Gender Difference in Earnings	277
10.12	Qualitative Right-Hand-Side Variables	278
10.13	A4 CASE STUDY – Understanding the Gender Difference in Earnings	279
10.14	Interactions: Uncovering Different Slopes across Groups	281
10.15	A5 CASE STUDY – Understanding the Gender Difference in Earnings	283
10.16	Multiple Regression and Causal Analysis	285
10.17	A6 CASE STUDY – Understanding the Gender Difference in Earnings	286
10.18	Multiple Regression and Prediction	289
10.19	B1 CASE STUDY – Finding a Good Deal among Hotels with Multiple Regression	291
10.20	Summary and Practice	293
10.20.1	Main Takeaways	293
10.20.2	Practice Questions	293
10.20.3	Data Exercises	294

10.21	Under the Hood: A Two-Step Procedure to Get the Multiple Regression Coefficient	295
10.22	References and Further Reading	296
11	Modeling Probabilities	297
11.1	The Linear Probability Model	298
11.2	Predicted Probabilities in the Linear Probability Model	299
11.3	A1 CASE STUDY – Does Smoking Pose a Health Risk?	301
11.4	Logit and Probit	307
11.5	A2 CASE STUDY – Does Smoking Pose a Health Risk?	308
11.6	Marginal Differences	309
11.7	A3 CASE STUDY – Does Smoking Pose a Health Risk?	311
11.8	Goodness of Fit: R-Squared and Alternatives	312
11.9	The Distribution of Predicted Probabilities	314
11.10	Bias and Calibration	314
11.11	B1 CASE STUDY – Are Australian Weather Forecasts Well Calibrated?	315
11.12	Refinement	317
11.13	A4 CASE STUDY – Does Smoking Pose a Health risk?	318
11.14	Using Probability Models for Other Kinds of y Variables	321
11.15	Summary and Practice	323
11.15.1	Main Takeaways	323
11.15.2	Practice Questions	323
11.15.3	Data Exercises	324
11.16	Under the Hood: Saturated Models	325
11.17	Under the Hood: Maximum Likelihood Estimation and Search Algorithms	326
11.18	Under the Hood: From logit and probit Coefficients to Marginal Differences	327
11.19	Under the Hood: Brier Score Decomposition	329
11.20	References and Further Reading	329
12	Regression with Time Series Data	330
12.1	Preparation of Time Series Data	331
12.2	Trend and Seasonality	333
12.3	Stationarity, Non-stationarity, Random Walk	334
12.4	A1 CASE STUDY – Returns on a Company Stock and Market Returns	336
12.5	Time Series Regression	339
12.6	A2 CASE STUDY – Returns on a Company Stock and Market Returns	340
12.7	Trends, Seasonality, Random Walks in a Regression	345
12.8	B1 CASE STUDY – Electricity Consumption and Temperature	347
12.9	Serial Correlation	350
12.10	Dealing with Serial Correlation in Time Series Regressions	351
12.11	B2 CASE STUDY – Electricity Consumption and Temperature	353
12.12	Lags of x in a Time Series Regression	357
12.13	B3 CASE STUDY – Electricity Consumption and Temperature	358
12.14	The Process of Time Series Regression Analysis	360
12.15	Summary and Practice	361
12.15.1	Main Takeaways	361

12.15.2	Practice Questions	362
12.15.3	Data Exercises	362
12.16	Under the Hood: Testing for Unit Root	363
12.17	References and Further Reading	363
III	PREDICTION	365
13	A Framework for Prediction	367
13.1	Prediction Basics	368
13.2	Various Kinds of Prediction	369
13.3	A1 CASE STUDY – Predicting Used Car Value with Linear Regressions	371
13.4	The Prediction Error and Its Components	371
13.5	A2 CASE STUDY – Predicting Used Car Value with Linear Regressions	373
13.6	The Loss Function	375
13.7	Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)	377
13.8	Bias and Variance of Predictions	378
13.9	The Task of Finding the Best Model	379
13.10	Finding the Best Model by Best Fit and Penalty: The BIC	381
13.11	Finding the Best Model by Training and Test Samples	382
13.12	Finding the Best Model by Cross-Validation	384
13.13	A3 CASE STUDY – Predicting Used Car Value with Linear Regressions	385
13.14	External Validity and Stable Patterns	386
13.15	A4 CASE STUDY – Predicting Used Car Value with Linear Regressions	388
13.16	Machine Learning and the Role of Algorithms	389
13.17	Summary and Practice	391
13.17.1	Main Takeaway	391
13.17.2	Practice Questions	391
13.17.3	Data Exercises	392
13.18	References and Further Reading	392
14	Model Building for Prediction	393
14.1	Steps of Prediction	394
14.2	Sample Design	395
14.3	Label Engineering and Predicting $\log y$	396
14.4	A1 CASE STUDY – Predicting Used Car Value: Log Prices	397
14.5	Feature Engineering: Dealing with Missing Values	399
14.6	Feature Engineering: What x Variables to Have and in What Functional Form	400
14.7	B1 CASE STUDY – Predicting Airbnb Apartment Prices: Selecting a Regression Model	401
14.8	We Can't Try Out All Possible Models	404
14.9	Evaluating the Prediction Using a Holdout Set	405
14.10	B2 CASE STUDY – Predicting Airbnb Apartment Prices: Selecting a Regression Model	406
14.11	Selecting Variables in Regressions by LASSO	409

14.12	B3 CASE STUDY – Predicting Airbnb Apartment Prices: Selecting a Regression Model	411
14.13	Diagnostics	412
14.14	B4 CASE STUDY – Predicting Airbnb Apartment Prices: Selecting a Regression Model	413
14.15	Prediction with Big Data	415
14.16	Summary and Practice	416
14.16.1	Main Takeaways	416
14.16.2	Practice Questions	416
14.16.3	Data Exercises	417
14.17	Under the Hood: Text Parsing	418
14.18	Under the Hood: Log Correction	418
14.19	References and Further Reading	419
15	Regression Trees	420
15.1	The Case for Regression Trees	421
15.2	Regression Tree Basics	422
15.3	Measuring Fit and Stopping Rules	423
15.4	A1 CASE STUDY – Predicting Used Car Value with a Regression Tree	424
15.5	Regression Tree with Multiple Predictor Variables	428
15.6	Pruning a Regression Tree	428
15.7	A Regression Tree is a Non-parametric Regression	429
15.8	A2 CASE STUDY – Predicting Used Car Value with a Regression Tree	430
15.9	Variable Importance	433
15.10	Pros and Cons of Using a Regression Tree for Prediction	434
15.11	A3 CASE STUDY – Predicting Used Car Value with a Regression Tree	435
15.12	Summary and Practice	437
15.12.1	Main Takeaways	437
15.12.2	Practice Questions	438
15.12.3	Data Exercises	439
15.13	References and Further Reading	439
16	Random Forest and Boosting	440
16.1	From a Tree to a Forest: Ensemble Methods	441
16.2	Random Forest	442
16.3	The Practice of Prediction with Random Forest	444
16.4	A1 CASE STUDY – Predicting Airbnb Apartment Prices with Random Forest	445
16.5	Diagnostics: The Variable Importance Plot	446
16.6	Diagnostics: The Partial Dependence Plot	447
16.7	Diagnostics: Fit in Various Subsets	448
16.8	A2 CASE STUDY – Predicting Airbnb Apartment Prices with Random Forest	448
16.9	An Introduction to Boosting and the GBM Model	451
16.10	A3 CASE STUDY – Predicting Airbnb Apartment Prices with Random Forest	452
16.11	A Review of Different Approaches to Predict a Quantitative y	454
16.12	Summary and Practice	456
16.12.1	Main Takeaways	456

16.12.2	Practice Questions	456
16.12.3	Data Exercises	457
16.13	References and Further Reading	458
17	Probability Prediction and Classification	459
17.1	Predicting a Binary y : Probability Prediction and Classification	460
17.2	A1 CASE STUDY – Predicting Firm Exit: Probability and Classification	461
17.3	The Practice of Predicting Probabilities	464
17.4	A2 CASE STUDY – Predicting Firm Exit: Probability and Classification	465
17.5	Classification and the Confusion Table	468
17.6	Illustrating the Trade-Off between Different Classification Thresholds: The ROC Curve	470
17.7	A3 CASE STUDY – Predicting Firm Exit: Probability and Classification	471
17.8	Loss Function and Finding the Optimal Classification Threshold	473
17.9	A4 CASE STUDY – Predicting Firm Exit: Probability and Classification	475
17.10	Probability Prediction and Classification with Random Forest	477
17.11	A5 CASE STUDY – Predicting Firm Exit: Probability and Classification	479
17.12	Class Imbalance	482
17.13	The Process of Prediction with a Binary Target Variable	483
17.14	Summary and Practice	484
17.14.1	Main Takeaway	484
17.14.2	Practice Questions	484
17.14.3	Data Exercises	485
17.15	Under the Hood: The Gini Node Impurity Measure and MSE	486
17.16	Under the Hood: On the Method of Finding an Optimal Threshold	487
17.17	References and Further Reading	488
18	Forecasting from Time Series Data	489
18.1	Forecasting: Prediction Using Time Series Data	490
18.2	Holdout, Training, and Test Samples in Time Series Data	491
18.3	Long-Horizon Forecasting: Seasonality and Predictable Events	493
18.4	Long-Horizon Forecasting: Trends	494
18.5	A1 CASE STUDY – Predicting Daily Ticket Volume for a Swimming Pool	496
18.6	Forecasting for a Short Horizon Using the Patterns of Serial Correlation	502
18.7	Modeling Serial Correlation: AR(1)	502
18.8	Modeling Serial Correlation: ARIMA	503
18.9	B1 CASE STUDY – Forecasting a House Price Index	505
18.10	VAR: Vector Autoregressions	507
18.11	B2 CASE STUDY – Forecasting a House Price Index	509
18.12	External Validity of Forecasts	511
18.13	B3 CASE STUDY – Forecasting a House Price Index	512
18.14	Summary and Practice	514
18.14.1	Main Takeaways	514
18.14.2	Practice Questions	514
18.14.3	Data Exercises	515
18.15	Under the Hood: Details of the ARIMA Model	516

18.16	Under the Hood: Auto-Arima	518
18.17	References and Further Reading	518
IV	CAUSAL ANALYSIS	519
19	A Framework for Causal Analysis	521
19.1	Intervention, Treatment, Subjects, Outcomes	522
19.2	Potential Outcomes	524
19.3	The Individual Treatment Effect	525
19.4	Heterogeneous Treatment Effects	526
19.5	ATE: The Average Treatment Effect	527
19.6	Average Effects in Subgroups and ATET	529
19.7	Quantitative Causal Variables	529
19.8	A1 CASE STUDY – Food and Health	530
19.9	Ceteris Paribus: Other Things Being the Same	532
19.10	Causal Maps	533
19.11	Comparing Different Observations to Uncover Average Effects	535
19.12	Random Assignment	537
19.13	Sources of Variation in the Causal Variable	538
19.14	A2 CASE STUDY – Food and Health	539
19.15	Experimenting versus Conditioning	541
19.16	Confounders in Observational Data	543
19.17	From Latent Variables to Measured Variables	545
19.18	Bad Conditioners: Variables Not to Condition On	546
19.19	A3 CASE STUDY – Food and Health	547
19.20	External Validity, Internal Validity	551
19.21	Constructive Skepticism	553
19.22	Summary and Practice	554
	19.22.1 Main Takeaway	554
	19.22.2 Practice Questions	554
	19.22.3 Data Exercises	556
19.23	References and Further Reading	556
20	Designing and Analyzing Experiments	557
20.1	Randomized Experiments and Potential Outcomes	558
20.2	Field Experiments, A/B Testing, Survey Experiments	559
20.3	A1 CASE STUDY – Working from Home and Employee Performance	560
20.4	B1 CASE STUDY – Fine Tuning Social Media Advertising	561
20.5	The Experimental Setup: Definitions	562
20.6	Random Assignment in Practice	562
20.7	Number of Subjects and Proportion Treated	564
20.8	Random Assignment and Covariate Balance	565
20.9	A2 CASE STUDY – Working from Home and Employee Performance	567
20.10	Imperfect Compliance and Intent-to-Treat	569
20.11	A3 CASE STUDY – Working from Home and Employee Performance	571

20.12	Estimation and Statistical Inference	572
20.13	B2 CASE STUDY – Fine Tuning Social Media Advertising	573
20.14	Including Covariates in a Regression	574
20.15	A4 CASE STUDY – Working from Home and Employee Performance	575
20.16	Spillovers	578
20.17	Additional Threats to Internal Validity	579
20.18	A5 CASE STUDY – Working from Home and Employee Performance	581
20.19	External Validity, and How to Use the Results in Decision Making	583
20.20	A6 CASE STUDY – Working from Home and Employee Performance	584
20.21	Summary and Practice	586
20.21.1	Main Takeaways	586
20.21.2	Practice Questions	586
20.21.3	Data Exercises	587
20.22	Under the Hood: LATE: The Local Average Treatment Effect	587
20.23	Under the Hood: The Formula for Sample Size Calculation	588
20.24	References and Further Reading	589
21	Regression and Matching with Observational Data	590
21.1	Thought Experiments	591
21.2	A1 CASE STUDY – Founder/Family Ownership and Quality of Management	592
21.3	Variables to Condition on, Variables Not to Condition On	593
21.4	A2 CASE STUDY – Founder/Family Ownership and Quality of Management	594
21.5	Conditioning on Confounders by Regression	597
21.6	Selection of Variables and Functional Form in a Regression for Causal Analysis	599
21.7	A3 CASE STUDY – Founder/Family Ownership and Quality of Management	600
21.8	Matching	603
21.9	Common Support	605
21.10	Matching on the Propensity Score	606
21.11	A4 CASE STUDY – Founder/Family Ownership and Quality of Management	607
21.12	Comparing Linear Regression and Matching	609
21.13	A5 CASE STUDY – Founder/Family Ownership and Quality of Management	610
21.14	Instrumental Variables	612
21.15	Regression-Discontinuity	614
21.16	Summary and Practice	616
21.16.1	Main Takeaways	616
21.16.2	Practice Questions	616
21.16.3	Data Exercises	617
21.17	Under the Hood: Unobserved Heterogeneity and Endogenous x in a Regression	617
21.18	Under the hood: LATE is IV	619
21.19	References and Further Reading	620
22	Difference-in-Differences	622
22.1	Conditioning on Pre-intervention Outcomes	623
22.2	Basic Difference-in-Differences Analysis: Comparing Average Changes	624

22.3	A1 CASE STUDY – How Does a Merger between Airlines Affect Prices?	627
22.4	The Parallel Trends Assumption	631
22.5	A2 CASE STUDY – How Does a Merger between Airlines Affect Prices?	633
22.6	Conditioning on Additional Confounders in Diff-in-Diffs Regressions	635
22.7	A3 CASE STUDY – How Does a Merger between Airlines Affect Prices?	637
22.8	Quantitative Causal Variable	639
22.9	A4 CASE STUDY – How Does a Merger between Airlines Affect Prices?	640
22.10	Difference-in-Differences with Pooled Cross-Sections	642
22.11	A5 CASE STUDY – How Does a Merger between Airlines Affect Prices?	645
22.12	Summary and Practice	647
22.12.1	Main Takeaways	647
22.12.2	Practice Questions	648
22.12.3	Data Exercises	649
22.13	References and Further Reading	650
23	Methods for Panel Data	651
23.1	Multiple Time Periods Can Be Helpful	652
23.2	Estimating Effects Using Observational Time Series	653
23.3	Lags to Estimate the Time Path of Effects	655
23.4	Leads to Examine Pre-trends and Reverse Effects	655
23.5	Pooled Time Series to Estimate the Effect for One Unit	656
23.6	A1 CASE STUDY – Import Demand and Industrial Production	658
23.7	Panel Regression with Fixed Effects	661
23.8	Aggregate Trend	663
23.9	B1 CASE STUDY – Immunization against Measles and Saving Children	664
23.10	Clustered Standard Errors	667
23.11	Panel Regression in First Differences	668
23.12	Lags and Leads in FD Panel Regressions	669
23.13	B2 CASE STUDY – Immunization against Measles and Saving Children	671
23.14	Aggregate Trend and Individual Trends in FD Models	673
23.15	B3 CASE STUDY – Immunization against Measles and Saving Children	674
23.16	Panel Regressions and Causality	676
23.17	First Differences or Fixed Effects?	677
23.18	Dealing with Unbalanced Panels	679
23.19	Summary and Practice	680
23.19.1	Main Takeaways	680
23.19.2	Practice Questions	680
23.19.3	Data Exercises	682
23.20	References and Further Reading	682
24	Appropriate Control Groups for Panel Data	683
24.1	When and Why to Select a Control Group in xt Panel Data	684
24.2	Comparative Case Studies	684
24.3	The Synthetic Control Method	685
24.4	A1 CASE STUDY – Estimating the Effect of the 2010 Haiti Earthquake on GDP	686

24.5	Event Studies	689
24.6	B1 CASE STUDY – Estimating the Impact of Replacing Football Team Managers	692
24.7	Selecting a Control Group in Event Studies	696
24.8	B2 CASE STUDY – Estimating the Impact of Replacing Football Team Managers	698
24.9	Summary and Practice	702
24.9.1	Main Takeaways	702
24.9.2	Practice Questions	703
24.9.3	Data Exercises	704
24.10	References and Further Reading	704
	<i>Bibliography</i>	705
	<i>Index</i>	710

WHY READ THIS BOOK

An applied data analysis textbook for future professionals

Data analysis is a process. It starts with formulating a question and collecting appropriate data, or assessing whether the available data can help answer the question. Then comes cleaning and organizing the data, tedious but essential tasks that affect the results of the analysis as much as any other step in the process. Exploratory data analysis gives context to the eventual results and helps deciding the details of the analytical method to be applied. The main analysis consists of choosing and implementing the method to answer the question, with potential robustness checks. Along the way, correct interpretation and effective presentation of the results are crucial. Carefully crafted data visualization help summarize our findings and convey key messages. The final task is to answer the original question, with potential qualifications and directions for future inquiries.

Our textbook **equips future data analysts with the most important tools, methods and skills** they need through the entire process of data analysis to answer data focused, real life questions. We cover all the fundamental methods that help along the process of data analysis. The textbook is divided into four parts covering **data wrangling and exploration, regression analysis, prediction with machine learning, and causal analysis**. We explain when, why, and how the various methods work, and how they are related to each other.

Our approach has a **different focus compared to the typical textbooks** in econometrics and data science. They are often excellent in teaching many econometric and machine learning methods. But they don't give much guidance about how to carry out an actual data analysis project from beginning to end. Instead, students have to learn all of that when they work through individual projects, guided by their teachers, advisors, and peers – but not their textbooks.

To cover all of the steps that are necessary to carry out an actual data analysis project, we **lean on a set of fully developed case studies**. While each case study focuses on the particular method discussed in the chapter, they illustrate all elements of the process from question through analysis to conclusion. We facilitate individual work by **sharing all data and code in Stata, R, and Python**.

Curated content and focus for the modern data analyst

Our textbook focuses on the most relevant tools and methods. Instead of dumping many methods on the students, we selected the most widely used methods that tend to work well in many situations. That choice allowed us to discuss each method in detail so students can gain a deep understanding of when, why, and how those methods work. It also allows us to compare the different methods both in general and in the course of our case studies.

The textbook is divided into four parts. The first part starts with data collection and data quality, followed by organizing and cleaning data, **exploratory data analysis** and data visualization, generalizing from the data, and hypothesis testing. The second part gives a thorough introduction to **regression analysis**, including probability models and time series regressions. The third part covers **predictive analytics** and introduces cross-validation, LASSO, tree-based machine learning methods such as random forest, probability prediction, classification, and forecasting from time series data. The fourth part covers **causal analysis**, starting with the potential outcomes framework and causal maps, then discussing experiments, difference-in-differences analysis, various panel data methods and the event study approach.

When deciding on which methods to discuss and in what depth, we drew on our own experience as well as the advice of many people. We have taught Data Analysis and Econometrics to students in Master's programs for years in Europe and the US, and trained experts in business analytics, economics, and economic policy. We used earlier versions of this textbook in many courses with students who differed in background, interest, and career plans. In addition, we have talked to many experts both in academia and in industry: teachers, researchers, analysts, and users of data analysis results. As a result, this textbook offers **a curated content that reflects the views of data analysts with a wide range of experiences.**

Real life case studies in a central role

A cornerstone of this textbook are 47 case studies spreading over one-third of our material. This reflects our view that working through case studies is the best way to learn data analysis. Each of our case studies starts with a relevant question and answers it in the end, using real life data and applying the tools and methods covered in the particular chapter.

Similarly to other textbooks, our case studies illustrate the methods covered in the textbook. In contrast with other textbooks, though, they are much more than that.

Each of our case studies is a fully developed story linking business or policy questions to decisions in data selection, application of methods and discussion of results. Each case study uses **real-life data** that is messy and often complicated, and it discusses data quality issues and the steps of data cleaning and organization along the way. Then, each case study includes **exploratory data analysis** to clarify the context and help choose the methods for the subsequent analysis. After carrying out the main **analysis**, each case study emphasizes the correct **interpretation** of the results, effective ways to present and visualize the results, and many include robustness checks. Finally, each case study **answers the question** it started with, usually with the necessary qualifications, discussing internal and external validity, and often raising additional questions and directions for further investigation.

Our case studies are different also because they cover a wide range of topics, with a potential appeal to a wide range of students. They cover **consumer decision, economic and social policy, finance, business and management, health, and sport**. Their regional coverage is also wider than usual: one third is from the U.S.A., one third is from Europe and the U.K., and one third is from other countries or includes all countries from Australia to Thailand.

Support material with data and code shared

We offer a truly comprehensive material with data, code for all case studies, 360 **practice questions**, 112 **data exercises**, derivations for advanced materials and reading suggestions. Each chapter ends with practice questions that help revise the material with a focus on theory. They are followed by data exercises that invite students to carry out analysis on their own, in the form of robustness checks or replicating the analysis using other data.

We share all raw and cleaned data we use in the case studies. We also share the codes that clean the data and produce all results, tables, and graphs in **Stata, R, and Python** so students can tinker with our code and compare the solutions in the different software.

All data and code are available on the textbook website:

<http://gabors-data-analysis.com>

Who is this book for?

This textbook was written to be a **complete course** in data analysis. It introduces and discusses the most important concepts and methods in exploratory data analysis, regression analysis, machine learning and causal analysis. Thus, readers don't need to have a background in those areas.

The textbook includes formulae to define methods and tools, but it **explains all formulae in plain English**, both when a formula is introduced and, then, when it is used in a case study. Thus, understanding formulae is not necessary to learn data analysis from this textbook. They are of great help, though, and we encourage all students and practitioners to work with formulae whenever possible. The mathematics background required to understand these formulae is quite low, at the level of basic calculus.

This textbook could be useful for university students in graduate programs as **core text** in applied statistics and econometrics, quantitative methods, or data analysis. The textbook is best used as core text for non-research degree Masters programs or part of the curriculum in a Phd or research Masters programs. It may also **complement online courses** that teach specific methods to give more context and explanation. Undergraduate courses can also make use of this textbook, even though the workload on students exceeds the typical undergraduate workload. Finally, the textbook can serve as a **handbook for practitioners** to guide them through all steps of real-life data analysis.

SIMPLIFIED NOTATION

Finally, a note for the instructors who plan to use our textbook.

We introduced some new notation in this textbook, to make the formulae simpler and more focused. In particular, our **formula for regressions is slightly different** from the traditional formula. In line with other textbooks, we think that it is good practice to write out the formula for each regression that is analyzed. For this reason, it is important to use a notation for the regression formula that is as simple as possible and focuses only on what we care about. Our notation is intuitive, but it's slightly different from traditional practice. Let us explain our reasons.

Our approach starts with the definition of the regression: it is a model for the conditional mean. The formulaic definition of the simple linear regression is $E[y|x] = \alpha + \beta x$. The formulaic definition of a linear regression with three right-hand-side variables is $E[y|x_1, x_2, x_3] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

The regression formula we use in the textbook is a simplified version of this formulaic definition. In particular, we have y^E on the left-hand-side instead of $E[y|\dots]$. y^E is just a short-hand for the expected value of y conditional on whatever is on the right-hand-side of the regression.

Thus, the formula for the simple linear regression is $y^E = \alpha + \beta x$, and y^E is the expected value of y conditional on x . The formula for the linear regression with three right-hand-side variables is $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, and here y^E is the expected value of y conditional on x_1 , x_2 , and x_3 . Having y^E on the left-hand-side makes notation much simpler than writing out the conditional expectation formula $E[y|\dots]$, especially when we have many right-hand-side variables.

In contrast, the traditional regression formula has variable y itself on the left-hand-side not its conditional mean. Thus, it has to involve an additional element, the error term, which is the difference between y and its conditional mean. For example, the traditional formula for the linear regression with three right-hand-side variables is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$.

Our notation is simpler, because it has fewer elements. More importantly, our notation makes it explicit that the regression is a model for the conditional mean. It focuses on what data analysts care about (the right-hand-side variables and their coefficients), without adding anything else.

ACKNOWLEDGMENTS

Let us first thank our students at Central European University, the University of Michigan, and at the University of Reading. The idea of writing a textbook was born out of teaching and mentoring them. We have learned a lot from teaching them, and many of them helped us writing code, collecting data, reading papers, and hunting for ideas.

Many colleagues helped us with their extremely valuable comments and suggestions. We thank *Eduardo Arino de la Rubia*, *Emily Blanchard*, *Alberto Cairo*, *Gergely Daróczy*, *János Divényi*, *Christian Fons-Rosen*, *Miklós Koren*, *Mike Luca*, *Róbert Lieli*, *László Mátyás*, *Tímea Laura Molnar*, *Jenő Pál*, and *Ádám Szeidl* and anonymous reviewers of the first draft of the textbook.

We have received help with our case studies from *Alberto Cavallo*, *Daniella Scur*, *Nick Bloom*, *John van Reenen*, *Anikó Kristof*, *József Keleti*, and *Emily Oster* and *MyChelle Andrews*. We have learned a lot from them.

Several people helped us a great deal with our manuscript. At Cambridge University Press, our commissioning editor, *Phil Good* encouraged us from the day we met. Our editors *Heather Brolly*, *Jane Adams*, *Nicola Chapman*, guided us with kindness and steadfastness from first draft to proofs. We are not native English speakers, and support from *Chris Cartwright*, and *Jon Billam* were very useful. We are grateful for *Sarolta Rózsás* who acted as development editor, reviewing endless versions of chapters, checking consistency, clarity and pushed us to make the text more coherent and accessible.

Creating the code base in Stata, R and Python was a massive endeavour. Both of us are native Stata speakers, and we needed R code that would be fairly consistent with Stata code. Plus, all graphs were produced in R. So we needed help to have all our Stata codes replicated in R, and a great deal of code writing from scratch. *Zsuzsa Holler* and *Kinga Ritter* have provided enormous development support, spearheading this effort for years. Additional code and refactoring in Python and R was created by *Maté Tóth*, *János Bíró*, and *Eszter Pázmándi*. Additional coding, data collection, visualization and editing were done by *Viktória Kónya*, *Zsófia Komuves*, *Dániel Bánki*, *Abuzar Ali*, *Endre Borza* and *Ahmed Al Shaibani*.

Several former students helped us by reading the book and giving feedback as well as editing and checking latex files. We are grateful to *Imre Boda*, *Bonnie Kavoussi*, *Olivér Kiss*.

The wonderful cover design is based on the work by *Ágoston Nagy*, his first, surely not the last. The website was designed by *Imola Csoka*.

Collaborating with many talented young people and bringing them together was one the joys of writing this book.

Let us also shout out to the fantastic R user community – both online and offline – from whom we learnt tremendously. Special thanks to the Rstats and Econ Twitter community – we received wonderful suggestions from tons of people we have never met.

We thank the Central European University for professional and financial support. *Julius Horvath* and *Miklos Koren* as department heads provided massive support form the day we shared our plans.

Finally, let us thank those who were with us throughout the long, and often stressful, process of writing a textbook. Békés thanks *Saci*; Kezdi thanks *Zsuzsanna*, *Anna*, *Dorka*, and *Bori*. We could not have been able to done it without their love and support.

The Gabors

2020 February, Budapest, Hungary and Ann Arbor, Mi, USA

PROOF