Introductory Econometrics
Description, Prediction, and Causality

Second edition

David M. Kaplan
To my past, present, and future students, including NLK and OAK.
—DMK
The chief difficulty Alice found at first was in managing her flamingo: she succeeded in getting its body tucked away, comfortably enough, under her arm, with its legs hanging down, but generally, just as she had got its neck nicely straightened out, and was going to give the hedgehog a blow with its head, it would twist itself round and look up in her face, with such a puzzled expression that she could not help bursting out laughing: and when she had got its head down, and was going to begin again, it was very provoking to find that the hedgehog had unrolled itself, and was in the act of crawling away...Alice soon came to the conclusion that it was a very difficult game indeed.

Lewis Carroll, Alice’s Adventures in Wonderland
(An allegory for econometrics)
Brief Contents

Contents vii
List of Discussion Questions xiv
Preface xvii
Textbook Learning Objectives xix
Notation xxi
1 Getting Started with R (or Stata) 1

I Analysis of One Variable 11
Introduction 13

2 One Variable: Population 15

3 One Variable: Sample 53

4 One Variable, Two Populations 93

5 Midterm Exam #1 117

II Regression 119
Introduction 121

6 Comparing Two Distributions by Regression 123

7 Simple Linear Regression 155

8 Nonlinear and Nonparametric Regression 173

9 Regression with Two Binary Regressors 195
10 Regression with Multiple Regressors 221
11 Midterm Exam #2 237
12 Internal and External Validity 239

III Time Series 257

Introduction 259
13 Time Series: One Variable 261
14 First-Order Autoregression 281
15 AR($p$) and ADL Models 297
16 Final Exam 307

Bibliography 309
Index 315
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>vi List of Discussion Questions</td>
<td>xiv</td>
</tr>
<tr>
<td>xvii Preface</td>
<td></td>
</tr>
<tr>
<td>xix Textbook Learning Objectives</td>
<td></td>
</tr>
<tr>
<td>xxiv Notation</td>
<td></td>
</tr>
<tr>
<td>1 Getting Started with R (or Stata)</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Comparison of R and Stata</td>
<td>1</td>
</tr>
<tr>
<td>1.2 R</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Accessing the Software</td>
<td>2</td>
</tr>
<tr>
<td>1.2.2 Installing Packages</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Stata</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1 Accessing the Software</td>
<td>5</td>
</tr>
<tr>
<td>1.3.2 Installing Additional Commands</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Optional Resources</td>
<td>6</td>
</tr>
<tr>
<td>1.4.1 R Tutorials</td>
<td>6</td>
</tr>
<tr>
<td>1.4.2 R Quick References</td>
<td>7</td>
</tr>
<tr>
<td>1.4.3 Running Code in This Textbook</td>
<td>7</td>
</tr>
<tr>
<td>1.4.4 Stata Resources</td>
<td>7</td>
</tr>
<tr>
<td>Empirical Exercises</td>
<td>8</td>
</tr>
<tr>
<td>I Analysis of One Variable</td>
<td>11</td>
</tr>
<tr>
<td>13 Introduction</td>
<td></td>
</tr>
<tr>
<td>2 One Variable: Population</td>
<td>15</td>
</tr>
<tr>
<td>2.1 The World is Random</td>
<td>16</td>
</tr>
<tr>
<td>2.1.1 Before and After: Two Perspectives</td>
<td>16</td>
</tr>
<tr>
<td>2.1.2 Before and After Sampling</td>
<td>17</td>
</tr>
<tr>
<td>2.1.3 Outcomes and Mechanisms</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Population Types</td>
<td>18</td>
</tr>
<tr>
<td>2.2.1 Finite Population</td>
<td>18</td>
</tr>
<tr>
<td>2.2.2 Infinite Population</td>
<td>19</td>
</tr>
</tbody>
</table>
10.2.3 Code .................................................. 224
10.3 Interaction Terms ....................................... 225
  10.3.1 Limitation of Linear-in-Variables Model ...... 225
  10.3.2 Interpretation of Interaction Term .......... 225
  10.3.3 Non-Binary Interaction ....................... 227
  10.3.4 Code ............................................. 228
10.4 Other Examples ......................................... 228
10.5 Assumptions for Linear Projection ................. 229
  10.5.1 Multicollinearity (Two Types) ............. 229
  10.5.2 Formal Assumptions and Results ............ 230
10.6 Structural Identification ........................... 231
  10.6.1 Linear Structural Model .................... 231
  10.6.2 General Structural Model ................... 231
  10.6.3 Conditional ATE .............................. 232
Empirical Exercises ...................................... 233

11 Midterm Exam #2 .......................... 237

12 Internal and External Validity .................... 239
  12.1 Terminology ........................................ 240
  12.2 Threats to External Validity ................. 240
    12.2.1 Different Place ........................... 241
    12.2.2 Different Time ................................ 241
    12.2.3 Different Population ...................... 242
  12.3 Threats to Internal Validity ................. 243
    12.3.1 Functional Form Misspecification ....... 243
    12.3.2 Measurement Error in the Outcome Variable . 243
    12.3.3 Measurement Error in the Regressors .... 246
    12.3.4 Non-iid Sampling and Survey Weights .... 248
    12.3.5 Missing Data .................................. 250
    12.3.6 Sample Selection ........................... 251
    12.3.7 Omitted Variable Bias and Collider Bias .. 252
    12.3.8 Simultaneity and Reverse Causality ....... 252
Empirical Exercises ...................................... 254

III Time Series .............................. 257

Introduction .......................................... 259

13 Time Series: One Variable ................. 261
  13.1 Terms and Notation ............................. 262
  13.2 Populations, Randomness, and Sampling ....... 263
  13.3 Stationarity ....................................... 264
  13.4 Autocovariance and Autocorrelation ......... 264
  13.5 Estimation ........................................ 266
    13.5.1 Mean ......................................... 266
    13.5.2 Autocovariances and Autocorrelations .... 266
13.6 Nonstationarity ........................................... 268
  13.6.1 Trends .............................................. 268
  13.6.2 Seasonality ....................................... 270
  13.6.3 Cycles ............................................. 271
  13.6.4 Structural Breaks .................................. 272
13.7 Decomposition .......................................... 272
13.8 Transformations ........................................ 274
  Empirical Exercises ...................................... 276

14 First-Order Autoregression 281
  14.1 Model ................................................... 282
  14.2 Description .......................................... 283
  14.3 Prediction (Forecasting): Optimality .................. 284
  14.4 Estimation ............................................ 285
    14.4.1 Code ............................................. 286
  14.5 Parameter Stability ................................... 287
  14.6 Multi-Step Forecast .................................. 287
    14.6.1 Intuition: Mean Zero (Special Case) ............... 288
    14.6.2 General Results with AR Parameters .............. 288
    14.6.3 Direct Approach .................................. 289
    14.6.4 Code ............................................. 289
  14.7 Interval Forecasts ................................... 289
    14.7.1 Goal and Sources of Uncertainty ................... 289
    14.7.2 Intervals Assuming Normality ..................... 290
    14.7.3 Code ............................................. 291
  14.8 More R Examples ..................................... 291
    14.8.1 AR(1) Multi-Step Forecast Intervals .............. 291
    14.8.2 General R Forecast Allowing Seasonality and Trend . 293
  Empirical Exercises ...................................... 295

15 AR(p) and ADL Models 297
  15.1 The AR(p) Model ..................................... 298
  15.2 Model Selection: How Many Lags? ....................... 298
    15.2.1 Difficulties and Intuition ........................ 298
    15.2.2 AIC and BIC Formulas ............................ 299
    15.2.3 Comparison of AIC and BIC ....................... 300
    15.2.4 Code ............................................. 301
  15.3 Autoregressive Distributed Lag Regression .......... 302
  Empirical Exercises ...................................... 304

16 Final Exam 307

Bibliography 309

Index 315
# List of Discussion Questions

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>web traffic</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>student data</td>
<td>21</td>
</tr>
<tr>
<td>2.3</td>
<td>banana loss function</td>
<td>49</td>
</tr>
<tr>
<td>2.4</td>
<td>optimal banana prediction</td>
<td>50</td>
</tr>
<tr>
<td>3.1</td>
<td>probability of positive mean</td>
<td>67</td>
</tr>
<tr>
<td>3.2</td>
<td>equal p-values, equal belief?</td>
<td>85</td>
</tr>
<tr>
<td>3.3</td>
<td>jellybean solution?</td>
<td>87</td>
</tr>
<tr>
<td>3.4</td>
<td>nova</td>
<td>87</td>
</tr>
<tr>
<td>3.5</td>
<td>Ebola drug</td>
<td>90</td>
</tr>
<tr>
<td>4.1</td>
<td>DPC with two populations</td>
<td>94</td>
</tr>
<tr>
<td>4.2</td>
<td>description, prediction, causality</td>
<td>97</td>
</tr>
<tr>
<td>4.3</td>
<td>cash transfer spillovers</td>
<td>105</td>
</tr>
<tr>
<td>4.4</td>
<td>breakfast effect?</td>
<td>111</td>
</tr>
<tr>
<td>6.1</td>
<td>logic with feathers</td>
<td>127</td>
</tr>
<tr>
<td>6.2</td>
<td>joint distribution and causality</td>
<td>130</td>
</tr>
<tr>
<td>6.3</td>
<td>ES habits and final scores</td>
<td>143</td>
</tr>
<tr>
<td>6.4</td>
<td>marriage and salary</td>
<td>146</td>
</tr>
<tr>
<td>7.1</td>
<td>Facebook</td>
<td>158</td>
</tr>
<tr>
<td>7.2</td>
<td>BLP</td>
<td>163</td>
</tr>
<tr>
<td>7.3</td>
<td>student-teacher ratio simple regression</td>
<td>170</td>
</tr>
<tr>
<td>8.1</td>
<td>pollution and house price</td>
<td>178</td>
</tr>
<tr>
<td>8.2</td>
<td>nonlinear OVB</td>
<td>180</td>
</tr>
<tr>
<td>8.3</td>
<td>nonlinear wage model interpretation</td>
<td>184</td>
</tr>
<tr>
<td>8.4</td>
<td>model evaluation</td>
<td>190</td>
</tr>
<tr>
<td>9.1</td>
<td>assessing OVB</td>
<td>199</td>
</tr>
<tr>
<td>9.2</td>
<td>OVB: ES habits</td>
<td>202</td>
</tr>
<tr>
<td>9.3</td>
<td>bad panel approach #1, for Mariel boatlift</td>
<td>211</td>
</tr>
<tr>
<td>9.4</td>
<td>bad panel approach #2, for Mariel boatlift</td>
<td>211</td>
</tr>
<tr>
<td>9.5</td>
<td>bad panel approach #1, for fracking</td>
<td>211</td>
</tr>
<tr>
<td>9.6</td>
<td>bad panel approach #2, for fracking</td>
<td>211</td>
</tr>
<tr>
<td>9.7</td>
<td>parallel trends skepticism</td>
<td>213</td>
</tr>
<tr>
<td>10.1</td>
<td>OVB with multiple regressors</td>
<td>223</td>
</tr>
<tr>
<td>10.2</td>
<td>sleep and interactions</td>
<td>226</td>
</tr>
<tr>
<td>10.3</td>
<td>wage and interactions</td>
<td>226</td>
</tr>
<tr>
<td>Section</td>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>10.4</td>
<td>linear-in-variables?</td>
<td>227</td>
</tr>
<tr>
<td>12.1</td>
<td>external validity: minimum wage</td>
<td>242</td>
</tr>
<tr>
<td>12.2</td>
<td>exercise error</td>
<td>244</td>
</tr>
<tr>
<td>12.3</td>
<td>measurement error: scrap rate</td>
<td>246</td>
</tr>
<tr>
<td>12.4</td>
<td>program attrition</td>
<td>250</td>
</tr>
<tr>
<td>12.5</td>
<td>missing salary data</td>
<td>251</td>
</tr>
<tr>
<td>12.6</td>
<td>health and medical expenditure</td>
<td>253</td>
</tr>
<tr>
<td>13.1</td>
<td>autocorrelation</td>
<td>266</td>
</tr>
<tr>
<td>13.2</td>
<td>nonstationarity</td>
<td>274</td>
</tr>
<tr>
<td>14.1</td>
<td>forecast and reality</td>
<td>285</td>
</tr>
<tr>
<td>14.2</td>
<td>recession-affected coefficient</td>
<td>287</td>
</tr>
<tr>
<td>14.3</td>
<td>long-horizon AR forecast</td>
<td>289</td>
</tr>
<tr>
<td>14.4</td>
<td>forecast sanity check</td>
<td>291</td>
</tr>
<tr>
<td>15.1</td>
<td>lag choice for forecasting</td>
<td>301</td>
</tr>
</tbody>
</table>
LIST OF DISCUSSION QUESTIONS
Preface

This text was prepared for the 15-week semester Introductory Econometrics course at the University of Missouri. The class focuses on statistical description, prediction, and “causality,” including both structural parameters and treatment effects. Description and prediction (forecasting) with time series are also covered. Students learn to think probabilistically, understand prediction and causality, judge whether various assumptions hold true in real-world examples, and apply econometric methods in R.

As usual, this text may be used to teach different types of classes. In full, the text provides a 15-week semester class that assumes a previous class in probability and statistics. That prerequisite could be skipped if more time is spent on the “review” material in the first few chapters. Calculus is avoided but could be added in the usual places. A shorter class could omit the time series material. Of course, any material may be expanded, condensed, or skipped, as the instructor desires.

Some complementary, complimentary texts and courses deserve mention. Econometrics professor Matt Masten has a “Causal Inference Bootcamp” video series, as well as some “Causal Inference with R” free courses on DataCamp. Relevant videos are linked at the beginning of each chapter in this textbook. Stanford statistics professors Trevor Hastie and Rob Tibshirani created a free introductory machine learning (statistical learning) course, focusing more on prediction and estimation. Their course uses their free textbook (James, Witten, Hastie, and Tibshirani, 2013) that includes R examples. Hastie, Tibshirani, and Friedman (2009) also provide their more advanced statistical learning text for free. For econometrics texts focused on prediction and time series, see Diebold (2018a,b,c). The forecasting book by Hyndman and Athanasopoulos (2019) is at https://.otexts.com/fpp2 and uses R. Finally, Hanck, Arnold, Gerber, and Schmelzer (2018) mirror the structure of the (expensive) textbook of Stock and Watson (2015), providing many R examples to illustrate the concepts they explain.

---

1 https://mattmasten.github.io/bootcamp
2 https://www.datacamp.com/community/open-courses
3 https://www.edx.org/course/statistical-learning
4 http://faculty.marshall.usc.edu/gareth-james/ISL
5 https://web.stanford.edu/~hastie/ElemStatLearn
6 http://www.ssc.upenn.edu/~fdiebold/Textbooks.html
7 https://www.econometrics-with-r.org
One distinguishing feature of this text is the development of the ideas of (and distinctions among) statistical description, prediction, causal inference, and structural estimation in the simplest possible settings. Other texts combine these with all the complications of regression from the beginning, often confusing students (like my past self).

A second distinguishing feature is that this text’s source files are freely available. Instructors may modify them as desired, or copy and paste \LaTeX{} code into their own lecture notes, subject to the Creative Commons license linked on the copyright page. I wrote the text in Overleaf, an online (free) \LaTeX{} environment that includes knitr support, so most of the R code and output is in the same .Rtex files alongside the \LaTeX{} code. Graphs are either generated from code in the .Rtex files or else from a single .R file also provided in the source material. You may see, copy, and download the entire project from Overleaf\(^8\) or from my website.\(^9\)

Third, I provide learning objectives for the overall book and for each chapter. This follows current best practices for course design. Upon request, I can provide a library of multiple choice questions, labeled by learning objective. (Empirical exercises are already at the end of each textbook chapter.)

Fourth, in-class (or online) discussion questions are included along the way. When I teach in person (30–40 students), I prefer to punctuate lectures with such questions every 20–30 minutes, where students first discuss them for a couple minutes in small groups of 2–3 students, and then volunteer to share their group’s ideas with the whole class for another couple minutes. This provides an active learning opportunity, a time for students to realize they don’t understand the lecture material (so they can ask questions), practice discussing econometrics with peers, and (if nothing else) a few minutes’ rest.

Thanks to everyone for their help and support: my past econometrics instructors, my colleagues and collaborators, my students (who have not only inspired me but alerted me to typos and other deficiencies in earlier drafts), and my family.

David M. Kaplan
Summer 2018 (edited Summer 2020)
Columbia, Missouri, USA

\(^8\)https://www.overleaf.com/read/fszrgmzwzfrk
\(^9\)http://faculty.missouri.edu/kaplandm/teach.html
Textbook Learning Objectives

For good reason, it has become standard practice to list learning objectives for a course as well as each unit within the course. Below are the learning objectives corresponding to this text overall. Each chapter lists more specific learning objectives that map to one or more of these overall objectives. The accompanying exercises are also classified by learning objectives. I hope you find these helpful guidance, whether you are a solo learner, a class instructor, or a class student.

The textbook learning objectives (TLOs) are the following.

1. Define terms from probability, statistics, and econometrics, both mathematically and intuitively.

2. Describe various econometric methods both mathematically and intuitively, including their objects of interest and assumptions, and the logical relationship between the assumptions and corresponding theorems and properties.

3. Interpret the values that could be estimated with infinite data, in terms of description, prediction, and causality (or economic meaning).

4. Explain the frequentist/classical statistical and asymptotic frameworks, including their benefits and limitations.

5. Provide multiple possible (causal) explanations for any statistical result, distinguishing between statistical and causal relationships.

6. For a given economic question, dataset, and econometric method, judge whether the method is appropriate and judge the economic significance and statistical significance of the results.

7. Using R (or Stata): manipulate and analyze data, interpreting results both economically and statistically.
Notation

Much of the notation below will not make sense until you get to the corresponding point in the text. The following is primarily for your reference later.

Variables

Usually, uppercase denotes a random variable, whereas lowercase denotes a non-random (fixed, constant) value. The primary exception is for certain counting variables, where uppercase indicates the maximum value and lowercase indicates a general value; e.g., time period $t$ can be $1, 2, 3, \ldots, T$, or regressor $k$ out of $K$ total regressors. Scalar, vector, and matrix variables are typset differently. For example, an $n$-by-$k$ random matrix with scalar (random variable) entries $X_{ij}$ (row $i$, column $j$) is written

\[
X = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1k} \\
X_{21} & X_{22} & \cdots & X_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nk}
\end{pmatrix}
\]

and a $k$-dimensional non-random vector is

\[
z = \begin{pmatrix}
z_1 \\
z_2 \\
\vdots \\
z_k
\end{pmatrix}
\]

Unless otherwise specified, vectors are column vectors (like above).

Both vectors and matrices can be transposed. The transpose of a column vector is a row vector. For example, the transpose of the $z$ defined above is

\[
z' = (z_1, z_2, \ldots, z_k)
\]

and the transpose of the $X$ defined above is

\[
X' = \begin{pmatrix}
X_{11} & X_{21} & \cdots & X_{n1} \\
X_{12} & X_{22} & \cdots & X_{n2} \\
\vdots & \vdots & \ddots & \vdots \\
X_{1k} & X_{2k} & \cdots & X_{nk}
\end{pmatrix}
\]
where the row $i$, column $j$ entry in $\mathbf{X}'$ is the row $j$, column $i$ entry in $\mathbf{X}$.

Greek letters like $\beta$ and $\theta$ generally denote non-random (fixed) population parameters.

Estimators usually have a “hat” on them. Since estimators are computed from data, they are random from the frequentist perspective. Thus, even if $\theta$ is a non-random population parameter, $\hat{\theta}$ is a random variable.

I try to put “hats” or bars on other quantities computed from the sample, too. For example, a $t$-statistic would be $\hat{t}$ (a random variable computed from the sample) instead of just $t$ (which looks like a non-random scalar). The sample average of $Y_1, \ldots, Y_n$ is $\bar{Y}$.

Estimators and other statistics (i.e., things computed from data) may sometimes have a subscript with the sample size $n$ to remind us that their sampling distribution depends on $n$. For example, $\hat{\theta}_n$, $\hat{t}_n$, and $\bar{Y}_n$.

The following is a summary.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>scalar fixed (non-random) value</td>
</tr>
<tr>
<td>$Y$</td>
<td>scalar random variable</td>
</tr>
<tr>
<td>$\theta$</td>
<td>scalar non-random value</td>
</tr>
<tr>
<td>$\hat{\theta}$</td>
<td>scalar random variable</td>
</tr>
<tr>
<td>$\mathbf{x}$</td>
<td>non-random column vector</td>
</tr>
<tr>
<td>$\mathbf{x}'$</td>
<td>transpose of $\mathbf{w}$</td>
</tr>
<tr>
<td>$\mathbf{X}$</td>
<td>random column vector</td>
</tr>
<tr>
<td>$\beta$</td>
<td>non-random column vector</td>
</tr>
<tr>
<td>$\hat{\beta}$</td>
<td>random column vector</td>
</tr>
<tr>
<td>$\mathbf{w}$</td>
<td>non-random matrix</td>
</tr>
<tr>
<td>$\mathbf{w}'$</td>
<td>transpose of $\mathbf{w}$</td>
</tr>
<tr>
<td>$\mathbf{W}$</td>
<td>random matrix</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>non-random matrix</td>
</tr>
<tr>
<td>$\hat{\Omega}$</td>
<td>random matrix</td>
</tr>
</tbody>
</table>

**Symbols**

In addition to the following symbols, vocabulary words and abbreviations (like “regression” or “OLS”) can be looked up in the Index in the very back of the text.

- $\implies$ implies; see Section 6.1
- $\Leftarrow$ is implied by; see Section 6.1
- $\iff$ if and only if; see Section 6.1
- $\lim_{n \to \infty}$ limit (like in pre-calculus)
- $\text{plim}_{n \to \infty}$ probability limit; see Section 3.7.3
- $\to$ converges to (like in pre-calculus)
- $p \to$ converges in probability to; see Section 3.7.3
- $\equiv$ is defined as
- $\approx$ approximately equals
- $\sim$ is distributed as
\( \sim \)
is distributed approximately (or asymptotically) as; see Section 6
\( X \perp Y \)
\( X \) and \( Y \) are statistically independent; see Section 6.2.6
\( N(\mu, \sigma^2) \)
\( \)normal distribution with mean \( \mu \) and variance \( \sigma^2 \)
\( N(0, 1) \)
\( \)standard normal distribution
\( F_Y(\cdot) \)
cumulative distribution function (CDF) of \( Y \); see Section 2.3
\( f_Y(\cdot) \)
PMF of \( Y \) (if \( Y \) is discrete); see Section 2.3
\( f_Y(\cdot) \)
PDF of \( Y \) (if \( Y \) is continuous); see Figure 2.3
\( \mathbb{1}\{\cdot\} \)
indicator function; see (2.3)
\( P(\cdot) \)
probability of event \( A \)
\( P(\cdot | \cdot) \)
conditional probability of \( A \) given \( B \); see Section 6.2.3
\( E(\cdot) \)
expectation (mean) of \( Y \); see Section 2.3
\( E(\cdot | \cdot) \)
CEF (a function of \( \cdot \)); see Section 6.3
\( \sum_{i=1}^{\cdot} \)
summation from \( i = 1 \) to \( i = n \)
\( \text{Var}(\cdot) \)
variance of \( Y \) (square of standard deviation); see Section 2.3
\( \text{Var}(\cdot | \cdot) \)
conditional variance (a non-random value); see Section 6.8.1
\( \text{Var}(\cdot | \cdot) \)
conditional variance (a random variable)
\( \text{Cov}(\cdot, \cdot) \)
covariance
\( \text{Corr}(\cdot, \cdot) \)
correlation
\( \{a, b, \ldots\} \)
a set (containing elements \( a, b, \) etc.)
\( i = 1, \ldots, n \)
same as \( i \in \{1, \ldots, n\} \) (integers from 1 to \( n \))
\( j = 1, \ldots, J \)
same as \( j \in \{1, \ldots, J\} \) (integers from 1 to \( J \))
\( s \in S \)
element \( s \) is in set \( S \)
\( \hat{E}(\cdot) \)
expectation for sample distribution; see Section 3.4.1
\( \bar{Y}_n \)
\( \frac{1}{n} \sum_{i=1}^{n} Y_i \); same as \( \hat{E}(\cdot) \); see Section 3.4.1
\( \hat{\theta} \)
estimator of population parameter \( \theta \); see Section 3.4
\( \hat{\text{SE}}(\hat{\theta}) \)
standard error of estimator \( \hat{\theta} \); see Section 3.8.1
\( \arg\min_{g} f(g) \)
the value of \( g \) that minimizes \( f(g) \)
\( \arg\max_{g} f(g) \)
the value of \( g \) that maximizes \( f(g) \)
\( \mathbf{v}^\prime, \mathbf{x}^\prime \)
transposes of matrix \( \mathbf{v} \) and vector \( \mathbf{x} \), respectively
\( \mathbf{v}^{-1} \)
inverse of matrix \( \mathbf{v} \)
Chapter 1

Getting Started with R (or Stata)

⇒ Kaplan video: Course Introduction

Depends on: no other chapters

Unit learning objectives for this chapter

1.1. Run statistical software (R/RStudio or Stata) [TLO 7]
1.2. Write code to do basic data manipulation, description, and display [TLO 7]

You will use R (or Stata) for the empirical exercises in this textbook. The code examples in the textbook are all in R.

No previous experience with any statistical software is assumed. Consequently, the primary goal of the empirical exercises is to develop your confidence and experience with statistical software, applying the text’s methods and ideas to real datasets. Toward this goal, there are lots of explicit hints about the code you need to write.

If you actually do have previous experience (or above-average interest), then the empirical exercises may feel too boring. You could try figuring out alternative ways to code the solution, or coding alternative analyses, etc. You can also explore other online resources like one of the free DataCamp courses.¹

Due to the many excellent resources online (see Section 1.4), there are many people who can write R code, but most do not understand how to properly interpret econometric results or judge which method is most appropriate. So, overall, this class/textbook focuses more on understanding econometrics than coding.

1.1 Comparison of R and Stata

I like both R and Stata statistical software, and I have used both professionally. They excel in different ways mentioned below.

¹https://www.datacamp.com/community/open-courses
For this textbook/class, I focus on R for the following reasons.

1. It’s widely used in the private sector, government, and academia alike, in many fields (including economics).

2. It’s free to download/use, and can even be used through a web browser.

3. It has many econometric/statistical functions available, and creators of new econometric/statistical methods often provide code in R.

4. There are many online resources for learning R and getting help.

In comparison, Stata:

1. is widely used in economics and certain social sciences, but less so in fields like data science and statistics.

2. is not free, and can’t be used in a browser; but is free to use in many campus computer labs.

3. is easier to use for standard econometric methods, and has some new econometric methods (while others take a few years to be implemented).

4. also has good help files (documentation) and online support.

1.2 R

1.2.1 Accessing the Software

There are three ways you could run R: downloaded onto your own computer, through a web browser (in the cloud), or on another computer like in a campus computer lab.

Other computers or web browser versions may have the core R software but lack certain packages needed for the empirical exercises. In some cases, you can simply install the necessary packages with a single command (e.g., in Mizzou computer labs). In other cases, you may be prohibited from installing packages, in which case you won’t be able to complete the exercises, so make sure to check this first.

Through a Web Browser

There are many free options for using R through a web browser, and they evolve quickly. This means both new and improved options becoming available, as well as existing options disappearing, even from major companies (e.g., Microsoft Azure Notebooks was “retired”).

Currently, I suggest you use RStudio Cloud. It’s free, reliable, and the same RStudio interface as if you downloaded RStudio, so you can learn from the latter half of my RStudio video. To get started:

1. Go to https://rstudio.cloud/ in any web browser
2. Click the GET STARTED FOR FREE button (or else “Log In” if you already have an account)
3. Click the “Sign Up” button (the free “Cloud Free” account is selected by default)
4. Enter your email, (new) password, and name, and click “Sign Up” (or else “Sign up with Google” if you prefer)
5. Start using RStudio like it were on your own computer
6. Install necessary packages like usual: see Section 1.2.2
7. After you log out and later log in, click “Untitled Project” (feel free to rename) to get back to where you were

At http://mybinder.org/v2/gh/binder-examples/r/master?urlpath=rstudio you can also use the RStudio interface through a web browser, without even making an account, but 1) it does not run the most current version of R, 2) it cannot save your files from one session to the next, 3) you have to install the packages every time (which takes many minutes to run). But these are not critical problems for this class: older R versions are fine, you can save your code/output in a text file on your own computer, and you can make some tea while the packages install.

Currently, the best other options use Jupyter Notebooks. In order of my preference (for this class):

- CoCalc: no account required; all needed packages already installed; go to https://cocalc.com and click “Run CoCalc Now” and wait for it to load, then click “R (system-wide)” under “Suggested kernels” and you can start typing R code
- Google CoLab: requires Google account; go to https://colab.research.google.com/drive/1BYmbqeyZAlYnxR9IHCTpW07EpD and then in the Edit file menu click “Select all cells” and (also in the Edit menu) “Delete selected cells” to get a blank notebook; then under Insert click “Code cell” and start typing code; can install all needed packages as in Section 1.2.2 (takes a few minutes to run)
- Gradient by Paperspace: I haven’t tried it but looked promising; requires free account at https://gradient.paperspace.com/

In a Mizzou Computer Lab

You can check which Mizzou computing sites/labs have your favorite software on the Computing Sites Software web page.2 Scroll down to RStudio to see where you can use R with RStudio. However, sometimes there are classes or other events in computer labs; you can check the weekly schedule posted near the door to find a free time, or you can check online.3

After you log into the computer in the computer lab, open RStudio from the Start menu. (RStudio calls R itself in the background; you don’t have to open R directly.) Then just start typing commands, and hit Enter to run them.

---

2https://doit.missouri.edu/services/computing-sites/sites-software
3https://doit.missouri.edu/services/computing-sites and click the lab name
The computer labs don’t currently have the necessary packages pre-installed, but you can easily install them. Note that you’ll have to do this every time you log in (since any files you download/save get deleted when you log out), but you can just run the same line of code when you start RStudio each time.

Also, make sure to email yourself your code (or otherwise save it, if you haven’t finished and uploaded to Canvas) before you log out, since your files get deleted when you log out.

### Downloading Software

You’ll download two pieces of software: R itself, and RStudio. Both are free. R has all the functions you need. RStudio makes the interface nicer and makes things easier for you.

#### On Windows:
- Download the .exe installer file for R: Google “R Windows” or try [https://cran.r-project.org/bin/windows/base](https://cran.r-project.org/bin/windows/base) and click the “Download...” link near the top.
- Open the downloaded .exe installer and follow the instructions.
- Download the .exe installer file for RStudio Desktop (free version): Google “RStudio download” or try [https://www.rstudio.com/products/rstudio/download/#download](https://www.rstudio.com/products/rstudio/download/#download)
- Open the downloaded .exe installer and follow the instructions.

#### On Mac:
- Download the .pkg file for R: Google “R Mac” or try [https://cran.r-project.org/bin/macosx](https://cran.r-project.org/bin/macosx)
- Open the file and follow the usual Mac installation procedure.
- Download the .dmg file for RStudio Desktop (free version): Google “RStudio download” or try [https://www.rstudio.com/products/rstudio/download/#download](https://www.rstudio.com/products/rstudio/download/#download)
- Open the file and follow the usual Mac installation procedure.

#### On Linux, etc.: if you can figure out how to run something besides Windows or Mac, you can probably figure out how to download a couple files by yourself, but please let me know if not.

Regardless of OS, after both are installed, you only ever need to open RStudio, never R. Once you open RStudio, just type a command and hit Enter to run it.

### 1.2.2 Installing Packages

You may need to install certain packages to do the empirical exercises. This can be done with a single command in R. You should double-check the package names required for each exercise, but it would be something like:
install.packages(c('wooldridge', 'lmtest', 'sandwich', 'forecast'))

With R on your own computer, you only need to run this once (not every time you use your computer), but with a web interface or computer lab, you may need to run this code every time you start a session in R. You can check which packages are already installed with `installed.packages()`

A bit about the packages:

- **wooldridge** (Shea, 2018) has datasets originally collected by Wooldridge (2020) from various sources.
- **lmtest** and **sandwich** (Zeileis, 2004; Zeileis and Hothorn, 2002) help construct confidence intervals (and other things) appropriate for economic data.
- **survey** (Lumley, 2004, 2019) has functions for dealing with complex survey sampling.

1.3 Stata

1.3.1 Accessing the Software

There are three ways you could run Stata: in a campus computer lab, through Mizzou’s Software Anywhere, or (if you purchase your own copy) downloaded onto your own computer.

Empirical exercises only require built-in commands. Stata has additional commands available for download, but none are needed for the exercises, so any (internet-connected) computer with Stata is sufficient.

In a Mizzou Computer Lab

You can check which Mizzou computing sites/labs have your favorite software on the Computing Sites Software web page. Scroll down to Stata to see where it’s available. However, sometimes there are classes or other events in computer labs; you can check the weekly schedule posted near the door to find a free time, or you can check online.

After you log into the computer in the computer lab, open Stata from the Start menu (the actual name is somewhat longer, like “StataSE 15 (64-bit)”). Ideally, you should open the do-file editor, and save a .do file, but for this class you could just type commands

---

4 [https://doit.missouri.edu/services/computing-sites/sites-software](https://doit.missouri.edu/services/computing-sites/sites-software)

5 [https://doit.missouri.edu/services/computing-sites](https://doit.missouri.edu/services/computing-sites) and click the lab name
into the short, horizontal space at the bottom labeled “Command.” You type a command and hit Enter to run it.

Also, make sure to email yourself your code (or otherwise save it, if you haven’t finished and uploaded to Canvas) before you log out, since your files get deleted when you log out.

**Purchasing and Downloading Software**

Student pricing is shown on the Stata website. Currently (Spring 2020), the cheapest option is the 6-month Stata/IC license. Other, more expensive licenses are fine, too.

The software is delivered via download. Follow instructions for installation, and contact Stata if you have any technical difficulties.

**Software Anywhere (Mizzou)**

From the Software Anywhere web page, click the “Getting Started” tab and follow the instructions. Once logged in, it’s the same as if you were sitting at a computer in a Mizzou computer lab (see above).

Technical assistance: MU Division of IT, techsupport@missouri.edu

1.3.2 **Installing Additional Commands**

Like in R, there are additional Stata commands that can be easily downloaded and installed. Commonly, this can be done with the command `ssc install` followed by the name of the command.

For the exercise sets, the only additional command you’ll need is `bcuse`. You can install this with the command `ssc install bcuse`. If you’re in a computer lab, you may need to run this command every time you start Stata; if you have it on your computer, just once is sufficient. This command makes it easy to load the datasets from Wooldridge (2020).

1.4 **Optional Resources**

If you only want to learn enough R (or Stata) to do well in this class, then you may skip this section. If you’d like to learn more on your own, these resources might help you get started in the right direction.

1.4.1 **R Tutorials**

Eventually, you will be able to simply Google questions you have about R. There are lots of people on the internet really excited about helping you figure stuff out in R, which is great.

---


7 [https://doit.missouri.edu/services/software/software-anywhere](https://doit.missouri.edu/services/software/software-anywhere)

8 Descriptions: [http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html](http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html)
However, when you are first getting started, it may help to go through a basic tutorial. You are welcome to Google “R basic tutorial” yourself, or you could try one of the following.

1. Section 2.3 (“Lab: Introduction to R”) in James, Witten, Hastie, and Tibshirani (2013)
2. Section 1.1 in Hanck et al. (2018)
3. Sections 1.1–1.3 in Heiss (2016)
4. Sections 2.1–2.5 in Kleiber and Zeileis (2008) [Chapter 2 is free on their website]
5. Chapter 2 in Kaplan (2020)
6. No longer free after first chapter: datacamp.com courses like Introduction to R

1.4.2 R Quick References
At first, it may help to have some quick reference “cheat sheets.”

1.4.3 Running Code in This Textbook
If you’d like, you should be able to copy code directly from the textbook .pdf file and paste it into R. Sometimes, you need to install a certain package first. This can be done either manually or with the R function install.packages(). For example, to install package mgcv, run the command install.packages('mgcv') within R.

1.4.4 Stata Resources
For Stata, helpful cheat sheets (quick references) are available for free as well as various tutorials.

---

9 https://www.datacamp.com/courses/free-introduction-to-r
11 https://cran.r-project.org/doc/contrib/Short-refcard.pdf
13 https://www.stata.com/links/resources-for-learning-stata
Empirical Exercises

Empirical Exercise EE1.1. In either R or Stata, create a script (a sequence of commands, with one command per line) to do the following. The data are from a New York Times article on December 28, 1994.

a. R: load (and install if necessary) package wooldridge:
   ```r
   if (!require(wooldridge)) {
     install.packages('wooldridge'); library(wooldridge)
   }
   ```
   Stata: run `ssc install bcuse` to ensure command `bcuse` is installed, and then load the dataset with `bcuse wine, clear`

b. View basic dataset info with R command `?wine` or Stata command `describe`

c. View the first few rows of the dataset with R command `head(wine)` or Stata command `list if _n<=5`

d. Rename the `alcohol` column, which measures liters of alcohol from wine (consumed per capita per year).
   R: ```r
   names(wine)[2] <- 'wine'
   ```
   Stata: `rename alcohol wine`

e. Add a column named `id` whose value is just 1, 2, 3, 4, 5, etc.
   R: ```r
   wine$id <- 1:nrow(wine)
   ```
   Stata: `generate id = _n`

f. Display the countries with fewer than 100 heart disease deaths per 100,000 people.
   R: ```r
   wine$country[wine$heart<1/zero.alt3/zero.alt3]
   ```
   Stata: `list country if heart<1/zero.alt3/zero.alt3`

g. Display the rows for the countries with the 5 lowest death rates, sorted by death rate.
   R: ```r
   wine[order(wine$deaths)[1:5],]
   ```
   Stata: `sort deaths` followed by `list if _n<=5`

h. Add a column with the sum of heart and liver disease deaths per 100,000.
   R: ```r
   wine$heart.plus.liver <- wine$heart + wine$liver
   ```
   Stata: `generate heart_plus_liver = heart + liver`

i. Generate a variable with the squared death rate.
   R: ```r
   wine$deaths.sq <- wine$deaths^2
   ```
   Stata: `generate deaths_sq = deaths^2`

j. Display the sorted death rates.
   R: `print(sort(wine$deaths))`
   Stata: `sort deaths` followed by `list deaths`
k. R: create a vector with the proportion of total deaths (per 100,000) caused by heart disease with command `heart.prop <- wine$heart/wine$deaths` and then name the entries by country with `names(heart.prop) <- wine$country` and print the named vector of heart disease death proportions, rounded to three decimal places, with `print(round(heart.prop, digits = 3))`

Stata: add a column with the proportion of heart deaths to total deaths with command `generate heart_prop = heart / deaths`

l. Create a histogram of liver deaths.
   R: `hist(wine$liver)`

   Stata: `histogram liver`

m. Create a scatterplot of liver death rates (vertical axis) against wine consumption (horizontal axis).
   R: `plot(x=wine$wine, y=wine$liver)`

   Stata: `scatter liver wine`

n. R only: make the same plot but with axes starting at zero, adding the arguments `xlim=c(0, max(wine$wine))` and `ylim=c(0, max(wine$liver))` to the previous `plot()` command
CHAPTER 1. GETTING STARTED WITH R (OR STATA)
Part I

Analysis of One Variable
Introduction

This text explores methods to answer three types of economic questions, each detailed in Part I:

1. Description (how things are/were: statistical properties and relationships)
2. Prediction (guessing an unknown value, without interfering)
3. Causality (how changing one variable would affect another, all else equal)

For example, imagine you are interested in income. Depending on your job, you may want to answer a different type of question, like:

1. **Description:** how many adults in the U.S. have an income below $20,000/yr? What’s the mean income among U.S. adults? What’s the difference in mean income between two socioeconomic or demographic groups, like those with and without a college degree?

2. **Prediction:** for advertising purposes, what’s the best guess of the income of an unknown person visiting your company’s website? What’s the best prediction if you also know their zip code (where they live)?

3. **Causality:** for a given individual, how much higher would her income be if she had a college degree than if she didn’t, keeping everything else about her (parents, height, social skills, etc.) identical? How much higher would her income be if she were a man, all else equal? If she were white?

Description helps us see. It summarizes an incomprehensible mass of numbers into specific, economically important features we can understand. By analogy: knowing the color of each of 40,000 pixels in a photograph is not as valuable as knowing it’s a cat.

Prediction aids decisions dependent on unknowns. The example questions above consider the purpose of advertising, where correctly guessing a person’s income helps decide which ad is most effective. In other private sector jobs, you may need to predict future demand to know how many self-driving cars to start producing, or predict future oil prices to aid a freight company’s decisions. In government or non-profit work, optimal policy may depend on predicting next year’s
unemployment rate. In each case, as detailed in Section 2.5, the “best” prediction depends on the consequences of the related decision.

Causality also aids decisions. The example question about the causal effect on income of a college degree matters for government policies to subsidize college (or not), as well as individual decisions to attend college. With business decisions, like changes to advertising or website layout, the causal effect on consumer behavior is what matters: does the change itself actually cause consumers to buy more? Among the three types, questions of causality are the most difficult to answer. Learning about causality from data has been a primary focus of the field of econometrics.

Of course, not all important questions concern description, prediction, and causality. Policy questions usually involve tradeoffs that ultimately require value judgments. For example, how much future wellbeing is worth sacrificing to be better off right now? How much GDP is worth sacrificing to decrease inequality? Should a school have honors classes that help the best students at the expense of the other students? Each of these policy questions requires a subjective value judgment that cannot be answered objectively from data.

That said, each policy question also depends on objectively quantified description, prediction, and causality. For example, the policy question about decreasing inequality depends on the current levels of GDP and inequality (description) as well as the causal effect of the policy (e.g., tax) change on GDP and inequality (causality). The future/present wellbeing tradeoff depends on the current level of wellbeing (description) as well as future levels (prediction). The honors class tradeoff depends on the causal effect of honors classes on different types of students (causality) as well as the current mix of student types (description) and future mix (prediction).
Chapter 2

One Variable: Population

⇒ Kaplan video: Chapter Introduction

Depends on: no other chapters

Unit learning objectives for this chapter

2.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]

2.2. Describe and distinguish among different types of populations, including which is most appropriate for answering a certain question [TLO 3]

2.3. Describe distributions in different ways, including units of measure [TLO 3]

2.4. Assess the most appropriate loss function and prediction in a real-world situation [TLO 6]

2.5. Compute mean loss and the optimal prediction in simple mathematical examples [TLO 2]

Optional resources for this chapter

- Basic probability: the Khan Academy AP Statistics unit includes instructional material and practice questions:
- Mean (expected value) (Lambert video)
- Probability distribution basics on Wikipedia (more than you need to know for this class)
- Optimal prediction: Hastie, Tibshirani, and Friedman (2009, §2.4)
- Section 2.1 (“Random Variables and Probability Distributions”) in Hanck et al. (2018)

Chapter 2 studies a single variable by itself. This setting’s simplicity helps us focus on the complexity of fundamental concepts in
probability, description, and prediction. This fundamental understanding will help you tackle more complex models later, in this class and beyond.

If you’ve previously had a probability or statistics class, then most of this chapter may be review for you, although the optimal prediction material is probably new. If you haven’t, then now is your opportunity to catch up.

2.1 The World is Random

⇒ Kaplan video: “Before” and “After” Perspectives of Data

2.1.1 Before and After: Two Perspectives

Consider a coin flip. The two possible outcomes are heads (h) and tails (t). After the flip, we observe the outcome (h or t). Before the flip, either h or t is possible, with different probabilities.

Let variable W represent the outcome. After the flip, the outcome is known: either \( W = h \) or \( W = t \). Before the flip, both \( W = h \) and \( W = t \) are possible. If the coin is “fair,” then possible outcome \( W = h \) has probability \( \frac{1}{2} \), as does \( W = t \). (Recall it is equivalent to write \( \frac{1}{2}, 0.5, \) or \( 50\% \).)

The “after” view sees \( W \) as a realized value (or realization). It is either heads or tails. Even if the actual “value” (heads or tails) is unknown to us, there is just a single value. For example, in physics the variable \( c \) represents the speed of light in a vacuum; you may not know the value, but \( c \) represents a single value.

Instead, the “before” view sees \( W \) as a random variable. That is, instead of representing a single (maybe unknown) value like in algebra, \( W \) represents a set of possible values, each associated with a probability. In the coin flip example, the possible outcomes are \( h \) and \( t \), and the associated probabilities are 0.5 and 0.5.

Other terms for \( W \) include a random draw (or just draw), or more specifically a random draw (or “randomly drawn”) from a particular probability distribution. Seeing the population as a probability distribution (see Section 2.2), we could say \( W \) is randomly sampled from its population distribution, or if there are multiple random variables \( W_1, W_2, \ldots \) (e.g., multiple flips of the same coin), we could say they are randomly sampled from the population or that they collectively form a random sample; see Section 3.2 for more about sampling.

Notationally, in this text, random variables are usually written uppercase (like \( W \) or \( Y \)), whereas realized values are usually written lowercase (like \( w \) or \( y \)). This notation is not unique to this textbook, but beware that other books use different notation. (For more on notation, see the Notation section in the front matter before Chapter 1.)
2.1.2 Before and After Sampling

Extending Section 2.1.1 are the before sampling and after sampling perspectives, or “before observation” and “after observation.” Similar to Section 2.1.1, “before” corresponds to random variables, whereas “after” corresponds to realized values.

For example, imagine you plan to record the age of one person living in your city. You take a blank piece of paper on which you’ll write the age. As in Section 2.1.1, after you choose a person and write their age (“after sampling”), that number can be seen as a realized value, like \( w \). Before sampling, there are many possible numbers that could end up on your paper. It’s not that your city’s citizens’ ages are undetermined; they each know their own age. But before you “sample” somebody, it’s undetermined whose age will end up on your paper. It could be your neighbor DeMarcus, age 88. It could be your kid’s friend Lucia, age 7. It could be your colleague Xiaohong, age 35. The random variable \( W \) is like your blank paper: it has many possible values, each with some probability of occurring, like \( P(\text{W} = 88) \) or \( P(\text{W} = 7) \).

Discussion Question 2.1 (web traffic). Let \( Y = 1 \) if you’re logged into the course website and \( Y = 0 \) if not.

a) From what perspective is \( Y \) a non-random value?

b) From what perspective is \( Y \) a random variable?

There is always a “before” view from which data samples (like ages) can be seen as random variables, although sometimes it requires some additional peculiar thought experiments, like imagining we first “sample” one universe out of many, like with the superpopulation in Section 2.2.

In Sum: Before & After

Before: multiple possible values \( \Rightarrow \) random variable
After: single observed value \( \Rightarrow \) realized value (non-random)

2.1.3 Outcomes and Mechanisms

Knowing everything about a coin does not fully determine the outcome of a single coin flip. For example, even if we flip two identical coins (i.e., same probability of heads) at the same time, one may get heads while the other gets tails. Mathematically, with two coins represented by \( W \) and \( Z \), even if they are “identical” in that \( P(\text{W} = h) = P(\text{Z} = h) \) and \( P(\text{W} = t) = P(\text{Z} = t) \), we could still sometimes observe \( W = h \) and \( Z = t \). More abstractly: knowing everything about random variable \( W \) does not fully determine any particular realization \( w \). Even if random variables \( W \) and \( Z \) have the same properties, specific realizations \( W = w \) and \( Z = z \) may differ.

Conversely, a single coin flip’s outcome does not tell us everything about the coin itself. For example, consider a “fair” coin \( W \) with
$P(W = h) = P(W = t) = 1/2$ (50% chance of either heads or tails), and biased coin $Z$ with $P(Z = h) = 0.99$ (99% heads). By chance, we may flip both and observe $W = h$ and $Z = h$. But the fact that they both came up heads once does not imply that the coins themselves are identical. More abstractly: observing a single realization $W = w$ does not tell us all the properties of random variable $W$.

We usually want to learn about the underlying mechanisms, like the coin itself. The “before” view in Section 2.1.1 lets us describe the underlying properties that we want to learn, like a coin’s probability of heads, $P(W = h)$.

The coin flip is a metaphor for more complex mechanisms. In economics, instead of learning how coin flip outcomes are determined, we care about the underlying mechanisms that determine a wide variety of outcomes like unemployment, wages, inflation, trade volume, fertility, and education. The underlying mechanism is often called the data-generating process (DGP).

### 2.2 Population Types

This section describes different population types and how to determine which is most appropriate for a particular economic question, which in turn helps determine which econometric method is most appropriate.

In this textbook, the population is modeled mathematically as a probability distribution. This is appropriate for the infinite population or superpopulation below, but not the finite population. Consequently, it is most important to distinguish between the finite population and the other two types.

The finite population cares more about the “after” view: which outcomes actually occurred? The other two population types care more about the “before” view, describing properties of the underlying mechanisms that generated the outcomes (the DGP).

#### 2.2.1 Finite Population

In English, “population” means all the people living in some area, like everybody living in Missouri. In econometrics, this type of population is called a **finite population**. Other examples of finite populations are all employees at a particular firm, all firms in a particular industry, all students in a particular school, or all hospitals of a certain size.

The finite population is appropriate when we only care about the outcomes of the population members, not the mechanisms that determine such outcomes. For example, if we want to know how many individuals in Missouri are currently unemployed, then our interest is in a finite population. That is, we don’t care why they’re unemployed, and we don’t care about the probability that they’re unemployed; we only care about whether or not they are currently unemployed.
2.2.2 Infinite Population

Sometimes a finite population is so large compared to the sample size (i.e., the number of population members we observe) that an infinite population is a reasonable approximation. For example, if we observe only 600 individuals out of the 6+ million in Missouri, econometric results based on finite and infinite populations are practically identical.

Although “infinite” sounds more complex than “finite,” it is actually simpler mathematically. Instead of needing to track every single member of a finite population, an infinite population is succinctly described by a probability distribution or random variable. For example, a finite population would need to consider the employment status of all 6+ million Missourians, because sampling somebody unemployed then reduces the number of unemployed individuals remaining in the population who could be sampled next. In contrast, an infinite population considers realizations of a random variable $W$ with some probability of having value “unemployed.” There is no effect of removing one individual from an infinite population since $1/\infty = 0$.

Besides this convenience, sometimes there is no finite population (however large) that answers your question. For example, imagine there’s a new manufacturing process for carbon monoxide monitors that should sound an alarm above 50ppm. Most work properly, but some are faulty and never alarm. Specifically, this manufacturing process corresponds to some probability of producing a faulty monitor. This is similar to the probability of the coin flipping process producing a “heads.” Mathematically, the manufacturing process can be modeled as random variable $W$ with some probability of the value “faulty.” If you want to learn this probability (i.e., this property of the manufacturing process), then there is no finite number of monitors that can exactly answer your question; no finite number of realizations exactly determines $P(W = \text{faulty})$. This is an infinite population question.

2.2.3 Superpopulation

One variation of the infinite population is the superpopulation (coined by Deming and Stephan, 1941). This imagines (infinitely) many possible universes; our actual universe is just one out of infinity. Thus, even if it appears we have a finite population, we could imagine that our universe’s finite population is actually a single sample from an infinite number of universes’ finite populations. The term “superpopulation” essentially means “population of populations.” Our universe’s finite population “is only one of the many possible populations that might have resulted from the same underlying system of social and economic causes” (Deming and Stephan, 1941, p. 45).

For example, imagine we want to learn the relationship between U.S. state-level unemployment rates and state minimum wage levels. It may appear we are stuck with a finite population because there are only 50 states, each of which has an observable unemployment rate and minimum wage. However, observing all 50 states still doesn’t
fully answer our question about the underlying mechanism that relates unemployment and minimum wage, so a finite population seems inappropriate. But we can’t just manufacture new states like we can manufacture new carbon monoxide monitors, so an infinite population also seems inappropriate. The superpopulation imagines manufacturing new entire universes, each with 50 states and the same economic and legal systems. Given these underlying systems and mechanisms, the states’ unemployment rates can be seen as random variables, with various probabilities of the possible values. To answer our economic question, we need to learn about the properties of these random variables, not merely the actual unemployment in our actual 50 states.

2.2.4 Which Population is Most Appropriate?

Practically, you need to decide which econometric method to use to answer a particular question. This decision depends partly on which population type is most appropriate. Specifically, finite-population methods differ from other methods that are appropriate for either superpopulations or infinite populations. Because they are less commonly used in econometrics, finite-population methods are not covered in this textbook.

Consequently, it is most important to judge whether or not a finite population is more appropriate than the other types. Which is most appropriate depends on your question (i.e., what you want to learn). The finite population is most appropriate if you could fully answer your question by observing every member of a finite population. If not, then a superpopulation or infinite population is more appropriate.

The distinction is described by Deming and Stephan (1941, p. 45). They say the finite population perspective is more appropriate for “administrative purposes” or “inventory purposes,” whereas the superpopulation perspective is more appropriate for “scientific generalizations and decisions for action [policy],” as well as “prediction” (assuming you want to predict values outside the finite population, like in the future).

In Sum: Population Type

Hypothetically, could a finite number of observations fully answer your question?

No $\implies$ superpopulation or infinite population, modeled as probability distribution (as in this textbook)

Yes $\implies$ finite population (use different methods unless sample is much smaller than population)

Example: Coin Flips

Imagine the president flips a coin 20 times and then randomly selects 10 observations to report to you; which population types is most
appropriate? It depends on your question.

The finite population is most appropriate if you only care about the outcomes of those 20 flips. For example, this may be true if the president was flipping the coin to make a major military decision that you care about (like, “invade if at least 10/20 heads”). Then, knowing the 20 flip outcomes is enough to learn the decision. Further, the sample size is a fairly large proportion (10/20), so approximating 20 as infinity seems inappropriate.

The infinite population is more appropriate if you care about the properties of the coin. For example, even with a fair coin \( p = 1/2 \), maybe only 5 of 20 flips came up heads. You don’t care that the finite-population proportion of heads was 1/4; you care about the \( p = 1/2 \) property of the coin itself. You still have uncertainty about \( p \) even after observing all 20 outcomes.

Other Examples

Consider the employment status of individuals in Missouri. A finite population is more appropriate if you want to document the actual percentage of Missouri individuals unemployed last week. A superpopulation is more appropriate if you want to learn about the underlying mechanism that relates education and unemployment. That is, knowing each individual’s employment status fully answers the first question, but not the second question.

Consider the productivity of employees at your company (you’re the CEO). If you want to know each employee’s productivity over the past fiscal quarter, then a finite population is more appropriate. If you want to learn how a particular company policy affects productivity, then a superpopulation is more appropriate. That is, knowing each employee’s productivity fully answers the first question, but not the second question.

Discussion Question 2.2 (student data). Imagine you’re a high school principal. You have data on every student, including their standardized test scores from last spring.

a) Describe a specific question for which the finite population is most appropriate, and explain why.

b) Describe a specific question for which an infinite population or superpopulation is most appropriate, and explain why.

2.3 Description of a Population

Like most econometrics textbooks, this textbook models the population as a probability distribution. Section 2.2 helps you distinguish when this is appropriate.

Description of a population is thus description of a probability distribution. Some distributions are completely described by a single number, like a coin’s probability of heads. Others are very complicated, so they are summarized by particular features like the mean and standard deviation.
Later, with regression, we’ll think about the relationship between the value of variable $X$ and the value of a summary feature of the $Y$ probability distribution. There some caveats to that statement, but the point is that you must understand the probability distribution of $Y$ by itself before understanding how its features depend on $X$.

Remember: there is no data yet. In practice (and starting in Section 3.4), you use data to learn about the population, to answer questions about description, prediction, or causality. Here, we consider what could possibly be learned, specifically for description.

The following subsections describe probability distributions for different types of variables, as well as appropriate summary features. First, a brief overview is given.

### 2.3.1 Overview of Distributions and Their Features

#### Complete Description

To completely describe a distribution requires a probability mass function or cumulative distribution function, depending on the type of variable (details below).

When appropriate, the **probability mass function** (PMF) gives the probability that random variable $Y$ is equal to any one of its possible values. Notationally, the PMF is usually a lowercase $f$, sometimes with the variable as a subscript, like $f_Y(\cdot)$. The $\cdot$ indicates that $f_Y(\cdot)$ is an entire function, not a scalar variable, nor a function evaluated at a particular point. Mathematically,

$$f_Y(y) \equiv P(Y = y). \tag{2.1}$$

If $y$ is not a possible value of $Y$, then $P(Y = y) = 0$.

In other cases, more appropriate is the **cumulative distribution function** (CDF) that gives the probability of all values less than or equal to $y$. (Less commonly: “distribution function” or DF.) Notationally, the CDF is usually an uppercase $F$, sometimes with the variable as a subscript, like $F_Y(\cdot)$. Again, the $\cdot$ indicates that $F_Y(\cdot)$ is an entire function. Mathematically,

$$F_Y(y) \equiv P(Y \leq y). \tag{2.2}$$

Either the PMF or CDF provides a full description of the distribution of $Y$. For some variable types, both are appropriate, in which case the PMF and CDF contain the same information (just represented differently). For other variable types, only the PMF is appropriate, or only the CDF.

If you are studying a single variable, graphing the PMF or CDF is helpful, and it shows all the available information about that variable’s distribution. Even if the PMF or CDF is a complicated function, humans are good at processing visual data. In practice, however, often you study many variables together, in which case even graphing becomes intractable (e.g., you can’t make a five-dimensional graph easily understood).
Distributions’ features like the mean are convenient summaries but lose information. There is a tradeoff. For some purposes, you may need all the information of the probability distribution. For other purposes, a few summary features may suffice and be easier to understand, compare, and communicate.

The **mean** is the main summary feature considered, for numeric variables. It provides a general idea of how high or low values are, weighted by probability. It provides some sense of the “center” or “location” of the distribution. The mean is used extensively in later chapters.

The **standard deviation** captures how spread out a distribution is, for numeric variables. If both very low and very high values have enough probability, then the standard deviation is high. Conversely, if possible values are all concentrated in a small range (with high probability), then the standard deviation is low.

Other summary features are mentioned briefly but not studied in future chapters. The **median** (a particular percentile) provides another way to define the “center” of a distribution. The **mode** is the single most likely value. The mode applies even to non-numeric variables, and the median also applies as long as the non-numeric values have a low-to-high order. An alternative spread measure is the **interquartile range**.

### In Sum: Random Variable Types & Features

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary</strong></td>
<td>(Section 2.3.2): P(Y = 1) = E(Y) is complete description</td>
</tr>
<tr>
<td><strong>Discrete</strong></td>
<td>(Section 2.3.3): mean E(Y) captures high/low; standard deviation ( \sigma_Y ) captures how spread out; PMF ( f_Y(y) = P(Y = y) ) shows probability of each possible value; CDF ( F_Y(y) \equiv P(Y \leq y) )</td>
</tr>
<tr>
<td><strong>Nominal</strong></td>
<td>(Section 2.3.4): PMF says probability of each category; mode is most likely category; no CDF, mean, standard deviation</td>
</tr>
<tr>
<td><strong>Ordinal</strong></td>
<td>(Section 2.3.4): similar to nominal but has CDF; use median instead of mean</td>
</tr>
<tr>
<td><strong>Continuous</strong></td>
<td>(Section 2.3.5): similar to discrete but PDF instead of PMF</td>
</tr>
</tbody>
</table>

### 2.3.2 Binary Variable

A **binary variable** has two possible values. Other terms for a binary variable are **dummy variable**, **indicator variable**, and **Bernoulli random variable**. In economics, “dummy” and “binary” are most common.

Unless otherwise specified, a binary variable’s two possible values are 0 and 1. For writing mathematical models, these values are usually more convenient than values like “heads” and “tails.” Mathematically, this can be indicated by \( Y \in \{0, 1\} \): the value of \( Y \) must be in the set that includes only the numbers 0 and 1. (The set \( \{0, 1\} \) is different...
Many important variables are binary. Examples include:

- whether the economy is in a recession (1) or not (0)
- whether somebody has a college degree (1) or not (0)
- whether a pharmaceutical drug is branded (1) or generic (0)
- whether somebody is employed (1) or not (0)
- whether a retailer is a franchise (1) or not (0)

Mathematically, binary variables are often defined using the indicator function. The indicator function $1\{\cdot\}$ equals 1 if the argument is true and 0 if false:

$$1\{A\} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases}.$$ (2.3)

For example, consider defining a binary random variable $Y$ based on the coin flip random variable $W$. Recall that the possible values of the flip are $W = h$ (heads) and $W = t$ (tails). We now want $Y = 1$ to indicate heads, and $Y = 0$ tails. Mathematically,

$$Y = 1\{\text{heads}\} = 1\{W = h\} = \begin{cases} 1 & \text{if } W = h \text{ (heads)} \\ 0 & \text{if } W = t \text{ (tails)} \end{cases}.$$ (2.4)

Other examples can also be written with an indicator function. For example, $Y = 1\{\text{recession}\}$, $Y = 1\{\text{branded}\}$, or $Y = 1\{\text{franchise}\}$.

**Probability Mass Function**

A binary random variable’s PMF is like in (2.1). Specifically, $f_Y(0) = P(Y = 0)$ and $f_Y(1) = P(Y = 1)$.

For example, consider the employment dummy $Y = 1\{\text{employed}\}$. That is, $Y = 1$ if the individual is employed, otherwise $Y = 0$. The PMF $f_Y(\cdot)$ is

$$f_Y(1) \equiv P(Y = 1) = P(\text{employed}), \quad f_Y(0) \equiv P(Y = 0) = P(\text{not employed}).$$ (2.5)

If in the population 80% of individuals are employed (and thus 20% not), then $f_Y(1) = 80\% = 0.8$ and $f_Y(0) = 20\% = 0.2$.

The binary PMF can actually be written in terms of one single parameter, $p \equiv P(Y = 1)$. Since $Y \in \{0, 1\}$, $P(Y = 0) + P(Y = 1) = 1 = 100\%$. By algebra, $P(Y = 0) = 1 - P(Y = 1) = 1 - p$. Thus, the PMF is

$$f_Y(1) \equiv P(Y = 1) = p, \quad f_Y(0) \equiv P(Y = 0) = 1 - p.$$ (2.6)

The probability distribution corresponding to (2.6) is called a **Bernoulli distribution**. That is, if random variable $Y$ has the PMF in (2.6), then we say $Y$ follows a Bernoulli distribution with parameter $p$. Mathematically,

$$Y \sim \text{Bernoulli}(p).$$ (2.7)
2.3. DESCRIPTION OF A POPULATION

Cumulative Distribution Function

Although there is no practical benefit of a binary cumulative distribution function (over a PMF), it may help you develop intuition. The CDF of binary $Y$ has a particular structure. If $r < 0$, then $P(Y \leq r) = 0$. If $r = 0$, then $P(Y \leq r) = P(Y = 0)$; the CDF jumps up (discontinuously) from 0 to $P(Y = 0)$ at $r = 0$. If $0 < r < 1$, then $P(Y \leq r) = P(Y = 0)$, too; the CDF is flat. If $r = 1$, then $P(Y \leq r) = P(Y \leq 1) = 1$; the CDF again jumps, now from $P(Y = 0)$ to 1. If $r > 1$, then $P(Y \leq r) = 1$, too; the CDF remains flat. Altogether, letting $p \equiv P(Y = 1)$ and $1 - p = P(Y = 0)$, the CDF of $Y$ is

$$F_Y(r) = (1 - p) \mathbb{1}\{r \geq 0\} + p \mathbb{1}\{r \geq 1\} = \begin{cases} 0 & \text{if } r < 0 \\ 1 - p & \text{if } 0 \leq r < 1 \\ 1 & \text{if } r \geq 1. \end{cases}$$

(2.8)

Summary Feature: Mean

Since a Bernoulli distribution is fully described by $p \equiv P(Y = 1)$, there is no need to summarize it further. However, the following result is helpful for interpretation of regressions with binary $Y$, and it helps develop intuition. A random variable’s mean is a probability-weighted average of its possible values. With binary $Y$, the possible values are 0 and 1, with respective weights $P(Y = 0)$ and $P(Y = 1)$. The mean $E(Y)$ is thus

$$E(Y) = \sum_{j=0}^{1} (j) P(Y = j) = (0) P(Y = 0) + (1) P(Y = 1) = P(Y = 1).$$

(2.9)

So for any binary $Y$, $E(Y) = P(Y = 1) = f_Y(1)$.

For terminology, the mean $E(Y)$ is also called the expected value or expectation. These names explain the letter $E$ in the mathematical notation.

However, the terms “expectation” and “expected value” cause much confusion. They are technical terms whose meaning differs greatly from the colloquial English meaning. For example, if you say in plain English, “I expect the value will be 0.5,” it means you think there’s a good chance (high probability) that the value will exactly equal 0.5. This is not what $E(Y) = 0.5$ means. In fact, with a binary $Y$, it is impossible to have $Y = 0.5$. We may expect (colloquially) $Y = 1$ if $P(Y = 1)$ is high, or we may expect (colloquially) $Y = 0$ if $P(Y = 0)$ is high, but it is impossible to have $Y = E(Y)$ (unless $p = 1$ or $p = 0$), which is very confusing. I suggest you think “mean” every time you see $E(Y)$ or read “expected value” or “expectation.”

Summary Feature: Standard Deviation

Again, since a Bernoulli distribution is fully described by $p \equiv P(Y = 1) = E(Y)$, there is no need to summarize it further. However, the
The standard deviation is one measure of how “spread out” or “dispersed” a distribution is. The standard deviation is defined as the square root of the variance. Most commonly, lowercase sigma is used for notation: \( \sigma_Y^2 \) is the variance, and \( \sigma_Y \) is the standard deviation. With this notation,

\[
\sigma_Y^2 = \text{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))^2], \quad \sigma_Y = \sqrt{\sigma_Y^2}. \tag{2.10}
\]

For \( Y \sim \text{Bernoulli}(p) \), the variance and standard deviation are \( \sigma_Y^2 = p(1 - p) \) and \( \sigma_Y = \sqrt{p(1 - p)} \). The derivation from (2.10) is not important (unless you want a PhD).

The formula \( \sigma_Y = \sqrt{p(1 - p)} \) has some intuition. If \( p = P(Y = 1) = 1 \), then \( Y = 1 \) always (never \( Y = 0 \)), so the distribution is not at all spread out; it is very concentrated, on one single value. This is reflected by \( \sigma_Y = \sqrt{1(1 - 1)} = \sqrt{0} = 0 \). Similarly, if \( p = P(Y = 1) = 0 \), then \( Y = 0 \) always (never \( Y = 1 \)), again not at all spread. This is reflected by \( \sigma_Y = \sqrt{0(1 - 0)} = \sqrt{0} = 0 \). In contrast, if \( p = P(Y = 1) = 1/2 \), then \( Y = 0 \) and \( Y = 1 \) are equally likely. This is as spread out as possible for a binary distribution. Then, \( \sigma_Y = \sqrt{(1/2)(1 - 1/2)} = \sqrt{1/4} = 1/2 \). You can graph \( \sqrt{p(1 - p)} \) over \( 0 \leq p \leq 1 \) to see that indeed \( \sigma_Y \) is highest at \( p = 1/2 \).

That said, for binary \( Y \), it is redundant to report both \( \mathbb{E}(Y) \) and \( \sigma_Y \), since \( \sigma_Y = \sqrt{\mathbb{E}(Y)[1 - \mathbb{E}(Y)]} \). Once you know \( p = \mathbb{E}(Y) \), or equivalently \( p = P(Y = 1) \), there is no new information in \( \sigma_Y \).

### 2.3.3 Discrete Variable

A binary variable is a special case of a discrete variable, which has any (countable) number of possible values. That is, all binary variables are discrete variables, but not all discrete variables are binary. Discrete variable examples include:

- an individual’s years of education
- number of children in a household
- the number of times a stock has split since its IPO
- the number of trading partners a country has
- number of students in a classroom

The units of measure are important for interpreting a discrete variable and its distribution. For most discrete variables, like number of children, the units are obvious. Sometimes it is not immediately obvious: number of students...per room or per grade? Number of bills passed...in one month or one year or one term?

### Probability Mass Function

A discrete PMF is similar to a binary PMF. It is again usually written like \( f_Y(\cdot) \) for the PMF of discrete random variable \( Y \). The PMF’s input is again a possible value, and it’s output is the corresponding...
Recall uppercase \( Y \) is a random variable, whereas \( y \) stands for one particular non-random value (like 41). If \( y \) is not one of the possible values of \( Y \), then \( f_Y(y) = P(Y = y) = 0 \).

The main difference is that a general discrete PMF cannot be fully described by a single parameter \( p \) like in (2.6). This additional complexity is a reason people look at summary features like the mean, standard deviation, and percentiles.

One dimension of added complexity is the possible values. There can be more than two possible values. Further, they are not always just 0, 1, 2, ... For example, if recessions are determined on a monthly basis, then the fraction of a year spent in recession could be 0, 1/12, 2/12, ..., 11/12, 1. Consequently, the possible values are often written as \( y_1, y_2, ..., y_J \), where \( J \) is the number of different values. Equivalently: the possible values are \( y_j \) for \( j = 1, 2, ..., J \). (Notation: this has the same meaning as \( j \in \{1, 2, ..., J \} \), but the convention is to write \( j = 1, 2, ..., J \), or simply \( j = 1, ..., J \).)

Having more possible values also means more probabilities to keep track of. Specifically, if there are \( J \) possible values, then there are \( J \) probabilities: \( P(Y = y_j) \) for \( j = 1, ..., J \). Mathematically, the PMF can be written as

\[
f_Y(y) = \sum_{j=1}^{J} 1\{y = y_j\} P(Y = y_j). \tag{2.12}
\]

The last \( P(Y = y_J) \) can be solved for using the fact that all \( J \) probabilities sum to 1, but that still leaves \( J - 1 \) probabilities.

Figure 2.1 shows two common ways of graphing the PMF \( f_Y(y) = 1/3 \) for \( y = 1, 2, 3 \), i.e., \( f_Y(1) = f_Y(2) = f_Y(3) = 1/3 \).

Cumulative Distribution Function

A discrete CDF is defined as in (2.2). It has a similar pattern as (2.8), where it is flat but then jumps up discontinuously at each possible
value $y_j$. Mathematically, it can be written similarly to (2.12),

$$F_Y(y) \equiv P(Y \leq y) = \sum_{j=1}^{J} 1\{y_j \leq y\} P(Y = y_j). \quad (2.13)$$

The CDF corresponding to the PMF in Figure 2.1 could be written two equivalent ways:

$$F_Y(y) = (1/3) \sum_{j=1}^{3} 1\{j \leq y\}, \quad (2.14)$$

$$F_Y(y) = \begin{cases} 
0 & \text{if } y < 1 \\
1/3 & \text{if } 1 \leq y < 2 \\
2/3 & \text{if } 2 \leq y < 3 \\
1 & \text{if } 3 \leq y 
\end{cases}. \quad (2.15)$$

Figure 2.2 plots this CDF.

Figure 2.2: Example discrete CDF from (2.14).

**Summary Feature: Mean**

Generalizing the binary mean in (2.9), the mean of discrete $Y$ can be written in terms of the $J$ possible values $y_j$ ($j = 1, \ldots, J$) and their probabilities:

$$E(Y) = \sum_{j=1}^{J} y_j P(Y = y_j) = y_1 P(Y = y_1) + \cdots + y_J P(Y = y_J), \quad (2.16)$$

which could also be written in terms of the PMF because $f_Y(y_j) = P(Y = y_j)$. If $Y$ is binary, then $J = 2$, $y_1 = 0$, and $y_2 = 1$, in which case (2.16) simplifies to (2.9).

The mean gives a rough sense of whether the distribution has high or low values, weighted by their probability. For example, consider random variables $W$ and $Z$, with $P(W = 0) = P(W = 2) = 1/2$ and $P(Z = 2) = P(Z = 4) = 1/2$. Then,

$$E(W) = (0)(1/2) + (2)(1/2) = 1, \quad E(Z) = (2)(1/2) + (4)(1/2) = 3, \quad (2.17)$$
reflecting that \( Z \) has higher values. As another example, imagine \( W \) and \( Z \) both have possible values \( j = 1, 2, 3, 4 \), but \( P(W = j) = j/10 \), whereas \( P(Z = j) = (5 - j)/10 \). Although the possible values are identical, \( W \) has higher weight for the higher values, which is reflected by its larger mean:

\[
E(W) = \sum_{j=1}^{4} (j)(j/10) = (1)(1/10) + (2)(2/10) + (3)(3/10) + (4)(4/10) = 1 + 0.4 + 0.9 + 1.6 = 3.9
\]

\[
E(Z) = \sum_{j=1}^{4} (j)(5-j)/10 = (1)(4/10) + (2)(3/10) + (3)(2/10) + (4)(1/10) = 0.4 + 0.6 + 0.6 + 0.4 = 2.0
\]

However, the mean is sensitive to very large values, so it does not reflect the value of the “average member of the population.” For example, let \( Y \) denote hourly wage (\$/hr) for a population with three equally-likely types of individuals. The possible values are \( y_1 = 10 \), \( y_2 = 20 \), and \( y_3 = 270 \). The probabilities are \( P(Y = y_j) = 1/3 \) for \( j = 1, 2, 3 \). The “average person” is the middle type, who gets paid $20/hr. (This is the median.) But the mean is, in \$/hr,

\[
E(Y) = (10)(1/3) + (20)(1/3) + (270)(1/3) = 300/3 = 100. \tag{2.18}
\]

This $100/hr mean wage is way higher than what the average person earns. The reason is that the extremely high value $270/hr brings the mean way up. A similar but more extreme example has \( P(Y = 10) = 0.99 \) and \( P(Y = 3010) = 0.01 \). The “average person” is one of the 99% who make $10/hr, but the mean is four times larger, $40/hr:

\[
E(Y) = (10)(0.99) + (3010)(0.01) = 9.9 + 30.1 = 40. \tag{2.19}
\]

The mean helps capture the aggregate earnings rate of the population as a whole, but it does not capture the typical wage of the average population member.

For practice with another example, consider years of education. That is, \( Y = 11 \) means 11 years of education, \( Y = 12 \) means 12 years of education (through high school), etc. For simplicity, imagine the only possible values are \( Y \in \{11, 12, 16, 18\} \). In the notation of (2.16), \( J = 4 \) (four possible values) with \( y_1 = 11 \), \( y_2 = 12 \), \( y_3 = 16 \), and \( y_4 = 18 \). Let \( P(Y = 11) = 0.2 \), \( P(Y = 12) = 0.3 \), \( P(Y = 16) = 0.4 \), and \( P(Y = 18) = 0.1 \). Applying (2.16),

\[
E(Y) = (11)(0.2) + (12)(0.3) + (16)(0.4) + (18)(0.1) = 2.2 + 3.6 + 6.4 + 1.8 = 14. \tag{2.20}
\]

The expectation operator \( E(\cdot) \) has a useful property called **linearity.** Formally: the mean of a linear combination of random variables equals the linear combination of the random variables’ means. For example, given two random variables \( Y \) and \( Z \) (of any type), and two non-random constants \( a \) and \( b \),

\[
E(aY + bZ) = a E(Y) + b E(Z). \tag{2.21}
\]
Here, $aY + bZ$ is a linear combination of random variables $Y$ and $Z$. Thus, the mean of the linear combination of $Y$ and $Z$ equals the linear combination of the means $E(Y)$ and $E(Z)$.

Equation (2.21) implies other identities. For example, in the special case when $b = 0$, it implies $E(aY) = aE(Y)$. As another example, if $a = 1$ and $Y = cW + dX$, then

$$E(cW + dX + bZ) = E(aY + bZ) = aE(Y) + bE(Z) = (1) E(cW + dX) = cE(W) + dE(X) + bE(Z).$$

Extending this further, if we have random variables $Y_i$ for $i = 1, \ldots, n$, and corresponding constants $c_i$, then

$$E\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i E(Y_i).$$

### Summary Feature: Standard Deviation

The standard deviation has the same definition and interpretation as before. It measures how “spread out” or “dispersed” a distribution is, with the same units of measure as the variable itself, and it is formally defined in (2.10).

Without worrying about calculating these by hand, consider the following examples for intuition. Imagine $W$ and $X$ both have possible values $v_1 = -1$, $v_2 = 0$, and $v_3 = 1$, but $P(X = v_j) = 1/3$ for each $v_j$, whereas $P(W = 0) = 1$. Clearly $X$ is more spread out, as reflected by the standard deviations: $\sigma_X = \sqrt{2/3}$, $\sigma_W = 0$. If we spread out the values $v_1$ and $v_3$ farther from zero, then the standard deviation should increase. Let $Y$ have $P(Y = -2) = P(Y = 0) = P(Y = 2) = 1/3$. Then $\sigma_Y = \sqrt{8/3}$, twice as big as $\sigma_X = \sqrt{2/3}$. Alternatively, we could “spread out” $X$ by defining $Z$ to have the same $v_j$ values as $X$, but with even more probability on the extreme values. Specifically, $P(Z = -1) = P(Z = 1) = 1/2$. Then, $\sigma_Z = 1$, bigger than $\sigma_X = \sqrt{2/3}$.

Like the mean, the standard deviation is sensitive to very large values. For example, let $P(Y = 0) = 0.98$, $P(Y = -100) = P(Y = 100) = 0.01$. Then, even though 98% of the population has $Y = 0$ (very concentrated, not spread out), $\sigma_Y = \sqrt{200} \approx 14.1$.

### Other Summary Features

Although beyond the scope of this text, I must briefly mention that percentiles (or quantiles) are also helpful summary features. They can capture aspects of a probability distribution that the mean and standard deviation do not. For example, the median captures the value of the “average member of the population” discussed above. High and low percentiles help capture the “tails” of a distribution, like people with the very highest (or lowest) income. Measures like the interquartile range capture the “spread” in a way that is not sensitive to very large outliers, complementing the standard deviation.
Quantile regression extends percentiles to regression parallel to what we’ll study with the mean.

### 2.3.4 Categorical or Ordinal Variable

A binary variable is usually a special case of a **categorical variable** whose possible values are “categories,” not numbers. This was true of most of the examples in Section 2.3.2, like whether a retailer is a franchise or not, or whether a pharmaceutical drug is branded or generic. Such values can be coded as 0 or 1 for convenience, but they lack any numeric meaning.

Categorical variables can have more than two possible values. For example, non-franchise retailers could be categorized further as national chain, regional chain, or independent. Other categorical variable examples include:

- **geographics region** (north, south, east, west)
- **mode of transportation** (car, bike, train, etc.)
- **industry** (like NAICS)
- **college major** (economics, English, ecology, electrical engineering, etc.)

The previous examples’ categories have no particular order to them, so they constitute **nominal variables** (or **nominal categorical variables**). Sometimes these are simply called categorical variables.

In contrast, there could be an **ordinal variable** (or **ordinal categorical variable**). An ordinal variable’s possible values have a natural order, usually from “low” to “high.” For example:

- **bond rating** (e.g., D, C, . . ., AA+, AAA)
- **self-reported health status** (poor, fair, good, excellent)
- **teaching evaluation responses** (disagree, neutral, agree)
- **letter grades** (F, C, B, A; although often A and C are 4.0 and 2.0, there is nothing intrinsic in the letter grade system that suggests A is exactly twice as good as C)

Some categorical variables are not clearly nominal or ordinal. For example, consider educational degree. Some degrees are higher than others (e.g., you need a bachelor’s degree before you get a master’s), but others are not (e.g., PhD and MD). As another example, consider sex. Neither male nor female is “higher” than the other, but arguably intersex is in between. As another example, some occupations are ordered (e.g., within the same consulting firm, junior analyst is lower than senior analyst, which is lower than associate, then manager, etc.), but others are not (e.g., painter, chef, carpenter). Such variables require careful thought but are beyond our scope.

In any case, categorical variables are often represented by dummy variables (binary variables). For example, consider the teaching evaluation response whose possible values are disagree, neutral, and agree. Using the indicator function from (2.3), we can define \( W = 1\{\text{disagree}\} \),
Then \( P(W = 1) = E(W), \) \( P(X = 1) = E(X), \) and \( P(Y = 1) = E(Y). \) In a way, \( Y \) is redundant since \( Y = 1 \) if \( W = X = 0 \), completely determined by \( W \) and \( X \).

### Probability Mass Function

A categorical or ordinal PMF is essentially the same as a discrete PMF. As before, it is usually written like \( f_Y(\cdot) \) for the PMF of \( Y \). The PMF’s input is again a possible value, and it’s output is the corresponding probability. The only difference is that the possible values are categories instead of numbers.

For example, consider the coin flip. Random variable \( W \) has two possible values: \( h \) (heads) or \( t \) (tails); mathematically, \( W \in \{h, t\} \). Its PMF is \( f_W(w) \equiv P(W = w) \). That is, \( f_W(h) \) is the probability of heads, and \( f_W(t) \) is the probability of tails.

### Cumulative Distribution Function

A nominal categorical variable does not have a CDF. Recall from (2.2) that the CDF of \( Y \) evaluated at value \( v \) is \( F_Y(v) \equiv P(Y \leq v) \). (It does not matter which lowercase letter we use; whether \( y \) or \( r \) or \( v \), it simply represents a non-random value.) If \( Y \) is nominal, then the inequality relationship \( \leq \) has no meaning. For example, we cannot evaluate if the value “painter” is less than or equal to “chef.” The PMF is still well-defined because it relies only on equality, and we can evaluate if “painter” and “chef” are equal. But the “cumulative” part of CDF has no meaning for nominal categorical variables.

In contrast, an ordinal variable can have a CDF. The values are ordered, so we can evaluate if one value is \( \leq \) another. (If there is a clear order but not a clear “low” and “high” ends, then one end can arbitrarily be picked as “low,” and the CDF shows the cumulative probability from the “low” end through a given category.)

For example, consider self-reported health status \( Y \). Any two values can be compared with \( \leq \): poor \( \leq \) good, good \( \leq \) excellent, etc. The CDF evaluated at “good” is the probability of health that is good or worse. There are three such possible values: poor, fair, and good. Thus, the CDF evaluated at good is the probability of poor, fair, or good health. Mathematically,

\[
F_Y(\text{good}) \equiv P(Y \leq \text{good}) = P(Y = \text{poor}) + P(Y = \text{fair}) + P(Y = \text{good})
\]

\[
= f_Y(\text{poor}) + f_Y(\text{fair}) + f_Y(\text{good}).
\]

### Summary Features: Mean and Standard Deviation

Categories cannot be summed or averaged, so categorical variables do not have a mean. You could arbitrarily assign numeric values to each
2.3. DESCRIPTION OF A POPULATION

category and then use the discrete variable formula in (2.16), but if I assigned different numeric values then I could get a very different result than you. Fundamentally, we cannot average “painter” and “chef,” or average “poor” and “good.” (E.g., what is poor plus good? What is 0.17 times poor?)

For the same reason, categorical variables do not have a standard deviation.

Other Summary Features

The **mode** is often useful for summarizing categorical variables. The mode is the single most likely value. Mathematically,

\[
\text{mode}(Y) \equiv \arg \max_y P(Y = y),
\]

which is read as “the value of \( y \) that maximizes \( P(Y = y) \).”

Nominal categorical variables do not have anything like a mean that can give a sense of whether values are generally high or low, because there is no sense of high or low for nominal variables. Similarly, there is no sense of “close” or “far,” so it is meaningless to ask how “spread out” a nominal distribution is. Only features of the PMF can be summarized, without accounting for the values themselves. So, we could measure whether the probability is spread more evenly across categories rather than very high in some categories (but not others), but not much else.

Ordinal variables do have a sense of high or low, so it is possible to summarize ordinal distributions analogous to a mean (overall high or low) or standard deviation (how spread out).

The median (and other percentiles) can summarize how generally high or low an ordinal distribution is. The median is the category for the “average member” of the population. That is, at least half the population has the same or lower value, and at least half the population has the same or higher value. Mathematically, the median of ordinal random variable \( Y \) is the value \( m \) such that

\[
P(Y \leq m) \geq 0.5 \quad \text{and} \quad P(Y \geq m) \geq 0.5.
\]

For example, consider self-reported health status \( Y \). The five possible values are poor, fair, good, great, and excellent. Imagine there is 1/5 probability of each value, i.e., \( f_Y(y) = 1/5 \) for any of the five possible \( y \) values. Because each category has the same probability, naturally the median is the middle category, “good.” Mathematically, to verify,

\[
P(Y \leq \text{good}) = \left(\frac{1}{5}\right) \left(\frac{1}{5}\right) + \frac{1}{5} = \frac{3}{5} > 1/2,
\]

\[
P(Y \geq \text{good}) = \frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{3}{5} > 1/2.
\]

Imagine a different population represented by random variable \( W \) that is generally healthier. To make it obvious, imagine nobody has poor
or fair health, and the other three categories have 1/3 probability each. Now “great” is the median, reflecting that $W$ represents a generally healthier population than $Y$, whose median was only “good.” Mathematically, to verify,

$$
P(W \leq \text{great}) = f_Y(\text{poor}) + f_Y(\text{fair}) + f_Y(\text{good}) + f_Y(\text{great}) = 2/3 > 1/2
$$

$$
P(W \geq \text{great}) = f_Y(\text{great}) + f_Y(\text{excellent}) = 2/3 > 1/2.
$$

(2.26)

There are other ways to compare whether one ordinal distribution is “higher” or “more spread out” than another (e.g., Kaplan and Zhuo, 2019), but they are beyond our scope.

2.3.5 Continuous Variable

A continuous variable differs from a discrete variable in some strange technical ways, but the intuition is the same. This textbook often uses discrete variables to build intuition since the math is simpler. You could imagine a continuous variable like a discrete variable with a very large number of possible values packed very tightly together. Indeed, many variables typically called “continuous” are actually discrete, like monetary values (like annual income or sales) that come in discrete units (like $0.01). Practically, the difference is negligible. Examples of other variables modeled as continuous are:

- market concentration measures (like market share of largest firm or HHI)
- a country’s per capita annual meat consumption
- percentage growth of GDP (or sales, or stock price, etc.)
- crime rates (e.g., a city’s number of property crimes per year per 10,000 people)

Always specify units of measure. For example, if $Y$ is the distance from an individual’s residence to their workplace, it is meaningless to say $Y = 15$ because 15 is just a number, not a measure of distance. It could be 15 km, but it could also be 15 mi, which is 24 km; or it could even be measured in meters or feet (or parsecs, though unlikely). The mean, standard deviation, median (and other percentiles), and interquartile range all share the same units as the variable itself, whereas the variance has squared units (which is harder to interpret: e.g., squared dollars?). Units always matter greatly, whether for description, prediction, or causality.

Probability Density Function

A truly continuous random variable does not have a PMF. It has an “uncountably infinite” number of possible values, implying each has zero probability. That is, if $Y$ is the random variable, then for any possible value $y$, $P(Y = y) = 0$. There is a difference between
2.3. DESCRIPTION OF A POPULATION

“possible” and “non-zero probability.” Every observed realization $y$ is clearly possible, yet had zero probability of occurring.

Although individual values have zero probability, ranges of values have non-zero probability. For example, even if $P(Y = y) = 0$ for each individual $0 \leq y \leq 1$, it’s possible that $P(0 \leq Y \leq 1) = 0.34$.

The probability density function (PDF) helps us see such probabilities for different ranges of value. The “bell curve” is an example of a PDF. Generally, the PDF is higher around more probable values.

If the CDF is also differentiable, the derivative is called a probability density function (PDF). PDFs are commonly denoted with lowercase $f$, sometimes with a subscript like $f_Y(\cdot)$ for the PDF of random variable $Y$. Similar to a histogram, a PDF shows the probability of a random variable taking a value in a certain interval, as the area under the PDF. Since $P(−\infty < Y < \infty) = 1$ for any random variable, the total area under any PDF must equal one.

Figure 2.3 shows an example PDF. No matter how big or small we draw it, the total area between the horizontal axis and the PDF is defined to be 1. The shaded area under the PDF between $y = 0$ and $y = 1$ shows $P(0 \leq Y \leq 1)$. That is, $P(0 \leq Y \leq 1)$ is the proportion of the total area under the PDF that is shaded. The fact that $P(Y = y) = 0$ for any $y$ can also be seen. For example, $P(Y = 0)$ is the “area” under the PDF between $y = 0$ and $y = 0$, but this is just a line, which has zero area, hence zero probability.

![Figure 2.3](image)

Cumulative Distribution Function

The CDF of a continuous random variable has the same definition as for a discrete or ordinal variable: $F_Y(y) \equiv P(Y \leq y)$.

Unlike the jumpy stair-step CDF of a discrete random variable, a continuous random variable’s CDF is a continuous function. (And if you know calculus: when it exists, the PDF is the first derivative of the CDF.)
CHAPTER 2. ONE VARIABLE: POPULATION

Summary Feature: Mean

The intuition for the mean and the linearity of the expectation operator apply equally to continuous random variables.

Computing the mean of a continuous random variable by hand requires calculus, so it is not covered in this textbook. (If you happen to know calculus and are curious: it’s extending the idea of probability-weighted average by replacing the sum with an integral; e.g., if the PDF $f_Y(\cdot)$ exists, $E(Y) = \int_{\mathbb{R}} y f_Y(y) dy$, analogous to $\sum_{j=1}^{J} y_j f_Y(y_j)$.)

Summary Feature: Standard Deviation

The intuition for the standard deviation is also the same for continuous distributions, but again requires calculus (integration) to compute.

Other Summary Features

The median and other percentiles complement the mean and standard deviation in summarizing continuous distributions. For example, if you’ve ever taken a standardized test before (like SAT, ACT, or GRE), your score report probably included your percentile. Your score percentile is the proportion of the population you scored better than; e.g., if you were in the 90th percentile, you scored better than 90% of other students. Despite their utility, percentiles (quantiles) are beyond our scope, but I hope you study econometrics further to learn more about them.

The Normal (Gaussian) Distribution

One particular distribution appears frequently in statistics and econometrics: the normal distribution (or Gaussian distribution). Without getting too detailed, some comments may help, especially when you read other books. (If you want, you can find plenty of details on Wikipedia.)

Although some variables are indeed approximately normally distributed, this is not “normal” in the common English sense: most variables are not normal. To start, any discrete or categorical random variable cannot be normal (Gaussian). Most continuous variables are also not normal.

None of the methods in this textbook require variables to be Gaussian. Historically, sometimes normality was assumed for certain results, but it is not necessary. You don’t need to worry about testing whether or not variables are normal. It doesn’t matter.

Normal distributions are convenient for educational examples because the mean and standard deviation uniquely characterize a normal distribution. Notationally, the normal distribution is written $N(\mu, \sigma^2)$ for a normal distribution with mean $\mu$ and variance $\sigma^2$, or sometimes equivalently $N(\mu, \sigma)$ with standard deviation $\sigma = \sqrt{\sigma^2}$. That is, if $Y \sim N(\mu, \sigma^2)$, then $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2$, so its standard deviation is $\sigma$. This is convenient for illustrative examples because, as
we’ve seen, the mean gives a general sense of values being high or low, and the standard deviation describes how spread out the distribution is. With \( \mu = 0 \) and \( \sigma = 1 \), \( N(0,1) \) is called the standard normal distribution.

However, this is not true of other distributions, so in general the mean and standard deviation do not fully summarize a distribution. For example, there could be two random variables with mean 0 and standard deviation 1 that are very different. One such random variable is \( W \) with \( P(W = -1) = P(W = 1) = 1/2 \). Another is the standard normal \( Z \sim N(0,1) \). If you only know the mean and standard deviation, then you cannot tell the difference between \( W \) and \( Z \). These are very different, and there are many other random variables with the same mean and standard deviation.

### 2.4 Prelude to Prediction: Precipitation

This section introduces prediction concepts through a simple example to develop intuition. The two main goals are 1) to show that there is no single best prediction because “best” depends on the ultimate purpose of the prediction, and 2) to begin translating intuition into formal mathematics. Further mathematical formalization and more complex examples are in Section 2.5.

Notationally (details below), \( g \) is your non-random guess of the realized value \( y \) of random variable \( Y \), and \( L(y,g) \) quantifies how bad it is to have guessed \( g \) when the realized value is \( y \). (Hypothetically, you could randomize your guess, but this is never optimal, so it is not considered.)

Throughout this section, imagine you want to predict whether or not it will rain tomorrow. Mathematically, random variable \( Y \) represents tomorrow’s weather: \( Y = 1 \) if it rains tomorrow, and \( Y = 0 \) if not. Assume you actually know the probability distribution of \( Y \) (you do not need to estimate it from data). Since \( Y \) is binary, \( Y \sim \text{Bernoulli}(p) \), so knowing the distribution is equivalent to knowing \( p = P(Y = 1) \), the probability that it rains. (This is usually what weather forecasts report, so you could actually do the following examples in real life.) Thus, equivalently: given the probability of rain, you want to predict the realized value (yes or no).

The distribution of \( Y \) alone is not enough to make a good prediction: you also need to know the consequences of correct and incorrect predictions in each case. Intuitively: if one outcome is really really bad, then you should prefer to avoid even a small risk of it, compared to a larger risk of a not-so-bad outcome.

Mathematically, consequences are formalized as a loss function. The loss function \( L(y,g) \) specifies how bad it is to have guessed \( g \) when the realized value is \( y \). (Other sources may switch the order of \( y \) and \( g \), so be careful.) In the rain example, \( L(0,1) \) represents how bad it is to guess rain when it does not rain. This may be different than \( L(1,0) \), how bad it is to guess no rain when in fact it rains. For
example, $L(0, 1) = 20$ and $L(1, 0) = 100$ means it is much worse to be wrong when it rains ($y = 1$) than when it doesn’t ($y = 0$).

It can be confusing to define “loss” when you’re correct ($g = y$). If something actually good happens, it can be represented by negative loss, like $L(0, 0) = -10$. If $L(1, 1) = -30$ (even more negative than $-10$), then it’s even better to guess right when it rains.

Even if there are good outcomes, loss values can be normalized to be non-negative without changing the best prediction. In the rain example, imagine the most negative loss possible is $L(1, 1) = -30$. Then, simply adding 30 to each loss makes them all non-negative without changing their relative values. This essentially sets $L(1, 1) = 0$ as the reference point, and the loss function’s interpretation is, “How much worse is this situation than $y, g = (1, 1)$?” It’s also possible to normalize $L(v, v) = 0$ for any $v$, meaning that guessing $g = y$ is not bad at all, by subtracting the original $L(v, v)$ from the original $L(v, g)$ for all $g$. Understanding the detailed reasoning is beyond our scope, but just be aware that $L(y, g)$ is not necessarily an absolute “how bad is $(y, g)$” but can also be “how bad relative to $(y, y)$” or “how bad relative to the best possible $(v, v)$.”

If you are more familiar with utility functions from economics, you can think of the loss function as essentially a negative utility function. Apparently economists are optimistic, modeling how good things are (utility), whereas statisticians are pessimistic, modeling how bad things are (loss). If you had a utility function $u(y, g)$ that says how good it is to have guessed $g$ when the truth is $y$, then you can just define $L(y, g) = -u(y, g)$.

Throughout this section, the consequences are the results of a bet with your friend. If you guess wrong, then you lose some money (positive loss). If you guess right, then you win some money (negative loss).

2.4.1 Easy: “Predict” Current Weather

Let’s start easy: you’re standing outside, and you want to predict whether or not it’s currently raining. Since you can observe this directly, this is like the “after” view of Section 2.1.1: instead of multiple possible values of random variable $Y$, you see the realized value $y$ with no uncertainty.

You make a simple $1 bet: if you guess right ($g = y = 0$ or $g = y = 1$) then you win $1, but if you guess wrong ($g \neq y$) then you lose $1. Recall that negative loss means winning. Formally, $L(y, g)$ is

$$L(0, 0) = L(1, 1) = -1, \quad L(0, 1) = L(1, 0) = 1 \implies L(y, g) = 2\cdot 1\{y \neq g\}.$$  (2.27)

If you remember your microeconomics classes, you may realize that the loss function in (2.27) implicitly assumes a linear utility function $u(x) = x$ for simplicity. That is, if you currently have $sx$ and you win $1, then your utility increases by $u(x + 1) - u(x)$. If you lose $1, then your utility decreases from $u(x)$ to $u(x - 1)$. If you are risk-averse, $u(\cdot)$ is concave, so even though the dollar amount
2.4. PRELUDE TO PREDICTION: PRECIPITATION

is the same, the potential utility increase is smaller than the potential utility decrease: $u(x + 1) - u(x) < u(x) - u(x - 1)$. Generally, your loss function should be $L(0, 0) = L(1, 1) = u(x) - u(x + 1)$ and $L(0, 1) = L(1, 0) = u(x) - u(x - 1)$. For simplicity, plugging in $u(x) = x$ yields $u(x) - u(x + 1) = x - (x + 1) = -1$ and $u(x) - u(x - 1) = x - (x - 1) = 1$, as in (2.27).

Obviously, you guess $g = y$. You are correct. You win $1$.

Mathematically, how can this intuition be formalized? If you know $y$, then you can compute both $L(y, 0)$ and $L(y, 1)$. If $y = 1$, so $L(y, 0) = 1$ and $L(y, 1) = -1$, then “guessing” $g = 1$ minimizes loss since $-1 < 1$. If $y = 0$, so $L(y, 0) = -1$ and $L(y, 1) = 1$, then “guessing” $g = 0$ minimizes loss since $-1 < 1$. Thus, the best “guess” is indeed $g = y$.

2.4.2 Minimizing Mean Loss

With the same loss function from (2.27), consider predicting tomorrow’s weather if $P(Y = 1) = 0.4$. That is, there’s a 40% probability of rain tomorrow (and 60% chance of no rain); for the purpose of your bet, should you predict rain?

Now we need some way to deal with uncertainty. Regardless of guessing $g = 0$ or $g = 1$, there is some chance of being right and some chance of being wrong.

In microeconomics, the typical approach is to choose the action that maximizes mean utility. The same could be done here. Equivalently, since the loss function is essentially a negative utility function, we could minimize mean loss. This doesn’t guarantee you’ll win the bet every time, but over the long-run (if you bet many times), it leads to the lowest total loss.

Mean loss is more commonly called expected loss, but this can be confusing. Again, “expected” is technical jargon that is unrelated to what “expected” means in colloquial English. Below, it is in fact impossible to actually receive the “expected” loss since it is a decimal value (whereas you can only win or lose $1$).

Mean loss is also sometimes called risk. Again, this has a precise technical meaning, but it is probably not how you would define “risk” colloquially.

Given the distribution of $Y$ and the loss function in (2.27), how can we pick $g$ to minimize mean loss? There are two values of $g$ to consider: $g = 0$ or $g = 1$. (Other values are allowed but would always be wrong, so you’d always lose; e.g., with $g = 0.4$, $g \neq y$ for both $y = 0$ and $y = 1$.) Given a particular guess like $g = 0$, the loss still depends on $y$, which has multiple possible values. Thus, the loss has multiple possible values. That is, given $g = 0$, the loss is a random variable. We can derive its distribution from the distribution of $Y$. Then we can compute the mean of the loss distribution. In this way, we can compute mean loss for each possible $g$. Finally, the best guess is the $g$ with the smallest mean loss. Details on these steps are given below.
To work through these steps mathematically, consider the loss when \( g = 0 \), and separately the loss when \( g = 1 \). Mathematically, define random variables

\[
L_0 \equiv L(Y, 0), \quad L_1 \equiv L(Y, 1),
\]

respectively representing the (distribution of) loss when \( g = 0 \) and when \( g = 1 \). When \( g = 0 \), the loss \( L(Y, 0) \) is either \( L(0, 0) = -1 \) if \( Y = 0 \) or else \( L(1, 0) = 1 \) if \( Y = 1 \). That is, \( L_0 = -1 \) when \( Y = 0 \), and \( L_0 = 1 \) when \( Y = 1 \). We know \( P(Y = 1) = 0.4 \), so

\[
P(L_0 = 1) = P(Y = 1) = 0.4, \quad P(L_0 = -1) = P(Y = 0) = 1 - P(Y = 1) = 0.6.
\]

(2.29)

Similarly, if you guess \( g = 1 \), then the loss \( L(Y, 1) \) is \( L(1, 1) = -1 \) when \( Y = 1 \), or else \( L(0, 1) = 1 \) when \( Y = 0 \), so

\[
P(L_1 = 1) = P(Y = 0) = 0.6, \quad P(L_1 = -1) = P(Y = 1) = 0.4.
\]

(2.30)

Given the loss distributions in (2.29) and (2.30), mean loss can be computed for each possible \( g \). If you guess \( g = 0 \), then using (2.16) and (2.29), mean loss (in $) is

\[
E(L_0) = (0.4)(1) + (0.6)(-1) = -0.2.
\]

(2.31)

If you instead guess \( g = 1 \), then using (2.16) and (2.30), mean loss (in $) is

\[
E(L_1) = (0.6)(1) + (0.4)(-1) = 0.2.
\]

(2.32)

The best prediction for your bet is the \( g \) that minimizes mean loss. Mathematically, \( E(L_0) < E(L_1) \); equivalently, using (2.28), \( E[L(Y, 0)] < E[L(Y, 1)] \). That is, \( g = 0 \) generates the smallest mean loss, so \( g = 0 \) is the best prediction to make for your bet. Even though it’s the best prediction, you’ll still be wrong and lose $1 if it rains tomorrow, which has a 40% probability of happening. But in the long run, you’ll win money if you always predict \( g = 0 \), whereas you’ll lose money if you always predict \( g = 1 \).

2.4.3 Different Probability

Imagine the same setup as in Section 2.4.2 but with a different distribution of \( Y \): \( P(Y = 1) = 0.7 \). Intuitively, rain being more likely might change the optimal prediction from \( g = 0 \) to \( g = 1 \). Mathematically, this intuition proves correct.

Following the same steps from Section 2.4.2, first compute the distribution of loss separately for \( g = 0 \) and \( g = 1 \). Define \( L_0 \) and \( L_1 \) as in (2.28). Parallel to (2.29) and (2.30) but with \( P(Y = 1) = 0.7 \),

\[
P(L_0 = 1) = P(Y = 1) = 0.7, \quad P(L_0 = -1) = P(Y = 0) = 1 - P(Y = 1) = 0.3,
\]

\[
P(L_1 = 1) = P(Y = 0) = 0.3, \quad P(L_1 = -1) = P(Y = 1) = 0.7.
\]

(2.33)
2.5. PREDICTION WITH A KNOWN DISTRIBUTION

Parallel to (2.31) and (2.32), using (2.33),
\[
\begin{align*}
\mathbb{E}[L(Y, 0)] &= \mathbb{E}(L_0) = (0.7)(1) + (0.3)(-1) = 0.4, \\
\mathbb{E}[L(Y, 1)] &= \mathbb{E}(L_1) = (0.3)(1) + (0.7)(-1) = -0.4.
\end{align*}
\]

(2.34)

Since \(\mathbb{E}[L(Y, 1)] < \mathbb{E}[L(Y, 0)]\), \(g = 1\) minimizes mean loss, so \(g = 1\) is the best prediction for your bet.

2.4.4 Different Loss Function

Now imagine the original setup in Section 2.4.2 but with a different loss function. Specifically, if you correctly predict rain, you win $10 (i.e., \(-10\) loss), but otherwise \(L(y, g)\) is the same as in (2.27):
\[
L(0, 0) = -1, \quad L(1, 1) = -10, \quad L(0, 1) = L(1, 0) = 1
\]

(2.35)

Intuitively, even though rain is less probable than no rain, the much larger payoff for correctly predicting rain might make us bet on rain. Mathematically, this intuition proves correct.

Following the same steps from Section 2.4.2, first compute the distribution of loss separately for \(g = 0\) and \(g = 1\). Define \(L_0\) and \(L_1\) as in (2.28). Parallel to (2.29) and (2.30) but with \(L(1, 1) = -10\),
\[
\begin{align*}
P(L_0 = 1) &= P(Y = 1) = 0.4, & P(L_0 = -1) &= P(Y = 0) = 0.6, \\
P(L_1 = 1) &= P(Y = 0) = 0.6, & P(L_1 = -10) &= P(Y = 1) = 0.4.
\end{align*}
\]

(2.36)

Parallel to (2.31) and (2.32), using (2.36),
\[
\begin{align*}
\mathbb{E}[L(Y, 0)] &= \mathbb{E}(L_0) = (0.4)(1) + (0.6)(-1) = -0.2, \\
\mathbb{E}[L(Y, 1)] &= \mathbb{E}(L_1) = (0.6)(1) + (0.4)(-10) = -3.4.
\end{align*}
\]

(2.37)

Since \(\mathbb{E}[L(Y, 1)] < \mathbb{E}[L(Y, 0)]\), \(g = 1\) minimizes mean loss, so \(g = 1\) is the best prediction for your bet. (As noted earlier, ideally you’d also allow for a utility function with some risk aversion, in which case the best prediction may additionally depend on the degree of risk aversion.)

2.5 Prediction with a Known Distribution

⇒ Kaplan video: Optimal Prediction

What does prediction mean? It may seem surprising to discuss prediction without any data, and with a completely known distribution. In English, usually prediction means using what you know now to “predict” what will happen in the future (e.g., “Beware the Ides of March!”). In econometrics and statistics, prediction shares the qualities of guessing something unknown using something known, but the details differ. (Predicting the future is a special case of prediction called forecasting; see Part III.)

Here, as in Section 2.4, the goal is to predict the value of a random draw from a known distribution. The distribution summarizes
“what you know”: different possible values and their probabilities. As in Section 2.1, the random draw need not occur in the future; indeed, it may have already happened, but we haven’t observed it yet. So, besides applications like predicting ridesharing demand tomorrow, “prediction” also includes guessing the income of a customer standing right in front of you (who hasn’t told you their income yet).

This section extends Section 2.4, including more complex examples. Further, close connections with description (Section 2.3) are shown.

Understanding the role of the loss function is particularly crucial. Even if you do some fancy machine learning prediction with cross-validation, you need to consider the loss function carefully. Just using whatever you find online may be inappropriate. I have seen PhD students puzzled by their results because they did not use an appropriate loss function. Loss functions are also central to Bayesian prediction, although that is beyond our scope.

2.5.1 Common Loss Functions

Ideally, a loss function reflects the real-world consequences of guessing \( g \) when the realized value is \( y \), like in Section 2.4. In practice, this may be infeasible. For example, maybe the consequences are not easily quantified. Or maybe you have to make a single prediction that will be used for multiple decisions with different consequences, or for multiple people whose utility functions differ. Or maybe you have a deadline and simply don’t have time to carefully construct a loss function.

There are infinitely many possible loss functions, but those below are used more commonly than others.

0–1 Loss

Define 0–1 loss as

\[
L_0(y, g) \equiv 1\{y \neq g\}. \tag{2.38}
\]

This equals 0 (which is good) if you guess correctly, and 1 (bad) if not. This reflects a case where it doesn’t matter “how wrong” you are; it only matters whether you’re right or wrong.

This 0–1 loss is often used when \( Y \) is a nominal categorical variable (Section 2.3.4) or binary. For example, if you need to predict occupation, and the true \( y \) is “painter,” it’s probably no worse to have incorrectly guessed “zookeeper” than “bookkeeper”; they’re just both wrong. This is not true of ordinal categorical variables; e.g., if somebody’s health is “excellent,” it’s probably worse (higher loss) to have predicted “poor” than “great.”

Although not obvious, 0–1 loss determines the same optimal prediction as the rain bet loss function in (2.27). (Although coded as binary, the rain \( Y \) could be considered nominal categorical with possible values “rain” and “no rain.”) The optimal prediction does not change if you add 1 to all losses. It also does not change if you divide
all losses by 2. Thus, given $L(y, g)$ in (2.27), it is equivalent to use the loss function $[L(y, g) + 1]/2 = I\{y \neq g\}$, which is 0–1 loss as in (2.38).

### Quadratic Loss

Define **quadratic loss** (or **squared loss** or **squared error loss** or **$L_2$ loss**) as

$$L_2(y, g) = (y - g)^2.$$  \hspace{1cm} (2.39)

This is zero when the guess is perfect ($g = y$) and larger when $g$ is farther from $y$, in either direction (higher or lower). Thus, unlike 0–1 loss, quadratic loss can differentiate between a slightly-wrong guess and a really-wrong guess.

For example, let the true $y = 100$. The guess $g = y = 100$ is best since $L_2(100, 100) = 0$ is the smallest possible loss. (Squaring produces non-negative numbers.) The guess $g = 99$ is worse: loss is $L_2(100, 99) = (100 - 99)^2 = 1$. The guess 90 is even worse since $L_2(100, 90) = (100 - 90)^2 = 100$. In fact, even though 90 is only 10 times farther from $y$ than 99 is, the loss is 100 times as big: much, much worse. The guess 110 is just as bad as 90 since they are both wrong by 10 (higher or lower doesn’t matter): $L_2(100, 110) = (100 - 110)^2 = 100$.

Quadratic loss is often used for discrete or continuous variables (Sections 2.3.3 and 2.3.5). Although ordinal loss functions should also differentiate between slightly-wrong and really-wrong, quadratic loss cannot be used because you cannot subtract two category values, or square them, as required by (2.39); e.g., you cannot square the difference between “excellent” and “great.”

Despite its common use, quadratic loss is not always sensible. For example, sometimes it may be much worse to over-predict ($g > y$) than under-predict ($g < y$), or vice-versa. Quadratic loss does not differentiate between over-prediction and under-prediction because $(y - g)^2 = (g - y)^2$; only the absolute error $|y - g|$ matters, not whether it’s positive or negative. For example, if $y = 100$, it may be much worse to predict $g = 110$ than $g = 90$, but $L_2(100, 110) = (100 - 110)^2 = 100$ is the same as $L_2(100, 90) = (100 - 90)^2 = 100$. As another example, sometimes it may be twice as bad to over-predict by 20 units ($g - y = 20$) than 10 units ($g - y = 10$), but quadratic loss treats it as four times worse since $10^2 = 100$ but $20^2 = 400$.

Nonetheless, quadratic loss is usually not crazy, especially if we need to make a prediction but don’t know how it will be used in decision-making.

### Other Loss Functions

Although beyond our scope, if you’re curious: there are other interesting, common loss functions. One is absolute loss (or $L_1$ loss), $L_1(y, g) = |y - g|$. Variations of absolute loss include asymmetric versions for which over-prediction is worse than under-prediction (or
vice-versa), as well as absolute percentage loss \(|(y - g)/y|\). Another common loss function is weighted 0–1 loss, which generalizes 0–1 loss by allowing \(L(0, 1) \neq L(1, 0)\). (This is related to hypothesis testing, where there are different consequences for rejecting a true hypothesis than not rejecting a false hypothesis.)

2.5.2 Optimal Prediction: Generic Examples

In Sum: Optimal Prediction

1. Choose appropriate loss function \(L(y, g)\): quantifies how bad it is to guess \(g\) when the true value is \(y\)
2. Optimal prediction: the value of \(g\) with smallest mean loss \(E[L(Y, g)]\)

The following generic examples show the procedure to find the prediction \(g\) that is “optimal” in the sense of minimizing mean loss. The procedure does not allow continuous \(Y\) since that would require calculus, but it works for any other variable type from Section 2.3.

Two Possible Values

This is the same procedure used in the rain example in Section 2.4.

Step 1 is to write out the loss function values for all combinations of \((y, g)\), where \(g\) is our guess and \(y\) is the true realized value. Assume two possible values: \(y = a\) or \(y = b\). (Possibly \(a = 0\) and \(b = 1\), but they could have any value, including non-numeric values like “cat” and “dog.”) Assume you have to guess either \(a\) or \(b\): \(g = a\) or \(g = b\). Thus, there are four possible combinations of \((y, g)\): \((a, a)\), \((b, b)\), \((a, b)\), and \((b, a)\). The four corresponding loss function values can be arranged in a matrix, where each column has the same \(y\) value and each row has the same \(g\) value:

\[
\begin{pmatrix}
L(a, a) & L(b, a) \\
L(a, b) & L(b, b)
\end{pmatrix}
= \begin{pmatrix}
L(y = a, g = a) & L(y = b, g = a) \\
L(y = a, g = b) & L(y = b, g = b)
\end{pmatrix}.
\tag{2.40}
\]

(Although we won’t use it, if you’re handy with matrix algebra, you’ll see that the vector of mean losses can be written as the product of...
the matrix in (2.40) with the column vector of probabilities \( P(Y = a) \) and \( P(Y = b) \); i.e., each mean loss is the dot product of a row in the loss matrix with the probability vector.)

Step 2 could also be interpreted in terms of two new random variables, like \( L_0 \) and \( L_1 \) in Section 2.4. Let \( L_a \equiv L(Y, a) \) be a random variable representing loss when \( g = a \). The distribution of \( L_a \) is \( P(L_a = L(a, a)) = P(Y = a) \) and \( P(L_a = L(b, a)) = P(Y = b) \). Similarly, let \( L_b \equiv L(Y, b) \) be a random variable representing loss when \( g = b \), with \( P(L_b = L(a, b)) = P(Y = a) \) and \( P(L_b = L(b, b)) = P(Y = b) \). Thus, yielding the same results as (2.41),

\[
\begin{align*}
E[L(Y, a)] &= E[L_a] = P(Y = a)L(a, a) + P(Y = b)L(b, a), \\
E[L(Y, b)] &= E[L_b] = P(Y = a)L(a, b) + P(Y = b)L(b, b). \\
\end{align*}
\]

Step 3 is to find the \( g \) that minimizes \( E[L(Y, g)] \). The \( g \) that minimizes mean loss is the optimal predictor. That is, if \( E[L(Y, a)] < E[L(Y, b)] \), then \( g = a \) is the optimal predictor; if \( E[L(Y, b)] < E[L(Y, a)] \), then \( g = b \) is the optimal predictor; or if \( E[L(Y, a)] = E[L(Y, b)] \), then \( g = a \) and \( g = b \) are equally good (or equally bad).

For example, let

\[
\begin{pmatrix}
L(a, a) & L(b, a) \\
L(a, b) & L(b, b)
\end{pmatrix}
= \\
\begin{pmatrix}
0 & 7 \\
5 & 0
\end{pmatrix},
\]

\( P(Y = a) = 0.7 \), \( P(Y = b) = 0.3 \).

Using (2.41),

\[
\begin{align*}
E[L(Y, a)] &= P(Y = a)L(a, a) + P(Y = b)L(b, a) = (0.7)(0) + (0.3)(7) = 2.1, \\
E[L(Y, b)] &= P(Y = a)L(a, b) + P(Y = b)L(b, b) = (0.7)(5) + (0.3)(0) = 3.5.
\end{align*}
\]

Since \( E[L(Y, a)] < E[L(Y, b)] \), the predictor \( g = a \) is better than \( g = b \) according to mean loss with this particular loss function.

**Many Possible Values**

Now let \( Y \) take \( J \) different possible values. Label these \( v_1, v_2, \ldots, v_J \). For example, if \( J = 3 \), we could have \( v_1 = 10, v_2 = 20, \) and \( v_3 = 22 \); or \( v_1 \) could be “cat,” \( v_2 \) “dog,” and \( v_3 \) “echidna.” Like before, assume \( g \) must take one of these same values.

The three steps are the same as with \( J = 2 \), just with more values to handle.

First, write out the loss function values in a matrix with the row \( m \), column \( k \) entry equal to \( L(Y = v_k, g = v_m) \):

\[
\begin{pmatrix}
L(v_1, v_1) & L(v_2, v_1) & \cdots & L(v_J, v_1) \\
L(v_1, v_2) & L(v_2, v_2) & \cdots & L(v_J, v_2) \\
\vdots & \vdots & \ddots & \vdots \\
L(v_1, v_J) & L(v_2, v_J) & \cdots & L(v_J, v_J)
\end{pmatrix}.
\]
Second, compute all the mean losses:

\[
E[L(Y, v_1)] = P(Y = v_1)L(v_1, v_1) + P(Y = v_2)L(v_2, v_1) + \cdots = \sum_{j=1}^{J} P(Y = v_j)L(v_j, v_1)
\]

\[
E[L(Y, v_2)] = P(Y = v_1)L(v_1, v_2) + P(Y = v_2)L(v_2, v_2) + \cdots = \sum_{j=1}^{J} P(Y = v_j)L(v_j, v_2)
\]

\[\vdots\]

\[
E[L(Y, v_J)] = P(Y = v_1)L(v_1, v_J) + P(Y = v_2)L(v_2, v_J) + \cdots = \sum_{j=1}^{J} P(Y = v_j)L(v_j, v_J)
\]

Third, find the \(g\) that minimizes \(E[L(Y, g)]\). That is, find the smallest of the \(J\) values computed in (2.47); the corresponding \(g\) is the optimal predictor.

For example, with \(J = 3\), let \(v_1 = -1\), \(v_2 = 0\), \(v_3 = 1\). Let

\[
\begin{pmatrix}
L(v_1, v_1) & L(v_2, v_1) & L(v_3, v_1) \\
L(v_1, v_2) & L(v_2, v_2) & L(v_3, v_2) \\
L(v_1, v_3) & L(v_2, v_3) & L(v_3, v_3)
\end{pmatrix} = \begin{pmatrix}
L(-1, -1) & L(0, -1) & L(1, -1) \\
L(-1, 0) & L(0, 0) & L(1, 0) \\
L(-1, 1) & L(0, 1) & L(1, 1)
\end{pmatrix}
\]

(2.48)

Let

\[
P(Y = v_1) = P(Y = -1) = 0.2,
\]

\[
P(Y = v_2) = P(Y = 0) = 0.3,
\]

\[
P(Y = v_3) = P(Y = 1) = 0.5 = 1 - P(Y = -1) - P(Y = 0).
\]

Then, the mean losses are

\[
E[L(Y, v_1)] = P(Y = v_1)L(v_1, v_1) + P(Y = v_2)L(v_2, v_1) + P(Y = v_3)L(v_3, v_1) = (0.2)(0) + (0.3)(2) + (0.5)(8) = 4.6,
\]

\[
E[L(Y, v_2)] = P(Y = v_1)L(v_1, v_2) + P(Y = v_2)L(v_2, v_2) + P(Y = v_3)L(v_3, v_2) = (0.2)(1) + (0.3)(0) + (0.5)(2) = 1.2,
\]

\[
E[L(Y, v_3)] = P(Y = v_1)L(v_1, v_3) + P(Y = v_2)L(v_2, v_3) + P(Y = v_3)L(v_3, v_3) = (0.2)(4) + (0.3)(1) + (0.5)(0) = 1.1.
\]

The smallest of these three values is \(E[L(Y, v_3)] = E[L(Y, 1)] = 1.1\), so the optimal predictor is \(g = v_3 = 1\).

### 2.5.3 Optimal Prediction: Specific Examples

**Example: Carnival Age Game**

Imagine predicting a person’s age, \(Y\). Imagine, when you were younger, you worked at a carnival in the summer, where people paid five tickets to see if you could guess their age. If you guessed correctly, they won nothing; if incorrect, they won the plush animal or fruit of their choice (which of course was still worth much less than five tickets).
Since they pay five tickets regardless of what you guess, that need not enter the loss function. The fact that it only matters if you guess correctly or not implies 0–1 loss: $L_0(y,g) = \mathbb{I}\{y \neq g\}$. (For added challenge, come back to this example and see what changes if they only win when you are more than three years off.)

For simplicity, let $P(Y = 20) = 0.6$ and $P(Y = 25) = 0.4$. The procedure in Section 2.5.2 can be used. Mean losses with 0–1 loss are

$$E(L_0(Y,20)) = E(\mathbb{I}\{Y \neq 20\}) = (0.6)(0) + (0.4)(1) = 0.4,$$
$$E(L_0(Y,25)) = E(\mathbb{I}\{Y \neq 25\}) = (0.6)(1) + (0.4)(0) = 0.6,$$

(2.50)

so $g_0^* = 20$ is the optimal prediction. (See also (2.56) for a different path to the same conclusion.)

Quadratic loss could lead to the wrong guess, depending how blindly we apply it. Comparing only $g = 20$ to $g = 25$,

$$E[L_2(Y,20)] = E[(Y - 20)^2]$$

$$= P(Y = 20)(20 - 20)^2 + P(Y = 25)(25 - 20)^2$$
$$= (0.6)(0) + (0.4)(5^2) = (0.4)(25) = 10,$$
$$E[L_2(Y,25)] = E[(Y - 25)^2]$$

$$= (0.6)(-5)^2 + (0.4)(0) = (0.6)(25) = 15.$$

Like with 0–1 loss, it is better to guess the more likely value 20 than the less likely 25. However, it is even better to guess something in between:

$$E[L_2(Y,22)] = E[(Y - 22)^2]$$

$$= P(Y = 20)(20 - 22)^2 + P(Y = 25)(25 - 22)^2$$
$$= (0.6)(-2)^2 + (0.4)(3)^2 = (0.6)(4) + (0.4)(9) = 6.$$

Some calculus shows $g = 22$ is actually optimal for $L_2$ loss. However, according to the rules of the carnival game, if we guess $g = 22$ when everyone has either $Y = 20$ or $Y = 25$, then we’ll lose every single time! This is the worst possible guess. This is one example where quadratic loss is not appropriate.

**Example: Advertising**

Later in life, well past your carnival days, you work in advertising. Coincidentally, your job is still to guess a person’s age, but with different consequences. If you guess somebody is 40 years old, then your client’s website shows an ad specifically designed for 40-year-olds. The loss function should capture how much worse it is to show an ad targeting the guessed age than to show the optimal ad for the individual’s true age.

Unlike at the carnival, some incorrect guesses are much worse than others. For example, it doesn’t matter much if you guess a person is 40 years old but really they’re 41. The optimal ad for the 40-year-old
is almost equally effective on the 41-year-old, so there is very little loss from guessing 40 instead of 41. However, guessing that the 41-year-old is 20 years old is much worse than guessing 40. Similarly, guessing that the 41-year-old is 60 is bad, but still better than guessing 80. Consequently, 0–1 loss is inappropriate because it treats guesses of 20, 40, 60, and 80 as equally bad when \( y = 41 \).

In this case, quadratic loss seems more appropriate than 0–1 loss, if not perfect.

Imagine the age distribution of \( Y \) is the same as in the carnival game. We saw that quadratic and 0–1 loss lead to different “optimal” predictions. That result depends only on the mathematical distribution of \( Y \), not on the real-world interpretation (carnival, advertising). So, the predictions are still different, but now we may prefer \( g = 22 \) over \( g = 20 \). That is, we might “lose” a lot by showing 25-year-olds an ad targeting 20-year-olds, but maybe both 20-year-olds and 25-year-olds respond to the ad targeting 22-year-olds.

Example: More Ages

Consider more complex versions of the carnival and advertising examples, with the following distribution of \( Y \). Now, any value \( Y \in \{20, 21, 22, 23, 24, 25\} \) is possible. Imagine young people are more likely; specifically,

\[
P(Y = j) = \frac{(26 - j)}{21}, \quad j = 20, 21, \ldots, 25.
\]

(2.51)

With 0–1 loss (for the carnival), mean loss when predicting \( g \) can be computed as follows. From (2.38), the loss function is \( L_0(Y, g) = \mathbb{1}_{\{y \neq g\}} \), which equals 0 if \( y = g \) but equals 1 otherwise. This can be rewritten as \( 1 - \mathbb{1}_{\{y = g\}} \). Also, \( \mathbb{1}_{\{y = g\}} P(Y = y) \) equals \( P(Y = g) \) when \( y = g \) but otherwise equals 0 when \( y \neq g \). Putting these pieces together, mean loss is

\[
E(L_0(Y, g)) = \sum_{y=20}^{25} \mathbb{1}_{\{y \neq g\}} P(Y = y) = \sum_{y=20}^{25} [1 - \mathbb{1}_{\{y = g\}}] P(Y = y)
\]

\[
= \sum_{y=20}^{25} P(Y = y) - \sum_{y=20}^{25} \mathbb{1}_{\{y = g\}} P(Y = y) = 1 - P(Y = g).
\]

Thus, the smallest possible mean loss is achieved by the largest possible \( P(Y = g) \),

\[
\arg \min_g E(L_0(Y, g)) = \arg \min_g [1 - P(Y = g)] = \arg \max_g P(Y = g) = 20,
\]

(2.52)

so the best prediction is \( g^*_0 = 20 \). (As defined in the Notation section before Chapter 1, \( \arg \min_g f(g) \) means “the value of \( g \) that minimizes \( f(g) \),” and \( \arg \max_g f(g) \) means “the value of \( g \) that maximizes \( f(g) \).”)

This result is intuitive because 0–1 loss only cares whether a prediction is right or wrong. Guessing \( g^*_0 = 20 \) gives you a 6/21 probability of being correct, the largest possible. Equivalently, it is the
2.5. PREDICTION WITH A KNOWN DISTRIBUTION

smallest possible probability of being wrong. More generally, \( g_0^* \) is always the single most likely value (the mode); see (2.56).

Like before, quadratic loss yields a different optimal prediction. If we guessed \( g = 20 \), then

\[
E[L_2(Y, g)] = \sum_{y=20}^{25} P(Y = y)(y-20)^2 = \sum_{y=20}^{25} [(26-y)/21](y-20)^2 = 5. 
\]  

(2.53)

(Link to calculation) But we can do better by guessing a value more toward the “middle” of the distribution. Although \( g = 20 \) is exactly correct sometimes, it’s bad when \( Y = 25 \). If we try \( g = 21 \),

\[
E[L_2(Y, g)] = \sum_{y=20}^{25} P(Y = y)(y-21)^2 = \sum_{y=20}^{25} [(26-y)/21](y-21)^2 = 8/3 \approx 2.67. 
\]  

(2.54)

(Link to calculation) Since 2.67 < 5, guessing \( g = 21 \) is better than \( g = 20 \) according to mean quadratic loss. (Can you do even better than \( g = 21 \)?)

Discussion Question 2.3 (banana loss function). Imagine you run a small banana shop. You buy bananas wholesale for 2 cents each ($0.02) and sell each for 40 cents ($0.40). The wholesaler delivers every Monday. Any bananas not sold by the next Monday spoil; you cannot sell them (they just go in the compost). Let \( y \) be the actual number of bananas that customers want to buy in some week. Let \( g \) be your guess, i.e., how many you bought wholesale on Monday.

a) Why isn’t 0–1 loss appropriate?

b) Why isn’t quadratic loss appropriate?

c) What might the loss function look like, if you only care about maximizing profit? Try to be as specific and mathematical as you can. In particular, consider the different consequences of over-buying \((g > y)\) versus under-buying \((g < y)\).

2.5.4 Mean and Mode as Optimal Predictions

Under quadratic loss, the mean is the optimal predictor that minimizes mean loss. Although the details are beyond our scope, calculus can be used to take the derivative of mean loss with respect to \( g \) and set it to zero (the “first-order condition”), yielding

\[
g_2^* \equiv \arg \min_g E[(Y - g)^2] = E(Y). 
\]  

(2.55)

This says the mean of a distribution has two interpretations. For description, the mean helps summarize the “center” of the distribution. For prediction, the mean is the prediction of an unknown value of \( Y \) that minimizes mean quadratic loss.
Under 0–1 loss, the optimal prediction is

\[ g^*_0 \equiv \arg\min_g \mathbb{E}(L_0(Y, g)) = \arg\min_g \mathbb{E}[\mathbf{1}\{Y \neq g\}] = \arg\min_g \mathbb{P}(Y \neq g) \]

\[ = \arg\min_g [1 - \mathbb{P}(Y = g)] = \arg\min_g -\mathbb{P}(Y = g) \]

\[ = \arg\max_g \mathbb{P}(Y = g). \]  

(2.56)

That is, the mode (the single most likely value of \( Y \)) is the optimal predictor. (This formula does not make sense if \( Y \) is not continuous because then \( \mathbb{P}(Y = g) = 0 \) for any \( g \).) Thus, like the mean, the mode has two interpretations: one for description, and one for prediction.

**Discussion Question 2.4** (optimal banana prediction). Consider the same setup as in DQ 2.3, and again assume you want to maximize (mean) profit. Imagine you know the distribution of \( Y \) (banana quantity demanded in one week).

a) Do you think the mean \( \mathbb{E}(Y) \) is a good “predicted” number of bananas to buy wholesale? Explain why or why not; if not, also explain why you think \( \mathbb{E}(Y) \) is too high or too low.

b) What if the retail price were $99 per banana, and the wholesale cost is still $0.02 per banana—would \( \mathbb{E}(Y) \) be good, or too high, or too low, and why?

c) What if the retail price were equal to the wholesale price?

### 2.5.5 Interval Prediction

Only point prediction has been discussed so far, i.e., the single number that provides the best guess of the unknown value. Alternatively, interval prediction lets the guess be a range of numbers called a prediction interval.

The disadvantage of point predictions is that they are usually wrong for discrete variables and always wrong for continuous variables. For example, if \( Y \sim \mathcal{N}(0, 1) \), then the best point prediction under quadratic loss is \( \mathbb{E}(Y) = 0 \). But this guess will be wrong 100% of the time since \( \mathbb{P}(Y = 0) = 0 \).

By guessing a range of numbers, a prediction interval can actually contain the true value with large probability. The length of the interval captures the level of uncertainty: with lots of uncertainty, the interval must be very long to have a high probability of containing the true value.

For example, let \( \mathbb{P}(Y = j) = 1/100 \) for \( j = 1, 2, \ldots, 100 \). The mean \( \mathbb{E}(Y) = 50.5 \) is the best prediction under quadratic loss, but it never actually happens: \( \mathbb{P}(Y = 50.5) = 0 \). Even \( \mathbb{P}(Y = 50) = 0.01 \), still very small. Alternatively, the prediction interval \([26, 75]\) has \( \mathbb{P}(26 \leq Y \leq 75) = 50/100 \): there is roughly a 50% probability that a randomly sampled \( Y \) value is inside the prediction interval.

Or, \( \mathbb{P}(6 \leq Y \leq 95) = 90/100 \), so the prediction interval \([6, 95]\) has around 90% probability of containing a randomly drawn \( Y \).
The interval length reflects amount of uncertainty. Above, the 90% prediction interval was [6, 95]. If instead \( P(Y = j) = 1/10 \) for \( j = 1, 2, \ldots, 10 \), then the much shorter interval [2, 10] has 90% probability of containing \( Y \). The values of \( Y \) are concentrated more closely together, so the prediction interval can be smaller but still have the same 90% probability.

Even for the same \( Y \) distribution and same interval probability (like 90%), there may be multiple possible prediction intervals. In the original example, [26, 75] was a 50% prediction interval, but so is [25, 74], or [1, 50], or [51, 100]. Other properties can be used to distinguish among these intervals, but such is beyond our scope.
Chapter 3

One Variable: Sample

⇒ Kaplan video: Chapter Introduction

Depends on: Chapter 2

Unit learning objectives for this chapter

3.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]
3.2. Describe and distinguish Bayesian and frequentist perspectives [TLO 4]
3.3. Identify and interpret properties of a sampling procedure or estimator [TLO 4]
3.4. Judge which estimator is better based on its properties [TLO 6]
3.5. Interpret different measures of statistical uncertainty [TLOs 6 and 7]
3.6. Assess the statistical significance and economic significance of empirical results [TLO 6]
3.7. In R (or Stata): compute estimates of a population mean along with measures of uncertainty [TLO 7]

Optional resources for this chapter

- Basic statistics: the Khan Academy AP Statistics unit includes instructional material and practice questions
- Quantifying uncertainty and statistical significance (Masten video)
- Estimator properties (Lambert video)
- Unbiasedness and consistency (Lambert video 1 of 2)
- Unbiasedness and consistency (Lambert video 2 of 2)
- iid sampling (Lambert video)
• Bayesian vs. frequentist cookie inference example (StackExchange)

• Section 2.8 (“Exploratory Data Analysis with R”) in Kleiber and Zeileis (2008) [Chapter 2 is available free on their website]

• Section 2.2 (“Random Sampling and the Distribution of Sample Averages”) and Chapter 3 (“A Review of Statistics Using R”) in Hanck et al. (2018)

• Sections 1.5.4 (“Fundamental Statistics”) and 1.9.3 (“Simulation of Confidence Intervals and t Tests”) in Heiss (2016)

• R package boot (Canty and Ripley, 2019; Davison and Hinkley, 1997)

Sections 2.3 and 2.5 considered only the population distribution, whereas Chapter 3 considers data sampled from that distribution. The words data, dataset, sample values, and sample all refer to the same thing: the set of values that the researcher actually sees. But, as in Chapter 2, this could be seen either from the “before” perspective as random variables, or from the “after” perspective as non-random realized values. Section 2.1 gave the general idea of seeing observations as random variables (the “before” view); here, specific details are provided on estimation and uncertainty.

Although long, this chapter is mostly review of material you should have seen already in an introductory statistics class.

3.1 Bayesian and Frequentist Perspectives

⇒ Kaplan video: Bayesian and Frequentist Perspectives

Two frameworks constitute econometrics and statistics: Bayesian and frequentist (or classical). These are cynically deemed “sects” by some, but outside the vocal extremes (and amusing webcomics: xkcd.com/1132), most econometricians appreciate and respect both frameworks (and the people who use them), sometimes working with both in turn.

This text uses the frequentist framework. Why? Mostly, that’s just how I wrote it; I’ll spare you post hoc rationalization.

There is little disagreement about the population and what we want to learn. Generally, both Bayesian and frequentist perspectives agree on everything in Chapter 2 about the population and how data are generated.

The disagreements are about how to use the sampled data to learn about the population. Frequentist and Bayesian approaches have different advantages, appropriate for different settings.

The goal of the remainder of Section 3.1 is to give you a very basic overview and comparison of Bayesian and frequentist approaches. At minimum, I hope you get a sense of their different ways of quantifying
uncertainty, and the different types of questions they can (and cannot) answer.

3.1.1 Very Brief Overview: Bayesian Approach

The Bayesian approach models your beliefs about an unknown population value $\theta$, like the mean $\theta = \mathrm{E}(Y)$. Your prior (or prior belief) is what you believe about $\theta$ before seeing the data. Your posterior (or posterior belief) is what you believe about $\theta$ after seeing the data. The Bayesian approach describes how to update your prior using the observed data, to get your posterior.

Mathematically, “belief” is a probability distribution. For example, let random variable $B$ represent your belief about the population mean. If you think there’s a 50% chance the mean is negative, then $\Pr(B < 0) = 50\%$. If you think there’s a $1/4$ probability that $B$ is below $-1$, then $\Pr(B < -1) = 1/4$. There are formal procedures for “prior elicitation,” i.e., quantifying your beliefs as a distribution.

Notationally, though it is confusing, usually the belief is represented by the (usually Greek) letter for the parameter, like $\theta$, rather than a separate variable (like $B$ above). This does not mean “there is no population mean”; it is purely a notational convention for representing beliefs.

For a concrete example, imagine you find an archaeological site in Missouri with many artifacts, but you are unsure of which people group had lived in that site. Based on its location, it was either the Missouria, Illini, or Osage tribe, which can be represented as $\theta = M$, $\theta = I$, or $\theta = O$, respectively. That is, $\theta$ is the unknown parameter of interest, with possible values $M$, $I$, or $O$. Before looking at the artifacts (the data), you believed there was equal chance of each tribe; i.e., your prior belief was $\Pr(\theta = j) = 1/3$ for each $j \in \{M, I, O\}$. After looking at the artifacts (the data) more closely, they look most similar to Missouria artifacts, but you are unsure. Quantitatively, you believe there’s a 50% chance they were Missouri, 40% chance Osage, and 10% chance Illini. This is your posterior distribution, i.e., your beliefs about $\theta$ after seeing the data. Mathematically, your posterior is $\Pr(\theta = M) = 0.5$, $\Pr(\theta = O) = 0.4$, $\Pr(\theta = I) = 0.1$.

The posterior distribution is the Bayesian way of quantifying uncertainty. It is arguably more intuitive; it is more similar to how people talk about uncertainty in daily life. The posterior distribution is often summarized by a credible interval, i.e., a range of values that you’re pretty sure contains the true $\theta$, like $\Pr(a \leq \theta \leq b) = 90\%$. Or in the above example with categorical $\theta$, the credible set $\{M, O\}$ has 90% posterior belief; you’d say, “I’m pretty sure it’s Missouria or Osage, although I think there’s a 10% chance I’m wrong.”

3.1.2 Very Brief Overview: Frequentist Approach

Other sections in Chapter 3 flesh out details, but the core of the frequentist approach is the “before” perspective, which can also be
described in terms of repeated sampling. Instead of the belief probabilities of a Bayesian posterior, frequentist probabilities are from the “before” view of what dataset (and thus value of estimator and such) could be randomly sampled. Equivalently, as a thought experiment, we can imagine many different random samples drawn from the same population; the “before” probabilities are then how often certain values occur in these many random datasets.

For intuition, imagine you could randomly sample 100 datasets from the same population. Then, the frequentist probability of an event says approximately how many times that event occurs among the 100 samples. (To replace “approximately” with “exactly,” replace 100 with $\infty$.) For example, we could compute the sample mean $\bar{Y}$ in all 100 samples; since the datasets are all different, the sample means $\bar{Y}$ are also all different. If $\bar{Y} \leq 0$ in 50 of the 100 hypothetical samples, then $P(\bar{Y}_n \leq 0) \approx 50/100 = 50\%$. Or, if $\bar{Y}$ is in the interval $[-0.4, 0.4]$ in 70 of 100 samples, then $P(-0.4 \leq \bar{Y} \leq 0.4) = P(\bar{Y} \in [-0.4, 0.4]) \approx 70\%$. A similar example is in Table 3.1.

### 3.1.3 Bayesian and Frequentist Differences

The following makes explicit some of the differences between the Bayesian and frequentist approaches described above.

First, the frameworks treat different objects as random or non-random. The frequentist framework treats the population mean and other population features as non-random values, whereas it treats the data as random. For example, the population mean $\mu = E(Y)$ is a non-random value, whereas an observation $Y$ is a random variable. In contrast, the Bayesian framework treats population features as random (to reflect your beliefs), whereas it treats the data as non-random values (the “after” view).

Second, due to this different treatment, the frameworks answer different types of questions, especially when quantifying uncertainty. For example, the Bayesian framework is designed to answer questions like, “Given the observed data, what do I believe is the probability that the population mean is above $1/2$?” Mathematically, if $y$ is the “observed data,” this is usually written $P(\mu > 1/2 \mid y)$, noting the confusing notation where $\mu$ represents beliefs. This question makes no sense from the frequentist perspective: either $\mu > 1/2$ or not; it cannot be “maybe,” with some probability. In contrast, the frequentist framework answers questions like, “Given the value of $\mu$, what’s the probability that the sample mean is above $1/2$?” Mathematically, this is usually written $P(\bar{Y} > 1/2)$, or $P_\mu(\bar{Y} > 1/2)$ to be explicit about the dependence on $\mu$. The sample mean $\bar{Y}$ is a function of data, so it is treated as a random variable. This question makes no sense from the Bayesian perspective: we can see the data, so we can see either $\bar{Y} > 1/2$ or not; it cannot be “maybe,” with some probability.

Interestingly, both frameworks can answer questions like $P(\bar{Y} < \mu)$, but with different interpretations. The Bayesian answer interprets $\bar{Y}$ as a number (that we see in the data) and $\mu$ as a random variable.
representing our beliefs. The frequentist answer interprets \( \bar{Y} \) as the random variable (from the “before” view) and \( \mu \) as the non-random population value.

Third, frequentist methods use only the data, whereas Bayesian methods can formally incorporate additional knowledge. In practice, though, even frequentist results should be interpreted in light of other knowledge. The difference is that this process is not formally within the frequentist methodology itself. Unfortunately, many people do not combine frequentist results with other knowledge, instead interpreting frequentist results as if one single dataset contains the full, absolute truth of the universe; please do not do this!

In Sum: Bayesian & Frequentist
Frequentist: “before” view of data (random variables); assess methods’ performance across repeated random samples from same population
Bayesian: “after” view of data (non-random); model beliefs (about population features) as random variables

3.2 Types of Sampling

In practice, judging which econometric method is most appropriate requires understanding different types of sampling procedures and sampling properties. Such judgment is mostly left to another textbook, but this section hopes to help your understanding.

Notationally, we observe the values from \( n \) units, which could be individuals, firms, countries, etc. Let \( i = 1 \) refer to the first unit, \( i = 2 \) to the second, etc., up to \( i = n \), where \( n \) is the sample size. The corresponding values are \( Y_1, Y_2, \ldots, Y_n \), with \( Y_i \) more generally denoting the observation for unit \( i \). A particular dataset may have specific values like \( Y_1 = 5, Y_2 = 8 \), etc., but to analyze statistical properties, each \( Y_i \) is seen as a random variable as in Section 2.1.

In this section, two important sampling properties are considered: “independent” and “identically distributed.” If both hold, then the \( Y_i \) are called independent and identically distributed (iid) random variables (or “sampled iid”), and “sampling is iid.” Sometimes the vague phrase random sample refers to iid sampling. This iid sampling is mathematically simplest but not always realistic. Although iid sampling is the focus here (like other introductory textbooks), weights are briefly mentioned, and Part III considers dependent (i.e., not independent) data.

Notationally, iid sampling is indicated by \( \sim^{iid} \). If \( F_Y(\cdot) \) is the population CDF,

\[
Y_i \overset{iid}{\sim} F_Y, \quad i = 1, \ldots, n. \tag{3.1}
\]

If there is only a population PMF \( f_Y(\cdot) \) and not a CDF (like with
nominal categorical variables), then $F_Y$ is replaced by $f_Y$ in (3.1). If the $Y_i$ follow a known distribution like N(0, 1), then $F_Y$ is replaced by N(0, 1), for example.

There are other sampling properties not considered in this section, like sampling bias. This is about whether we observe a “representative sample” of the population we want to learn about (the population of interest). Sometimes sampling bias is our fault (for using the wrong dataset for our economic question), but sometimes we try to get the right data and people refuse to answer our survey, or we can’t get access to certain confidential data, etc. This is discussed more in Chapter 12, in terms of “missing data” and “sample selection.”

After introducing “independent” and “identically distributed” sampling, examples are discussed in Section 3.2.3.

### 3.2.1 Independent

Qualitatively, in the context of sampling, independence (or independent sampling) means that from the “before” view, any two observations are unrelated. For example, the value of $Y_2$ is unrelated to $Y_1$: we are not any more likely to see a high $Y_2$ if we see a high $Y_1$ in the sample.

Mathematically, independence means

$$Y_i \perp \perp Y_k \text{ for any } i \neq k,$$

where $\perp \perp$ denotes statistical independence. That is, $Y_1 \perp \perp Y_2, Y_1 \perp \perp Y_5, Y_6 \perp \perp Y_4$, etc. For any $i \neq k$, independent sampling implies

$$\text{Cov}(Y_i, Y_k) = 0, \quad \text{Var}(Y_i + Y_k) = \text{Var}(Y_i) + \text{Var}(Y_k), \quad \text{E}(Y_i | Y_k) = \text{E}(Y_i)$$

among other things.

### 3.2.2 Identically Distributed

The identically distributed property means that from the “before” view, the distribution of $Y_i$ is the same for any $i$. Qualitatively, all units are sampled from the same population. Mathematically, if $F_Y(\cdot)$ is the population CDF, $Y_i \sim F_Y$ for all $i = 1, \ldots, n$. (Or $Y_i \sim f_Y(\cdot)$.) Note $\sim$ and not $\overset{iid}{\sim}$; it only claims the distribution of $Y_i$ in isolation, not independence from the other observations.

Mathematically, identically distributed $Y_i$ means that for any $i$ and $k$, $Y_i$ and $Y_k$ have the same distribution. Thus, any feature of their distributions is also identical. For example, $\text{E}(Y_i) = \text{E}(Y_k)$ and $\text{Var}(Y_i) = \text{Var}(Y_k)$.

**Practice 3.1** (i/id sampling). You are planning to sample to values, $Y_1$ and $Y_2$, but you have not sampled them yet. The following four statements correspond to the four sampling properties (or their implications): 1) independent, 2) not independent (i.e., dependent), 3) identically distributed, 4) not identically distributed. Which is which?

a) You are just as likely to get $Y_1 = 3$ as $Y_2 = 3$. 
b) If you get a negative $Y_1$, then you’ll probably get a negative $Y_2$; but if you get a positive $Y_1$, then you’ll probably get a positive $Y_2$.

c) Separately and simultaneously, you will randomly sample $Y_1$ and $Y_2$ using the same exact procedure from the same population.

d) For $Y_1$ you are going to get the salary of somebody with an economics degree, and $Y_2$ will be the salary of somebody with an art history degree.

3.2.3 Examples

Consider the following sampling procedures and their properties. Each example has 4 observations of Mizzou students. You can imagine 4 buckets (or pieces of paper), initially empty, that will eventually contain information from 4 observations. The sampling procedure does not determine the specific numeric values that end up in the buckets, but it determines how the buckets get filled.

**Random Student ID**

Imagine randomly picking a Mizzou student ID number, then randomly picking a 2nd, then 3rd, then 4th. These $Y_i$ are both independent and identically distributed (iid). They are independent because each ID number is randomly drawn without any consideration of how the other numbers are drawn, and without any consideration of the other observed $Y_i$ values. They are identically distributed because each ID number is drawn from the same population (anyone who has a Mizzou student ID).

**Random Economics and Philosophy Students**

Imagine randomly picking 2 economics major students and 2 philosophy major students. That is, buckets 1 and 2 say “econ,” while buckets 3 and 4 say “philosophy.” Observations $Y_1$ and $Y_2$ are from economics students, while $Y_3$ and $Y_4$ are from philosophy students. Assigning buckets to different groups (“strata”) before sampling is called **stratified sampling**.

Observations are probably not identically distributed, although it depends on the variable. Specifically, they are not identically distributed for variables whose distributions differ for economics and philosophy majors. For example, if philosophy students have higher college entrance exam scores (like ACT or SAT) than economics students, then $Y_3$ and $Y_4$ will probably be higher than $Y_1$ and $Y_2$. We know this even before observing the specific values; it is not about the values, but about how the buckets are filled. Mathematically, this could be expressed as something like $E(Y_1) = E(Y_2) < E(Y_3) = E(Y_4)$ if the population mean score is higher for philosophy students.

But, couldn’t there be econ and philosophy students in the first example (random ID numbers), too, which was claimed to be identically distributed? Yes, but “could be” is a different procedure than
“must be.” In the ID example, students’ majors are only known after sampling (after filling the bucket), whereas now the buckets themselves specify “econ” or “philosophy.”

Observations are independent. That is, sampling is independent but not identically distributed (inid). Although the buckets are filled from different populations (econ, philosophy), each student is still picked randomly, without consideration of the other students in the sample. For example, if $Y_1$ is a very high ACT score, this does not make $Y_2$ any more likely to be high (or low).

**Students in Same Class**

Imagine randomly picking a class (like Intro Econometrics) at Mizzou, then filling the first two buckets ($Y_1$ and $Y_2$) with two random students from that class, and then randomly picking another class and another two students for the other buckets ($Y_3$ and $Y_4$). This is an example of clustered sampling, where each class is a “cluster.” (This is different than “clustering” in the sense of cluster analysis.)

Observations are identically distributed because each $Y_i$ has the same probability of getting any particular student. For example, $Y_2$ is just as likely to be a random student from Intro Econometrics as is $Y_3$, and similarly for any other class.

Observations may be dependent (i.e., not independent), depending on the variable. The dependence may come from students in the same class being similarly affected by their shared experience. Here, buckets 1 and 2 are correlated, and 3 and 4 are correlated, but not 1 and 3, nor 2 and 4, etc. For example, if we observe starting salary after graduation, it is more likely that $Y_1$ and $Y_2$ are both high (if their shared class was very helpful), or both low (if it wasn’t), than one high and one low. In the extreme where the class fully determines salary, we’d see $Y_1 = Y_2$ and $Y_3 = Y_4$, though the specific values are not known before sampling. In contrast, there is no such correlation between $Y_1$ and $Y_3$.

But, couldn’t the first two buckets be from an econ class, and the others from a philosophy class, so it should be independent like in the previous example? The difference is whether the bucket labels (econ, philosophy) are random or non-random. If they are non-random like in the prior example, then there is stratified sampling and independence. The first two values may tend to be higher, but they are not more likely to be both-high or both-low than high-low or low-high. If the bucket labels are random (like this example), then the first two values don’t tend to be higher or lower than the other values, but they may tend to be both-high or both-low more than high-low or low-high.

**Two Students, Two Semesters**

Imagine randomly picking 2 students (iid, like with random ID numbers), then observing them this semester and next semester. This
3.3. THE EMPIRICAL DISTRIBUTION

is another type of clustered sampling that usually violates independence. Further, observations may not be identically distributed, again depending on the variable.

For example, imagine the variable is semester GPA. Bucket 1 contains the first student’s GPA this semester, bucket 2 contains the same student’s GPA next semester, and buckets 3 and 4 contain the other student’s GPAs from this semester and next semester. Buckets 1 and 2 ($Y_1$ and $Y_2$) are probably both high or both low, rather than one high and one low, and similarly for buckets 3 and 4 ($Y_3$ and $Y_4$). That is, buckets 1 and 2 are correlated, and 3 and 4 are correlated.

Observations may not be identically distributed if fall GPA and spring GPA do not have the same distribution. For example, perhaps spring GPAs tend to be higher.

From a different perspective, sampling is actually iid. The students themselves are sampled iid. Observations are iid if we see $(Y_1, Y_2)$ as a single observation, and see $(Y_3, Y_4)$ as a second observation. That is, $(Y_1, Y_2)$ is randomly sampled from the same population as $(Y_3, Y_4)$.

One Student, Four Semesters

Similar to above, if you randomly pick one student but then observe the same student over four consecutive semesters, there is probably dependence, and possibly not identical distributions (e.g., if GPA tends to increase over time). This is time series data; see Part III.

Practice 3.2 (rural household sampling). You want to learn about household consumption in rural Indonesia. In an area with 100 villages, you either i) pick 5 villages at random, then survey every household in each of the 5 villages; or ii) make a list of all households in all 100 villages, then randomly pick 5% of them. Explain why each approach is or isn’t iid.

3.3 The Empirical Distribution

The empirical distribution is a probability distribution that reflects the sample data. It can be confusing at first, but it unifies many approaches in this class and beyond, helping them seem less ad hoc and mysterious. Qualitatively, the empirical distribution treats the sample as if it were the population.

Mathematically, first consider a binary variable. The population is represented by binary random variable $Y$ with some $P(Y = 1) = p$. The sample of size $n$ can be represented by binary random variable $S$ with

$$P(S = 1) = \hat{p} = \frac{\text{how many } Y_i = 1}{n} = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i = 1\}, \quad (3.4)$$
the sample proportion of observations with $Y_i = 1$. The distribution of $S$ is the empirical distribution.

The **plug-in principle** or **analogy principle** suggests we compute whatever features of $S$ we want to learn about $Y$. For example, if we want to learn $E(Y)$, then compute $E(S)$. With enough data, $S$ is usually very similar to $Y$, so features of $S$ should usually be very similar to those of $Y$.

Mathematically, consider now a categorical or discrete variable. The population is represented by random variable $Y$ with PMF $f_Y(\cdot)$. Imagine there are $J$ categories with values $(v_1, \ldots, v_J)$, so $Y$ is fully described by $f_Y(v_j)$ for each $j = 1, \ldots, J$. The sample is represented by random variable $S$ with

$$f_S(v_j) = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i = v_j\}, \quad j = 1, \ldots, J.$$  \hfill (3.5)

That is, $f_S(v_j)$ is the sample proportion of observations with $Y_i = v_j$.

Mathematically, consider finally a continuous variable. The population is represented by random variable $Y$ with continuous CDF $F_Y(\cdot)$. However, even with an infinite number of possible values for $Y$ in the population, there are only $n$ possible values of $Y_i$ observed. With a continuous random variable, each observed $Y_i$ value is unique, so there are exactly $n$ different observed $Y_i$ values. The sample is thus represented by random variable $S$ with PMF

$$f_S(Y_i) = 1/n, \quad i = 1, \ldots, n.$$  \hfill (3.6)

Even though $Y$ is continuous, $S$ is discrete.

Notationally, instead of $f_S(\cdot)$, more common is $\hat{f}_Y(\cdot)$ if $Y$ has a PMF. If $Y$ has a CDF, then more often the empirical CDF (ECDF) is used. It is simply the CDF of $S$. The ECDF (or just EDF) is usually written $\hat{F}_Y(\cdot)$, or just $\hat{F}(\cdot)$, and defined as

$$\hat{F}_Y(y) = F_S(y) = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i \leq y\},$$  \hfill (3.7)

i.e., the sample proportion of observations less than or equal to $y$, the point of evaluation.

Notationally, a **hat** (circumflex) often denotes a sample analog, i.e., a feature of $S$ analogous to a population feature of $Y$. Above in (3.7), for the population $F_Y(y)$, the sample analog is $\hat{F}_Y(y) = F_S(y)$. As another example, for the population $P(Y = y)$, the sample analog is $\hat{P}(Y = y) = P(S = y)$. For the population $E(Y)$, the sample analog is $\hat{E}(Y) = E(S)$. The “hat” may indicate another value computed from the sample data (i.e., a statistic), usually an estimator, even if it is not a sample analog.

### 3.4 Estimation of the Population Mean

Sections 2.3 and 2.5 helped us think about which features of the population are useful for description and prediction. Such a population
3.4. ESTIMATION OF THE POPULATION MEAN

This section specifically considers estimating the population mean. This is most directly useful in later chapters. For other population features, the same concepts apply, though details differ.

Recall from Section 2.3.1 that the mean is directly useful for description and prediction with discrete and continuous variables, and it can have a useful probability interpretation for binary and even categorical variables. For binary $Y$, $E(Y) = P(Y = 1)$. Further, categorical variables can be turned into binary variables, like $Z = 1\{Y = v_2\}$. For example, if $Y$ is a 2-digit NAICS industry code whose possible values are “utilities,” “construction,” or “information,” then defining $Z = 1\{Y = \text{information}\}$ ($Z = 1$ for “information,” $Z = 0$ otherwise) implies $E(Z) = P(Y = \text{information})$.

The focus of this section is **point estimation**, as opposed to interval estimation. A point estimate is a single number, representing our best guess of the unknown population value. In contrast, an interval estimate is a range of numbers, most commonly a confidence interval; see Section 3.8.

3.4.1 “Description”: Sample Mean

As alluded to in Section 3.3, at least with identically distributed sampling, the population mean can be estimated by its sample analog, the mean of the empirical distribution. This is called the **sample mean**. It is also called the **sample average** because it averages the sample $Y_i$ values. The sample average is usually denoted $\bar{Y}$ (or $\bar{Y}_n$). Mathematically, for continuous or discrete (including binary) $Y$, using the notation of Section 3.3,

$$\bar{Y} = \hat{E}(Y) = E(S) = \frac{1}{n} \sum_{i=1}^{n} Y_i. \quad (3.8)$$

These expressions are equivalent, just emphasizing different interpretations.

3.4.2 “Prediction”: Least Squares

Section 2.5.4 showed that the population mean $E(Y)$ also solves an optimal population prediction problem. Specifically, (2.55) shows

$$E(Y) = g_2^* \equiv \arg\min_g E[(Y - g)^2].$$

From Section 3.3, the analogy principle suggests estimating $E(Y)$ by solving the same optimal prediction problem for the empirical distribution, i.e., replacing $Y$ with $S$. Mathematically, let

$$\hat{g}_2^* \equiv \arg\min_g E[(S - g)^2] = \arg\min_g \frac{1}{n} \sum_{i=1}^{n} (Y_i - g)^2. \quad (3.9)$$
The hope is that the sample analog $\hat{g}_2^*$ is close to the population $g_2^*$, which in turn equals $E(Y)$.

Skipping the derivation, it turns out

$$\hat{g}_2^* = \bar{Y}. \quad (3.10)$$

The prediction-motivated estimator equals the description-motivated estimator. This makes sense because in the population represented by $Y$, the mean equals the optimal prediction (under quadratic loss), and $S$ is simply another random variable like $Y$.

Rewriting (3.9) allows the introduction of some terms and concepts used in later chapters. In (3.9), the $1/n$ has no effect on the minimization problem because it is unaffected by $g$. Consequently, it is equivalent to write

$$\hat{g}_2^* = \arg\min_g n \sum_{i=1}^n (Y_i - g)^2. \quad (3.11)$$

To dissect the right-hand side of (3.11), imagine any estimate $\hat{g}$. Since $\hat{g}$ can be seen as trying to predict $Y$, sometimes $\hat{g}$ is called the predicted value of $Y_i$, which in this simple setting is the same for all $i$. However, the observed value of $Y_i$ is used to compute $\hat{g}$, so it seems misleading to say $Y_i$ was “predicted”: usually we assume the true value is not known when we discuss prediction. Instead, calling $\hat{g}$ the fitted value is more appropriate. Either way, the difference $\hat{U}_i = Y_i - \hat{g}$ is called the residual for observation $i$, i.e., the difference between the observed value $Y_i$ and the fitted value $\hat{g}$. The squared residuals are then $\hat{U}_i^2 = (Y_i - \hat{g})^2$. The sum of squared residuals (SSR) is then

$$\sum_{i=1}^n \hat{U}_i^2 = \sum_{i=1}^n (Y_i - \hat{g})^2. \quad (3.12)$$

Consequently, (3.10)–(3.12) together say that $\bar{Y}$ minimizes the SSR. For this reason, $\bar{Y}$ is a least squares estimator: “least” referring to minimization, and “squares” referring to the second S in SSR.

3.4.3 Non-iid Sampling: Weights

If your dataset has weights, then you should use them. Using weights adjusts the sample to be more representative of the population. Conversely, ignoring weights often produces misleading results because the sample is not representative.

There are multiple types of weights, although the distinction is beyond our scope. Generally, any type of weight is treated the same for estimation, but different for “inference” (confidence intervals, etc.). Types like survey weights (also called sampling weights) indicate non-iid sampling, whereas other types like frequency weights may simply allow more compact storage of iid sampled data.
Example

Skipping the theory, an example is shown in the following code. There are two subpopulations (two types of individuals) in the population: one with mean 0, one with mean 1. Each subpopulation forms half the overall population, so the overall population mean is \((1/2)(0) + (1/2)(1) = 1/2\). However, the second subpopulation forms much more than half of the sample because it is **oversampled**: each observation has a 2/3 probability (instead of 1/2) of coming from the second subpopulation. Thus, it’s like sampling from a different population whose mean is \((1/3)(0) + (2/3)(1) = 2/3\). Without weighting, the unweighted sample mean estimates this 2/3 value, not 1/2.

Sampling weights can be used to adjust the sample to represent the population. Specifically, the weights are the inverse of the sampling probabilities: \(1/(2/3) = 3/2 = 1.5\) for observations from the second subpopulation, and \(1/(1/3) = 3\) for individuals from the first subpopulation. This counteracts the fact that there are more individuals from the second subpopulation by weighting them less. Alternatively, instead of the inverse sampling probabilities, the inverse sample proportions of each type could be used. Another option is to use function \texttt{svymean()} in the R package \texttt{survey} Lumley (2004, 2019).

```r
set.seed(112358)
n <- 567
itype <- sample(x=0:1, size=n, replace=T, prob=1:2/3)
Y <- itype + rnorm(n)
mean(Y)  # without weights: near 2/3=0.67, not representative
## [1] 0.708

# with weights: should be closer to true 0.50
weighted.mean(x=Y, w=1/((itype+1)/3))
## [1] 0.551

weighted.mean(x=Y, w=ifelse(itype, n/sum(itype), n/(n-sum(itype))))
## [1] 0.54
```

### 3.5 Sampling Distribution of an Estimator

--- Kaplan video: Sampling Distribution of an Estimator

There are two goals of this section. The primary goal is to understand what it means for an estimator to have a probability distribution. The secondary goal is to observe some patterns that are formalized in Section 3.6.

At a high level, the **sampling distribution** is the probability distribution of an estimator, treated as a random variable in the “before”
view. Equivalently, from the repeated sampling perspective, the sampling distribution imagines computing the estimator in a large number of randomly sampled datasets from the same population, and seeing which values occur with what probability.

To develop intuition, the remainder of this section contains more specific examples using the sample mean. The examples include mathematical calculations, graphs, and tables. The details are not themselves important, but rather how they manifest the deeper concepts; i.e., there is no value in memorizing the examples, but rather using them to assess and develop your understanding of the fundamental ideas.

### 3.5.1 Some Mathematical Calculations

To be concrete, consider the sample mean as an estimator of the population mean, with iid sampling. Here, the $n$ subscript is added to $\bar{Y}_n$ because the sampling distribution depends on $n$. For example, the sampling distribution of $\bar{Y}_1 = Y_1$ differs from that of $\bar{Y}_2 = (Y_1 + Y_2)/2$.

From the “before” view, the sample mean $\bar{Y}_n$ is a random variable. The $Y_i$ are all random variables, so their average is also a random variable. That is, the $Y_i$ have multiple possible values, so the sample mean also has multiple possible values.

To develop intuition, consider some simple examples. The simplest example is $n = 1$, so $\bar{Y}_1 = Y_1$. The distribution of $\bar{Y}_1$ is the same as the population distribution that $Y_1$ follows.

Simpler still, imagine the population is binary with mean $p$. Then $Y_1 \sim \text{Bernoulli}(p)$, meaning $P(Y_1 = 1) = p$ and $P(Y_1 = 0) = 1 - p$. Since $\bar{Y}_1 = Y_1$, $P(\bar{Y}_1 = 1) = p$ and $P(\bar{Y}_1 = 0) = 1 - p$. If you imagine 100 randomly sampled datasets (i.e., 100 randomly sampled values of $Y_1$), then approximately $100p$ datasets would have $\bar{Y}_1 = 1$, while approximately $100(1 - p)$ would have $\bar{Y}_1 = 0$. For example, if $p = 0.7$, then around 70 datasets would have $\bar{Y}_1 = 1$ and around 30 datasets would have $\bar{Y}_1 = 0$. With infinite datasets, the words “approximately” and “around” are no longer needed, but at least for me it is easier to develop intuition by imagining 100 datasets than $\infty$.

Consider $n = 2$ with binary $Y$. Despite the simplicity, it takes some work to derive the sampling distribution of $\bar{Y}_2$. Let

$$Y_i \overset{iid}{\sim} \text{Bernoulli}(p) \implies Y_1 \perp \perp Y_2, P(Y_1 = 1) = P(Y_2 = 1) = p.$$  

(3.13)

There are four possible values of $(Y_1, Y_2)$: $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$. This makes three possible values of $\bar{Y}_n = (Y_1 + Y_2)/2$: 0, 1/2, or 1. The corresponding probabilities can be calculated using (3.13) since $Y_1 \perp \perp Y_2$ implies $P((Y_1, Y_2) = (a, b)) = P(Y_1 = a) P(Y_2 = b)$. (That is, due to independence, the probability that both $Y_1 = a$ and $Y_2 = b$
3.5. SAMPLING DISTRIBUTION OF AN ESTIMATOR

equals the product of the individual probabilities.) Thus,

\[ P(\bar{Y}_n = 0) = P(Y_1 = 0 \text{ and } Y_2 = 0) = P(Y_1 = 0)P(Y_2 = 0) = (1 - p)^2, \]

\[ P(\bar{Y}_n = 1) = P(Y_1 = 1 \text{ and } Y_2 = 1) = P(Y_1 = 1)P(Y_2 = 1) = p^2, \]

\[ P(\bar{Y}_n = 1/2) = 1 - P(\bar{Y}_n = 0) - P(\bar{Y}_n = 1) = 1 - (1 - p)^2 - p^2 = 2p(1 - p), \]  

(3.14)

To make (3.14) more concrete, imagine again \( p = 0.7 \) and 100 randomly sampled datasets, i.e., 100 randomly sampled pairs of \((Y_1, Y_2)\).

With \( p = 0.7 \), \( (1 - p)^2 = 0.09 \), \( p^2 = 0.49 \), and \( 2p(1 - p) = 0.42 \). If we compute \( \bar{Y}_2 = (Y_1 + Y_2)/2 \) for each of the 100 datasets (pairs), then there are approximately 9 with \( \bar{Y}_2 = 0 \), 49 with \( \bar{Y}_2 = 1 \), and 42 with \( \bar{Y}_2 = 1/2 \). That is, the sampling distribution of \( \bar{Y}_n \) shows us how frequently each possible value occurs in the long-run (when repeatedly sampling many datasets from the same population).

With larger \( n \) and/or non-binary \( Y \) and/or estimators more complicated than the sample mean, the calculations become much more complex. Further, they depend on knowing the true population distribution of the \( Y_i \), which in practice is unknown. Consequently, the sampling distribution is often approximated, as in Section 3.6.

Discussion Question 3.1 (probability of positive mean). After seeing the data, you want to know the probability that the true mean is strictly positive, \( E(Y) > 0 \). Does the frequentist sampling distribution help? If yes, explain how; if no, explain why not. Hint: recall Section 3.1.

3.5.2 Graphs: Binary Population

Consider a binary population with \( p = 0.7 \), and datasets with \( n \) observations sampled iid.

Figure 3.1 graphs the sampling distribution (PMF) of \( \bar{Y}_n \) for various \( n \). That is, for a given \( n \), \( \bar{Y}_n \) is a random variable, whose PMF is shown. The horizontal axis shows possible values of \( \bar{Y}_n \). The vertical axis can be interpreted in two ways. First, it shows the probability of each possible value as a percentage; e.g., if the bar at horizontal value 0.5 has height 42, then \( P(\bar{Y}_n = 0.5) = 42\% \). Second, you could imagine randomly sampling 100 datasets, and the vertical axis shows the number of datasets in which a particular value of \( \bar{Y}_n \) occurs.

Figure 3.1 first shows the sampling distribution of \( \bar{Y}_1 = Y_1 \) (the \( n = 1 \) graph) and then the sampling distribution of \( \bar{Y}_2 \) from (3.14). Figure 3.1 then shows the sampling distribution of \( \bar{Y}_n \) for larger \( n \) values that would be very tedious to compute by hand.

Figure 3.1 helps show how \( \bar{Y}_n \) can be seen as a random variable with a probability distribution, i.e., its sampling distribution. For different \( n \), different values are possible. For different \( n \), each value has a different probability of occurring. That is, if we randomly sample many datasets from the same population with the same sample size \( n \),
some datasets have one value of $\bar{Y}_n$, while other datasets have another, or another, and each possible value of $\bar{Y}_n$ occurs with the probability shown in the graphs.

Figure 3.1 also shows two interesting patterns when comparing small $n$ and larger $n$. First, with larger $n$, the sampling distribution of $\bar{Y}_n$ is more closely concentrated around the mean $E(Y) = 0.7$. In the extreme when $n = 1$, there can only be $\bar{Y}_1 = 0$ or $\bar{Y}_1 = 1$, neither of which is very close to the mean 0.7. With larger $n$, even if $\bar{Y}_n = 0.7$ exactly is unlikely (or impossible), the probability of being close to 0.7 is higher. For example, the probability of $0.6 \leq \bar{Y}_n \leq 0.8$ is zero when $n = 1$, but it is relatively high for the largest $n$. Second, the shape of the sampling distribution differs for large and small $n$. With $n = 1$, $\bar{Y}_n$ has a Bernoulli (binary) sampling distribution. With the largest $n$ shown, although it is still discrete, the shape looks like the “bell curve” shape of a normal distribution’s PDF. Both these observations are formalized in Section 3.6.

### 3.5.3 Graphs: Continuous Population

Now consider a continuous variable whose population distribution is uniformly distributed over all real (decimal) numbers between 0 and 1. There is again iid sampling, so $Y \sim iid \ Unif(0, 1)$.

Figure 3.2 again shows the sampling distribution of $\bar{Y}_n$, but now it is a PDF instead of PMF since $\bar{Y}_n$ is a continuous random variable (see Section 2.3.5). The horizontal axis again shows possible values of $\bar{Y}_n$. The vertical axis shows the probability density. The area under
the PDF over any range of horizontal values shows the corresponding probability of that range, as in Figure 2.3.

Figure 3.2 shows that for any \( n \), different values of \( \bar{Y}_n \) are possible given different datasets. Some datasets are more likely than others, so some ranges of \( \bar{Y}_n \) values are more likely than others.

Figure 3.2 shows two patterns when comparing the graphs with small \( n \) and larger \( n \). First, the distribution is more spread out with small \( n \), and more concentrated around the mean \( E(Y) = 0.5 \) for larger \( n \). For example, consider \( P(0.45 \leq \bar{Y}_n \leq 0.55) \), the area under the PDF between horizontal values 0.4 and 0.6. This probability is positive even with \( n = 1 \), but relatively small (it is only 20%, i.e., the area between 0.4 and 0.6 is only 20% of the total area under the PDF in the \( n = 1 \) graph). With the largest \( n \), it is relatively high. Second, the shape differs by \( n \). With \( n = 1 \), the PDF is flat, reflecting values uniformly spread between 0 and 1. With \( n = 2 \), there is a single peak at the population mean \( E(Y) = 0.5 \), but the PDF has straight lines and a sharp corner. With larger \( n \), the PDF looks like the “bell curve” shape of a normal distribution’s PDF. See Section 3.6 for more.

### 3.5.4 Table: Values in Repeated Samples

Table 3.1 records values and events across 100 datasets randomly sampled from the same population. The population is discrete, with \( P(Y = j) = 1/5 \) for \( j = -0.2, -0.1, 0, 0.1, 0.2 \). Sampling is iid, so each \( Y_i \) has the same distribution as the population \( Y \), and all \( Y_i \) are mutually independent. Let \( n = 10 \). The population mean is
Table 3.1: Example estimates and event probabilities.

<table>
<thead>
<tr>
<th>Sample</th>
<th>$\bar{Y}_n$</th>
<th>$1{\bar{Y}_n \leq 0}$</th>
<th>$1{\bar{Y}_n - 0.4 \leq 0 \leq \bar{Y}_n + 0.4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>0.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#2</td>
<td>0.20</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>#3</td>
<td>0.00</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>#4</td>
<td>-0.10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>#5</td>
<td>-0.50</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>#100</td>
<td>0.30</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Average 0.01 52/100 67/100

**Note:** $P(Y = j) = 0.2$ for $j = -2, -1, 0, 1, 2$, iid, $n = 10$.

Table 3.1 shows the value of $\bar{Y}_n$ computed from each sample (dataset). It shows that $\bar{Y}_n = 0.5$ in the first sample, $\bar{Y}_n = 0.2$ in the second sample, etc. This reflects the sampling distribution. A histogram of these values (not shown) would produce a graph with similar interpretation to Figures 3.1 and 3.2.

Table 3.1 shows for each sample whether or not the sample mean $\bar{Y}_n$ is less than or equal to the population mean $E(Y) = 0$, in the column labeled with $1\{\bar{Y}_n \leq 0\}$. That is, 1 indicates that it does, 0 indicates that it doesn’t. For example, in Sample #1, $\bar{Y}_n = 0.5$, which is not negative, so $1\{\bar{Y}_n \leq 0\} = 0$. In Sample #4, $\bar{Y}_n = -0.1$, which is negative, so $1\{\bar{Y}_n \leq 0\} = 1$. From the frequentist view, the event $\bar{Y}_n \leq E(Y)$ is “random” in that it could occur or not occur, with some probability for each possibility. The $E(Y)$ is non-random, but $\bar{Y}_n$ is random, hence the event is random. The event’s probability is the probability of randomly sampling a dataset in which the event occurs. The bottom row of the table says the event occurred 52 times out of 100 samples (52% of the time). Since there are only 100 samples and not $\infty$, this is not the exact probability, but it reflects that the event occurs slightly more than half the time.

Table 3.1 also shows for each sample whether or not the random interval $[\bar{Y}_n - 0.4, \bar{Y}_n + 0.4]$ contains the population mean $E(Y) = 0$, i.e., whether or not $\bar{Y}_n - 0.4 \leq E(Y) \leq \bar{Y}_n + 0.4$. The interval is “random” in the frequentist sense that it has different possible probabilities in different datasets (since it depends on $\bar{Y}_n$). In Sample #1, the interval does not contain $E(Y)$: $\bar{Y}_n = 0.5$, so the interval is $[0.5 - 0.4, 0.5 + 0.4] = [0.1, 0.9]$, which does not contain $E(Y) = 0$. In Sample #2, the interval does contain $E(Y)$: $\bar{Y}_n = 0.2$, so the interval is $[-0.2, 0.6]$, which contains $E(Y) = 0$. The bottom row of the table says this event occurred 67 times out of 100 samples (67% of the time). This is the “coverage probability” of a “confidence interval” described in Section 3.8.2.
3.6 Sampling Distribution Approximation

Because of the difficulties mentioned in Section 3.5, usually an estimator’s sampling distribution is approximated.

Most common is a particular type of approximation called an asymptotic approximation. All else equal, this type of approximation is better when $n$ is larger. For example, comparing Figures 3.1 and 3.2 with (3.15) below, the approximation is very bad with $n = 1$, but very good for the largest $n$ in each figure. Unfortunately, there is no general magic threshold for $n$ because the approximation’s accuracy also depends on certain (unknown) population features. In practice, people usually just hope $n$ is large enough that the approximation is reasonable.

With iid sampling, the approximate distribution of $\bar{Y}_n$ is

$$\bar{Y}_n \sim N(\mu_Y, \sigma_Y^2 / n),$$

where the $a$ over $\sim$ stands for “approximately” (or “asymptotically”), $\mu_Y = E(Y)$ is the population mean, and $\sigma_Y^2 \equiv \text{Var}(Y)$ is the population variance. This reflects the three patterns seen in the examples in Sections 3.5.2 and 3.5.3. First, the standard deviation is $\sigma_Y / \sqrt{n}$, which (given the same $\sigma_Y$) is smaller for larger $n$. Second, the mean of the distribution is the population mean. Third, the shape is normal (Gaussian). A normal approximation of a sample mean’s sampling distribution is often called a central limit theorem (CLT).

For technical mathematical reasons, you may often see a variation of (3.15):

$$\sqrt{n}(\bar{Y}_n - \mu_Y) \sim N(0, \sigma_Y^2),$$

Practically, this says the same thing as (3.15); it just moves the $\mu_Y$ and $n$ to the left-hand side from the right-hand side.

In practice, $\sigma_Y^2$ is unknown but can be estimated from data (i.e., the sample variance).

Figure 3.3 shows the same sampling distribution (PDF) from Figure 3.2, along with the normal approximation. The horizontal axis has been rescaled to see the shape more easily; the horizontal values are like the left-hand side of (3.16). The approximation is bad with $n = 1$ but very good for the largest $n$.

3.6.1 Non-iid Sampling

With non-iid sampling (see Section 3.2), an estimator’s sampling distribution is different, so the approximation also differs. Usually the sampling distribution is still normal, but with a different standard deviation. In practice, ideally you should understand the type of sampling enough to know which R functions can provide an appropriate approximation. For now, the goal is only to be able to assess the type of sampling and understand that non-iid sampling leads to a different sampling distribution.
Figure 3.3: Sampling distribution (solid PDF) of $\sqrt{n}(\bar{Y}_n - \mu_Y)$ with normal approximation (dashed).

### 3.7 Quantifying Accuracy of an Estimator

From the frequentist perspective, an estimator’s accuracy can be quantified by comparing features of its sampling distribution to the true population value. Bias is an important, commonly mentioned property, but it is not sufficient. Mean squared error better quantifies accuracy. Bias and mean square error are finite-sample properties that derive from the estimator’s sampling distribution for a finite sample size $n$. Approximate ("large-sample" or "asymptotic") versions of these properties are also discussed below.

Throughout, let $\theta$ be the population parameter estimated by $\hat{\theta}_n$. This includes $\theta = \text{E}(Y)$ and $\hat{\theta}_n = \bar{Y}_n$ but is more general.

#### 3.7.1 Bias

Recall from Section 3.5 the frequentist perspective that an estimator is a random variable whose probability distribution is called its sampling distribution. The sampling distribution differs with $n$, hence the subscript $n$ on the estimator $\hat{\theta}_n$.

**Definitions**

The bias of $\hat{\theta}_n$ compares the mean of its sampling distribution to the true population $\theta$. Mathematically,

$$\text{Bias}(\hat{\theta}_n) \equiv \text{E}(\hat{\theta}_n) - \theta. \quad (3.17)$$
The bias captures if the estimator systematically differs from \( \theta \) in a particular direction, i.e., how wrong the average \( \hat{\theta}_n \) is.

There are four types of bias:

- **upward bias (positive bias):** \( E(\hat{\theta}_n) > \theta \),
- **downward bias (negative bias):** \( E(\hat{\theta}_n) < \theta \),
- **attenuation bias (bias toward zero):** \( 0 < \frac{E(\hat{\theta}_n)}{\theta} < 1 \), so \( |E(\hat{\theta}_n)| < |\theta| \),
- **bias away from zero:** \( \frac{E(\hat{\theta}_n)}{\theta} > 1 \), so \( |E(\hat{\theta}_n)| > |\theta| \).

An estimator is **unbiased** if its bias is zero. Using (3.17),

\[
\text{Bias}(\hat{\theta}) = 0 \iff E(\hat{\theta}) = \theta, \tag{3.18}
\]

where symbol \( \iff \) can be read as “is equivalent to” (see Section 6.1).

For example, with iid sampling, the sample mean is unbiased. With \( n = 1 \), \( \bar{Y}_1 = Y_1 \), so \( E(\bar{Y}_1) = E(Y_1) = \mu_Y \). With \( n = 2 \),

\[
E[\bar{Y}_2] = E[(1/2)Y_1 + (1/2)Y_2] = (1/2)E(Y_1) + (1/2)E(Y_2) = \mu_Y, \tag{3.19}
\]

using the linearity property of \( E(\cdot) \) from (2.21). Similar derivations hold for any \( n \), so \( E(\bar{Y}_n) = \mu_Y \), thus the bias is zero given (3.18).

### Insufficiency of Bias to Quantify Accuracy

Bias alone does not fully quantify accuracy. That is, if you only consider bias when choosing between two possible estimators, then you may be fooled into choosing the worse estimator.

Let \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) be two estimators of unknown parameter \( \theta \). Here, the subscripts 1 and 2 do not indicate \( n \) but just that the estimators are different. For simplicity, let \( \theta = 0 \). The first estimator’s distribution is

\[
P(\hat{\theta}_1 = -100) = P(\hat{\theta}_1 = 100) = 1/2. \tag{3.20}
\]

The second estimator’s distribution is

\[
P(\hat{\theta}_2 = 1) = 1. \tag{3.21}
\]

The first estimator has smaller bias. The mean of each estimator is

\[
E(\hat{\theta}_1) = (1/2)(-100) + (1/2)(100) = 0, \quad E(\hat{\theta}_2) = (1)(1) = 1. \tag{3.22}
\]

Thus, recalling \( \theta = 0 \), the bias of each estimator is

\[
\text{Bias}(\hat{\theta}_1) = E(\hat{\theta}_1) - \theta = 0 - 0 = 0, \quad \text{Bias}(\hat{\theta}_2) = E(\hat{\theta}_2) - \theta = 1 - 0 = 1. \tag{3.23}
\]

Estimator \( \hat{\theta}_1 \) is unbiased, while \( \hat{\theta}_2 \) has upward bias.

But intuitively, \( \hat{\theta}_2 \) is much better. It is always wrong by 1, but \( \hat{\theta}_1 \) is always wrong by 100, which is much worse. Regardless of the dataset, \( \hat{\theta}_2 \) is 100 times closer to the true \( \theta = 0 \). Bias alone does not properly quantify our preferences: it tells us to prefer \( \hat{\theta}_1 \) when in fact we strongly prefer \( \hat{\theta}_2 \).
3.7.2 Mean Squared Error

The **mean squared error** (MSE) is a more complete measure of “how bad” an estimator is. The idea is analogous to using quadratic loss for prediction (e.g., Sections 2.5.1 and 2.5.4). Among other possible loss functions, this is most common and is generally reasonable. MSE is mean quadratic loss:

\[ \text{MSE}(\hat{\theta}) \equiv \mathbb{E}[L_2(\hat{\theta}, \theta)] = \mathbb{E}[(\hat{\theta} - \theta)^2]. \]  
(3.24)

Continuing the example, our intuitive preference for \( \hat{\theta}_2 \) over \( \hat{\theta}_1 \) is supported by MSE. Since MSE measures “how bad” an estimator is, \( \hat{\theta}_2 \) being “better” means it has lower MSE. Specifically,

\[ \text{MSE}(\hat{\theta}_1) = \mathbb{E}[(\hat{\theta}_1 - \theta)^2] = (1/2)(-100 - 0)^2 + (1/2)(100 - 0)^2 = 10,000, \]
\[ \text{MSE}(\hat{\theta}_2) = \mathbb{E}[(\hat{\theta}_2 - \theta)^2] = (1)(1 - 0)^2 = 1. \]  
(3.25)

This matches our intuition: \( \hat{\theta}_2 \) is much better than \( \hat{\theta}_1 \) because it has much lower MSE.

MSE can also be decomposed into variance plus squared bias. The variance can also be seen as the squared “standard error” (see Section 3.8.1). Skipping the algebra,

\[ \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var} \hat{\theta} + [\text{Bias} \hat{\theta}]^2. \]  
(3.26)

All else equal, larger bias is bad, but it’s also bad to have very high and very low estimates across datasets (large variance and “standard error”) even if they happen to average to \( \theta \).

Other MSE Examples

More generally, instead of assuming \( \theta = 0 \), let

\[ P(\hat{\theta}_1 = \theta - 100) = P(\hat{\theta}_1 = \theta + 100) = 1/2, \quad P(\hat{\theta}_2 = \theta + 1) = 1. \]  
(3.27)

The MSEs are the same as before since the \( \theta \) cancels out:

\[ \text{MSE}(\hat{\theta}_1) = \mathbb{E}[(\hat{\theta}_1 - \theta)^2] = (1/2)(\theta - 100 - \theta)^2 + (1/2)(\theta + 100 - \theta)^2 = 1 \]
\[ \text{MSE}(\hat{\theta}_2) = \mathbb{E}[(\hat{\theta}_2 - \theta)^2] = (1)(\theta + 1 - \theta)^2 = 1. \]  
(3.28)

As another example, imagine we know the bias and variance of two estimators, but not the full sampling distributions. This is still sufficient to compute MSE using (3.26). For example, let

\[ \text{Bias}(\hat{\beta}_1) = 1, \text{Var}(\hat{\beta}_1) = 16, \quad \text{Bias}(\hat{\beta}_2) = 10, \text{Var}(\hat{\beta}_2) = 9. \]  
(3.29)

Plugging these into (3.26),

\[ \text{MSE}(\hat{\beta}_1) = 1^2 + 16 = 17, \quad \text{MSE}(\hat{\beta}_2) = 10^2 + 9 = 109. \]  
(3.30)

According to MSE, \( \hat{\beta}_1 \) is better because it has lower MSE (“less bad”) than \( \hat{\beta}_2 \). In this case, although \( \hat{\beta}_1 \) has larger variance, its bias is enough smaller than its overall MSE is also smaller.
Practice 3.3 (estimator MSE). Consider three estimators of the population mean $\mu = E(Y)$, and their three sampling distributions: $\hat{\mu}_1 \sim N(\mu, 25)$, $\hat{\mu}_2 \sim N(\mu + 3, 16)$, and $\hat{\mu}_3 \sim N(\mu + 2, 9)$, i.e., all normal distributions with respective means $\mu$, $\mu + 3$, and $\mu + 2$, and respective variances 25, 16, and 9. (Hint: for MSE, does it matter than the distributions are normal?)

a) Compute the MSE of each estimator.

b) Rank the three estimators from best to worst, in terms of MSE.

3.7.3 Consistency

Analogous to how bias compares the sampling distribution’s mean to the true population $\theta$, consistency compares the approximate (asymptotic) sampling distribution’s mean to $\theta$. However, in addition to the approximate distribution having mean $\theta$, consistency also requires that the approximate distribution’s standard deviation is smaller for larger $n$.

For example, consider the approximate distribution of the sample mean $\bar{Y}_n$ in (3.15). The mean is $\mu_Y$. Further, the standard deviation is proportional to $1/\sqrt{n}$ (since the variance is proportional to $1/n$), i.e., smaller for larger $n$. Thus, $\bar{Y}_n$ is consistent.

Visually, the consistency of $\bar{Y}_n$ was seen in Figures 3.1 and 3.2. In each case, with larger $n$, the sampling distribution concentrated probability around $\mu_Y$.

Intuitively, consistency means that in “large” samples (large $n$), there is a “high” probability of the estimator being “close” to the true value. This is similar to the idea of “probably approximately correct” in computer science: estimator $\hat{\theta}_n$ is “consistent” if with large $n$ it is “probably approximately correct.” Unfortunately, there are usually no precise, quantitative definitions of “large,” “high,” and “close.” Still, the qualitative idea is that the sampling distribution of $\hat{\theta}_n$ is converging to the true $\theta$ if we imagine larger and larger $n$, which is represented notationally by

$$\hat{\theta}_n \xrightarrow{p} \theta \text{ as } n \to \infty \text{ or } \lim_{n \to \infty} \hat{\theta}_n = \theta.$$  \hspace{1cm} (3.31)

If $\hat{\theta}_n$ is not consistent, its asymptotic bias can be defined as

$$\text{AsyBias}(\hat{\theta}_n) \equiv \lim_{n \to \infty} \hat{\theta}_n - \theta.$$  \hspace{1cm} (3.32)

(Other definitions have the same practical meaning, though the technical details differ.) Similar to “unbiasedness” being “zero bias,” here “consistency” is “zero asymptotic bias.” There are the same four types of asymptotic bias as bias (upward, downward, attenuation, away from zero).

3.7.4 Asymptotic MSE

It is possible to compare approximate (asymptotic) MSE, but details are omitted.
3.8 Quantifying Uncertainty: Conventional Frequentist Approaches

The point estimates in Section 3.4 provide our best guesses about unknown population values, but they offer no sense of our uncertainty. Here, we consider only statistical uncertainty (or sampling uncertainty), i.e., the uncertainty due to observing only a random sample of data instead of knowing the true population distribution. Although the term is ambiguous, inference often refers to the types of methods in this section (i.e., statistical methods other than point estimation).

This section concerns only the conventional frequentist approaches to quantifying uncertainty. Section 3.9 provides warnings about misinterpretation and misuse.

The general consensus among econometricians and statisticians is that confidence intervals are more informative and easier to interpret than p-values and hypothesis tests. You should focus on confidence intervals when producing your own empirical analysis, but you may need to understand p-values and hypothesis tests to understand others.

3.8.1 Standard Errors

Empirical economics results almost always report standard errors alongside estimates. Standard errors are commonly used to compute confidence intervals, as well as p-values and hypothesis tests.

Definition and Terminology

The standard error (SE) of estimator \( \hat{\theta} \) the standard deviation of its sampling distribution:

\[
SE(\hat{\theta}) \equiv \sqrt{\text{Var}(\hat{\theta})}. \tag{3.33}
\]

Recall from Section 2.3 that the standard deviation has the same units as the variable itself, so the SE has the same units as \( \hat{\theta} \).

Unfortunately, people may say “SE” to refer to either (3.33) or an estimate of (3.33), so its meaning is ambiguous. In this textbook at least, “estimated SE” and notation \( \hat{SE}(\hat{\theta}) \) refer to an estimate of (3.33). Causing yet more confusion, \( \hat{SE}(\bar{Y}_n) \) is often called the standard error of the mean, whereas personally I would call it the estimated standard error of the sample mean.

Interpretation

The SE helps quantify uncertainty due to random sampling, or “statistical uncertainty.” Larger SE means more uncertainty.

For example, consider estimators \( \hat{\theta}_1, \hat{\theta}_2, \) and \( \hat{\theta}_3 \) that all estimate \( \theta \). If \( \hat{\theta}_1 = \theta \), then \( SE(\hat{\theta}_1) = 0 \), reflecting zero uncertainty. If \( P(\hat{\theta}_2 = \theta) = 0 \)
\[ \theta + 1 = P(\hat{\theta}_2 = \theta - 1) = 1/2, \text{ then (skipping the math) } \text{SE}(\hat{\theta}_2) = 1. \]

If \( P(\hat{\theta}_2 = \theta + 10) = P(\hat{\theta}_2 = \theta - 10) = 1/2, \text{ then SE}(\hat{\theta}_2) = 10. \)

Unfortunately, it’s possible to be very certain about the wrong value. Consider the very bad estimator \( \hat{\theta} = 4. \) Since it’s a constant, \( \text{SE}(\hat{\theta}) = 0. \) This may appear to be great (no uncertainty!), but we should feel very uncertain about the methodology of “just guess 4.” Our uncertainty about appropriate methodology is not captured by the SE.

### 3.8.2 Confidence Intervals

A **confidence interval** (CI) helps quantify statistical uncertainty, with a longer CI indicating more statistical uncertainty. A CI does not capture any other source of uncertainty, so small values can be misleading if there is still uncertainty about certain assumptions or methodological choices.

Essentially, a CI is a range of values that tries to include the true population value with high probability, like 90% or 95%. Again, “probability” means frequentist probability from the “before” view, or equivalently over many repeated samples from the same population, like in Table 3.1.

For example, recall the last column in Table 3.1. In each of 100 random samples, it showed whether or not the true mean \( E(Y) = 0 \) was inside the interval \([\bar{Y}_n - 0.4, \bar{Y}_n + 0.4]\). This CI contained the true population mean in 67 of the 100 datasets. From the “before” view, the probability of randomly sampling a dataset in which the CI contains the true value is around 67%.

A 90% CI does not mean, “I believe there’s a 90% chance that the true value is in this range.” That is the interpretation of a Bayesian credible interval; see Section 3.1.

The actual probability that a CI contains the true value often differs from the desired probability. In practice, when you ask R to compute a CI, you specify your desired probability (like 90% or 95%), called the **confidence level** or **nominal coverage probability** (or “nominal level” or other variations). The actual probability is the **coverage probability**. There are three possibilities.

1. Ideally, a CI’s coverage probability is close to the nominal level.
2. Sometimes, a CI is too long and has coverage probability above what you requested. This is bad because it does not help you narrow down the possible values of the population parameter well (since the CI is longer than necessary).
3. Sometimes, a CI is too short and has coverage probability below what you requested, as low as 80%, 50%, or even close to 0%. This is bad because you think think the true value is inside the CI, but actually in many datasets (more than you realized) the CI does not contain the true value.

The levels 90% and 95% are most common, but sometimes you may desire 99% or even higher, if it is particularly important that the
true value be in the interval (or if you have a very large sample with very short CIs).

Formally, coverage probability is defined as follows. Consider a **two-sided** confidence interval of the form $[\hat{L}, \hat{U}]$, where the hats remind us that the interval endpoints are computed from the data. For example, Table 3.1 had $\hat{L} = \bar{Y}_n - 0.4$ and $\hat{U} = \bar{Y}_n + 0.4$. A **one-sided** confidence interval would set $\hat{L} = -\infty$ or $\hat{U} = \infty$. The coverage probability of this CI for the parameter $\theta$ is

$$P(\text{CI contains true value}) = P(\theta \in [\hat{L}, \hat{U}]) = P(\hat{L} \leq \theta \leq \hat{U}). \quad (3.34)$$

Sometimes a CI is written in terms of a **critical value**. The critical value depends on the confidence level and comes from the standard normal distribution $N(0, 1)$. It is used when an estimator’s approximate sampling distribution is normal (Gaussian). For example, if $c$ is the critical value for a two-sided 95% CI, and $\hat{SE}$ is the estimated standard error of estimator $\hat{\theta}$, then the conventional CI is $[\hat{\theta} - c\hat{SE}, \hat{\theta} + c\hat{SE}]$.

**Practice 3.4** (CI interpretation). Imagine you have a CI with 95% nominal coverage probability for the true $\theta$, $[1.4, 2.9]$.

a) Explain why this does **not** mean, “I think there’s a 95% chance that $1.4 \leq \theta \leq 2.9$.”

b) Explain why it’s still possible that the true value is $\theta = 0$.

c) Explain why if the true coverage probability is also 95% and you had 99 other randomly sampled datasets, then around 95 of the 100 total datasets would have a CI containing the true $\theta$.

d) Despite the 95% confidence level, imagine the actual coverage probability of your CI is 75%. Would a CI with actual 95% coverage probability be longer or shorter than yours? Explain.

**Example in R**

The following R example constructs two-sided 95% confidence intervals for the mean, from simulated iid standard normal data (so the true population mean is zero). One CI uses `t.test()`, a standard $t$-test; the other CIs use nonparametric bootstrap methodology from the `boot` package, though details are beyond our scope.

```r
library(boot)
set.seed(112358) #for replicability
Y <- rnorm(n=50, mean=0, sd=1) # iid N(0,1)
CIttest <- t.test(x=Y, conf.level=0.95, alternative='two.sided')$conf.int
ret <- boot(data=Y, statistic=function(x,i) mean(x[i]), R=10)
tmp <- boot.ci(boot.out=ret, conf=0.95, type=c('basic', 'bca'))
out.table <- rbind(CIttest, tmp$basic[4:5], tmp$bca[4:5])
rownames(out.table) <- c('Normality', 'Boot.basic', 'Boot.BCa')
colnames(out.table) <- c('Lower', 'Upper')
print(round(out.table, digits=3))
```
3.8.3 \( p \)-values

Frequentist \( p \)-values are precisely defined, strange, common, and commonly misunderstood.

A \( p \)-value measures how unlikely the observed data would be if a certain hypothesis were true. Notationally, a \( p \)-value is conventionally just denoted as \( p \), but since it is computed from data, I usually write \( \hat{p} \) (with a hat) for clarity. The range is \( 0 \leq \hat{p} \leq 1 \). Small values closer to \( \hat{p} = 0 \) indicate such a dataset would be unlikely if the hypothesis were true. In that case, either the hypothesis is true and we just happened by chance to observe an unlikely dataset, or else the hypothesis is false.

For example, consider the \( p \)-value for the hypothesis \( H_0: \mu_Y = 0 \). Values near \( \hat{p} = 0 \) indicate that the observed dataset would be unlikely if actually \( \mu_Y = 0 \). As usual, \( \bar{Y}_n \) is a random variable. Here, the observed sample mean \( \bar{Y}_o \) is treated as non-random. The \( p \)-value is then

\[
\hat{p} = P(\left| \bar{Y}_n \right| \geq \left| \bar{Y}_o \right| | \mu_Y = 0). \tag{3.35}
\]

That is, the \( p \)-value is the probability of observing a sample mean even farther away from zero than the observed sample mean, given a population with \( \mu_Y = 0 \).

More generally,

\[
\hat{p} = P(\text{estimate magnitude at least as big as observed} | H_0 \text{ is true}). \tag{3.36}
\]

3.8.4 Statistical Significance

Results with low \( p \)-values are often called statistically significant, or having statistical significance. These terms are usually used when trying to estimate an effect (or difference), where the null hypothesis is zero effect. A statistically significant effect estimate means the \( p \)-value is low, meaning an estimate that large would be unlikely if the true effect were zero, which provides some evidence of a non-zero effect.

Conceptually, statistical significance is not a yes/no property, but a continuum; i.e., not “if” but “how much?” Results can be somewhat statistically significant, or extremely statistically significant, or lacking statistical significance, etc. Confusingly, a lower \( p \)-value means greater statistical significance.

In practice, often people say a result is statistically significant at a particular level. For example, if the \( p \)-value is below 0.05, then the result is “statistically significant at a 5% level”; if below 0.01, then

<table>
<thead>
<tr>
<th>Method</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normality</td>
<td>-0.213</td>
<td>0.370</td>
</tr>
<tr>
<td>Boot.basic</td>
<td>-0.234</td>
<td>0.387</td>
</tr>
<tr>
<td>Boot.BCa</td>
<td>-0.233</td>
<td>0.388</td>
</tr>
</tbody>
</table>
it is “statistically significant at a 1% level”; etc. Generally, there is statistical significance at a 100c% level if the p-value is below c.

Why 5%? Indeed, 5% is arbitrary. Its origin seems to be from Ronald Fisher, who wrote in 1926, “We shall not often be astray if we draw a conventional line at 0.05.” Recently, 72 prominent researchers from many fields (including statistics, econometrics, and economics) wrote a piece simply titled, “Redefine statistical significance” (Benjamin, Berger, Johannesson, Nosek, Wagenmakers, Berk, Bollen, Brembs, Brown, Camerer, Cesarini, Chambers, Clyde, Cook, De Boeck, Dienes, Dreber, Easwaran, Efferson, Fehr, Fidler, Field, Forster, George, Gonzalez, Goodman, Green, Green, Greenwald, Hadfield, Hedges, Held, Ho, Hoijtink, Hruschka, Imai, Imbens, Ioannidis, Jeon, Jones, Kirchler, Laibson, List, Little, Lupia, Machery, Maxwell, McCarthy, Moore, Morgan, Munafó, Nakagawa, Nyhan, Parker, Pericchi, Perugini, Rouder, Rousseau, Savalei, Schönbrodt, Sellke, Sinclair, Tingley, Van Zandt, Vazire, Watts, Winship, Wolpert, Xie, Young, Zinman, and Johnson, 2018). The suggestion was to reduce the conventional level for statistical significance from 5% to 0.5%. Indeed, it is already (much) lower in some fields like genetics and high-energy physics. However, they also agree that there may be very important empirical results with \( p = 0.05 \) or even larger. They simply advocate calling such results “suggestive evidence” rather than treating them as conclusive. They also note that it may be better to focus on confidence intervals than statistical significance.

3.8.5 Hypothesis Testing

In the scientific method, theories imply certain hypotheses that can be tested empirically (with data). A theory is maintained until it is disproved; then a new theory replaces it and is tested, etc.

In economics, more often hypothesis testing is used like the \( p \)-value, to provide evidence against a statement like “the true effect is zero.”

Notation and Terminology

Notationally, \( H_0 \) denotes the null hypothesis,\(^1\) while \( H_1 \) (sometimes \( H_a \)) denotes the alternative hypothesis. A specific null hypothesis is a statement about a parameter, written after a colon, like \( H_0 : \mu_Y = 0 \). The alternative hypothesis is usually just that the null is false, like \( H_1 : \mu_Y \neq 0 \), so they are mutually exclusive (cannot both be true).

Ostensibly, the goal of hypothesis testing is use data to decide whether \( H_0 \) or \( H_1 \) is true, but many caveats apply.

Much jargon accompanies hypothesis testing. A hypothesis test has two possible results: either reject \( H_0 \), or do not reject \( H_0 \). (“Fail to reject” means the same as “do not reject”; it does not reflect a “failure” in the colloquial English sense.) Sometimes “do not reject” is replaced by “accept,” but “do not reject” emphasizes the asymmetry between

\(^1\)Pedagogical criticism duly noted: xkcd.com/892
Table 3.2: Terms for hypothesis test outcomes.

<table>
<thead>
<tr>
<th>$H_0$ true</th>
<th>don’t reject $H_0$</th>
<th>reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>correct</td>
<td>type I error (false positive)</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>type II error</td>
<td>correct</td>
</tr>
</tbody>
</table>

$H_0$ and $H_1$. That is, although rejection of $H_0$ indicates evidence against it, lack of rejection only indicates a lack of evidence against $H_0$, not necessarily strong evidence supporting it. The test’s result is either correct or an error, with terms shown in Table 3.2. Related probabilities are:

- **rejection probability**: $P(\text{reject } H_0)$
- **type I error rate**: $P(\text{reject true } H_0)$
- **type II error rate**: $P(\text{don’t reject false } H_0)$
- **power**: $P(\text{reject false } H_0)$.

There is yet more jargon. When $H_0$ includes multiple values of a parameter $\theta$, the largest possible type I error rate is called the **size** of the test. When $H_0$ is just a single value like $H_0: \theta = 0$, then the size is just the type I error rate. A test’s **level** is what it claims to be the maximum type I error rate (size). Like a confidence interval’s nominal coverage probability, the level could be above, below, or equal to the true type I error rate. Usually it is close for large $n$, but it may be very different with small $n$, and there is (as usual) no quantitative threshold for “large.”

**Practice 3.5** (testing terms). State the technical term(s) associated with each of these.

a) Your hypothesis test did not reject the **permanent income hypothesis** even though it’s false.

b) For a type of lab experiment where people do not behave according to expected utility maximization, when repeatedly running the experiment on different randomly sampled groups of people, your hypothesis test rejects the null hypothesis of expected utility maximization 80% of the time.

c) You want to see if there is enough empirical evidence to reject the **efficient market hypothesis**.

d) Your hypothesis test rejected the **permanent income hypothesis** even though it’s true.

e) In your test of $H_0: \theta \leq 0$, the type I error rate is very low when $\theta$ is very negative, but the type I error rate can be as high as 7% when $\theta = 0$.

**Computation**

Computationally, the hypothesis test for $H_0: E(Y) = 0$ can be computed using the $p$-value. In this sense, the test is strictly less informative than the $p$-value: the $p$-value takes any number between 0 and 1, whereas the test can only reject or not. Specifically, the level $\alpha$
test rejects when \( p \leq \alpha \), so the test essentially just reports whether \( 0 \leq p \leq \alpha \) or \( p > \alpha \). In fact, the function `t.test()` in R does not even report “reject” or “do not reject”: it instead reports a \( p \)-value.

Alternatively, a hypothesis test compares a test statistic to a critical value. This is the same critical value from Section 3.8.2 that depends on the level \( \alpha \) and the standard normal distribution \( N(0,1) \), for use when the estimator is approximately normal. For example, if \( c \) is the two-sided level 5% critical value and \( t \) is the \( t \)-statistic, then the test rejects \( H_0 \) when \( |t| > c \). In R, for a two-sided level \( \text{ALPHA} \) test, the critical value is \( \text{qnorm}(1-\text{ALPHA}/2) \), but usually you do not need to compute it manually.

### 3.8.6 Mental Math for Statistical Uncertainty

For a quick approximation, you can use a critical value of 2. Specifically, if a point estimate’s absolute value \( |\hat{\theta}| \) is more than two SEs away from zero, then it is statistically significantly different from zero at a 5% level (actually 4.55%). That is, it is statistically significant at 5% when \( \hat{\theta} \geq 2 \hat{\text{SE}}(\hat{\theta}) \). You can also take \( \hat{\theta} \) and add and subtract two SEs to get an approximately 95% CI (actually 95.45%): the CI is \( \left[ \hat{\theta} - 2 \hat{\text{SE}}(\hat{\theta}), \hat{\theta} + 2 \hat{\text{SE}}(\hat{\theta}) \right] \).

This is useful for a few reasons. First, it’s often easy enough to do in your head. Second, 5% is arbitrary anyway, so 4.55% is equally good (or equally bad!). Third, confidence intervals and \( p \)-values are already based on (asymptotic) approximations, whose approximation error is almost always bigger than the difference between 5% and 4.55%.

Consider an example. Imagine you estimate \( \hat{\theta} = -3.2 \) and \( \hat{\text{SE}}(\hat{\theta}) = 1.5 \). Then, \( |\hat{\theta}| = 3.2 \) and \( 2 \hat{\text{SE}}(\hat{\theta}) = 3.0 \). Since \( 3.2 > 3.0 \), the result is statistically significant at a 5% level (\( p < 0.05 \)). A 95% CI is \( [-3.2 - 3.0, -3.2 + 3.0] = [-6.2, -0.2] \).

**Practice 3.6** (mental math). Let \( \hat{\theta} = -28 \), \( \hat{\text{SE}}(\hat{\theta}) = 19 \); in your head, assess statistical significance and compute an approximate 95% CI.

### 3.9 Quantifying Uncertainty: Misinterpretation and Misuse

This section addresses misinterpretations and misuse of frequentist inference. Some of the most common problems are discussed below, as well as on the (pretty good) Wikipedia page devoted to the topic.

**Practice 3.7** (frequentist or Bayesian?). For each of the following, say whether it is a frequentist question, Bayesian question, neither,

---

2For example, a famous econometrician said, “I tell my students, if you can get a 5% test that controls the actual type I error rate below 10%, that’s pretty good” (Jerry Hausman, April 6, 2019, keynote talk at Chinese Economists Society conference in Lawrence, KS).

3https://en.wikipedia.org/wiki/Misunderstandings_of_p-values
or both; if both, explain the two possible interpretations. Hint: use Section 3.1 as well as Section 3.8.

a) What’s the probability that the current natural unemployment rate in the U.S. is between 4.5% and 7.5%?

b) Can we create a diagnostic tool for our company’s daily website traffic data to identify whether it’s normal or has been hacked, limiting the rate of falsely reporting “hacked” on normal days to only 1% of normal days?

c) What is the probability that the true unemployment rate is within 1 percentage point of the estimated unemployment rate?

d) Is the positive estimate \( \hat{\theta} > 0 \) primarily due to the income effect or substitution effect?

3.9.1 Perils of Ignoring Non-iid Sampling

A CI justified by iid sampling may perform poorly when sampling is not iid. (Similar problems befall \( p \)-values and hypothesis tests.) Stratified sampling that is independent but not identically distributed (inid) is not too problematic since it tends to make coverage probability higher than requested; CIs are “too long” but still have high probability of containing the true population value. In contrast, dependent sampling can make the actual coverage probability much lower than the requested confidence level. For example, maybe you ask for a 90% CI, but the CI’s coverage probability is actually only 75%. Unfortunately, you cannot simply ask the computer for the true coverage probability, so you must carefully consider whether you have (in)dependent sampling.

Example

Consider the following example with dependent sampling, with results simulated in R below. Imagine a “time use survey” that includes a question about watching television (TV). In hours per day, \( P(Y = j) = (4 - j)/10 \) for \( j = 0, 1, 2, 3 \). Imagine each household contains two adults who only watch TV together. That is, if individuals \( i \) and \( k \) live together, then \( Y_i = Y_k \).

For comparison, imagine two samples are collected: one iid, one with clustered sampling. Sample A is collected iid, randomly sampling individuals from the population. Sample B is collected by randomly visiting households and surveying both individuals within the household. Sample A contains observations \( Y_i \) for \( i = 1, \ldots, n_A = 10 \). Sample B contains observations \( Y_i \) for \( i = 1, \ldots, n_B = 20 \), where \( Y_k \) and \( Y_{k+10} \) live together (\( k = 1, \ldots, 10 \)). Since \( Y_k = Y_{k+10} \) in Sample B, there are really only 10 observations; the other 10 are literally duplicates.

Samples A and B contain the same amount of information, so they should lead to the same amount of uncertainty (same CI). However, if (incorrectly) both are assumed iid, the larger sample size \( n_B > n_A \) is incorrectly interpreted as greater certainty and leads to incorrectly smaller CIs.
CHAPTER 3. ONE VARIABLE: SAMPLE

This TV example is shown in the following simulation. Many datasets are simulated. In each, a 90% CI is constructed for

\[ \mu_Y = E(Y) = (4/10)(0) + (3/10)(1) + (2/10)(2) + (1/10)(3) = 10/10 = 1. \]

Finally, the code reports the proportion of simulated datasets in which the CI contained the true value, i.e., the simulated coverage probability.

```r
set.seed(112358) # for replicability
vY <- 0:3  # possible values of Y
pY <- (4:1)/10  # P(Y=0), P(Y=1), P(Y=2), P(Y=3)
muY <- sum(pY*vY)  # E(Y)
na <- 10  # sample size for A; B's is twice this
cl <- 0.90  # 90% CI
nrep <- 1000  # number of simulated datasets
tmp <- data.frame(lo=rep(NA,nrep),hi=NA)
CIs <- list(A=tmp, B=tmp)  # store both CIs for each dataset
for (irep in 1:nrep) {
  sampleA <- sample(x=vY, size=na, replace=TRUE, prob=pY)
sampleB <- rep(x=sample(x=vY, size=na, replace=TRUE, prob=pY), each=2)
  CIs[["A"]][irep,] <- t.test(x=sampleA, conf.level=cl, alternative="two.sided")$conf.int
  CIs[["B"]][irep,] <- t.test(x=sampleB, conf.level=cl, alternative="two.sided")
}
CP.A <- mean(CIs$A[,1]<=muY & muY<=CIs$A[,2])
CP.B <- mean(CIs$B[,1]<=muY & muY<=CIs$B[,2])
data.frame(conf.level=cl, CP.A=CP.A, CP.B=CP.B)

# conf.level CP.A CP.B
# 1 0.9 0.895 0.749
```

The Sample A CI is much better than the Sample B CI that incorrectly assumes iid sampling. “Better” means coverage probability is closer to the desired confidence level of 90%. Specifically, the simulated coverage probability (CP) of the Sample A CI is 89.5%, whereas the CP of the Sample B CI is 74.9%.

3.9.2 Not a Bayesian Belief

The p-value is often interpreted as the probability that the hypothesis \( H_0 \) is true, but this is wrong. While intuitive, such an interpretation could only be possible in a Bayesian framework, not frequentist. For example, in the frequentist framework, either \( \mu_Y \leq 0 \) or not, whereas the Bayesian posterior describes our belief about the probability that \( \mu_Y \leq 0 \).

For example, a p-value of 0.08 does not mean, “I believe there’s an 8% chance that \( H_0 \) is correct.”
The example in Section 3.9.4 shows how your belief about $H_0$ depends on more than just a frequentist hypothesis test or $p$-value, which only account for what happens if $H_0$ is true.

### 3.9.3 Unlikely Events Happen (or: Use Common Sense)

As pointed out in an insightful webcomic ([xkcd.com/1132](http://xkcd.com/1132)), common sense and outside knowledge should be used when interpreting $p$-values and hypothesis tests. A small $p$-value alone (e.g., rejecting $H_0$ at a 5% level) does not mean $H_0$ is definitely false. It does not even mean that $H_0$ is probably false. A $p$-value below 0.05 is observed 5% of the time even if $H_0$ is false. A 5% chance is somewhat unlikely but far from rare, especially considering the thousands and thousands of $p$-values being computed every day. A small $p$-value just means it is unlikely to occur if $H_0$ is true, but common sense sometimes suggest that it is even less likely that $H_0$ is false, as in the comic.

#### Discussion Question 3.2 (equal $p$-values, equal belief?)

Consider three examples from Berger (1985, p. 2), which he attributes to L. J. Savage. First: a person claims to be able to tell whether milk was poured into a cup of tea or tea was poured into milk; in ten trials, the person guessed correctly each time. Second: a music expert claims to be able to tell whether a page of sheet music was written by Mozart or Haydn; in ten trials, the expert guessed correctly each time. Third: your drunk friend claims to be able to predict the outcome of a coin flip (heads or tails) from a fair coin (50% probability of each outcome); in ten trials, your friend is correct each time. Note that each “experiment” has a $p$-value of $2^{-10} \approx 0.001$, since guessing randomly could only get all correct with that probability (around 1/1000, 0.1%). After seeing all this data, do you have the same belief about whether each claim is true (i.e., do you think there’s the same chance that each claim is true)? Why not?

### 3.9.4 Example of Ignoring Outside Knowledge

The following example illustrates ideas from Sections 3.9.2 and 3.9.3.

Table 3.3 shows the setup for a classic example in which frequentist hypothesis testing is misleading. It shows the disease status and test results for 1,000,000 people randomly sampled from the population. Dividing everything by 1,000,000 would yield a joint probability distribution table, but intuition is easier with people instead of probabilities. The table shows that the disease is uncommon since only 1000 people have it, i.e., 1/1000 or 0.1%.

Table 3.3 can be interpreted in hypothesis testing terms. The null hypothesis $H_0$ is that somebody does not have the disease. A positive test result means rejecting $H_0$. Table 3.3 shows the type II error rate is zero: when $H_0$ is false, the test is always correct (+). The type I error rate is 5%: when $H_0$ is true, the test is wrong (+) with rate $49,950/999,000 = 5%.$
Table 3.3: Disease status and test result for 1,000,000 random people.

<table>
<thead>
<tr>
<th></th>
<th>(Don’t reject)</th>
<th>(Reject)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test −</td>
<td>Test +</td>
<td></td>
</tr>
</tbody>
</table>
| \(H_0 \text{ true})
No disease        | 949,050        | 49,950   | 999,000 |
| \(H_0 \text{ false})
Disease            | 0              | 1000     | 1000   |
| Total             | 949,050        | 50,950   | 1,000,000 |

The frequentist properties make the test sound very reliable: the type I error rate is controlled at the conventional 5% level, and the type II error rate is zero.

However, if you’re a random person in the population, a positive test result should not make you think you have the disease. Given that you tested positive, you could be in one of two boxes in the second column of the table. That is, either you’re one of the 49,950 people who tested positive but didn’t have the disease, or one of the 1000 people who tested positive and did have the disease. Clearly 49,950 is much larger than 1000, so it’s much more likely that you don’t have the disease, even though you tested positive. That is, conditional on having tested positive, the probability of actually having the disease is still only \(1000/50,950 = 1.96\%\), a very low probability.

Put in more general terms: even though our test rejected \(H_0\) at a 5% level, it is still much (much) more likely that \(H_0\) is true than false.

On the other hand, you may have gotten tested because you had all 17 characteristic symptoms of the disease. In that case, you’re not exactly “a random person in the population.” Your conclusion would be very different.

The conclusion is not “don’t believe positive results” or “don’t believe negative results,” but rather, “think about what else you know to critically think about results.”

### 3.9.5 Multiple Testing (Multiple Comparisons)

Another insightful comic (xkcd.com/882) illustrates the multiple testing problem (or multiple comparisons problem). Essentially, the scientists keep testing whether a different color jelly bean (a candy) causes acne (a skin condition), until they finally find \(p < 0.05\) and reject the null hypothesis of “no effect” at a 5% level. Since jelly beans do not actually cause acne (\(H_0\) is actually true), this “false positive” should happen roughly 5% of the time, or 1 in 20; knowing this, the comic shows them testing 20 different colors.

The multiple testing problem is essentially that if you keep testing and testing, eventually you’ll get a false positive with \(p < 0.05\) and rejecting \(H_0\). As an analogy: even though there’s a full moon less than 5% of all nights, as long as you keep looking up at the sky every night, eventually you’ll see one.
Practice 3.8 (research assistants). Imagine you’re a powerful profes-
or with a cadre of 100 research assistants (post-docs, grad students,
dergrads, your neighbor’s precocious high-schooler, etc.). You as-
sign each research assistant (RA) one of 100 variables characterizing
different counties in the U.S.: number of tennis courts, average tem-
perature, per capita income, etc. Each RA collects a dataset with
their particular variable and computes the correlation with county-
level May 2020 COVID-19 rates. Each RA then computes a $p$-value
for the null hypothesis that the correlation is zero. Of the 100 RAs,
5 report statistical significance at a 5% level ($p < 0.05$), including
a significantly positive correlation with tennis courts and negative
correlation with temperature. In light of multiple testing, how do
interpret these results?

Discussion Question 3.3 (jellybean solution?). Consider the jelly
bean comic from xkcd.com/882. In the comic, they essentially do
hypothesis tests with a 5% level, rejecting $H_0$ (no effect) if the $p$-
value is below 0.05. Since they do 20 hypothesis tests, even if each
individual test is unlikely to make an error, it’s likely that at least
one of the 20 tests makes an error.

a) Assuming jelly beans have zero effect, what type of error (I or
II) is made by the green jelly bean test?
b) Would it help (i.e., make such an error less likely) to use 1%
level tests instead of 5%? Explain why, why not, or how much
it might help.
c) Would it be even better to use 0% level hypothesis tests? Ex-
plain why or why not.

Discussion Question 3.4 (nova). Consider again the comic from
xkcd.com/1132 about the machine that detects if the sun has gone
nova (exploded). The null hypothesis $H_0$ is that the sun has not
exploded.

a) Does the frequentist statistician correctly compute the $p$-value
and correctly reject $H_0$ at a 5% level? Why/not?
b) What type of error (I or II) does the Bayesian statistician bet
has been made?
c) In this example, explain why it’s almost certainly incorrect
whenever the machine reports that the sun has exploded.

3.9.6 Publication Bias and Science

The jelly bean comic’s final panel illustrates publication bias: the
newspaper only reports the exciting positive result, omitting the 19
negative results for the other 19 colors. Not only popular media but
even academic journals are more likely to publish positive results, so
reading only published results gives a biased perspective.

The jelly bean experiments also illustrate the importance of re-
membering what “science” means. The result of a single study (even
a good one) by itself is not science. The scientific method is a process
of replication and repeated testing of hypotheses. If you ever hear,
CHAPTER 3. ONE VARIABLE: SAMPLE

“There was this one new study that found [crazy result]!!” you can ignore it and wait till it gets replicated at least a few times.  

3.9.7 Ignoring Point Estimates (Economic Significance)

Sometimes there is too much emphasis on p-values while the point estimates are ignored. This problem is mostly solved by simply looking at confidence intervals instead of only p-values and hypothesis tests.

While statistical significance (Section 3.8.4) assesses if the effect is statistically distinguishable from zero, economic significance assesses if the effect is “economically” distinguishable from zero. “Economically” just means “for real-world purposes,” like whether it is important to consider for policy purposes. One way to think about this is: would you personally care about the difference? For example, imagine \( \hat{\theta} \) estimates the effect on your final exam score of studying an additional hour per week. Would you (yes, you) care about having a final exam score that’s \( \hat{\theta} \) percentage points higher? If \( \hat{\theta} = 0.01 \), then no; if \( \hat{\theta} = 50 \), then yes. Some other examples: would you care if you had two additional years of education? Would you care if your annual salary were increased by five dollars?

Conceptually, like statistical significance, economic significance is not a binary yes/no but a continuum of “how much” economic significance. A result could have low economic significance, or extremely high economic significance, or moderate economic significance, etc.

In practice, unlike with statistical significance, there is no conventional level (like 5%) to mindlessly apply, so you are forced to think critically. (This is good.)

It is important to consider units of measure. For example, imagine the estimated effect on income is \( \hat{\theta} = 10 \); is that economically significant? If the units are dollars per hour, then yes; if it’s dollars per year, then no; if it’s thousands of dollars per month, then yes; etc.

It is also important to consider realistic policy changes. For example, imagine your estimated \( \hat{\theta} \) is the effect of a one-unit increase in the proportion of the state budget allocated to higher education. If the current proportion is 0.08 (meaning 8%), then a realistic policy change would be something like 0.02 units. A one-unit increase would mean changing from 0% to 100% of the budget spent on higher education. Possibly \( \hat{\theta} \) looks economically significant, but 0.02\( \hat{\theta} \) does not.

Examples

Simplifying “how much” significance to just low/high, there are four general possibilities: 1) both statistical and economic significance, 2) just statistical, 3) just economic, or 4) neither.

---

The following are examples of each possibility in the example where \( \hat{\theta} \) is an effect on annual income, measured in dollars per year.

- Both: \( \hat{\theta} = 20,000, p = 0.001 \)
- Only statistical: \( \hat{\theta} = 10, p = 0.001 \)
- Only economic: \( \hat{\theta} = 20,000, p = 0.43 \)
- Neither: \( \hat{\theta} = 10, p = 0.43 \)

**Practice 3.9** (significance: distance and education). You observe a sample of married couples; for each, you observe the difference in their years of education, divided by the difference in the distance between their childhood homes and the nearest college or university. That is, if \( E_1 \) and \( E_2 \) are the years of education, and \( D_1 \) and \( D_2 \) are the distances, you observe \( Y = (E_2 - E_1)/(D_2 - D_1) \). Distance is measured in kilometers (1 km = 0.600 mi). You estimate \( \bar{Y} = -0.03 \).

The \( p \)-value for testing \( H_0: E(Y) = 0 \) is \( p = 0.03 \).

a) Is this estimate economically significant? Hint: consider the units of \(-0.03\).

b) Is this statistically significant? Be precise.

### 3.9.8 Other Sources of Uncertainty

In practice, there are many sources of uncertainty, only one of which is the “statistical uncertainty” due to having a random sample. For example, there may be uncertainty about assumptions (in later chapters) required to interpret the population parameter in a certain way. There may be uncertainty about how reliable the data is. There may be uncertainty about whether sampling is iid.

The confidence intervals and other inference methods in Section 3.8 only quantify the statistical uncertainty from random sampling, not any other type of uncertainty. Thus, even if you have many other types of uncertainty, a CI could be very short, incorrectly suggesting you should feel very confident in the estimates.

For example, imagine you want to learn the mean household income in Kansas, but you only have data from Missouri. You decide to assume that Kansas has the same mean as Missouri. The Missouri dataset has very large \( n \), so \( \text{SE}(\bar{Y}_n) \approx 0 \). This makes it seem like we have zero uncertainty, but we may be very uncertain about the assumption that Missouri and Kansas are identical. Indeed, if the Kansas mean is very different, then our CI could have near 0% actual coverage probability.

**Practice 3.10** (uncertainty in Kansas). Imagine again \( \theta \) is the mean household income in Kansas, but this time you have data from Kansas. Which of the following could make the actual coverage probability much lower than the nominal level, i.e., make the CI shorter than it would be to fully capture all your uncertainty? Why?

a) You have data from state tax returns, but only 70% of households filed a state tax return, and you worry these 70% may not be fully representative of the population.

b) Your sample size is only \( n = 10,000 \), so a different random sample could have resulted in different observed income values.
c) You have survey data from a representative sample, but you doubt people accurately report their household income.
d) You use a CI based on iid sampling, but honestly you just found the data online somewhere and didn’t really understand if maybe it was clustered sampling.
e) You have data on individual adult income, which you then multiply by the average number of adults per household, but you only have the average number of adults per household from the year 2013 and think it might have changed.

3.10 Statistical Decision Theory

If you want to incorporate data more formally into your decisions, then you should learn more about statistical decision theory, which is beyond our scope. It is related to the optimal prediction material in Sections 2.4 and 2.5. There’s both frequentist and Bayesian statistical decision theory. Either way, hypothesis testing is basically never the best way to make a decision using data. For example, see Berger (1985).

Discussion Question 3.5 (Ebola drug). Imagine you have data for a new drug that tries to cure Ebola, a disease with a high mortality rate. Assume that there are no other treatments available, and that without the drug, an infected individual will die 100% of the time. With the drug, there is a possible side effect of occasional sneezing, and it possibly cures the disease (so the person does not die from it). You have a sample of 10 individuals infected with Ebola, and randomly picked 5 to take the experimental new drug and 5 to have no treatment. Of course, of the 5 without the drug, all 5 die. Of the 5 treated, 2 live, and 3 die. You input your data into R and run a t-test with command t.test(x=c(0,0,0,0,0), y=c(1,1,0,0,0)) where 0 means dead and 1 means alive. R says the two-sided p-value for testing the null hypothesis that the drug has zero effect on mortality is $p = 0.1778$; $H_0$ is not rejected even at a 10% level (let alone 5% or 1%), i.e., the result is not statistically significant at a 10% (or 5% or 1%) level.

a) If you then discovered that you had Ebola, would you take the drug? Why/not? Hint: did R compute the right p-value? What’s the probability of 2 people living if the drug actually has zero effect on mortality?
b) What if not everybody died without the drug: the untreated group had 1 person live (among 5), and the treated group had 3 live (among 5), yielding a p-value of 0.2429. Would you take the drug if you were infected? Why/not? Hint: what does your loss function look like?
Empirical Exercises

Empirical Exercise EE3.1. The data are originally from Card (1995), with individual-level observations of wages, years of education, and other variables.

a. R only: run `install.packages(c('wooldridge', 'survey'))` to download and install those packages (if you have not already)

b. Load the `card` dataset.

R: load package `wooldridge` with command `library(wooldridge)` and a `data.frame` variable named `card` becomes available; the command `?card` then shows you details about the dataset.

Stata: run `ssc install bcuse` to ensure command `bcuse` is installed, and then load the dataset with `bcuse card, clear`

c. Compute the sample average of variable `wage`.

R: `mean(card$wage)`

Stata: `mean wage` (which also computes a 95% confidence interval)

d. Estimate the population mean accounting for the sampling weights.

R: `weighted.mean(x=card$wage, w=card$weight)`

Stata: `mean wage [pweight=weight]` (also computes a 95% CI)

e. R only (since Stata reported this already): compute a two-sided 95% CI for the mean ignoring weights with `t.test(x=card$wage, conf.level=0.95)`

f. R only (since Stata reported this already): compute a two-sided 95% confidence interval for the mean accounting for weights, first loading the `survey` package with `library(survey)` and then with commands

```r
carddes <- svydesign(data=card, weights=~weight, id=~1)
svyret <- svymean(x=~wage, design=carddes)
c(w.mean=coef(svyret), SE=SE(svyret),
   CI=confint(svyret, level=0.95))
```

g. Compute a weighted, 90% confidence interval for wage.

R: replace `level=0.95` with `level=0.90`

Stata: add “option” `level(90)` to get `mean wage [pweight=weight] , level(90)`

h. Repeat computation of a point estimate and 95% confidence interval (without and with weights) for the mean of a different variable in the dataset.
R: part (c) computes the unweighted point estimate, part (d) computes the weighted point estimate, part (e) computes the unweighted CI, and part (f) computes the weighted CI.

Stata: part (c) computes both the unweighted point estimate and unweighted CI, and part (d) computes both the weighted point estimate and weighted CI.
Chapter 4

One Variable, Two Populations

⇒ Kaplan video: Chapter Introduction
Depends on: Chapters 2 and 3

Unit learning objectives for this chapter

4.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]

4.2. Describe and distinguish among descriptive, predictive, and causal questions, and among different approaches to learning about causality from data in economics [TLOs 3, 5, and 6]

4.3. Describe and interpret the elements of the primary statistical framework for understanding causality [TLO 3]

4.4. Assess whether a mean difference can be interpreted with causal meaning in a real-world example [TLO 6]

4.5. In R (or Stata): compute estimates of mean differences, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

Optional resources for this chapter

- Structural and reduced form approaches: Lewbel (2019)
- Potential outcomes and SUTVA (Wikipedia)
- Causal inference intro (Masten video)
- Correlation vs. causation (Masten video)
- ATE (Masten video)
- Individual causal effects (Masten video)
- Potential outcomes example (Masten video)
With two populations, we can discuss not only description and prediction, but also causality. Foundational ideas introduced here are extended to regression in Part II.

**Discussion Question 4.1** (DPC with two populations). Let \( Y \) denote the hourly wage of an individual in the U.S. Let \( Y^A \) be the wage of an individual without a college degree in the U.S., and \( Y^B \) the wage of an individual with a college degree.

a) How are means \( E(Y^A) \) and \( E(Y^B) \) more helpful for description than only \( E(Y) \)?

b) How could \( E(Y^A) \) and \( E(Y^B) \) be used to make better predictions than only \( E(Y) \)?

c) Why can’t we interpret \( E(Y^B) - E(Y^A) \) as the causal effect of a college degree on wage? Hint: what other factors might make \( E(Y^B) - E(Y^A) \) large, even if the effect of a college degree itself is small?

### 4.1 Description

#### 4.1.1 Population Mean Difference

Let \( Y^A \) and \( Y^B \) be random variables representing \( Y \) (e.g., income) for two populations (labeled \( A \) and \( B \)). For example, if \( Y \) is income, \( A \) is the population of individuals without a high-school degree, and \( B \) is the population of individuals with a high-school degree, then \( Y^A \) is income for individuals who do not have a high-school degree, and \( Y^B \) is income for those who do.

The difference of means is \( E(Y^B) - E(Y^A) \). It describes how much higher (or lower, if negative) is the mean in population \( B \) than in population \( A \).

For example, let \( Y \in \{0, 1, 2\} \) be the number of kids per family. Let the distributions in populations \( A \) and \( B \) be, respectively,

\[
P(Y^A = 0) = 0.8, \ P(Y^A = 1) = 0.2, \ P(Y^A = 2) = 0, \\
P(Y^B = 0) = P(Y^B = 1) = P(Y^B = 2) = 1/3, \tag{4.1}
\]

where \( Y^A \) represents the number of kids per family in population \( A \), and \( Y^B \) represents the number of kids per family in population \( B \).
\[
\begin{align*}
E(Y^B) - E(Y^A) &= \left[ \sum_{y=0}^{2} y P(Y^B = y) \right] - \left[ \sum_{y=0}^{2} y P(Y^A = y) \right] \\
&= [(0)(1/3) + (1)(1/3) + (2)(1/3)] - [(0)(0.8) + (1)(0.2)] \\
&= [(1/3) + (2/3)] - 0.2 = 0.8.
\end{align*}
\]

Always clarify whether you are subtracting the mean of population \(A\) from that of population \(B\), or \(B\) from \(A\). Saying, “The difference in mean number of children between the populations is 0.8,” it is unclear which population has a larger mean. Instead say, “The mean number of children in population \(B\) is 0.8 higher than the mean in population \(A\).”

The difference of means is also the mean of the differences. “Mean difference” could mean either; they’re equal anyway. Because of the linearity of the expectation operator as in (2.21),

\[
E(Y^B - Y^A) = E(Y^B) - E(Y^A). \tag{4.2}
\]

Despite mathematical equality, the interpretation differs. For example, the expression \(Y^B - Y^A\) is the number of children difference between a family from population \(B\) and a family from population \(A\). Seeing \(Y^B\) and \(Y^A\) as random variables, the difference \(Y^B - Y^A\) is itself a random variable. Thus, \(E(Y^B - Y^A)\) is the population mean of the child number difference \(Y^B - Y^A\), whereas \(E(Y^B) - E(Y^A)\) is the difference between the mean number of children in \(B\) and the mean number of children in \(A\). Generally, due to (4.2), either interpretation of the mean difference is correct; the same population value has two interpretations. It’s like if one person says, “The glass is half full of water,” and a second person says, “The glass is half empty”; both are correct interpretations of the same thing.

### 4.1.2 Estimation

Simply estimate the means separately (see Section 3.4) and take the difference, like \(\bar{Y}^B - \bar{Y}^A\) with iid data. If each individual estimator is consistent, then this is a consistent estimator of \(E(Y^B) - E(Y^A)\), and thus a consistent estimator of \(E(Y^B - Y^A)\) due to (4.2).

The following code estimates \(E(Y^B) - E(Y^A)\) from simulated iid samples, suggesting population \(B\) has a lower mean.

```r
set.seed(112358)
YA <- 0 + sample(x=1/zero.alt3:2/zero.alt3, size=4/zero.alt3, replace=TRUE, prob=rep(1/11,11))
YB <- 2 + sample(x=1/zero.alt3:25, size=3/zero.alt3, replace=TRUE, prob=26:1/sum(1:26))
mean(YB) - mean(YA)
```

```
## [1] -2.99
```
4.1.3 Quantifying Uncertainty

The same approaches (and warnings) from Sections 3.8 and 3.9 apply to \( \theta = E(Y^B) - E(Y^A) \).

The following R code shows 95% confidence intervals for the mean difference with iid data.

```r
set.seed(112358)
YA <- 0 + sample(x=10:20, size=40, replace=TRUE, prob=rep(1/11, 11))
YB <- 2 + sample(x=1:25, size=30, replace=TRUE, prob=26:1/sum(1:26))
# 95% CI for mean diff
round(t.test(x=YA, y=YB, alternative='two.sided',
mu=0, conf.level=0.95)$conf.int[1:2], digits=2)
## [1] 0.44 5.55
```

4.2 Prediction

Prediction is essentially the same as with one population. Given a loss function, an optimal predictor can be defined to minimize mean loss in the population, and this optimal predictor can be estimated from data. For example, mean quadratic loss is minimized by the population mean, and the means \( E(Y^A) \) and \( E(Y^B) \) can be estimated by (weighted) sample means.

Prediction accuracy improves by distinguishing between individuals (or firms, etc.) from population \( A \) and those from population \( B \). For example, at your carnival job, imagine you now guess people’s height instead of age. In Chapter 2, you make the same guess for everybody. Now, we consider two populations, like child and adult (assuming this is observable). Now, we can make a different prediction for each population, like 165 cm for adults and 105 cm for children. This should perform better than guessing 135 cm for every individual.

Part II extends this idea, exploring how regression models can incorporate additional information to improve prediction accuracy.

4.3 Causality: Overview

The concepts in the remainder of this chapter appear often in later chapters.

First: when is causality important, rather than description or prediction? We each have an innate sense of cause and effect. Trying to articulate it in language sometimes creates more confusion than understanding.\(^1\) For example, start reading the Wikipedia page on causality and see how you feel in 10 minutes. Unlike description and prediction, causality is about “why.” A “cause” is the “because” of the

\(^1\)Some of my failed attempts include: “causality is about what will happen if a policy changes” (but isn’t “what will happen” prediction?) and “description is seeing how things are” (but aren’t causal relationships also “how things are”?).
4.3. CAUSALITY: OVERVIEW

effect. Description helps us see which variables tend to have high or low values together. Prediction helps us guess one variable’s value based on other information. But only causality concerns why. Why do these two variables tend to have similar values? Only causality (not description or prediction) helps inform policy decisions: we want to know how a policy change itself influences other variables, causing them to change.

Discussion Question 4.2 (description, prediction, causality). Which type of question (description, prediction, causality) is each of the following? Explain why. Hint: there’s one of each.

a) If you only know whether an individual is from Canada or the U.S., what is your best guess of their income?

b) You are currently working in the U.S. but considering moving to Canada. How will your income change if you do?

c) Which country’s population has higher income: Canada or the U.S.?

4.3.1 Correlation Does Not Imply Causation

Generally, imagine $E(Y_B) > E(Y_A)$. This shows a clear descriptive relationship: population $B$ has a higher mean. The implication for prediction is clear: under quadratic loss, the optimal prediction is higher for population $B$ than $A$. In contrast, the implication for causality is not clear. It’s possible that being in population $B$ has a positive causal effect on the outcome variable. But it’s also possible that people with large $Y$ choose to join population $B$. Or maybe there is something else altogether that separately causes people to join population $B$ and have high $Y$. Or maybe all of these. The causal interpretation of $E(Y_B) > E(Y_A)$ is ambiguous.

For example, consider rainfall and umbrellas. Let $Y_A$ denote rainfall when nobody is carrying an umbrella, and $Y_B$ rainfall when everybody is carrying an umbrella. For description, it rains more on days when everyone carries an umbrella than on days when nobody does; e.g., $E(Y_B) > E(Y_A)$. For prediction, it’s better to predict a higher rainfall value if you see everyone carrying an umbrella than if you see no umbrellas; e.g., under quadratic loss, the optimal predictions are $E(Y_B)$ and $E(Y_A)$. For causality, if there’s a drought and we want rain, should we all walk around with umbrellas to cause it to rain? No: rain causes umbrella-carrying, not vice-versa.

Consider another example with a different type of causal relationship. Let $Y_A$ be my commute time when nobody is carrying umbrellas, and let $Y_B$ be my commute time when everyone is carrying umbrellas. Descriptively, $E(Y_B) > E(Y_A)$, and you should predict a longer commute time if you see everybody has an umbrella. But causally, this doesn’t mean that you can make me late for class by opening lots of umbrellas. Rain is a confounder that has a causal effect on both umbrella-carrying and commute time, as depicted in

$\implies$ Kaplan video: Correlation Does Not Imply Causation
These examples illustrate the famous saying, “correlation does not imply causation.” The saying is a bit imprecise: correlation does indeed imply some sort of causal relationship, just not any one particular type of causal relationship. In the first example, “correlation does not imply causation” means that “higher rainfall when people carry umbrellas” (rain is correlated with umbrellas) does not imply “carrying umbrellas causes rain.” But, the correlation is ultimately driven by a causal relationship: rain causes umbrella-carrying. In the second example, “correlation does not imply causation” means that “longer commute when people carry umbrellas” (commute time is correlated with umbrellas) does not imply “carrying umbrellas causes longer commutes.” But, the correlation is ultimately drive by causal relationships: rain causes both umbrella-carrying and longer commutes.

Although common sense helps us see that umbrellas cannot cause longer commutes, similar arguments are often made. For example, in the 2018 August election in Missouri, a “right-to-work” proposition appeared on the ballot. Very roughly speaking, such laws restrict the power of unions to collect certain fees from certain employees, but the following discussion about causality does not depend on the details.

Before the election, one mailer ad opposing right-to-work said something like, “Do you want $8000 less in your pocket each year?” The implication is that, were the law to pass, the causal effect would be a decrease in annual income of $8000/yr. According to the ad’s footnote, this $8000/yr was computed as the difference in workers’ mean annual income between states that had a right-to-work law and those that did not, i.e., an estimate of \( \mathbb{E}(Y^B) - \mathbb{E}(Y^A) \). Recall that \( \mathbb{E}(Y^B) - \mathbb{E}(Y^A) \neq 0 \) in the example with umbrellas and commute time, too, but we did not conclude that umbrellas have a causal effect on commute time. For example, maybe having lower income causes states to pass such laws, i.e., causality is in the opposite direction.

---

2 https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation
3 But if you’re curious: https://en.wikipedia.org/wiki/Right-to-work_law
4 Just to clarify, this is not a discussion of whether the law itself is good or bad, or whether the groups supporting or opposing the law are good or bad; ads endorsing right-to-work also made errors, just not illustrative econometric errors.
4.3 CAUSALITY: OVERVIEW

4.3.2 Structural and Reduced Form Approaches

There are two econometric approaches to learning about causality: the reduced form approach, and the structural approach. Confusingly, the reduced form approach is sometimes called causal inference even though the structural approach also aims to learn about causality. Also confusingly, the reduced form approach is commonly associated with program evaluation, like assessing the effects of a job training program or welfare program, but the structural approach could also be used.

The reduced form approach tries to isolate causal effects by using comparisons that are either randomized or “as good as randomized.” In our current context of populations A and B, randomized would mean that units (e.g., individuals, firms, hospitals) are randomly assigned to a population, without regard to the units’ characteristics. The “treated” population would receive some special treatment that the “untreated” (“control”) population does not. “As good as randomized” would mean that although we did not explicitly randomly assign units to each population, the actual assignment mechanism did not depend on units’ characteristics anyway. These situations are often called natural experiments; they often arise from unexpected weather or disease outbreaks (like COVID-19), or capricious political decisions. Sometimes other variables are used to help find these “as good as randomized” comparisons or to reduce statistical uncertainty, but the core methodology remains the comparison of treated and untreated units, to estimate the causal effect of a particular “treatment” (a variable or policy). The actual underlying causal mechanisms that produce the effect are not modeled, like a black box.

In contrast, the structural approach tries to explicitly model the inner workings of causal systems. Some “structural” models are not particularly detailed, but some involve models of decision-making (like expected utility maximization), models of market equilibria, models from game theory, and other economic theory. Consequently, the structural approach tries to estimate structural parameters (or “deep parameters”) that govern economic behavior and equilibria, like elasticities, discount factors, risk aversion, and demand curves. The hope is that all this “structure” would not change under the set of policies being considered, i.e., that the policies may change the values of certain variables but not change these underlying economic direction (reverse causality). Or maybe there is a third, unobserved characteristic that causes states to pass such laws and causes lower income, i.e., a confounder, like rain in the commute example. Of course, it is also possible that $8000/yr really is the causal effect. The point is not that the number is right or wrong (or that the law is good or bad), but that the econometric argument is incomplete. Additional assumptions are required to interpret a mean difference as a causal effect. Such assumptions are discussed more in Section 4.6.
The structural and reduced form approaches have complementary advantages, and often both are helpful; e.g., see the survey by Lewbel (2019). Both have contributed to our understanding of economics. Often, structural models require stronger (less realistic) assumptions, but in return they can analyze a wider variety of possible policies. Conversely: the reduced form approach often has the advantage analyzing the effects of existing policies, but it is more difficult to extrapolate to hypothetical policies.

For example, consider the relationship between a woman’s education and fertility (number of children born). A reduced form approach might try to find a group of women with college degrees and a group without, where it seemed “as good as random” who had a degree. This is difficult because usually going to college is a carefully considered decision, and not one that we can force others to make in a randomized fashion, but there may be peculiar situations like a need-based college scholarship that had to be randomized because too many people applied, or the Cultural Revolution that suddenly shut down universities in 1966. A structural approach might try to model the different factors affecting fertility choice and the different channels through which education could have an effect. This is difficult because it is a very complex decision with many variables involved. The benefit is the ability to consider more possible policies with more nuance; for example, the effect on fertility of more women attending college may depend on whether the increased education is due to a reduced price of college (e.g., due to government subsidy) or greater incentive due to higher salaries for jobs requiring college education, or some other reason.

In Sum: Structural & Reduced Form Approaches

**Reduced form**: randomized or “as good as randomized” comparisons to isolate causality

**Structural**: more explicit economic models of causal relationships

### 4.3.3 General Equilibrium and Partial Equilibrium

Besides structural vs. reduced form, another dichotomy is between *general equilibrium* (GE) and *partial equilibrium* (PE) analysis. GE more ambitiously tries to model entire markets, sometimes multiple markets, whereas PE takes current market equilibria as given. Similar to the tradeoff between the structural and reduced form approaches, the tradeoff is that the GE framework can analyze policies that change equilibria (i.e., that have *general equilibrium effects*), but it requires stronger assumptions to do so.

For example, imagine you were analyzing the impact of free public childcare on mothers’ employment. A PE analysis would consider how mothers might respond to different childcare policies given the...
current prices of private childcare, current wages, etc. A GE analysis might further model the childcare and labor markets, to allow for the possible general equilibrium effects of public childcare policy on the prices in those markets. If there is a big expansion of free public childcare, then private childcares may indeed change their prices. If the expansion allows many mothers to enter the workforce, then the labor supply curve shifts out, which could lower wages. However, if the proposed changes to childcare policy are relatively small, then such GE effects may be negligible, and PE analysis may suffice.

The famous Lucas critique (Lucas, 1976) argues in part that macroeconomic policy analysis requires structural, GE models. Lucas writes (p. 41), “Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models.” If we want to guess how people and firms will behave in the future, under new macroeconomic policies, we have to account for GE effects, which requires deeper, structural understanding and modeling of economic behavior.

In Sum: General & Partial Equilibrium Models
Partial equilibrium models treat prices and other market equilibria as fixed, whereas general equilibrium models allow markets to change.

4.4 Potential Outcomes Framework

The reduced form approach uses the potential outcomes framework, also called the Neyman–Rubin causal model after its two earliest contributors (although sometimes Neyman’s name is dropped). It is popular not only in economics, but statistics, medicine, political science, and other fields.

The terms treatment and treatment effect just refer to any variable and its causal effect on another variable. In English, usually “treatment” makes us think narrowly about medicine (or lumber...and facials?), but it can be anything. For example, the “treatment” could be a job training program, and the “treatment effect” is the causal effect of the program on a person’s wage. Or, a treatment could be going to a charter school (instead of public school). Another treatment could be a policy or law, like a higher sales tax, or a certain labor law.

This section says “individual” to be more concrete, but you can also imagine a firm, county, school, etc.
4.4.1 Potential Outcomes

Imagine two parallel universes. The universes are identical except for one difference: whether or not an individual is treated. The individual’s outcome in the universe without treatment is their untreated potential outcome, and the individual’s outcome in the universe with treatment is their treated potential outcome.

Notationally, in this chapter, $Y_T$ represents the treated potential outcome and $Y_U$ the untreated potential outcome. Elsewhere, often $Y_1$ and $Y_0$ represent the treated and untreated potential outcomes, or $Y(1)$ and $Y(0)$.

For example, consider parallel universes identical except for whether a particular student takes introductory econometrics (this class) or Intro Stat II (STAT 3500). Literally everything else in each universe is identical: the student’s parents, her other classes, her height, her DNA, the weather on October 14, etc. (For now, some difficulties with “everything” are glossed over; e.g., what if econometrics is required for her degree?) The “treatment” is taking econometrics (instead of statistics). The outcome variable is the student’s annual income five years after graduation, in thousands of U.S. dollars per year (e.g., $Y = 70$ is $70,000/yr$). Let $Y_U$ denote her outcome in the universe without treatment (STAT 3500), and $Y_T$ her outcome in the universe with treatment (econometrics). That is, $Y_T$ is her treated potential outcome, and $Y_U$ is her untreated potential outcome.

Unlike in Section 4.1, potential outcomes $Y_U$ and $Y_T$ are not always observable. In the above example, if a student takes STAT 3500, then we can observe her untreated potential outcome $Y_U$, but not $Y_T$; conversely, if she takes econometrics, then her treated potential outcome $Y_T$ is observable but not $Y_U$. This partial observability makes causal inference more difficult than description or prediction.

Consider some other potential outcomes examples. In the right-to-work example, $Y_T$ is an individual’s income in the universe where the individual’s state has a right-to-work law, and $Y_U$ is their income in the universe that’s identical except there is no such law. In our universe, either the individual’s state does or does not currently have such a law; it cannot be both, so we cannot observe both potential outcomes. (Perhaps the state did not have the law last year and does this year, but the universe “last year” was different in many ways than the universe “this year”; much more than one single law has changed.)

As another example, imagine universe $B$ is where a student wins the lottery to enter a popular charter school, and universe $A$ is where the student remains in the conventional public school. Potential outcomes $Y_T$ and $Y_U$ are dummy (binary) variables for whether or not
the student eventually graduated from college in each universe. Again, in our universe, we can observe $Y^T$ if the students wins the lottery and $Y^U$ if not, but we cannot observe both.

### 4.4.2 Treatment Effects

The difference $Y^T - Y^U$ between an individual’s two potential outcomes is their **treatment effect**. Just as different individuals can have different $(Y^U, Y^T)$, individuals can have different treatment effects $Y^T - Y^U$; i.e., individuals can be affected differently by the same treatment.

In the intro econometrics example, the student’s treatment effect $Y^T - Y^U$ has the following interpretation. Recall $Y^T$ is their income after taking econometrics and $Y^U$ after instead taking STAT 3500. Thus, that particular student’s treatment effect is how much higher (or lower, if negative) their income is in the parallel universe that is identical other than taking econometrics instead of STAT 3500.

In the right-to-work example, $Y^T - Y^U$ is the treatment effect of the law on an individual’s income. The interpretation now is the difference between their income in the universe with the law and the universe without the law, with everything else held constant. The treatment effect can be big or small, positive or negative (or zero). A numerical example is shown later in Table 4.2.

In the charter school example, $Y^T - Y^U$ is the treatment effect of the charter school on college graduation. That is, it is the difference between the college graduation outcomes in the charter school universe and the public school universe. Since the outcome is binary (1 if graduate college, 0 if don’t), there are only four possible values of $(Y^U, Y^T)$ (student types) and only three possible treatment effect values: $Y^T - Y^U = 1$ if the student graduates in the charter school universe $(Y^T = 1)$ but not the public school universe $(Y^U = 0)$; $Y^T - Y^U = -1$ if they only graduate in the public school universe $(Y^U = 1)$ but not the charter school universe $(Y^T = 0)$; and $Y^T - Y^U = 0$ if they graduate either in both universes $(Y^T = Y^U = 1)$ or neither $(Y^T = Y^U = 0)$. This is seen in the later example of Table 4.1.

In all examples, the potential outcomes and treatment effects may be different for different individuals. For example, econometrics may be much better for some students but only a little better for others; right-to-work may help certain workers, but hurt others; the charter school may make the difference for some students to graduate college, but others would have graduated either way. The fancy term for people being different is **heterogeneity**, more specifically here “treatment effect heterogeneity.”

In economics, where many systems are interrelated, sometimes it’s difficult just to specify what “effect” we care about. For example, consider racial differences in salary. In the parallel universe that’s “identical” except for the individual’s race, does “identical” include having the same job at the same firm? Or does it allow for an effect...
of race on hiring? Does it allow for an effect on educational opportunities, or an effect on family background (parents’ education, wealth, etc.)? There is no “right” or “wrong” specification, but each answers a different question.

4.4.3 SUTVA

SUTVA Definition

The potential outcomes definition of causality relies critically on the stable unit treatment value assumption (SUTVA), which has two parts.

The first part of SUTVA is that every treated individual receives the same treatment. This seems true in the right-to-work example: the same law applies (or doesn’t) to everybody equally. This also seems true in the charter school example, but with more nuance. Two students may go to the same school but have very different experiences, like different teachers, different classmates, different electives, and different extra-curricular activities. Even if we say these two students nominally have the “same treatment,” we should expect a lot of heterogeneity, and we should expect treatment effects to change every year as the school adds or removes (or changes) its teachers, its students, its elective class offerings, and its extra-curricular activities. As another ambiguous example, if there’s a one-on-one mentoring program to help teen parents, but of course there are many different mentors, is every teen parent receiving the “same treatment”?

The second part of SUTVA is the no interference assumption. This assumes that one person’s treatment (or non-treatment) does not affect the potential outcomes of any other person. This often makes sense for medical treatments (e.g., doing surgery on me doesn’t affect your health), but it requires careful thought in economics, where often individuals are interacting either personally or through markets. In the charter school example, if a student’s success depends on being surrounded by other highly motivated students (or not), then SUTVA (specifically no interference) is violated. That is, one student’s outcome depends on whether the other motivated students are in the same school (whether charter or not), i.e., depends on the other students’ “treatment.”

The “same treatment” ambiguity also relates to the structural and reduced form differences in Section 4.3.2. In the charter school example, the structural critique would be that learning “the effect” of “going to charter school B” last year is not particularly helpful for guiding educational policy if we can’t confidently extrapolate from charter school B last year to charter school B next year, let alone extrapolate to other charter schools, let alone understand why the effects are positive or negative (e.g., is it because of teachers, or because of better classmates, or more electives and activities). The reduced form rebuttal would be that at least they can (sometimes) be pretty confident about their assessment of a particular school in a particular year, whereas trying to explicitly model the effects of teachers
and classmates and classes and activities will result in models nobody believes anyway. Hopefully, we could learn more by trying both approaches than giving up and trying neither.

**SUTVA Violations**

SUTVA can be violated in many ways, especially in economics. This is not about sampling, or randomization, or data; it is about the potential outcomes framework itself. Even if SUTVA is satisfied and treatment effects are well-defined, it is possible to have problems with randomization that make it impossible to actually learn about treatment effects. Conversely, even if there is a perfectly designed randomized experiment, SUTVA could be violated, in which case it may be unclear what “treatment effect” even means.

One violation of SUTVA is from **spillover effects**. For example, if the treatment provides helpful information (e.g., about financial planning, or social services, or risk probabilities), treated individuals may share such information with their untreated friends. That is, the benefit of the treatment “spills over” into untreated individuals. This could be true even if the treatment (information or otherwise) isn’t directly shared. For example, if the provided information leads to less **binge drinking** among treated individuals, this may reduce social pressure that results in less binge drinking among untreated individuals, even if they did not receive the “treatment” information. Or, if some treatment helps half the students in a classroom, their improvement itself may benefit their untreated classmates.

Another violation of SUTVA is from **general equilibrium effects** (Section 4.3.3). For example, maybe the treatment is a new agricultural technology hoping to increase cacao farmers’ earnings. If only one farmer gets this treatment (technology), then she benefits from increased production, selling more cacao at the current global price. But if all farmers in the world get the technology, then the global cacao supply curve shifts and the price drops. Thus, each farmer’s untreated and treated potential outcomes (earnings) are affected by all other farmers’ treatment, which affects the market equilibrium price. Other general equilibrium effects could come through other markets. For example, a treatment affecting workers might affect the labor market as a whole (and thus wages). Or, subsidies for housing or education could affect supply and demand (and thus prices) in those markets.

There can be yet other ways SUTVA is violated, either from not having the same treatment or from interactions that violate “no interference.”

**Discussion Question 4.3** (cash transfer spillovers). Consider the effect of income on food consumption \( (Y) \) in a rural village. Consider an “unconditional cash transfer” program (like GiveDirectly) that (potentially) gives the equivalent of $1000 to a treated individual. Describe different possible spillover effects that would violate SUTVA.
Sometimes one perspective of a treatment leads to SUTVA violations, but another does not. In the classroom example, SUTVA was violated by spillover effects from treated students to untreated students in the same class. Alternatively, if each classroom is treated or untreated (i.e., all students treated, or all not), then there is less possibility of spillover. In principle, even entire schools could be assigned as treated or untreated, further reducing spillovers.

In other cases, you may actually want to learn about spillover effects as part of the overall effect of a policy. For example, if the student-level treatment is specifically for students with certain special needs, then we probably care about its affect on both the treated and untreated students. (Further, it would be impossible to treat everyone in a school since the treatment is only appropriate for certain types of student.)

In deciding which perspective is best, it is helpful to think about the actual policy question: what is the potential policy that could actually be adopted in reality?

### 4.5 Average Treatment Effect

⇒ Kaplan video: Potential Outcomes and the ATE (again)

Although the full distribution of potential outcomes \((Y^U, Y^T)\) contains the most information, usually only certain summary features are studied. Although summary features like standard deviations and percentiles are interesting, we’ll focus on means.

#### 4.5.1 Definition and Interpretation

The average treatment effect (ATE) is \(E(Y^T - Y^U)\). “Average” refers to the population mean, while “treatment effect” refers to \(Y^T - Y^U\). Thus, the ATE may be interpreted as the probability-weighted average (mean) of all possible individual treatment effects in the population. Another name for the ATE is the average causal effect (ACE), but I use ATE to emphasize that this concept is from the potential outcomes framework.

The ATE has another interpretation. Using linearity as in (2.21),

\[
E(Y^T - Y^U) = E(Y^T) - E(Y^U).
\]

(4.3)

Here, \(E(Y^T)\) is the mean treated potential outcome, and \(E(Y^U)\) is the mean untreated potential outcome. Similar to Section 4.1, \(E(Y^T) - E(Y^U)\) is a mean difference, here between the treated and untreated potential outcome distributions. This could be rephrased as “the treatment effect on the mean outcome”: treatment causes the mean outcome to change from \(E(Y^U)\) to \(E(Y^T)\).

#### 4.5.2 ATE Examples

Table 4.1 shows a numerical version of the charter school example. The four student “types” refer to the four possible values of \((Y^U, Y^T)\),...
and each type has its own probability. Given the probabilities, the mean untreated outcome $E(Y_U)$, mean treated outcome $E(Y_T)$, and ATE $E(Y_T - Y_U)$ are computed using (2.16):

$$E(Y_U) = (0.3)(0) + (0.3)(0) + (0.1)(1) + (0.3)(1) = 0.4,$$  
(4.4)

$$E(Y_T) = (0.3)(0) + (0.3)(1) + (0.1)(0) + (0.3)(1) = 0.6,$$  
(4.5)

$$E(Y_T - Y_U) = (0.3)(0) + (0.3)(1) + (0.1)(-1) + (0.3)(0) = 0.2.$$  
(4.6)

To verify (4.3),

$$E(Y_T - Y_U) = 0.2 = 0.6 - 0.4 = E(Y_T) - E(Y_U).$$  
(4.7)

Table 4.1: Charter school example population of potential outcomes and ATE.

<table>
<thead>
<tr>
<th>Student type</th>
<th>Probability</th>
<th>$Y_U$</th>
<th>$Y_T$</th>
<th>$Y_T - Y_U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>0.3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>0.4</strong></td>
<td><strong>0.6</strong></td>
<td><strong>0.2</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Right-to-work example population of potential outcomes and ATE.

<table>
<thead>
<tr>
<th>Worker type</th>
<th>Probability</th>
<th>$Y_U$ ($/yr$)</th>
<th>$Y_T$ ($/yr$)</th>
<th>$Y_T - Y_U$ ($/yr$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>40,000</td>
<td>41,000</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>40,000</td>
<td>38,000</td>
<td>-2000</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>50,000</td>
<td>51,000</td>
<td>1000</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>50,000</td>
<td>47,000</td>
<td>-3000</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>43,000</strong></td>
<td><strong>43,000</strong></td>
<td><strong>0</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2 shows a numerical version of the right-to-work example. Each worker “type” corresponds to a different value of $(Y_U, Y_T)$, each type with its own probability. Given the probabilities, the mean untreated outcome $E(Y_U)$, mean treated outcome $E(Y_T)$, and ATE $E(Y_T - Y_U)$ are, in dollars per year,

$$E(Y_U) = (0.5)(40,000) + (0.2)(40,000) + (0.2)(50,000) + (0.1)(50,000) = 43,000,$$  
(4.8)

$$E(Y_T) = (0.5)(41,000) + (0.2)(38,000) + (0.2)(51,000) + (0.1)(47,000) = 43,000,$$  
(4.9)

$$E(Y_T - Y_U) = (0.5)(1000) + (0.2)(-2000) + (0.2)(1000) + (0.1)(-3000) = 0.$$  
(4.10)
Again, to verify (4.3),
\[ E(Y^T - Y^U) = 0/\text{yr} = \$43,000/\text{yr} - \$43,000/\text{yr} = E(Y^T) - E(Y^U). \]

(4.11)

4.5.3 Limitation of ATE

Figure 4.2 shows the ATE does not fully capture the effect of treatment on the distribution. The figure plots PDFs of three hourly wage distributions with identical means. Picking any two distributions to represent potential outcomes \( Y^U \) and \( Y^T \), \( E(Y^U) = E(Y^T) \), so the ATE is \( E(Y^T) - E(Y^U) = 0/\text{hr} \). However, zero ATE does not mean zero effect: the distributions are all different. For example, their standard deviations differ, and one distribution is right-skewed with a lower median wage. We may disagree about which is “better” or “worse,” but we can agree the differences are important.

This idea is also memorable in joke form; as retold by Hansen (2020, p. 29):

An economist was standing with one foot in a bucket of boiling water and the other foot in a bucket of ice. When asked how he felt, he replied, “On average I feel just fine.”

To address the limitations of the ATE, one approach is to examine effects on percentiles (“quantile treatment effects”), but these are beyond our scope.

Practice 4.1 (unrepresentative ATE). Describe a population in which the ATE is zero but every individual is affected by the treatment (i.e., all treatment effects are non-zero). For simplicity, assume there are only two types of individual. For each type, state the probability, potential outcomes \( Y^U \) and \( Y^T \), and causal effect \( Y^T - Y^U \), which must be non-zero. Then compute the ATE to verify it’s zero.

4.6 ATE: Identification

\( \Rightarrow \) Kaplan video: ATE Identification
Generally, **identification** is a concept central to econometrics that appears throughout this textbook. A parameter is **identified** if it equals a summary feature of the population distribution of observable variables. Identification requires certain conditions known as **identifying assumptions**.

Specifically, the ATE is identified when it equals a mean difference. (It may equal another summary feature in more complex settings not discussed here.) The required identifying assumptions are discussed later in this section. In practice, if the identifying assumptions are true, then the mean difference can be interpreted as the ATE, but if they are false, then it cannot.

### 4.6.1 Setup and Identification Question

For each individual, a single value is observed. If the individual was actually treated (in our universe), then treated potential outcome $Y^T$ is observed; otherwise, $Y^U$ is observed.

Consider actually treated individuals to be population $B$, and consider actually untreated individuals to be population $A$. The two populations are represented by random variables $Y^B$ and $Y^A$, respectively. For a population $B$ individual, $Y^B$ is always observable, with $Y^B = Y^T$. Similarly, for a population $A$ individual, $Y^A$ is always observable, with $Y^A = Y^U$. A random sample of $Y^B$ can be taken from actually treated individuals, and a random sample of $Y^A$ can be taken from actually untreated individuals. For example, $Y^B$ could be the graduation outcome for a student who actually attended the charter school (in our universe), while $Y^A$ is the outcome for a student who did not.

For the ATE, the question of **identification** is whether the ATE equals the mean difference between the actually treated and actually untreated populations. Mathematically, using the $E(Y^T) - E(Y^U)$ form of the ATE from (4.3), the identification question is whether or not

$$E(Y^T) - E(Y^U) = E(Y^B) - E(Y^A). \quad (4.12)$$

We know how to learn about the descriptive mean difference $E(Y^B) - E(Y^A)$ from data, as in Section 4.1. If (4.12) holds, then this is equivalent to learning about the ATE.

Consider (4.12) in the charter school and right-to-work examples. For the charter school example, if the ATE is identified, then it equals the college graduation probability of charter school students minus the college graduation probability of conventional public school students, $E(Y^B) - E(Y^A)$. (Recall from (2.9) that for binary $Y$, $E(Y) = P(Y = 1)$.) In that case, the ATE is estimated simply by comparing college graduation rates between the charter school and public school students. For the right-to-work example, if the ATE is identified, then it equals mean income in right-to-work states minus mean income in other states, $E(Y^B) - E(Y^A)$. In that case, the ATE is estimated simply by comparing average income between right-to-work states and other states. However, if the ATE is not identified
and (4.12) is false, then these comparisons do not estimate the ATE, i.e., the mean differences do not have a causal interpretation.

### 4.6.2 Randomization

Randomized experiments are often used to estimate the ATE. Ideally, in a randomized experiment, also called a randomized controlled trial (RCT), the experimenter can control who is treated and who is not (but see comments below). Mathematically, the experimenter gets to decide whether to observe $Y^U$ or $Y^T$ for each individual. “Randomized” means this decision is made without regard to the individual’s characteristics.

In practice, there are many complications; see Section 4.6.3 for some examples.

For intuition, consider the following experimental strategy. First, imagine we only want to estimate $E(Y^T)$. We could take a random sample of individuals from the population and treat each one, allowing us to observe their $Y^T$. That is, we have a random sample from the population distribution of $Y^T$. As in Chapter 3, we can estimate $E(Y^T)$ by the sample mean. Here, our $Y^B$ is $Y^T$, so $E(Y^B) = E(Y^T)$. Second, we can repeat the process for a second random sample but force everyone to be untreated. The key is the ability to force anyone to be either treated or untreated; this allows us to take random samples of $Y^T$ and $Y^U$. Although treatment may not seem “random,” it is assigned without consideration of any individual’s characteristics.

Section 6.6.2 contains more formal arguments for why randomization can help identify the ATE.

### 4.6.3 Reasons for Identification Failure

Generally (beyond only experiments), ATE identification fails when SUTVA fails (Section 4.4.3) or when treatment is not random.

Outside of experiments, random or “as good as random” treatment is rare. For example, in the right-to-work example, treatment is probably not random. Hopefully, state legislatures do indeed consider the characteristics of individuals when deciding whether or not to pass a right-to-work law; i.e., laws are not passed randomly. Specifically, legislatures may consider the distribution of $Y^U$ when deciding whether or not to pass the law (which would switch everyone’s income from their $Y^U$ to their $Y^T$). Further, just looking at a map, it is notable that (as of 2019) zero U.S. states in the Northeast census region have right-to-work laws, whereas almost all states in the South census region have right-to-work laws (the exceptions being Delaware and Maryland, which are not really “Southern” culturally or politically). Thus, it seems likely that the treatment decisions were related to other policy decisions that would in turn affect the income distribution.

Even with randomized treatment assignment, treatment itself may not be random. For example, imagine you randomly assign individuals to attend a job training program, but some assigned individuals
never attend. This is called non-compliance, i.e., not complying with the treatment assignment. This is a type of self-selection, meaning individuals decide which group to join. People who skip the program may also skip work regularly, which results in lower income. Thus, many low-income individuals who should have been in the treatment group (if we could force them) are now in the control group; i.e., they should have been $Y^B$ but are now $Y^A$. This decreases the control group’s average income and raises the treatment group’s average income, which falsely makes the treatment seem more effective than it is. Even if the training program has zero ATE, so $E(Y^T) = E(Y^U)$, this non-compliance makes it look like the treatment has a positive effect because $E(Y^A) < E(Y^U)$ and $E(Y^B) > E(Y^T)$.

One way to avoid this incorrect conclusion is to change perspective: compare groups based on treatment assignment rather than actual treatment. In the above example, the “treatment” is defined as being assigned to attend the job training program, rather than actually attending the program. The resulting ATE is called the intention-to-treat effect because it measures the mean change in $Y$ corresponding to the intention to treat (i.e., assignment to treatment). Sometimes this is more directly relevant for policy anyway, if the actual policy would not force people to be treated.

Attrition is another problem that can arise even if SUTVA and random treatment are satisfied. Attrition refers to individuals dropping out of the study after it starts. For example, maybe everyone comes to the first job training, but then some people move to a different state and disappear from your data. People leaving randomly is fine, but non-random attrition is problematic. For example, maybe the training program is so good that people get higher-paying jobs in other states. You only see data for individuals who didn’t move, who generally have lower-paying jobs. Then, even though the training program worked really well, it doesn’t look like it in the data because you don’t see all the highest-earning treated individuals who moved.

Other concerns are introduced later, especially in Section 12.3.

Discussion Question 4.4 (breakfast effect?). Schools with a high enough percentage of low-income students are eligible for a federally-funded free breakfast program for all students. Although the program is not mandatory, all eligible schools choose to have it. You compute a 95% CI for the mean math test score of the “breakfast” schools minus the mean of the other schools, and it is $[-32, -17]$ points. (The test is out of 100 points; most scores are in the 60 to 100 range.) How do you interpret this result? Think about ATE identification, statistical uncertainty, and frequentist vs. Bayesian perspectives.

4.7 ATE: Estimation and Inference

There is nothing new for estimation and inference. It is identical to Sections 4.1.2 and 4.1.3. Generally, the point of causal “identification”
is not to propose a new statistical object, but rather to imbue an existing descriptive statistical object with causal meaning. Here, (4.12) gives the descriptive mean difference a causal interpretation (ATE). The interpretation does not affect how we estimate or quantify statistical uncertainty about the mean difference.

However, recall from Section 3.9.8 that conventional methods for quantifying uncertainty (Section 3.8) only quantify statistical uncertainty, not uncertainty about identification. For example, if identification fails, a 95% CI for the mean difference may only contain the ATE with 80% probability, or even 50% or near 0%. There are some proposals for quantifying the sensitivity of results to violations of identification in various settings, but these are beyond our scope.
Empirical Exercises

Empirical Exercise EE4.1. You will analyze the effects of being assigned to a job training program, where assignment was randomized. The specific program was the National Supported Work Demonstration in the 1970s in the U.S. Data are originally from LaLonde (1986), via Wooldridge (2020). You will look at effects on earnings (re78) and unemployment (unem78), both overall and for different subgroups (e.g., married or not). The train variable indicates (randomized) assignment to job training if it equals 1, and 0 otherwise.

For now, we focus on computing various estimates; in later chapters we’ll think more critically about what could go wrong even with randomized assignment.

a. R only: run install.packages('wooldridge') to download and install that package (if you have not already)

b. Load the jtrain2 dataset.

R: load package wooldridge with command library(wooldridge) and a data.frame variable named jtrain2 becomes available; the command ?jtrain2 then shows you details about the dataset.

Stata: run ssc install bcuse to ensure command bcuse is installed, and then load the dataset with bcuse jtrain2, clear

c. R only: separate the data into “treatment” and “control” groups (depending on the value of train, the job training variable) with

trt <- jtrain2[jtrain2$train==1 , ]
ctl <- jtrain2[jtrain2$train==0 , ]

d. Estimate the mean 1978 earnings (in thousands of dollars) for the treatment group minus that of the control group, along with a 95% CI for the mean difference.

R:

mean(trt$re78) - mean(ctl$re78)
t.test(x=trt$re78, y=ctl$re78)

Stata: ttest re78 , by(train) unequal (also estimates the mean difference)

e. R only: separate out the data for treated, married individuals and untreated, married individuals, with

trt.mar1 <- trt[trt$married==1 , ]
ctl.mar1 <- ctl[ctl$married==1 , ]

f. Compute the mean difference estimate and 95% CI for the 1978 earnings outcome variable, comparing treated and untreated married individuals.

R:
CHAPTER 4. ONE VARIABLE, TWO POPULATIONS

mean(trt.mar1$re78) - mean(ctl.mar1$re78)
t.test(x=trt.mar1$re78, y=ctl.mar1$re78)

Stata: ttest re78 if married==1, by(train) unequal or alternatively bysort married: ttest re78, by(train) unequal

g. Repeat your above analysis in parts (d)–(f), but first create a variable where earnings are in dollars (instead of thousands of dollars).

R: jtrain2$re78USD <- 1000*jtrain2$re78
Stata: generate re78USD = re78*1000

h. Repeat your analysis in parts (e) and (f) for unmarried (instead of married) individuals.

i. Repeat your analysis in parts (d)–(f) but for unemployment (unem78) instead of earnings. For interpretation: note that unem78 equals 1 if unemployed all of 1978, and equals 0 otherwise, so the population mean is the probability of being unemployed all year (a value between 0 = 0% and 1 = 100%), and the sample average is the fraction of the sample thus unemployed. So, a value like 0.14 means 14%, and a difference of 0.14 − 0.11 = 0.03 is a difference of 3 percentage points, etc.

Empirical Exercise EE4.2. You will analyze data from an "audit study" that attempts to measure the effect of race on receiving a job offer. The Urban Institute found pairs of seemingly equally qualified individuals (one black, one white) and had them interview for a variety of entry-level jobs in Washington, DC in 1988. See Siegelman and Heckman (1993) for details and critique, and the raw data in their Table 5.1 (p. 195). In the data, each row (observation) corresponds to one job, to which one pair applied. Value \( w=1 \) indicates that the white applicant in the pair got a job offer, while \( b=1 \) if the black applicant got an offer.

a. R only: run install.packages('wooldridge') to download and install that package (if you have not already)

b. Load the audit dataset.

R: load package wooldridge with command library(wooldridge) and a data.frame variable named audit becomes available; the command ?audit then shows you details about the dataset.

Stata: run ssc install bcuse to ensure command bcuse is installed, and then load the dataset with bcuse audit, clear

c. Compute the difference (white minus black) in the sample fraction of job offers.

R: mean(audit$w) - mean(audit$b)
Stata: ttest w==b (which also computes a 95% CI)
d. Compute the sample mean of all the pairs’ white-minus-black difference. Note that \( w-b \) equals 1 if the white individual got a job offer but the black individual did not, equals \(-1\) if the black but not white individual got an offer, and equals 0 if both or neither of the pair got an offer.

R: `mean(audit$w - audit$b)`

Stata: `generate wminusb = w-b` then `ttest wminusb==0` (also computes 95% CI; see row labeled `diff` for both).

e. R only (since Stata already reported this in the row labeled `diff`): compute a 95% CI for the population mean difference with either `t.test(x=audit$w, y=audit$b, paired=TRUE)` or `t.test(x=audit$w-audit$b)`
Chapter 5

Midterm Exam #1

⇒ Kaplan video: Chapter Introduction

When I teach this class, the first midterm exam is this week. This “chapter” makes the chapter numbers match the week of the semester. The midterm covers Chapters 2–4, i.e., everything up till now except R/Stata coding.
Part II

Regression
Introduction

Part II concerns regression. Regression is the workhorse of empirical economics (and many other fields), for description, prediction, and causality alike.

Part II extends the concepts and methods of Part I to the regression setting. In the population, the concepts of description, prediction, and causality from Part I are extended to regression models. In the data, estimation and inference methods extend those of Part I.

More flexible regression is also considered, including different models, interpretation, and a glimpse of nonparametric regression and machine learning.
Chapter 6

Comparing Two Distributions by Regression

⇒ Kaplan video: Chapter Introduction

Depends on: Chapter 4 (which depends on Chapters 2 and 3)

Unit learning objectives for this chapter

6.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

6.2. Describe different ways of thinking about two distributions, both mathematically and intuitively [TLO 3]

6.3. Describe, interpret, identify, and distinguish among different population models and their parameters and estimators [TLO 3]

6.4. Judge which interpretation of a regression slope is most appropriate in a real-world example [TLO 6]

6.5. Interpret logical relationships and form appropriate logical conclusions [TLO 2]

6.6. In R (or Stata): estimate the parameters in a simple regression model, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

Optional resources for this chapter

- Conditional probability (Khan Academy)
- Basic joint, marginal, and conditional distributions (Khan Academy)
- James et al. (2013, §3.1)
- Covariance and correlation (Lambert video)
- Overlap assumption (Masten video)
- Correlation vs. causation (Masten video)
Assumptions for randomized experiment validity (Masten video)

Structural vs. causal/reduced form approach (Masten video)

OLS computation (Masten video)

Sections 2.1 (“Simple OLS Regression”) and 2.2 (“Coefficients, Fitted Values, and Residuals”) in Heiss (2016)

Section 5.3 (“Regression When X is a Binary Variable”) in Hanck et al. (2018)

R packages \texttt{lmtest} and \texttt{sandwich} (Zeileis, 2004; Zeileis and Hothorn, 2002)

Chapter 6 revisits Chapter 4 from the perspective of regression. The concepts of description, prediction, and causality are translated into regression language and regression models in the population. Estimation and quantifying uncertainty are also discussed.

The term \textit{regression} has different meanings in different contexts (and by different people). In the population, it usually refers to how the mean of a random variable $Y$ depends on the value of another random variable(s), as in Section 6.3. In the sample, as in Section 6.7, it usually refers to a particular estimation technique. But, beware of other (or ambiguous) uses of the word “regression,” especially in online resources.

\section{Logic}

Some basic logic is useful for understanding certain parts of econometrics. Theoretically, logic helps you understand the relationships among different conditions, like assumptions for theorems. Practically, logic helps you interpret results.

The following may not be fully technically correct from a philosopher’s perspective (e.g., perhaps I conflate logical implication with the material conditional), but it suffices for econometrics.

\subsection{Terminology}

Many words and notations can refer to the same logical relationship. Let $A$ and $B$ be two statements that can be either true or false. For example, maybe $A$ is “$Y \geq 10$” and $B$ is “$Y \geq 0$.” Or, $A$ is “this animal is a cat,” and $B$ is “this animal is a mammal.” The following ways of describing the logical relationship between $A$ and $B$ all have the same meaning.

1. If $A$ is true, then $B$ is true (often shortened: “if $A$, then $B$”)
2. $A \implies B$
3. $A$ implies $B$
4. $B \iff A$
5. $B$ is **implied by** $A$
6. $B$ is true **if** $A$ is true
7. $A$ is true **only if** $B$ is true
8. $A$ is a sufficient condition for $B$ (shorter: “$A$ is **sufficient** for $B$”)
9. $B$ is a necessary condition for $A$ (shorter: “$B$ is **necessary** for $A$”)
10. $A$ is **stronger** than $B$
11. $B$ is **weaker** than $A$
12. It is impossible for $B$ to be false when $A$ is true (but it is fine if both are true, or both are false, or $A$ is false and $B$ is true)
13. The **truth table** (T=true, F=false):

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$A \implies B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

14. The diagram (everything in $A$ is also in $B$):

![Diagram](image)

To state equivalence of $A$ and $B$, opposite statements can be combined. Specifically, any of the following have the same meaning.

1. $A \iff B$ (meaning both $A \implies B$ and $A \iff B$)
2. $A$ is true **if and only if** $B$ is true (meaning $A$ is true if $B$ is true and $A$ is true only if $B$ is true)
3. $B$ is true if and only if $A$ is true
4. $A$ is necessary and sufficient for $B$
5. $B$ is necessary and sufficient for $A$
6. $A$ and $B$ are equivalent
7. It is impossible for $A$ to be false when $B$ is true, and impossible for $A$ to be true when $B$ is false.
8. The **truth table** (T=true, F=false):

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$A \iff B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

Variations of $A \implies B$ have the following names. Read $\neg A$ as “not $A$”: $\neg A$ is false when $A$ is true, and $\neg A$ is true when $A$ is false.

- $\neg A \implies \neg B$ is the **inverse** of $A \implies B$.
- $B \implies A$ is the **converse** of $A \implies B$. 
• \( \neg B \implies \neg A \) is the contrapositive of \( A \implies B \).

The statement \( A \implies B \) is logically equivalent to its contrapositive. That is, statements “\( A \implies B \)” and “\( \neg B \implies \neg A \)” can be both true or both false, but it’s impossible for one to be true and the other false.

The statement \( A \implies B \) is not logically equivalent to either its inverse or converse. (The inverse and converse are equivalent to each other because the inverse is the contrapositive of the converse.)

For example, let \( A \) be “\( X \leq 0 \)” and let \( B \) be “\( X \leq 10 \).”
• \( A \implies B \): any number below 0 is also below 10.
• The contrapositive is \( X > 10 \implies X > 0 \), which is also true: any number above 10 is also above 0.
• The inverse is \( X > 0 \implies X > 10 \), which is false: e.g., if \( X = 5 \), then \( X > 0 \) but not \( X > 10 \).
• The converse is \( X \leq 10 \implies X \leq 0 \), also false: again if \( X = 5 \), then \( X \leq 10 \) but not \( X \leq 0 \).

6.1.2 Theorems

Theorems all have the same logical structure: if assumption \( A \) is true, then conclusion \( B \) is true. Sometimes \( A \) and \( B \) have multiple parts, like \( A \) is really “\( A_1 \) and \( A_2 \).” The theorem’s practical use is: if we can verify that \( A \) is true, then we know \( B \) is also true.

What if we think \( A \) is false? Then, \( B \) could be false, or it could be true. This may be seen most readily from the picture version of the \( A \) and \( B \) relationship in Section 6.1.1: we could be somewhere inside \( B \) but outside \( A \) (i.e., \( B \) true, \( A \) false); or we could be outside both (both false). That is, as in Section 6.1.1, the theorem \( A \implies B \) is not equivalent to its inverse.

Also from Section 6.1.1, a theorem is equivalent to its contrapositive. That is, if the theorem’s conclusion is false, then we know its assumption is false. (If the assumption has multiple parts like “both \( A_1 \) and \( A_2 \) are true,” then being false means either \( A_1 \) is false or \( A_2 \) is false or both are false.)

6.1.3 Comparing Assumptions

To compare assumptions, the terms “stronger” and “weaker” are most commonly used. Let \( A_1 \) and \( A_2 \) denote different assumptions. Per Section 6.1.1, “\( A_1 \) is stronger than \( A_2 \)” is equivalent to \( A_1 \implies A_2 \), which is also equivalent to “\( A_2 \) is weaker than \( A_1 \).”

All else equal, it is more useful to have a theorem with weaker assumptions because it applies to more settings. That is, if \( A_1 \implies A_2 \), then we prefer a theorem based on \( A_2 \), the weaker assumption. A theorem based on \( A_1 \) can only be used when \( A_1 \) is true. In contrast, a theorem based on \( A_2 \) can be used not only when \( A_1 \) is true (because \( A_1 \implies A_2 \)), but also sometimes when \( A_1 \) is false (but \( A_2 \) is still true).
For example, let assumption $A_1$ be, “a city is in Missouri,” and let assumption $A_2$ be, “a city is in the United States.” Consider the theorems $A_1 \implies B$ and $A_2 \implies B$. (The conclusion is irrelevant here, but to be concrete you could imagine $B$ is “the city is in the northern hemisphere.”) Since Missouri is part of the United States, $A_1 \implies A_2$, i.e., $A_1$ is the stronger assumption and $A_2$ is the weaker assumption. We prefer the theorem based on the weaker assumption because it applies to more cities. For example, only the theorem $A_2 \implies B$ applies to Houston; $A_1$ is false, but $A_2$ is true. (And recall that when $A_1$ is false, the theorem $A_1 \implies B$ does not conclude that $B$ is false; it just says, “I don’t know if $B$ is true or false,” i.e., it is useless.)

**Practice 6.1** (median theorem logic). Consider the theorem, “If sampling is iid, then the sample median consistently estimates the population median.” Hint: draw a picture and/or write it as $A \implies B$.

a) What does this tell us about consistency of the sample median when sampling is not iid?
b) What does this tell us about sampling when the sample median is not consistent?

**Practice 6.2** (mean theorem logic). Consider the theorem, “If sampling is iid and the population mean is well-defined, then the sample mean consistently estimates the population mean.” Hint: there may be multiple possible pictures that show this relationship among $A_1$ (iid), $A_2$ (well-defined), and $B$ (consistency).

a) What does this tell us about consistency of the sample mean when sampling is not iid?
b) What does this tell us about sampling when the sample mean is not consistent?

**Discussion Question 6.1** (logic with feathers). Consider two theorems. Theorem 1 says, “If $X$ is an eagle, then it has feathers.” Theorem 2 says, “If $X$ is a bird, then it has feathers.”

a) Describe each theorem logically: what’s the assumption ($A$), what’s the conclusion ($B$), what’s the relationship?
b) State Theorem 1’s contrapositive; is it true?
c) Compare: does Theorem 1 or Theorem 2 have a stronger assumption? Why?
d) Compare: which theorem is more useful? (Which applies to more situations?)

### 6.2 Preliminaries

$\implies$ Kaplan video: Joint, Marginal, and Conditional Distributions

Before getting to regression, some simpler material may provide intuition. (If it is not familiar to you from a previous statistics class, then you may want to consult additional resources for a deeper understanding; or you may not.) In Section 6.2, there is no data; only the population is considered.
6.2.1 Population Mean Model in Error Form

To help understand the conditional mean model, we start with an unconditional mean model. That is, interest is in $\mu_Y \equiv E(Y)$ for a single random variable $Y$, as in Chapter 2.

There are two ways to write the unconditional mean “model.” Both look silly and over-complicated, but they help bridge Chapter 4 and Chapter 6. First, the mean can be written directly:

$$E(Y) = \mu_Y. \quad (6.1)$$

Second, in terms of an error term $U$, the error form of this model is

$$Y = \mu_Y + U, \quad E(U) = 0, \quad (6.2)$$

Models (6.1) and (6.2) are equivalent. Taking the mean of both sides of (6.2), using the linearity property from (2.21),

$$E(Y) = E(\mu_Y + U) = E(\mu_Y) + E(U) = \mu_Y + 0 = \mu_Y. \quad (6.3)$$

The error term $U$ has a precise statistical definition and meaning, but no causal or economic meaning. Defining the mean error term as

$$U \equiv Y - E(Y) \quad (6.4)$$

always implies

$$E(U) = E[Y - E(Y)] = E(Y) - E(Y) = 0. \quad (6.5)$$

Thus, the property $E(U) = 0$ in (6.2) is true essentially by definition, not an assumption that can be false. By analogy, this is like defining $U$ to be an equilateral triangle, in which case the property “all angles are equal” is always true, not an additional assumption. However, $U$ has no causal or economic meaning: it is simply the difference between an individual’s $Y$ and the population mean $E(Y)$.

The error form often facilitates theoretical analysis of estimators, but in practice the more direct model may be easier to interpret.

6.2.2 Joint and Marginal Distributions

To understand regression, you must understand conditional distributions. To understand conditional distributions, it helps to understand joint distributions and marginal distributions.

The joint distribution is the distribution of values of $(X,Y)$ together, which can be any combination of the variable types in Section 2.3. For example, $X$ could be categorical and $Y$ continuous, or $X$ and $Y$ both discrete, or $X$ continuous and $Y$ ordinal, etc. Since there are so many combinations, they are not all enumerated here. Further, eventually we’ll focus on conditional distributions, in which case the variable type of $X$ does not matter as much for interpretation. For regression, the focus is on numeric (discrete or continuous) $X$ and $Y$. Implicitly, this also applies to categorical variables that
have been turned into dummy variables with the indicator function, like \( X = 1_{\text{cat}} \) or \( Y = 1_{\text{employed}} \).

For \((X, Y)\) with non-continuous variable types, the joint distribution can be described by a PMF. Like before, the PMF states the probability of each possible value. The difference is that “possible values” are now pairs of values \((x, y)\) instead of single values. For example, a possible value could be \((-5, 7)\), or \((\text{cat, dog})\). (It is more difficult to gain intuition with continuous variable types, but the idea of a PDF can be extended to multiple variables.)

Each joint probability in the PMF can be written multiple equivalent ways:

\[
P((X, Y) = (x, y)) = P(X = x, Y = y) = P(X = x \text{ and } Y = y).
\]  

(6.6)

For example, consider the joint distribution of dummy variables for employment and marital status. Let \( Y = 1 \) if somebody is employed, and \( Y = 0 \) if not. Let \( X = 1 \) if somebody is married, and \( X = 0 \) if not. The joint distribution of employment and marital status describes the probabilities of each possible value of the vector \((X, Y)\), i.e., the PMF of the vector \((X, Y)\). There are four possible values: unmarried and unemployed \((0, 0)\); unmarried and employed \((0, 1)\); married and unemployed \((1, 0)\); and married and employed \((1, 1)\). Since these categories are mutually exclusive and exhaustive, the four probabilities must sum to 1 (i.e., 100\%). Table 6.1 shows an example.

Table 6.1: Joint distribution of marital status \((X)\) and employment status \((Y)\).

<table>
<thead>
<tr>
<th></th>
<th>(Y = 0)</th>
<th>(Y = 1)</th>
<th>Marginal for (X) (row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X = 0)</td>
<td>0.10</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>(X = 1)</td>
<td>0.20</td>
<td>0.60</td>
<td>0.80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Marginal for (Y) (column sum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y = 0)</td>
<td>0.30</td>
</tr>
<tr>
<td>(Y = 1)</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 6.1 shows both joint and marginal probabilities. Here, the joint probability values can be written as \(p_{xy} \equiv P(X = x, Y = y)\), or equivalently \(P((X, Y) = (x, y))\). This is analogous to the scalar (one variable) PMF that described \(P(Y = y)\) for different values \(y\), but replacing \(Y\) with \((X, Y)\) and \(y\) with \((x, y)\). The joint probabilities shown inside the box are \(p_{00} = 0.10\) (i.e., 10\%), \(p_{01} = 0.10, p_{10} = 0.20\), and \(p_{11} = 0.60\). These sum to 1.

A marginal probability (or unconditional probability) considers just one of the random variables, ignoring the other. Specifically, the outer values in Table 6.1 show the marginal probabilities to be \(P(X = 0) = 0.20\) (at the right end of the \(X = 0\) row), \(P(X = 1) = 0.80, P(Y = 0) = 0.30\) (at the bottom of the \(Y = 0\) column), and \(P(Y = 1) = 0.70\). These probabilities describe the marginal distribution of each random variable, or more specifically the marginal PMFs. That is, \(X\) by itself is a random variable with \(P(X = 0) = 0.20\) and \(P(X = 1) = 0.80\): the population proba-
bility of an individual being married is 0.8 (80%). Similarly, by itself, $Y$ is a random variable with $P(Y = 0) = 0.30$ and $P(Y = 1) = 0.70$.

**Discussion Question 6.2** (joint distribution and causality). Consider two binary random variables, $X$ and $Y$, whose joint distribution is described by the probabilities $P(X = 0, Y = 0) = 0.4$, $P(X = 0, Y = 1) = 0.1$, $P(X = 1, Y = 0) = 0.1$, and $P(X = 1, Y = 1) = 0.4$. Note $P(X = Y) = 0.8 > 0.2 = P(X \neq Y)$. Hint: think about some concrete examples of $X$ and $Y$ (marital status, employment, rain, long commute time, etc.); to prove something is “possible” only requires a single example where it is true.

a) Explain why this joint distribution suggests some type of relationship between $X$ and $Y$.

b) Given the joint distribution, is it possible that $X$ has a causal effect on $Y$? Why/not?

c) Given the joint distribution, is it possible that $X$ does not have a causal effect on $Y$? Why/not?

### 6.2.3 Conditional Distributions

For non-continuous variables, the **conditional distribution** consists of the conditional probabilities of all different possible values. The conditional distribution of $Y$ given $X = x$ consists of the conditional PMF, i.e., the conditional probability of each possible value $y$ given $X = x$.

The **conditional probability** of one event (like $Y = 1$) given another event (like $X = 1$) considers only the times when the conditioning event (like $X = 1$) occurs, and then takes the proportion of those times that the first event (like $Y = 1$) occurs. Mathematically, the conditional probability $P(Y = 1 \mid X = 1)$ can be read as “the probability that $Y$ equals one conditional on $X$ equal to one” or “the probability of $Y$ being one given $X$ equals one” or other variations. More generally, $P(Y = y \mid X = x)$ is “the probability that $Y$ equals $y$ conditional on $X$ equal to $x$.”

For non-continuous variables, a conditional probability can be written in terms of joint and marginal probabilities. Specifically,

$$P(Y = y \mid X = x) = \frac{P(Y = y, X = x)}{P(X = x)}.$$  \hspace{1cm} (6.6)

(This doesn’t apply to continuous $X$ since the denominator would be $P(X = x) = 0$.)

There is nothing mathematically special about the labels $X$ and $Y$ here. However, conventional regression notation corresponds to conditioning on the variable named $X$. To examine $X$ conditional on $Y$, we could just switch the labels and then examine $Y$ conditional on $X$.

For example, consider the probability of employment ($Y = 1$) conditional on being married ($X = 1$). Applying (6.6),

$$P(Y = 1 \mid X = 1) = \frac{P(Y = 1, X = 1)}{P(X = 1)}.$$  \hspace{1cm} (6.7)
The denominator is the proportion of the population that’s married. The numerator is the proportion of the population that’s both married and employed.

Examples

For intuition, you can imagine the population is actually 100 people rather than abstract probabilities. You may have unknowingly computed (sample) conditional probabilities in grade school if you ever answered questions like “What proportion of the boys in our class are wearing glasses?” or “What proportion of students with black hair are wearing a sweater?”

In Table 6.1, multiplying values by 100 gives the number of people in each of the four cells: 10 people are unmarried and unemployed with \((X = 0, Y = 0)\); another 10 are unmarried but employed with \((X = 0, Y = 1)\); 20 people are married but not employed with \((X = 1, Y = 0)\); and 60 people are married and employed, \((X = 1, Y = 1)\). Probabilities are proportions; e.g., \(P(X = 1, Y = 1) = 0.60\), so the proportion of married and employed individuals in the 100-person population is \(60/100 = 0.60 = 60\%\).

Table 6.2 shows the number of people with different values, parallel to Table 6.1.

<table>
<thead>
<tr>
<th></th>
<th>not employed</th>
<th>employed</th>
<th>Marginal (sum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>not married</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>married</td>
<td>20</td>
<td>60</td>
<td>80</td>
</tr>
<tr>
<td>Marginal (sum)</td>
<td>30</td>
<td>70</td>
<td>100</td>
</tr>
</tbody>
</table>

Using Table 6.2 to continue with the 100-person population, the conditional probability \(P(Y = 1 \mid X = 1)\) asks: within the group of married individuals \((X = 1)\), what proportion of them are employed? There are 60 married and employed individuals, and 20 married who are not employed, so 80 total. This 80 is 100 times the marginal probability \(P(X = 1) = 0.80\). Out of those 80, 60 are employed. Thus, the proportion of married individuals who are employed is \(60/80 = 0.75 = 75\%\). That is, to compute the conditional probability, we take the “joint” number of individuals who are both married and employed (both \(X = 1\) and \(Y = 1\)), and divide by the “marginal” number of married individuals \((X = 1)\). Similarly, the proportion of married individuals who are not employed is \(20/80 = 0.25 = 25\%\). For the unmarried group (20 individuals total), the proportion who are employed is \(10/20 = 0.5 = 50\%\), which is also the proportion who are not employed.
6.2.4 Conditional Mean

The **conditional mean** is just the mean of a conditional distribution. Conditional on a particular value \( X = x \), like \( X = 1 \), there is a conditional distribution of \( Y \). The mean of that conditional distribution is written

\[
E(Y \mid X = x).
\]  

(6.9)

To read (6.9) aloud, you could say, “the conditional mean of \( Y \) given \( X = x \),” or “the mean of \( Y \) conditional on \( X = x \).”

**Examples**

From Table 6.1, we can compute a conditional mean. We already computed the conditional distribution of employment status (\( Y \)) conditional on being married (\( X = 1 \)): \( P(Y = 1 \mid X = 1) = 0.75 \) and \( P(Y = 0 \mid X = 1) = 0.25 \). The mean of that conditional distribution is written \( E(Y \mid X = 1) \). We can use the usual expected value formula, plugging in conditional probabilities. For comparison, the unconditional and conditional (on \( X = 1 \)) means of \( Y \) are, respectively,

\[
E(Y) = (0) P(Y = 0) + (1) P(Y = 1) = (0)(0.3) + (1)(0.7) = 0 + 0.7 = 0.7.
\]  

(6.10)

\[
E(Y \mid X = 1) = (0) P(Y = 0 \mid X = 1) + (1) P(Y = 1 \mid X = 1)
\]

\[
= (0)(0.25) + (1)(0.75) = 0 + 0.75 = 0.75.
\]  

(6.11)

Since \( Y \) is binary (0 or 1), the (conditional) mean is the (conditional) probability of \( Y = 1 \): \( E(Y) = P(Y = 1) = 0.7 \) and \( E(Y \mid X = 1) = P(Y = 1 \mid X = 1) = 0.75 \).

Conditional means can be computed similarly for non-binary \( Y \) and \( X \). For example, imagine \( Y \) is hours worked per week, which is either 0, 20, or 40, and \( X \) is years of education, which is either 11, 12, or 16. The conditional mean is

\[
E(Y \mid X = x) = \sum_{j \in \{0, 20, 40\}} (j) P(Y = j \mid X = x)
\]

\[
= (0) P(Y = 0 \mid X = x) + (20) P(Y = 20 \mid X = x) + (40) P(Y = 40 \mid X = x).
\]  

(6.12)

Table 6.3: Joint distribution of education (\( X \)) and weekly hours worked (\( Y \)).

<table>
<thead>
<tr>
<th></th>
<th>( X = 11 )</th>
<th>( X = 12 )</th>
<th>( X = 16 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = 0 )</td>
<td>0.10</td>
<td>0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>( Y = 20 )</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>( Y = 40 )</td>
<td>0.10</td>
<td>0.15</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 6.3 shows an example joint distribution of such an \( X \) and \( Y \), from which conditional means can be computed. The values in the
table are joint probabilities; e.g., the entry in the last column of the second row shows \( P(X = 12, Y = 40) = 0.15 \). Consider the conditional mean \( E(Y \mid X = 16) \). To apply (6.12) requires the conditional probabilities, which can be computed using (6.7). First, the marginal probability sums all entries in the row:

\[
P(X = 16) = 0.10 + 0.10 + 0.30 = 0.5. \tag{6.13}
\]

Second, plugging this into (6.7),

\[
\begin{align*}
P(Y = 20 \mid X = 16) &= \frac{P(Y = 20, X = 16)}{P(X = 16)} = \frac{0.10}{0.50} = 0.2, \\
P(Y = 40 \mid X = 16) &= \frac{P(Y = 40, X = 16)}{P(X = 16)} = \frac{0.30}{0.50} = 0.6. \tag{6.14}
\end{align*}
\]

Third, plugging these into (6.12),

\[
E(Y \mid X = 16) = 0 + (20)(0.2) + (40)(0.6) = 4 + 24 = 28. \tag{6.15}
\]

As a sanity check, note that the probability of \( Y = 40 \) is higher than that of \( Y = 0 \), so it makes sense that the conditional mean is above 20. The specific value \( E(Y \mid X = 16) \) says that within the part of the population with \( X = 16 \) years of education, the mean weekly hours worked is 28.

### 6.2.5 Comparison of Joint, Marginal, and Conditional Distributions

The joint distribution has all the possible information about the distribution of random variables \((X,Y)\). Rearranging (6.7), each joint probability can be reconstructed by multiplying the appropriate conditional and marginal probabilities:

\[
P(X = x, Y = y) = P(Y = y \mid X = x) P(X = x). \tag{6.16}
\]

Thus, knowing the joint distribution has the same information as knowing both the conditional (of \( Y \) given \( X = x \)) and marginal (of \( X \)) distributions. However, the conditional distributions alone (without the marginals) contain less information than the joint distribution; i.e., there are multiple possible joint distributions that would be consistent with a single set of conditional distributions. Similarly, the marginal distributions of \( X \) and \( Y \) alone (without the conditionals) contain less information than the joint distribution.

Going into regression, we’ll focus on the conditional means \( E(Y \mid X = x) \), which are summary features of conditional distributions, which in turn are summaries of the full joint distribution. That is, conditional means (and regression) only learn one particular feature about the population joint distribution of \((X,Y)\). As discussed in Section 2.3.1, there is a tradeoff between learning more information and having a summary that is more easily understood and communicated.
6.2.6 Independence and Dependence

If random variables $X$ and $Y$ are independent, then they are completely unrelated, statistically speaking. Notationally, independence is usually written as $X \perp \perp Y$, which is equivalent to $Y \perp \perp X$.

Independence implies equality of marginal and conditional distributions. Mathematically, the marginal (unconditional) distribution of $Y$ is the same as the conditional distribution of $Y$ given $X = x$, for any $x$. Intuitively, if $X$ is unrelated to $Y$, then knowing the value of $X$ has no information about the value of $Y$.

This characterization of independence can be written in terms of a PMF or CDF. If $Y$ is not continuous and thus has a PMF, then independence implies the marginal PMF equals the conditional PMF: for any possible $y$ and $x$ values,

$$Y \perp \perp X \implies P(Y = y) = P(Y = y \mid X = x). \quad (6.17)$$

If $Y$ is not a nominal categorical variable and thus has a CDF, then for any possible $y$ and $x$:

$$Y \perp \perp X \implies P(Y \leq y) = P(Y \leq y \mid X = x). \quad (6.18)$$

Consequently, independence implies equality of marginal and conditional means, known as mean independence. That is, for any possible $x$ value,

$$Y \perp \perp X \implies E(Y) = E(Y \mid X = x). \quad (6.19)$$

Independence implies many other properties, too, like $\text{Cov}(X, Y) = \text{Corr}(X, Y) = 0$ and $P(X = x, Y = y) = P(X = x) P(Y = y)$.

The opposite of independence is dependence. If any condition implies by independence does not hold, then the variables are dependent, written $X \not\perp \not\perp Y$. For example, if $\text{Corr}(X, Y) \neq 0$, then $X \not\perp \not\perp Y$. Or if $E(Y \mid X = 1) \neq E(Y \mid X = 0)$, then $X$ and $Y$ are neither independent nor mean independent.

6.3 Population Model: Conditional Expectation Function

This and the following sections consider what we want to learn about the population, and how we can write it mathematically. There is no data, no estimation, no uncertainty.

A model describes the relationship between two (or more) variables, like education and income. If it describes how income changes with education, then income is the usually written as $Y$ and called the outcome variable, regressand, dependent variable, left-hand side variable, or response variable, while education is written as $X$ and called the regressor, independent variable, right-hand side variable, predictor, covariate, or conditioning variable.
Like before, these variables are treated mathematically as random variables. The “population” is a joint probability distribution of the observable variables.

There are different models for different types of relationships between two variables. Section 6.3 models a statistical relationship with interpretations for description or prediction, whereas Sections 6.4 and 6.5 model causal relationships. Sometimes the descriptive and causal models coincide, but generally they differ.

This section combines Sections 6.2.1 and 6.2.3 to get a conditional mean regression model.

**In Sum: Conditional Expectation Function**

| Description: shows mean \( Y \) for each subpopulation with same \( X \), \( E(Y \mid X = x) \) |
| Prediction: with quadratic loss, optimal prediction of \( Y \) given \( X = x \) is \( E(Y \mid X = x) \) |
| Causality: CEF difference sometimes has causal interpretation (Section 6.6) |

### 6.3.1 Conditional Expectation Function

Using (6.9), let \( m(\cdot) \) be the **conditional expectation function** (CEF) of \( Y \) given \( X \):

\[
m(x) \equiv E(Y \mid X = x). \tag{6.20}
\]

That is, the CEF \( m(\cdot) \) takes a value of \( x \) as input, like \( x = 1 \), and tells us the corresponding conditional mean of \( Y \), like \( E(Y \mid X = 1) = 7 \).

It helps to remember what’s random and what’s non-random. The CEF \( m(\cdot) \) is a non-random function, just as \( E(Y) \) is non-random. For any \( X = x \), \( Y \) has a conditional distribution whose mean is \( m(x) \), a non-random value. You can draw a graph of a CEF just like you graphed any other (non-random) function in high school. In contrast, \( m(X) \) is a random variable. That is, there are multiple possible values of \( m(X) \) because there are multiple possible values of \( X \).

If \( X \) is binary as in this chapter, then there are two conditional means of interest:

\[
m(0) = E(Y \mid X = 0), \quad m(1) = E(Y \mid X = 1). \tag{6.21}
\]

There are two possible approaches. First, these two conditional means could be studied directly, similar to Chapter 4. That is, \( Y^A \) has the distribution of \( Y \) given \( X = 0 \), and \( Y^B \) has the distribution of \( Y \) given \( X = 1 \). Second, the conditional means can be captured in a CEF regression model.

For example, consider Table 6.1. From (6.11), \( m(1) \equiv E(Y \mid X = 1) = 0.75 \). Also,

\[
m(0) \equiv E(Y \mid X = 0) = (0) \frac{P(Y = 0 \mid X = 0)}{P(X = 0)} + (1) \frac{P(Y = 1 \mid X = 0)}{P(X = 0)} = 0.1 \cdot 0.2 = 0.02.
\]
Thus, the CEF is \( m(0) = 0.5, m(1) = 0.75 \). Also, from Table 6.1, the marginal distribution of \( X \) is \( P(X = 0) = 0.2, P(X = 1) = 0.8 \). Thus, \( m(X) \) is a random variable with
\[
P(m(X) = 0.5) = P(X = 0) = 0.2, \quad P(m(X) = 0.75) = P(X = 1) = 0.8.
\]

\[\text{(6.22)}\]

6.3.2 CEF Error Term

Extending (6.4), the **CEF error term** is defined as
\[
V \equiv Y - m(X),
\]
\[\text{(6.23)}\]
i.e., the difference between an individual’s actual outcome \( Y \) and the CEF evaluated at her \( X \) value, \( m(X) \). As always, other letters could be used besides \( V \), like \( U \) or \( W \); in other textbooks, you may see \( u \) or \( e \) or \( \epsilon \). Since \( Y \) and \( X \) are random variables, so is \( V \); e.g.,
\[
P(V = 0) = P(Y - m(X) = 0).
\]

For example, let \( X = 1 \) indicate a college degree (and \( X = 0 \) otherwise), and \( Y \) is income. Then, \( m(0) \) is the mean income among the no-college population and \( m(1) \) is mean income among college degree holders. If you are a successful tech company CEO who went to college (\( X = 1 \)), then your \( Y \) is high above \( m(1) \), so your CEF error in (6.23) is very large and positive. Or, if you didn’t go to college (\( X = 0 \)) and make exactly the mean income for that group, your \( Y \) equals \( m(0) \), so your CEF error is \( V = 0 \). Or, if you went to a fancy college but decided to live off your parents’ wealth and earn no income, then your \( Y = 0 \), so your \( V = Y - m(1) = 0 - m(1) \), very negative.

The CEF error has conditional mean zero. Extending (6.5), for any \( X = x \),
\[
E(V \mid X = x) = E(Y - m(X) \mid X = x) = E(Y \mid X = x) - E(m(X) \mid X = x) = m(x) - m(x) = 0.
\]
Equivalently,
\[
E(V \mid X) = 0.
\]
\[\text{(6.24)}\]
That is, \( E(V \mid X) \) is a random variable depending on \( X \), but it equals zero for every possible value of \( X \); or, just imagine “\( E(V \mid X = x) = 0 \) for all \( x \)” every time you see “\( E(V \mid X) = 0 \).”

As in Section 6.2.1, this conditional mean zero property is not an assumption: it is true by definition for any CEF error defined as in (6.23). By analogy, if we define \( V \) as a square, then it always has the property of having four equal sides and four equal angles; such properties are not additional assumptions.

6.3.3 CEF Model in Error Form

Given (6.23), extending (6.2), the CEF model in error form is
\[
Y = m(X) + V, \quad E(V \mid X) = 0.
\]
\[\text{(6.25)}\]
The statement $E(V \mid X) = 0$ is equivalent to saying $m(x) = E(Y \mid X = x)$, i.e., that $m(\cdot)$ is the CEF. Again, it is not an assumption about $V$; it is just stating what type of model this is. Equation (6.25) can apply to non-binary $X$, too, as in later chapters.

### 6.3.4 Linear CEF Model

With binary $X$, the model in (6.25) is equivalent to

$$Y = m(0) \mathbb{1}\{X = 0\} + m(1) \mathbb{1}\{X = 1\} + V$$

$$= m(0)(1 - X) + m(1)(X) + V$$

$$= m(0) + [m(1) - m(0)]X + V.$$  

To double-check: when $X = 0$, then $[m(1) - m(0)]X = 0$, so $Y = m(0) + V$, as in the original (6.25). When $X = 1$, then $m(0) + [m(1) - m(0)]X = m(0) + m(1) - m(0) = m(1)$, so $Y = m(1) + V$, also as in (6.25). Thus, (6.25) and (6.27) are equivalent for binary $X$.

The CEF model in (6.27) can be rewritten yet again to yield a more familiar, conventional structure. Following conventional notation, let $\beta_0 \equiv m(0)$ and $\beta_1 \equiv m(1) - m(0)$. Plugging these definitions into (6.27),

$$Y = \beta_0 + \beta_1 X + V,$$

$$E(V \mid X) = 0. \quad (6.28)$$

In (6.28), $\beta_0$ and $\beta_1$ are called the parameters. Greek letters like $\beta$ are commonly used to denote unknown parameters in a population model. In the frequentist framework, these are seen as unknown but fixed (non-random) values, whereas $Y$, $X$, and $V$ are random variables. In (6.28) specifically, $\beta_0$ is the intercept, and $\beta_1$ is the slope. Sometimes regression parameters are called coefficients; $\beta_1$ is the slope coefficient or the coefficient on $X$.

Model (6.28) is a linear CEF model. It is a “CEF” model because $E(Y \mid X = x) = \beta_0 + \beta_1 x$, or $E(Y \mid X) = \beta_0 + \beta_1 X$. The “linear” part is explained in Section 8.2.1; for now, it suffices to recall that a graph of $\beta_0 + \beta_1 x$ is a straight line.

Since $X$ is binary, no assumptions were required to write (6.28) given (6.25). However, when $X$ has more than two possible values, it is more complicated, as discussed in Chapter 7. For now, with binary $X$, the CEF model can always be written as in (6.28).

### 6.3.5 Interpretation, Description, and Prediction

**Practice 6.3 (regression parameter units).** Let $Y$ be salary, measured in $\$$/yr, and let $X$ be the number of college degrees an individual has, either $X = 0$ or $X = 1$. In (6.28), what are the units of measure for $\beta_0$ and $\beta_1$, respectively?

To interpret (6.28), first consider the units of measure. The left-hand side is just $Y$, the outcome variable. Since they are equal, the right-hand side must have the same units. Thus, each of the three right-hand side terms must have the same units as $Y$:

1. $\beta_0$ has the same units as $Y$.  

2. $\beta_1 X$ has the same units as $Y$, so the units of $\beta_1$ are the units of $Y$ divided by the units of $X$.

3. $V$ has the same units as $Y$.

For example, if $Y$ is measured in $$/yr and $X$ is the number of college degrees, then the units of $\beta_0$ are $$/yr and the units of $\beta_1$ are $$(/yr)/(# degrees)$$.

For description, the model in (6.28) is the CEF, a summary of the conditional distribution. As seen earlier, $\beta_0 = m(0) = E(Y | X = 0)$, the mean outcome among all individuals with $X = 0$. Also, $\beta_1 = m(1) - m(0) = E(Y | X = 1) - E(Y | X = 0)$ (6.29) is the difference between the mean outcome for the $X = 1$ subpopulation and the mean outcome for the $X = 0$ subpopulation.

A common phrase to describe such statistical (but maybe not causal) differences is associated with. For example, if individuals with a college degree have a mean annual income that is $20,000/yr higher than the mean annual income of non-college individuals, then $\beta_1 = 20,000/yr and you could say, “On average, having a college degree is associated with having a $20,000/yr higher annual income.” This does not claim that going to college has such a causal effect on income, only a statistical association.

For prediction, the model in (6.28) is also helpful. Section 2.5.4 says the mean is the best predictor if the loss function is quadratic. This continues to be true conditional on $X$: the conditional mean of $Y$ given $X = x$ is the best predictor given quadratic loss. Formally, letting $g(\cdot)$ denote any possible guess (of $Y$, as a function of $X$),

$$m(\cdot) = \arg\min_{g(\cdot)} E[(Y - g(X))^2].$$ (6.30)

In terms of the model parameters in (6.28), the best predictor of $Y$ given $X = 0$ is $\beta_0$, and the best predictor of $Y$ given $X = 1$ is $\beta_0 + \beta_1$. Combining these, the best predictor of $Y$ given $X$ is $\beta_0 + \beta_1 X$.

### 6.3.6 Interpretation with Values Besides 0 and 1

What if $X$ has only two possible values, but they aren’t 0 and 1? (Technically, such $X$ is still “binary,” but usually people mean 0 and 1 when they say “binary.”)

For example, let $Y$ again be income, and let $X$ be education, but now measured in years. Instead of comparing individuals with a college degree to those without, imagine comparing individuals with $X = 12$ years of education to $X = 13$. By convention, “years of education” is measured starting in grade 1. In the U.S., the last year of high school is grade 12, so completing high school means $X = 12$; taking a year of college classes leads to $X = 13$.

The fundamental conditional means are $m(12) = E(Y | X = 12)$ and $m(13) = E(Y | X = 13)$, but $\beta_0$ and $\beta_1$ in (6.28) have different meanings than before. Even the units of $\beta_1$ are different, because the units of $X$ have changed. Recall that the units of $\beta_1$ are units
of $Y$ divided by units of $X$. In both cases, $Y$ is measured in $$/yr. Before, $X$ was number of college degrees, so the units of $\beta_1$ were ($$/yr)/(# \text{degrees})$. Now, $X$ is measured in years (of education), so the units of $\beta_1$ are ($$/yr)/(yr)$. By this alone, $\beta_1$ must now have a different interpretation; it turns out $\beta_0$ does, too.

The parameters $\beta_0$ and $\beta_1$ from the CEF model in (6.28) can now be written in terms of $m(12)$ and $m(13)$. Writing the CEF as $m(X) = \beta_0 + \beta_1 X$ as in (6.28),

$$
m(13 \text{ yr}) = \beta_0 + (13 \text{ yr})\beta_1, \quad m(12 \text{ yr}) = \beta_0 + (12 \text{ yr})\beta_1,
$$

$$
m(13 \text{ yr}) - m(12 \text{ yr}) = (\beta_0 - \beta_0) + (13 \text{ yr} - 12 \text{ yr})\beta_1 = \beta_1,
$$

$$
\beta_1 = [m(13 \text{ yr}) - m(12 \text{ yr})]/(13 \text{ yr} - 12 \text{ yr}) = [m(13 \text{ yr}) - m(12 \text{ yr})]/m(13 \text{ yr})
$$

The meaning of $\beta_1$ is qualitatively similar to before: it is the difference in mean income between the high and low education subpopulations. However, instead of “per degree,” the units are now “per year” (one year of college education).

In contrast, the interpretation of $\beta_0$ is very different and not very meaningful. Before, $\beta_0 = m(0)$ was the mean of the low education subpopulation. Here, $\beta_0$ takes the low education mean, $m(12)$, and then subtracts 12 times the mean difference, $12[m(13) - m(12)]$. That is, $\beta_0$ is trying to extrapolate from the means for $X = 13$ and $X = 12$ all the way down to $X = 0$. This is not the true mean income for individuals with zero years of education, $m(0)$: only the means for $X = 12$ and $X = 13$ are known. Rather, it is just a guess of $m(0)$ based on $m(12)$ and $m(13)$, and probably a very poor guess. Further, individuals with zero years of education may be very rare or even nonexistent in the larger population, in which case it is not even interesting to guess. So, while the slope $\beta_1$ continues to have meaning, the intercept $\beta_0$ may not, unless $X = 0$ is a possible and interesting value.

Adding another twist: what if instead of $X = 12$ and $X = 13$ (years of education), we compare $X = 12$ and $X = 16$? That is, instead of comparing high school ($X = 12$) to one year of college ($X = 13$), it is compared to a typical four-year college degree ($X = 12 + 4 = 16$), more similar to our initial inquiry. With $m(X) = \beta_0 + \beta_1 X$ as in (6.28),

$$
m(16 \text{ yr}) = \beta_0 + (16 \text{ yr})\beta_1, \quad m(12 \text{ yr}) = \beta_0 + (12 \text{ yr})\beta_1,
$$

$$
m(16 \text{ yr}) - m(12 \text{ yr}) = (\beta_0 - \beta_0) + (16 \text{ yr} - 12 \text{ yr})\beta_1 = (4 \text{ yr})\beta_1,
$$

$$
\beta_1 = [m(16 \text{ yr}) - m(12 \text{ yr})]/(4 \text{ yr}) = (1/4)[m(16 \text{ yr}) - m(12 \text{ yr})]/m(16 \text{ yr})
$$

$$
\beta_0 = m(12 \text{ yr}) - (12 \text{ yr})\beta_1 = m(12 \text{ yr}) - 3[m(16 \text{ yr}) - m(12 \text{ yr})]/m(16 \text{ yr})
$$

Again, $\beta_0$ tries to extrapolate from $m(16 \text{ yr})$ and $m(12 \text{ yr})$ to guess $m(0 \text{ yr})$. Again, such a guess is both inaccurate and irrelevant.

The slope parameter $\beta_1$ is different than before, but still meaningful. It takes the mean income difference $m(16 \text{ yr}) - m(12 \text{ yr})$ and then divides by 4yr. This computes a per-year average difference.
This idea is similar to describing a 400 km car trip that took 5 hr by saying the average speed was \( \frac{400 \text{ km}}{5 \text{ hr}} = 80 \text{ km/hr} \). This may be easier to interpret since we more commonly think of \( \text{km/hr} \), but it doesn’t mean that the speed was constant during the whole trip; e.g., it may have been slower in the first half due to traffic. Similarly, \( \beta_1 \) does not mean each of the four college years is associated with the same increase in mean income. For example, the fourth year may have the biggest increase if the college degree (not the education itself) matters most.

### 6.4 Population Model: Potential Outcomes

In contrast to a purely “statistical” model like the CEF model, we could imagine a causal model that shows causal relationships between variables. One way to do this is with potential outcomes, as introduced in Section 4.4. Again, let \( Y^U \) and \( Y^T \) denote the untreated and treated potential outcomes, respectively.

Here, the two observable variables are the observed outcome \( Y \) and the treatment indicator (or treatment dummy) \( X \). That is, \( X = 1 \) if an individual was “treated” and \( X = 0 \) if not. As before, “treatment” is interpreted very broadly, including things like going to a charter school, a right-to-work law, a tax policy, or even personal characteristics.

The observed outcome is

\[
Y = (Y^U)(1-X) + (Y^T)(X). \tag{6.31}
\]

Plugging in \( X = 0 \) yields \( Y = Y^U \), whereas plugging in \( X = 1 \) yields \( Y = Y^T \). So, we observe \( Y = Y^U \) if the individual is untreated and \( Y = Y^T \) if treated.

Notationally, potential outcomes notation writes \( Y \) as a function of \( X \). This is what (6.31) shows: the treated potential outcome is \( Y(1) = Y^T \), and the untreated potential outcome is \( Y(0) = Y^U \).

Equation (6.31) can be rearranged to look more like a linear model in error form. First,

\[
Y = Y^U + X(Y^T - Y^U). \tag{6.32}
\]

This is a random coefficient model with intercept \( Y^U \) and slope \( Y^T - Y^U \). The intercept and slope coefficients are “random” in that they can have different possible values for different individuals.

Second, the random coefficients can be turned into non-random coefficients by adding an error term. Define \( \beta_0 \equiv \text{E}(Y^U) \) and \( \beta_1 \equiv \text{E}(Y^T - Y^U) \), i.e., the ATE. Then, (6.32) becomes

\[
Y = [Y^U + \beta_0 - \beta_0] + X[Y^T - Y^U + \beta_1 - \beta_1] \equiv U
= \beta_0 + X\beta_1 + Y^U - \beta_0 + X[Y^T - Y^U - \beta_1]. \tag{6.33}
\]

Equation (6.33) has the same structure as (6.28), but very different meaning. The parameter \( \beta_1 \) is the ATE of \( X \) on \( Y \); it has a
causal meaning, not just statistical meaning. However, we don’t know if \( E(U \mid X) = 0 \). The error term \( U \) might have such statistical properties, but the model itself does not define \( U \) in terms of statistical properties, but rather in terms of potential outcomes.

### 6.5 Population Model: Structural

A **structural model** also captures causal relationships. The assumption is that the model itself does not change even when variable values and policies change. (‘‘Policy’’ has a broad meaning here: policies of countries, firms, schools, etc., or even just personal decisions.) More specifically, if we want to assess the causal effect of a certain policy, then the structural model should be **invariant** to that particular policy. That is, the policy may change the population distribution of variables, but it cannot change the structural model itself, otherwise the model is not useful.

#### 6.5.1 Linear Structural Model

Consider the linear structural model

\[
Y = \beta_0 + \beta_1 X + U. \tag{6.34}
\]

Unlike in a CEF, the structural model’s \( \beta_1 \) and \( U \) have economic and/or causal meaning by definition. In (6.34), \( \beta_1 \) is called a **structural parameter** (as is \( \beta_0 \)). It has some economic or causal interpretation, like an elasticity or demand curve slope. Similarly, \( U \) is called the **structural error term**. This \( U \) can be interpreted as the aggregation of all other variables that causally determine \( Y \). We can think about \( U \) economically, not just statistically. It’s possible \( E(U \mid X) = 0 \), but usually not. Conversely, the CEF error \( Y - m(X) \) usually does not have causal or economic meaning.

Consider a structural model like (6.34) for the example where \( Y \) is income and \( X \) is a college degree dummy. Then, \( U \) contains everything else that helps determine a person’s income: their occupation, their different skill levels (human capital), where they live (city/country), etc.

**Warning on notation:** I used \( V \) in the CEF model in (6.25) and \( U \) here to help you avoid confusion, but they are simply letters. I could have used \( U \) in both models, or \( V \) in both, or \( \epsilon \), or anything else. So, don’t think \( V \) always means CEF and \( U \) means not CEF.)

With only a single binary \( X \), (6.33) and (6.34) seem very similar. Still, as seen in Section 6.6, the reduced form approach to identifying the ATE involves assumptions about potential outcomes, whereas the structural approach involves assumptions about \( X \) and \( U \).

Superficially, it appears (6.34) claims the causal effect of \( X \) on \( Y \) is the constant \( \beta_1 \), the same for everybody, but see Section 6.5.2.

**Warning:** if you see a model \( Y = \beta_0 + \beta_1 X + U \), make sure you know whether it’s a CEF model or a structural model, or yet another type of model (like in Chapter 7). The equation by itself only shows a
linear relationship; it does not tell us the meaning of the parameters or the error term $U$. This is something to be very wary of when you look at econometric resources online or in other books; they may have models that look identical but are interpreted very differently.

Practice 6.4. Let $X = 1$ if an individual’s body mass index (BMI) is 30 or greater (the technical definition of obesity) and $X = 0$ otherwise, and let $Y$ denote hourly wage. Consider the model $Y = \delta_0 + \delta_1 X + W$, where $\delta_0$ and $\delta_1$ are unknown, non-random parameters, and $W$ is the unobserved error term. What is the interpretation of $\delta_1$ and $W$? Explain. (Hint: yes, this is a “trick” question with a very short answer.)

6.5.2 General Structural Model and ASE

The linear model is better for building intuition, but to expand your mind, consider the structural model

$$Y = h(X, U), \quad (6.35)$$

where $Y$ is the outcome, $X$ is a binary regressor, $U = (U_1, U_2, \ldots)$ is a vector containing all causal determinants of $Y$ besides $X$, and $h(\cdot)$ could be any (non-random) function.

The linear structural model $Y = \beta_0 + \beta_1 X + U$ is a special case of (6.35). If $h(a, b) = \beta_0 + \beta_1 a + g(b)$, then (6.35) is

$$Y = h(X, \tilde{U}) = \beta_0 + \beta_1 X + g(\tilde{U}).$$

For a single individual, the **structural effect** on $Y$ of changing $X = 0$ to $X = 1$ is

$$s(U) \equiv h(1, U) - h(0, U). \quad (6.36)$$

This is the causal effect on $Y$ when $X$ increases from 0 to 1, all else equal (ceteris paribus), i.e., holding the value of $U$ fixed. Different individuals have different unobserved $U$, so they may have different structural effects of $X$ on $Y$. For example, getting a college degree may increase income a lot for some individuals, but not others.

It’s very difficult to learn about $s(U)$ since it depends on things we can’t observe. Instead, we can try to learn about its mean. As usual, there is a tradeoff: the mean effect is not as informative and helpful for policy, but it is easier to learn about.

The **average structural effect** (ASE) is a weighted average of individual structural effects. The weights depend on the population distribution of $U$. Mathematically, this weighted average is the population mean:

$$\text{ASE} \equiv \mathbb{E}[s(U)] = \mathbb{E}[h(1, U) - h(0, U)] = \mathbb{E}[h(1, U)] - \mathbb{E}[h(0, U)]. \quad (6.37)$$
In this special case with binary $X$, the final expressions looks like the ATE, where $Y^U = h(0, U)$ and $Y^T = h(1, U)$. That is, we can imagine two parallel universes where everything besides $X$ (specifically, $U$) is the same. But, $U$ may contain things that would violate SUTVA, like other people’s $X$.

The ASE is sometimes called the average causal effect (e.g., Hansen, 2020, Def. 2.7), but so is the ATE, so I avoid “average causal effect” to avoid confusion.

**Discussion Question 6.3 (ES habits and final scores).** Let $Y$ be a student’s final semester score in this class, $0 \leq Y \leq 100$, and $X = 1$ if the student starts each exercise set well ahead of the due date (and $X = 0$ if not). Consider the structural model $Y = a + bX + U$ and the CEF model $Y = c + dX + V$.

a) What does $U$ represent? Give some specific examples of what $U$ includes here. (Hint: imagine two students with the same $X$ but different $Y$; what causes them to have different $Y$?)

b) Do you think $\mathbb{E}(U \mid X = 0) = \mathbb{E}(U \mid X = 1)$? Why/not?

c) Do you think $b = d$, $b < d$, or $b > d$? Why?

**Practice 6.5 (ES habits: parameters).** In DQ 6.3, what would you guess are reasonable possible values of the parameters $a$, $b$, $c$, and $d$? Explain.

### 6.6 Identification

This section focuses on identification of parameters with causal meaning. In particular: when does the slope parameter $\beta_1$ in the CEF model also have a causal interpretation? With binary $X$, $\beta_1$ is a (conditional) mean difference, so intuition is similar to Section 4.6, but many more details and formal results are provided here. As in Section 4.6, the conditions required for identification are called identifying assumptions.

#### In Sum: When Does a CEF Difference Have a Causal Interpretation?

If $X$ “exogenous” (unrelated to other determinants of $Y$; Sections 6.6.1 and 6.6.3)

If $X$ (as good as) randomized (unrelated to potential outcomes; Section 6.6.2)

### 6.6.1 Linear Structural Model

Under certain conditions (identifying assumptions), the structural slope $\beta_1$ in (6.34) is identified, equal to the CEF slope $\gamma_1$ in

$$ \mathbb{E}(Y \mid X = x) = \gamma_0 + \gamma_1 x. $$ (6.38)

This is equivalent to (6.28), just with $\gamma$ to avoid confusion with (6.34).
Identifying Assumptions and Formal Results

Qualitatively, the structural slope is identified if \( X \) and \( U \) are “unrelated.” That is, the regressor \( X \) must be unrelated to the unobserved determinants of \( Y \) (that comprise \( U \)); \( U \) cannot be systematically higher or lower for certain \( X \) values. If true, then \( X \) is called **exogenous** (link to pronunciation). If not, then \( X \) is called **endogenous** (link to pronunciation). The precise mathematical condition for a regressor’s exogeneity (or endogeneity) depends on the model.

Quantitatively, there are a few ways to describe “exogeneity” of \( X \) in (6.34). Specifically, each of Assumptions A6.1–A6.3 is a sufficient condition for \( \beta_1 = \gamma_1 \).

**Assumption A6.1** (independent error). \( U \) is independent of \( X \):
\[
X \perp U
\]

**Assumption A6.2** (mean independent error). \( U \) is mean independent of \( X \):
\[
E(U \mid X) = E(U).
\]
For binary \( X \), equivalently,
\[
E(U \mid X = 0) = E(U \mid X = 1).
\]

**Assumption A6.3** (uncorrelated error). \( U \) is uncorrelated with \( X \):
\[
\text{Corr}(U, X) = 0,
\]
or equivalently
\[
\text{Cov}(U, X) = 0.
\]

Some of Assumptions A6.1–A6.3 are stronger than others. Independence is stronger than mean independence:
\[
\text{A6.1} \implies \text{A6.2}
\]

In general, mean independence is stronger than zero correlation (which is equivalent to zero covariance):
\[
\text{A6.2} \implies \text{A6.3}
\]

If \( X \) has only two possible values, then mean independence is actually equivalent to ( \( \iff \) ) zero correlation.

Theorem 6.1 formally states the identification theorem. You do not need to write (or even fully understand) proofs for this class, but the proof may help deepen understanding and appreciation for some students.

**Theorem 6.1** (linear structural identification). Consider the linear structural and CEF models in (6.34) and (6.38), respectively. Assume \( X \) has only two possible values, \( x_1 \) and \( x_2 \); as a special case, \( x_1 = 0 \) and \( x_2 = 1 \). If any one of Assumptions A6.1–A6.3 is true, then the structural slope is identified and equal to the CEF slope, i.e., \( \beta_1 = \gamma_1 \). If additionally \( E(U) = 0 \), then the structural intercept is also identified, with \( \beta_0 = \gamma_0 \).

**Proof.** It is sufficient to prove the result for A6.2 because it is weaker than A6.1 and equivalent to A6.3 (given only two possible \( X \) values).
Starting from the structural model,

\[ Y = \beta_0 + \beta_1 X + U \]

\[ = \beta_0 + \beta_1 X + U + E(U) - E(U) \]

\[ = \beta_0 + E(U) + \beta_1 X + U - E(U). \]

As labeled, the CEF intercept is \( \gamma_0 = \beta_0 + E(U) \) and the CEF slope \( \gamma_1 = \beta_1 \) because \( V \equiv U - E(U) \) is a CEF error:

\[ E(U - E(U) \mid X) = E[U \mid X] - E[E(U) \mid X] = E(U) - E(U) = 0. \] (6.41)

Thus, mean independence implies \( \gamma_1 = \beta_1 \), and \( E(U) = 0 \) further implies \( \gamma_0 = \beta_0 \).

In Practice

In practice, Assumptions \textit{A6.1–A6.3} are usually difficult to justify. Recall the example where \( Y \) is income and \( X \) is having a college degree. Imagine \( U \) includes something called “ability” that includes all skills not gained directly from college (e.g., skills learned from a parent). To have \( \beta_1 = \gamma_1 \), \( U \) would have to satisfy \( E(U \mid X = 0) = E(U \mid X = 1) \), i.e., the non-college and college subpopulations have the same mean ability. Of course, there are many types of ability, but it seems likely that in general college graduates are higher ability. In fact, the most famous of Michael Spence’s Nobel Prize-winning work\footnote{See Spence (1973) or the very brief overview of his signaling model at \url{https://en.wikipedia.org/wiki/Michael_Spence#Career}.} provides a more formal economic model of why the college graduates should have higher ability, i.e., why \( E(U \mid X = 1) > E(U \mid X = 0) \).

Returning to an even simpler example, let \( Y \) be commute time and \( X = 1 \) if people are carrying umbrellas, with \( X = 0 \) otherwise. Since the umbrellas themselves have no effect on \( Y \), the structural \( \beta_1 = 0 \). Since rain affects \( Y \), rain is part of \( U \), although \( U \) may also include traffic conditions and such. When \( X = 0 \), there is probably no rain, whereas when \( X = 1 \), there probably is rain; thus, \( E(U \mid X = 1) > E(U \mid X = 0) \). Again, the structural error \( U \) is clearly not a CEF error. Consequently, the CEF slope has only statistical meaning, not causal meaning. Indeed, here the CEF slope is the difference in mean commute time between days when people carry umbrellas and days they don’t, which should be a substantial positive difference (longer commutes on days people carry umbrellas, because those are rainy days). However, the causal effect of umbrellas is zero, so the CEF’s slope is bigger than the structural \( \beta_1 \).

If we could also observe weather conditions, then it might be plausible that the remaining parts of \( U \) are unrelated to \( X \). This identification approach is considered in Chapters 9 and 10.
Discussion Question 6.4 (marriage and salary). Let \( X = 1 \) if married and otherwise \( X = 0 \). Let \( Y \) be annual salary. Consider the structural model \( Y = \beta_0 + \beta_1 X + U \).

a) Explain why probably \( E(U \mid X = 1) \neq E(U \mid X = 0) \), and say which you think is higher. (Hint: first think about what else is in \( U \), i.e., what determines someone’s salary; or think about variables that differ on average between married and unmarried individuals, and whether any of those help determine salary.)

b) Does the average salary difference between married and unmarried individuals have a structural meaning? Why/not?

6.6.2 Average Treatment Effect

Identification of the average treatment effect (ATE) was initially discussed in Section 4.3. Here, identifying assumptions and results are formally stated in regression notation.

Updating (4.12) to this chapter’s notation, the ATE is identified when

\[
E(Y^T) - E(Y^U) = E(Y \mid X = 1) - E(Y \mid X = 0), \tag{6.42}
\]

where \( Y^T \) and \( Y^U \) are (still) the potential treated and untreated outcomes, respectively. The important feature of (6.42) is that the right-hand side contains only observable variables, \( Y \) and \( X \). Usually, with enough data, we can learn the population joint probability distribution of \( (Y, X) \), which in turn determines the conditional means \( E(Y \mid X = 1) \) and \( E(Y \mid X = 0) \). If (6.42) is true, then learning about these conditional means (on the right-hand side) helps us learn the ATE (left-hand side).

Identifying Assumptions and Formal Results

Assumption A6.4 is SUTVA, as discussed in Section 4.4.

Assumption A6.5 is related to the discussion of randomized treatment in Section 4.3. Mathematically, the key is that randomization satisfies statistical independence between the treatment assignment and the individual’s pair of potential outcomes: \( X \perp \perp (Y^U, Y^T) \).

Assumption A6.6 was not discussed before, but it is intuitive: if everybody (or nobody) is treated, then it’s impossible to compare treated and untreated outcomes. For example, if \( P(X = 1) = 0 \), then nobody is treated, so we only observe \( Y = Y^U \) for everybody. We can learn about \( E(Y^U) \), but it’s impossible to learn about \( E(Y^T) \) since \( Y^T \) is literally never observed. Although obvious here, a more general overlap assumption may not always hold in more complex models.

The following identifying assumptions combined together are sufficient, but not necessary. That is, if they are all true, then the ATE is identified, but there may be other ways to identify the ATE even if they are violated.

The assumptions have various names. Assumption A6.4 is usually just called SUTVA, but the main part of it is often called no interference (or non-interference). Assumption A6.5 has many names:
identification, ignorability, or unconfoundedness. The combination of A6.5 and A6.6 is called strong ignorability. For more detail, history, and discussion, see Imbens and Wooldridge (2007).

**Assumption A6.4** (SUTVA). Everyone with \( X = 1 \) receives the same treatment, and one individual’s treatment does not affect any other individual’s potential outcomes.

**Assumption A6.5** (unconfoundedness). Treatment is independent of the potential outcomes: \( X \perp \perp (Y_U, Y_T) \).

**Assumption A6.6** (overlap). There is strictly positive probability of both treatment and non-treatment: \( 0 < P(X = 1) < 1 \).

Formally, identification of the ATE is shown as follows. The key is that A6.5 allows us to observe representative samples of both \( Y_U \) and \( Y_T \). Mathematically, this independence assumption implies that the means of the potential outcomes do not statistically depend on the treatment \( X \):

\[
E(Y_T) = E(Y_T \mid X = 1), \quad E(Y_U) = E(Y_U \mid X = 0). \tag{6.43}
\]

From (6.31), \( Y = Y_T \) when \( X = 1 \) and \( Y = Y_U \) when \( X = 0 \), so

\[
E(Y_T \mid X = 1) = E(Y \mid X = 1), \quad E(Y_U \mid X = 1) = E(Y \mid X = 0). \tag{6.44}
\]

Combining (6.43) and (6.44), this says that the population mean of the treated potential outcome, \( E(Y_T) \), equals the mean of the observed outcome in the treated population, \( E(Y \mid X = 1) \), which in Section 4.6 was \( E(Y_B) \). Since \( E(Y \mid X = 1) \) is a feature of the joint distribution of \((Y, X)\), it is identified. Since \( E(Y_T) = E(Y \mid X = 1) \), it is also identified. Similarly, \( E(Y_U) = E(Y \mid X = 0) \) is identified, so \( E(Y_T) - E(Y_U) \) is identified.

**Theorem 6.2** (ATE identification). Under A6.4–A6.6, the ATE is identified:

\[
E(Y_T - Y_U) = E(Y_T) - E(Y_U) = E(Y \mid X = 1) - E(Y \mid X = 0),
\]

which is the slope \( \beta_1 \) in the linear CEF model in (6.28).

**Proof.** The constructive proof of ATE identification links the unobservable ATE with the observable mean difference. Specifically,

\[
\begin{align*}
E(Y_T - Y_U) &= E(Y_T) - E(Y_U) \\
&= E(Y_T \mid X = 1) - E(Y_U \mid X = 0) \\
&= E(Y \mid X = 1) - E(Y \mid X = 0).
\end{align*}
\]

This equals the linear CEF slope coefficient, as shown in (6.29). \(\square\)
Imagine a knee surgery treatment ($X$) to help arthritis, where $Y$ is knee-specific pain (between 0 and 100). For each individual, we can imagine two parallel universes, identical except for whether the individual gets the treatment (surgery) or not. It is the same surgery for everybody, and naturally one person’s surgery cannot affect another person’s pain, so SUTVA is satisfied. Half of patients are randomly assigned the treatment, so $X \perp \perp (Y^U, Y^T)$ and $0 < P(X = 1) < 1$. Thus, Assumptions A6.4–A6.6 are all satisfied, and Theorem 6.2 says the ATE equals the CEF slope, which we can estimate by OLS.

Surgery seems like a very straightforward example, but there can still be problems. For example, maybe the people who volunteer to participate in the randomized experiment are not representative of the general population; e.g., maybe they are feeling very desperate because they have particularly severe arthritis. Another issue is that some people may be hurt by the treatment even if the overall ATE seems helpful. Perhaps the biggest issue in real life was that “surgery” was treated as a black box without understanding the particular mechanism that reduced pain. It turned out that placebo (fake) surgery was equally effective.²

Consider Theorem 6.2 when $X$ is rain and $Y$ is commute time. In Columbia, MO, there is much less traffic in the “summer” (mid-May to mid-August) when most students are gone, meaning both $Y^T$ and $Y^U$ are lower. There is also more rain ($X = 1$). That is, $X$ and $(Y^U, Y^T)$ are related, violating Assumption A6.5. Intuitively, the problem is we’d see more short rainy commutes in the summer and long dry commutes during the academic year, which makes it seem like rain causes short commutes; but correlation does not imply causation.

Practice 6.6. Discuss the right-to-work example from Sections 4.3–4.5 in terms of Assumptions A6.4–A6.6.

### 6.6.3 Average Structural Effect

The ASE in (6.37) is identified if it equals the CEF slope. A sufficient condition for this is Assumption A6.7, which is qualitatively similar to Assumption A6.5.

**Assumption A6.7 (independence).** The unobservable determinants of $Y$ are independent of $X$; e.g., in the notation of (6.35), $U \perp \perp X$.

**Theorem 6.3 (ASE identification).** Consider the general structural model in (6.35) and the ASE defined in (6.37). If Assumption A6.7 holds, then the ASE is identified and equal to the slope of the linear CEF in (6.28).

²https://doi.org/10.1056/NEJMoa13259
Proof. Using $Y = h(X, U)$,
\[
E(Y \mid X = 1) - E(Y \mid X = 0) = E[h(X, U) \mid X = 1] - E[h(X, U) \mid X = 0] = E[h(1, U) \mid X = 1] - E[h(0, U) \mid X = 0] = E[h(1, U)] - E[h(0, U)],
\]
which equals the ASE as in (6.37).

6.7 Estimation: OLS

This section considers estimation of the CEF model (6.28) when $X$ has only two possible values. The interpretation (description, prediction, causality) does not matter for estimation.

One approach is to define $Y^A$ as the $X = 0$ subpopulation and $Y^B$ as the $X = 1$ subpopulation. Then $\beta_0 = E(Y^A)$ and $\beta_1 = E(Y^B) - E(Y^A)$, so Sections 4.1.2 and 4.1.3 can be used to estimate $E(Y^A)$ and $E(Y^B)$.

Though simple, that approach does not generalize as well as ordinary least squares (OLS). The intuition behind the least squares approach comes from the characterization of the conditional mean as the best predictor of $Y$ given $X = x$ with quadratic loss. The idea extends (3.11) for estimating the unconditional mean of $Y$. In the population, if $E(Y \mid X = x) = \beta_0 + \beta_1 x$, then
\[
(\beta_0, \beta_1) = \arg\min_{b_0, b_1} E[L_2(Y, b_0 + b_1 X)] = \arg\min_{b_0, b_1} E[(Y - b_0 - b_1 X)^2],
\]
where $L_2(y, g)$ is the quadratic loss function from (2.39). This shows that the CEF provides the best (with quadratic loss) predictor of $Y$ given $X$. In the sample, replacing the population mean ($E$) with the sample mean ($\frac{1}{n} \sum_{i=1}^n$), the minimization problem analogous to (6.45) is
\[
\text{OLS: } (\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2. \tag{6.46}
\]
The estimated CEF is thus
\[
\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{6.47}
\]

Extending Sections 3.3 and 3.4.2, the OLS regression estimator can be explained in terms of the empirical distribution. Instead of a single $S$ representing $Y$, now we have $(S_Y, S_X)$ representing $(Y, X)$. If all the sample values $(Y_i, X_i)$ are unique, then the empirical distribution has $P(S_Y = Y_i, S_X = X_i) = 1/n$ for all $i = 1, \ldots, n$. Thus, replacing the population mean in (6.45) with the sample average in (6.46) can be seen as using the empirical distribution. That is,
\[
(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{b_0, b_1} E[(S_Y - b_0 - b_1 S_X)^2] = \arg\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2. \tag{6.48}
\]
Notationally, as usual, the “hats” on $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{m}(x)$ indicate that they are computed from the sample, whereas the true population values $\beta_0$, $\beta_1$, and $m(x)$ lack hats. The form of (6.46) explains the L and S in OLS. “Least” (L) refers to minimization, and “squares” (S) refers to squaring $Y_i - b_0 - b_1X_i$. (Explaining the O is a special treat reserved for econ PhD students.)

Equation (6.46) can be described with the terms introduced around (3.12). Given any estimates $(\hat{\beta}_0, \hat{\beta}_1)$, the fitted values are

$$\hat{Y}_i \equiv \hat{\beta}_0 + \hat{\beta}_1X_i = \hat{m}(X_i).$$

(6.49)

Given $\hat{Y}_i$, the residual is defined as

$$\hat{U}_i \equiv Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1X_i.$$ (6.50)

Consequently, (6.46) can be interpreted as saying that the OLS estimates $(\hat{\beta}_0, \hat{\beta}_1)$ make the sum of squared residuals $\sum_{i=1}^{n} \hat{U}_i^2$ as small as possible.

The OLS estimator is consistent under fairly general conditions; see Section 7.7.2. In that case,

$$\hat{\beta}_0 \overset{p}{\to} \beta_0, \quad \hat{\beta}_1 \overset{p}{\to} \beta_1, \quad \hat{m}(0) \overset{p}{\to} m(0), \quad \hat{m}(1) \overset{p}{\to} m(1).$$ (6.51)

6.7.1 Code

The following code runs OLS. It also computes the sample means of $Y_i$ for the group with $X_i = 0$ and the group with $X_i = 1$, separately. This verifies that with a single, binary regressor, $\hat{\beta}_0$ is the sample average $Y_i$ for the $X_i = 0$ group, while $\hat{\beta}_0 + \hat{\beta}_1$ is the sample average $Y_i$ for the $X_i = 1$ group, so $\hat{\beta}_1$ is the difference.

```r
df <- data.frame(Y=c(1,4,0,2,3, 8,7,6,9,5), X=c(0,0,0,0,0, 1,1,1,1,1))
ret <- lm(formula=Y~X, data=df)
coef(ret)
## (Intercept) X
##  2  5
c( mean(df$Y[df$X==0]), mean(df$Y[df$X==1]) )
## [1] 2 7
```

6.8 Quantifying Uncertainty

The ways to quantify uncertainty in Section 3.8 also apply to $\beta_0$ and $\beta_1$ in the linear CEF model (6.28). The same interpretations and misinterpretations apply. In particular, these methods do not reflect uncertainty about identifying assumptions. For example, a CI that
contains the CEF slope with 95% probability does not contains the structural slope with 95% probability if it is not identified; it could be only 80%, or 50% or near 0%.

One new consideration is discussed in Section 6.8.1, followed by sample code in Section 6.8.2.

### 6.8.1 Heteroskedasticity

Different methods for quantifying uncertainty make different assumptions about the conditional variance. Whereas the conditional mean \( E(Y \mid X = x) \) is the mean of the conditional distribution of \( Y \) given \( X = x \), the conditional variance

\[
\sigma^2_Y(x) \equiv \text{Var}(Y \mid X = x)
\]

(6.52)
is the variance of the conditional distribution of \( Y \) given \( X = x \). The term **homoskedasticity** means \( \sigma^2_Y(x) = \sigma^2_Y \), a constant not depending on \( x \), whereas **heteroskedasticity** means \( \sigma^2_Y(x) \) is not constant. Equivalently, we could write \( Y = \beta_0 + \beta_1 X + U \) and consider the conditional variance of \( U \) since \( \text{Var}(Y \mid X) = \text{Var}(U \mid X) \), so often homoskedasticity and heteroskedasticity are thought of as properties of the error term.

Always use methods that are **robust to heteroskedasticity** (or **heteroskedasticity-robust**). This means they’re valid with homoskedasticity or heteroskedasticity, whereas other methods only work with homoskedasticity. Logically, the heteroskedasticity-robust methods have weaker assumptions, so they work more often. Besides, heteroskedasticity is very common in real economic data.

The term “robust” by itself is ambiguous. You should always ask: robust to what? Methods can be robust to heteroskedasticity, robust to clustered sampling, robust to measurement error, robust to infinite variance, etc.

**Practice 6.7** (heteroskedasticity). Let \( Y = 1 \) if employed (and \( Y = 0 \) if not), and let \( X = 1 \) if female (and \( X = 0 \) if not). Explain why there is probably heteroskedasticity. (Hint: if \( p = P(Y = 1) \), then \( \text{Var}(Y) = p(1-p) \). If \( p_x = P(Y = 1 \mid X = x) \), then what’s \( \text{Var}(Y \mid X = x) \)?)

### 6.8.2 Code

Unfortunately, the default in R is to use homoskedasticity-based standard errors, so you have to make an extra effort to get heteroskedasticity-robust results. The below code does this. Since \( X \) is binary, the same results can be obtained with a two-sample unpaired \( t \)-test with “unequal variances,” as shown.

The below code quantifies uncertainty about the CEF slope in a regression with a single, binary regressor. Using a variety of methods, the code computes a standard error (SE), 95% confidence interval, and \( t \)-statistic and two-sided \( p \)-value for testing the null hypothesis \( H_0: \beta_1 = 0 \).
In the table of output at the very end, the first two rows assume homoskedasticity, whereas the remaining four rows do not. The first row is a two-sample t-test assuming equal variances; the second row is the default results based on `lm()` output. The third row is a two-sample t-test allowing for unequal variances. The remaining rows use more general, regression-based methodology allowing for heteroskedasticity, based on the `lmtest` and `sandwich` packages in R (Zeileis, 2004; Zeileis and Hothorn, 2002). The first two rows are identical, and the following four rows are very similar to each other, but there is a big difference between the first two rows and the next four rows. This shows the (potentially) big difference between assuming homoskedasticity (as in the first two rows) and allowing for heteroskedasticity (as in the last four). There are multiple ways to allow for heteroskedasticity, like the HC0, HC1, and HC3 shown in the table. The differences are beyond our scope, but as the table suggests, the differences are often very small in practical terms.

In practice, you should use `coeftest` and `coefci` to allow for heteroskedasticity like below.

```r
library(lmtest); library(sandwich)
set.seed(112358)
n <- 1000
df <- data.frame(Y=rnorm(n=n/4,mean=0,sd=1),
                 rnorm(n=3*n/4,mean=0.2,sd=2)),
                 X=c(rep(0,n/4), rep(1,3*n/4)))
ret <- lm(formula=Y~X, data=df)
# Store results for slope in sl.out
rn <- c('ttest.eq','Homosk.', 'ttest.uneq', 'HC0', 'HC1', 'HC3')
sl.out <- data.frame(row.names=rn, SE=rep(NA,6), CI.lower=NA, CI.upper=NA, t.stat=NA, p.value=NA)
# HC/zero.alt3: original from Hal White (1980)
retVC<-
    vcovHe(ret, type="HC/zero.alt3")
sl.out[,'HC/zero.alt3',2:3] <- sqrt(retVC[2,2])
# HC1: matches Stata default, and two-sample t.test below
retVC1 <- vcovHe(ret, type="HC1")
# HC3: recommended/default (and larger SE than HC0, HC1)
retVC3 <- vcovHe(ret, type="HC3")
# Default homoskedastic results
sl.out['Homosk.',['SE','t.stat','p.value']] <-
    summary(ret)$coefficients['X',2:4]
# Heteroskedasticity-robust tests/p-values
sl.out[,'HC0',c(1,4,5)] <- coeftest(ret, vcov.=retVC0)['X',2:4]
sl.out[,'HC1',c(1,4,5)] <- coeftest(ret, vcov.=retVC1)['X',2:4]
sl.out[,'HC3',c(1,4,5)] <- coeftest(ret, vcov.=retVC3)['X',2:4]
# Heteroskedasticity-robust CIs (shortest to longest)
sl.out[,'HC0',2:3] <- coefci(ret, vcov. = retVC0)[,'X',]
sl.out[,'HC1',2:3] <- coefci(ret, vcov. = retVC1)[,'X',]
sl.out[,'HC3',2:3] <- coefci(ret, vcov. = retVC3)[,'X',]
sl.out['Homosk.',2:3] <- confint(ret, level=0.95)[,'X',]
```
## For comparison: `t.test()` results for slope

t.sl <- \( t.test( x=df$Y[ df$X==1 ], y=df$Y[ df$X==0 ], \mu=0, \text{conf.level}=0.95, \text{alternative}=\text{\textquote{two.sided}}, \text{paired}=FALSE, \text{var.equal}=FALSE) \)

sl.out['ttest.uneq',-1] <- \( c( \text{t.sl$conf.int}, \text{t.sl$statistic}, \text{t.sl$p.value}) \)

t2 <- \( t.test( x=df$Y[ df$X==1 ], y=df$Y[ df$X==0 ], \mu=0, \text{conf.level}=0.95, \text{alternative}=\text{\textquote{two.sided}}, \text{paired}=FALSE, \text{var.equal}=TRUE) \)

sl.out['ttest.eq',-1] <- \( c( \text{t2$conf.int}, \text{t2$statistic}, \text{t2$p.value}) \)

sl.out['ttest.uneq',1] <- \( (\text{t.sl$conf.int} \times c(-1,1)) / (2 \times \text{qt}( \text{p}=1-0.95/2, \text{df}=\text{t.sl$parameter} )) \)

sl.out['ttest.eq',1] <- \( (\text{t2$conf.int} \times c(-1,1)) / (2 \times \text{qt}( \text{p}=1-0.95/2, \text{df}=\text{t2$parameter} )) \)

print(\( \text{round}(\text{sl.out}, \text{digits}=3) \))

\[
\begin{array}{lccccc}
\text{##} & \text{SE} & \text{CI.lower} & \text{CI.upper} & \text{t.stat} & \text{p.value} \\
\text{ttest.eq} & 0.128 & -0.026 & 0.476 & 1.76 & 0.079 \\
\text{Homosk.} & 0.128 & -0.026 & 0.476 & 1.76 & 0.079 \\
\text{ttest.uneq} & 0.095 & 0.038 & 0.412 & 2.36 & 0.018 \\
\text{HC0} & 0.095 & 0.039 & 0.412 & 2.37 & 0.018 \\
\text{HC1} & 0.095 & 0.038 & 0.412 & 2.37 & 0.018 \\
\text{HC3} & 0.095 & 0.038 & 0.412 & 2.36 & 0.018 \\
\end{array}
\]

### Practice 6.8 (regression significance)

Consider the setup of the “audit study” from Bertrand and Mullainathan (2004). Resumes were fabricated that were identical except for the name: Emily (suggesting a white female), Greg (white male), Lakisha (black female), or Jamal (black male). The resumes were then submitted to job openings, and it was recorded whether or not an in-person interview for the job was then offered. Here, let \( Y = 1 \) if an interview was offered and \( Y = 0 \) if not; let \( X = 1 \) if the name is “black” and \( X = 0 \) if not. Note that \( \text{E}(Y \mid X = x) = \text{P}(Y = 1 \mid X = x) \), i.e., the conditional probability of an interview. A regression of \( Y \) on \( X \) (including an intercept, as always) is run, and heteroskedasticity-robust standard errors are computed. Consider both economic significance and statistical significance in the following possible results. (Economic and statistical significance were introduced in Section 3.9.7.) Hint: to quickly assess statistical significance, \( |\hat{\beta}_1 / \hat{\text{SE}}| \geq 2 \) means statistical significance at a 5% level, with higher values being more statistically significant.

a) Slope estimate \( \hat{\beta}_1 = 0.00001, \hat{\text{SE}} = 0.000001 \).
b) \( \hat{\beta}_1 = -0.1, \hat{\text{SE}} = 0.1 \).
c) \( \hat{\beta}_1 = -0.2, \hat{\text{SE}} = 0.02 \).
d) \( \hat{\beta}_1 = -0.01, \hat{\text{SE}} = 0.01 \).
Empirical Exercise EE6.1. You will essentially replicate EE4.1 but with regression commands.

a. R only: load the needed packages and look at a description of the dataset:
   ```
   library(wooldridge); library(sandwich); library(lmtest)
   ?jtrain2
   ```

b. Stata only: run `ssc install bcuse` if necessary, then load the data with
   ```
   bcuse jtrain2, nodesc clear
   ```

c. Run a regression of 1978 earnings (re78) on the job training assignment indicator (train).
   ```
   R: ret <- lm(re78~train, data=jtrain2)  
   Stata: regress re78 train , vce(robust)
   ```
   in which `vce(robust)` requests heteroskedasticity-robust standard errors

d. R only (since already reported in Stata): output the estimates along with heteroskedasticity-robust standard errors and two-sided 95% confidence intervals with the code
   ```
   coeftest(ret, vcov.=vcovHC(ret, type='HC1'))
   coefficient(ret, vcov.=vcovHC(ret, type='HC1'))
   ```
   where argument `type='HC1'` refers to one specific type (among multiple) of heteroskedasticity-robust standard error estimator (HC stands for “heteroskedasticity-consistent”)

e. R only: create a subset of the data including only married individuals, with code
   ```
   jt2.mar1 <- jtrain2[jtrain2$married ==1 , ]
   ```
f. Run your previous analysis for the subset of married individuals.
   ```
   R: replace data=jtrain2 with data=jt2.mar1  
   Stata: regress re78 train if married==1 , vce(robust)
   ```
g. Repeat your analysis, but for unmarried individuals

h. Repeat your analysis on the full sample of individuals, but for the outcome variable `unem78` (1978 unemployment indicator) instead of `re78` (and remember unemployment is bad, so negative coefficient is good).
Chapter 7
Simple Linear Regression

⇒ Kaplan video: Chapter Introduction

Depends on: Chapter 6 (which depends on Chapters 2–4)

Unit learning objectives for this chapter

7.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

7.2. Interpret what a linear regression estimates, in multiple ways, mathematically and intuitively [TLOs 2 and 3]

7.3. Assess whether certain assumptions for linear regression seem true or not in real-world examples [TLOs 2 and 6]

7.4. In R (or Stata): estimate a simple linear regression, along with measures of statistical uncertainty, and judge economic and statistical significance [TLO 7]

Optional resources for this chapter

- Regression as description (Masten video)
- James et al. (2013, §3.1)
- Sections 4.1–4.2 (“Simple Linear Regression” and “Estimating the Coefficients of the Linear Regression Model”) in Hanck et al. (2018)
- Sections 2.1 (“Simple OLS Regression”) and 2.2 (“Coefficients, Fitted Values, and Residuals”) in Heiss (2016) [repeated from Chapter 6]

Surprisingly, many critical issues arise with three (instead of two) possible X values. With two, the regression modeled conditional means, useful for description, prediction, and (sometimes) causality. However, with three (or more) X values, we may fail to model the conditional means. In simple cases, this can be solved with a more flexible model; in other cases, we need to reinterpret what OLS actually estimates in practice.
Generally, OLS estimates something called a linear projection. This can also be interpreted as a “best” linear approximation of the CEF (for description) or a “best” linear predictor of $Y$ given $X$ (for prediction). These interpretations are discussed along with statistical properties of OLS as an estimator of the linear projection (not CEF).

### 7.1 Misspecification

Consider the linear population model

$$Y = \beta_0 + \beta_1 X + U,$$

where supposedly $E(U \mid X) = 0$, and this time $X$ has three possible values: 0, 1, and 2.

Intuitively, you should worry already: there are now three conditional means, but still only two parameters. That is, we want to learn the three values

$$m(0) \equiv E(Y \mid X = 0), \quad m(1) \equiv E(Y \mid X = 1), \quad m(2) \equiv E(Y \mid X = 2),$$

but (7.1) has only two parameters, $\beta_0$ and $\beta_1$. That’s like trying to serve three dinners on only two plates. That does not sound like enough flexibility.

Mathematically, the question is whether

$$m(x) = \beta_0 + \beta_1 x, \quad x = 0, 1, 2,$$

i.e., whether $m(x)$ is a straight line.

For example, let $Y$ be income, and let $X$ be number of siblings. Maybe there is a big income gap between only children ($X = 0$) and individuals with one sibling ($X = 1$), but having a second sibling ($X = 2$) does not change much. To simplify this, let $m(0) > m(1) = m(2)$. From $m(1)$ and $m(2)$ alone, the CEF appears flat (zero slope), in which case $\beta_0 = m(1)$ and $\beta_1 = 0$ fits these two points. But from $m(0)$ and $m(1)$, the slope appears negative, $\beta_1 < 0$, and the intercept is $\beta_0 = m(0)$. There is no $(\beta_0, \beta_1)$ that can make $\beta_0 + \beta_1 x$ go through all three points $m(0), m(1),$ and $m(2)$ if $m(0) > m(1) = m(2)$.

Figure 7.1 shows the impossibility of a linear CEF in this previous example. In the figure, $m(0) = 60$ (in thousands of $/yr) and m(1) = m(2) = 40$. The line with $\beta_0 = m(0)$ and $\beta_1 = m(1) - m(0) < 0$ fits the first two CEF values but not the third. The line with $\beta_0 = m(1)$ and $\beta_1 = 0$ fits the second two CEF values but not the first. It is impossible to draw a straight line $(\beta_0 + \beta_1 x)$ through all three points on this CEF, as Euclid could tell us.

A wrong model is euphemistically termed **misspecified**. That is, the model assumes something that is not actually true. For the siblings and income example, the linear CEF model in (7.1) is misspecified. The model incorrectly assumes that the conditional mean
of $Y$ is linear in $X$ (i.e., an affine function of $X$). Equivalently, it assumed $m(1) - m(0) = m(2) - m(1)$, which is not true in the example. More specifically, this type of misspecification is called **functional form misspecification** since it is the linear functional form that is wrong. That is, even though any values of $(\beta_0, \beta_1)$ are allowed, $\beta_0 + \beta_1 x$ is always a straight-line function of $x$, so it has a linear functional form (the general “shape” of the function).

**Practice 7.1** (misspecification). Investigate whether the problem with the sibling example was that $X = 0$ was a possible value (so that the intercept had to be $\beta_0 = m(0)$), as follows. Consider the same example but with $X = 1, 2, 3$ instead of $X = 0, 1, 2$, so $m(1) = 60$, $m(2) = m(3) = 40$. Is it possible to write $m(x) = \beta_0 + \beta_1 x$ now? Why or why not?

It’s technically possible that a CEF is linear, though extremely unlikely in practice. Continuing the example, if $m(2) = 20$ exactly, then $m(x) = 60 - 20x$, i.e., $\beta_0 = 60$ and $\beta_1 = -20$. In that case, the linear CEF is **properly specified** (or **correctly specified**). If instead $m(2) = 20.001$, the linear CEF model is misspecified. However, with such a small amount of misspecification, a linear model is a very good approximation.

Arguably, we must learn how best to cope with misspecification since we cannot truly avoid it. As Box (1979, p. 2) famously wrote, “All models are wrong but some are useful.” With reference to Box’s quote, Section 8.3 essentially tries to maximize a model’s usefulness by choosing the optimal amount of “how wrong” it is (misspecification).

## 7.2 Coping with Misspecification

There are two ways to cope with misspecification: change the model, or reinterpret it. The first way is now discussed for (7.1), while rein-

---

1See [https://en.wikipedia.org/wiki/All_models_are_wrong](https://en.wikipedia.org/wiki/All_models_are_wrong) for additional discussion, including the analogous quote about art from Pablo Picasso, “We all know that art is not truth. Art is a lie that makes us realize truth, at least the truth that is given us to understand. The artist must know the manner whereby to convince others of the truthfulness of his lies.”
CHAPTER 7. SIMPLE LINEAR REGRESSION

7.2.1 Model of Three Values

To fix the misspecification, the model needs to be more flexible. Continuing with $X = 0, 1, 2$ for simplicity, there are three conditional means, so the model should have three parameters to be flexible enough to avoid misspecification.

One way to add another parameter is to use a dummy variable for each possible value of $X$. (See Section 2.3.2 to review dummy variables.) Recall the indicator function from (2.3). Here,\[ \mathbb{1}\{X = j\} = \begin{cases} 1 & \text{if } X = j \\ 0 & \text{otherwise} \end{cases}, \quad j = 0, 1, 2. \] (7.2)

Since only three values of $X$ are possible,\[ \mathbb{1}\{X = 0\} = 1 - \mathbb{1}\{X = 1\} - \mathbb{1}\{X = 2\}. \] Thus, extending (6.26),

\[ m(x) = m(0) \mathbb{1}\{x = 0\} + m(1) \mathbb{1}\{x = 1\} + m(2) \mathbb{1}\{x = 2\} \] (7.3)

\[ = m(0)[1 - \mathbb{1}\{x = 1\} - \mathbb{1}\{x = 2\}] + m(1) \mathbb{1}\{x = 1\} + m(2) \mathbb{1}\{x = 2\} \]

\[ = m(0) + [m(1) - m(0)] \mathbb{1}\{x = 1\} + [m(2) - m(0)] \mathbb{1}\{x = 2\} \]

\[ = \beta_0 + \beta_1 \mathbb{1}\{x = 1\} + \beta_2 \mathbb{1}\{x = 2\}, \]

\[ \beta_0 \equiv m(0), \quad \beta_1 \equiv m(1) - m(0), \quad \beta_2 \equiv m(2) - m(0). \] (7.4)

Although the structure of (7.3) is easier to interpret, the structure of (7.4) is more common and can be interpreted as follows. The parameter $\beta_0 = m(0)$ is the conditional mean for some base category $X = 0$. The other parameters show how other conditional means differ from this base category. Specifically, $\beta_1 = m(1) - m(0)$ is the conditional mean difference between the $X = 1$ and $X = 0$ subpopulations, and $\beta_2 = m(2) - m(0)$ is the conditional mean difference between the $X = 2$ and $X = 0$ subpopulations.

This interpretation can be applied to the income and siblings example. The parameter $\beta_0$ is the population mean income among individuals with zero siblings. Zero siblings is the base category. Then, $\beta_1$ is the difference in mean income between the 1-sibling and 0-sibling subpopulations. Earlier, $m(0) = 60$ (thousands of $/yr) and m(1) = 40$, so $\beta_1 = m(1) - m(0) = -20$. Finally, $\beta_2$ is the mean income difference between the 2-sibling and 0-sibling (not 1-sibling) subpopulations, $m(2) - m(0) = 40 - 60 = -20$.

Discussion Question 7.1 (Facebook). Let $X = 0, 1, 2$ be the number of Facebook accounts somebody has, and $Y$ is hours of social media consumption per week.

a) Explain what it means for a CEF model $E(Y \mid X = x) = \beta_0 + \beta_1 x$ to be misspecified.

b) Describe a specific real-world reason to suspect misspecification in this example.
7.2.2 More Than Three Values

More generally, even if $X$ has more than three possible values, dummy variables could be used similarly to avoid CEF misspecification. Extending (7.3), there can be a dummy variable for each possible value of $X$, and a corresponding parameter for each. Any such model allowing an arbitrarily different conditional mean of $Y$ for each possible value of $X$ is called **fully saturated**. A fully saturated CEF model cannot be misspecified. (But, it may not have any causal meaning and may be practically impossible to estimate.)

In more complex settings, it is impossible to fix misspecification completely. For example, if $X$ could be any real number between 0 and 1, then an infinite number of parameters is required to model the conditional expectations for the infinite number of $X$ values; this is impossible in practice.

In such settings where misspecification is unavoidable, how can we interpret the model and its parameters? There are three interpretations of a more general linear model that includes the linear CEF model as a special case. These are discussed next.

---

**In Sum: Interpretations of What OLS Estimates**

- Linear projection: gets $\beta_0 + \beta_1 X$ “closest to” $Y$, probabilistically (Section 7.3)
- “Best” linear approximation (BLA) of CEF: “best” (smallest mean quadratic loss) approximation of $E(Y \mid X)$ with linear form $\beta_0 + \beta_1 X$ (Section 7.4)
- “Best” linear predictor (BLP): “best” (smallest mean quadratic loss) prediction of $Y$ given $X$ with linear form $\beta_0 + \beta_1 X$ (Section 7.5)

---

7.3 Linear Projection

⇒ Kaplan video: Linear Projection and “Best” vs. “Good”

The linear projection model is important because it is what OLS actually estimates. Two additional interpretations of the linear projection are described in Sections 7.4 and 7.5.

7.3.1 Geometric Intuition

You may have seen orthogonal **projection** in geometry or linear algebra. There is some shape (or vector space), and there is a point outside it. Projecting the point onto the shape consists of finding the point within the shape that is closest to the outside point.

Figure 7.2 illustrates projection. There is a large gray circle shape, and two points outside of it (small triangle, dot). The small triangle on the border of the large circle is the “closest” point to the outside small triangle, as measured by Euclidean distance. That is, the dashed line connecting the small triangles is just barely long enough...
to reach the gray circle from the outside triangle point; if it were any shorter, it could not reach any point in the gray circle. Similarly, the dot on the border of the gray shape is the projection of the outside dot onto the shape: of all the points in the gray space, it is closest to the outside dot (by Euclidean distance).

This idea can be written mathematically. Let $d_E(w, z)$ denote the Euclidean distance between points $w$ and $z$. Let $S$ denote a shape, which is a set of points. Let $y$ denote the outside point, and $p$ the projection. In Figure 7.2, the gray circle is $S$, the outside small triangle (or dot) is $y$, and the small triangle (or dot) on the circle’s border is $p$. The projection of point $y$ onto shape $S$ is the point inside $S$ that’s closest to $y$, i.e., that minimizes the distance to $y$.

Mathematically,

$$p = \arg\min_{s \in S} d_E(y, s). \quad (7.5)$$

### 7.3.2 Probabilistic Projection

Linear projection with random variables is the same idea, but with a different definition of distance and a different “shape” to search over.

Notationally, let $\text{LP}(Y \mid 1, X)$ denote the linear projection (LP) of $Y$ onto $(1, X)$. The $(1, X)$ specifies the “shape” that we search over: random variables that can be written as $a + bX$ for constants $a$ and $b$, i.e., linear combinations of $(1, X)$. (Linear combinations and linearity are detailed in Section 8.2.1.) Without the 1, $\text{LP}(Y \mid X)$ would only consider $bX$ with no intercept.

The closest “point” inside the “shape” is usually written $\beta_0 + \beta_1 X$. Mathematically, parallel to (7.5),

$$\text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X = \arg\min_{a+bX} d(Y, a+bX) = \arg\min_{a+bX} \sqrt{E[(Y - a - bX)^2]}, \quad (7.6)$$

where Euclidean distance $d_E(\cdot, \cdot)$ has been replaced by a probabilistic “distance” measure

$$d(A, B) \equiv \sqrt{E[(A - B)^2]]. \quad (7.7)$$

That is, linear projection gets $\beta_0 + \beta_1 X$ as “close” to $Y$ as possible, in a probabilistic sense.
7.3.3 Formulas and Interpretation

Some calculus (omitted) yields a formula for each linear projection coefficient (LPC), $\beta_0$ and $\beta_1$. In this special case with a single regressor $X$ and an intercept,

$$\beta_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}, \quad \beta_0 = \text{E}(Y) - \beta_1 \text{E}(X). \quad (7.8)$$

Writing $\sigma_Y^2 = \text{Var}(Y)$ and $\sigma_X^2 = \text{Var}(X)$, $\beta_1$ can be rewritten in terms of correlation:

$$\beta_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \frac{\text{Cov}(Y, X) \sigma_Y}{\sigma_X \sigma_Y} = \frac{\text{Cov}(Y, X)}{\sigma_X} \text{Corr}(Y, X) \sigma_Y. \quad (7.9)$$

Either version of the formula shows how the linear projection slope $\beta_1$ is related to the linear dependence (covariance or correlation) between $Y$ and $X$. Once the slope is determined, the intercept $\beta_0$ simply moves the linear projection line up or down so that $\text{E}(Y) = \beta_0 + \beta_1 \text{E}(X)$. That is, the linear projection always goes exactly through the point $(x, y) = (\text{E}(X), \text{E}(Y))$.

People often interpret the linear projection coefficients less precisely. For the slope, a common phrase is, “A one-unit increase in $X$ is associated with a $\beta_1$ change in $Y$.” The intercept is often not mentioned since $\beta_0 = \text{E}(Y) - \beta_1 \text{E}(X)$ is not easy to interpret, except when the regressor has been demeaned so that $\text{E}(X) = 0$, in which case $\beta_0 = \text{E}(Y)$. In this case, $\beta_0$ is called the “centercept” instead of intercept; but despite the better interpretation, it is rarely seen in economics.

For description, (7.8) shows that the LPCs summarize the joint probability distribution of $(Y, X)$. The joint distribution of $(Y, X)$ determines $\text{E}(Y)$, $\text{E}(X)$, $\text{Cov}(Y, X)$, and $\text{Var}(X)$, which then determine $\beta_0$ and $\beta_1$. Although a two-number summary of a complicated joint distribution is very convenient, clearly much information is lost in such a summary. Just as percentiles (quantiles) complement the mean in describing $Y$, quantile regression complements the LPCs in describing $(Y, X)$, though it is beyond our scope.

Although you don’t need to know it for this class, the LPCs can be written in matrix form. This generalizes more easily. Define vector

$$X \equiv (1, X)' = \begin{bmatrix} 1 \\ X \end{bmatrix}, \quad (7.10)$$

where $(1, X)'$ indicates the transpose of the row vector $(1, X)$. Then,

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = [\text{E}(XX')^{-1}] \text{E}(XY) = \begin{bmatrix} 1 \\ \text{E}(X) \end{bmatrix}^{-1} \begin{bmatrix} \text{E}(Y) \\ \text{E}(XY) \end{bmatrix}. \quad (7.11)$$

7.3.4 Linear Projection Model in Error Form

Analogous to (6.25) for the CEF, the linear projection model can be written in error form. Analogous to defining the CEF error as
\[ Y - m(X), \text{ the linear projection error is defined as} \]
\[ U \equiv Y - \text{LP}(Y \mid 1, X) = Y - (\beta_0 + \beta_1 X). \]  
(7.12)

Notationally, as usual, there is nothing special about the letter \( U \) (or \( \beta \), or even \( X \) or \( Y \)); e.g., it is mathematically equivalent to define \( V \equiv W - \text{LP}(W \mid 1, Z) \) and use \((\gamma_0, \gamma_1)\) for the LPCs. Given the definition of \( U \) in (7.12), it is always true that \( E(U) = \text{Cov}(X, U) = 0 \). Thus, the model
\[ Y = \beta_0 + \beta_1 X + U, \quad E(U) = \text{Cov}(X, U) = 0 \]  
(7.13)
is equivalent to \( \text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X \).

As with the CEF, the meaning of \( \text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X \) is more clearly explicit, but sometimes the error for (7.13) is more convenient mathematically, as in Section 12.3.2.

### 7.4 “Best” Linear Approximation

#### 7.4.1 Definition and Interpretation

For description, the linear projection can be interpreted as the **best linear approximation** (BLA) of the true CEF. “Best” here assumes quadratic loss, similar to how the mean \( E(Y) \) is the “best” predictor of \( Y \) with quadratic loss. “Linear” refers to a function of the form \( a + bX \) (see Section 8.2.1). Mathematically,
\[
\text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X = \text{BLA} \arg \min_{a + bX} E\{ [m(X) - (a + bX)]^2 \}, \quad m(X) \equiv Y - \beta_0 - \beta_1 X
\]  
(7.14)
That is, among all possible \( a + bX \), the linear projection \( \beta_0 + \beta_1 X \) is the function of \( X \) that best approximates \( E(Y \mid X) \).

This implies that if the CEF is linear in \( X \), then the linear projection equals the CEF. That is, if \( m(X) = \beta_0 + \beta_1 X \), then \( m(X) - (\beta_0 + \beta_1 X) = 0 \). Since this term is squared in (7.14), zero is the smallest possible value, so \( \beta_0 + \beta_1 X \) is the BLA.

Otherwise, the BLA treats more probable \( X \) as more important when trying to get the linear approximation “close” to the true CEF. The mean \( E\{ \cdot \} \) in (7.14) is a weighted average with more weight on more probable \( X \), so it is more important to make \( m(X) - (a + bX) \) close to zero for such \( X \) values.

#### 7.4.2 Limitations

Unfortunately, “best” does not always mean “good.” Sometimes, the CEF is so highly nonlinear that even the best linear approximation
is still a very poor approximation. By analogy: “Among all cities in Missouri, St. Louis is closest to Kuwait” does not mean “St. Louis is close to Kuwait.” Here, Kuwait is the true CEF, Missouri is the set of all functions linear in $X$, and St. Louis is the BLA. Sometimes the best (closest) is still not good (not close).

The following example of a “bad” BLA is from Hansen (2020, §2.28). Let $Y = X + X^2$, with no error term, so $m(x) = x + x^2$, too. If $X \sim N(0,1)$, then the BLA/LP turns out to be $LP(Y \mid 1, X = x) = 1 + x$. The function $1 + x$ is a bad approximation of $x + x^2$ (try graphing it).

Further, the distribution of $X$ can greatly affect the BLA of a nonlinear CEF. For example, Figure 7.1 shows two possible BLA lines for the same nonlinear CEF. One line is the BLA when the distribution of $X$ satisfies $P(X = 2) = 0$. The other line is the BLA when $P(X = 0) = 0$. The two lines are very different.

However, the BLA interpretation does at least assure us that when the CEF is approximately linear, the linear projection approximates the CEF well.

### 7.5 “Best” Linear Predictor

For prediction, the linear projection can be interpreted as the best linear predictor (BLP) of $Y$ given $X$. As with the BLA, “best” assumes quadratic loss. “Linear” again refers to the form $a + bX$. As in (2.55), the optimal predictor minimizes mean quadratic loss. Mathematically,

$$LP(Y \mid 1, X) = \beta_0 + \beta_1 X = \arg\min_{a+bX} \mathbb{E}\{L_2(Y, a + bX)\} = \arg\min_{a+bX} \mathbb{E}\{(Y - (a + bX))^2\}.$$  

(7.15)

That is, among all possible $a + bX$, the linear projection $\beta_0 + \beta_1 X$ is precisely the function of $X$ that “best” predicts $Y$ given knowledge of $X$.

Mathematically, (7.15) is the same as (7.6) but without the $\sqrt{\cdot}$. Although phrased differently, the linear projection goal of getting $\beta_0 + \beta_1 X$ “closest” to $Y$ is essentially the same as prediction: we want a predictor $\beta_0 + \beta_1 X$ that is “closest” to $Y$.

Unfortunately, as with BLA, “best” does not mean “good.” However, as with BLA, this means the CEF does not need to be exactly linear in order for the linear projection to make good predictions.

As in Section 2.5, “prediction” here is defined entirely within the population. It does not refer to using data to guess the future; there is no data here. Instead, the BLP is an ideal predictor; it is the (linear) predictor we would use if we fully knew everything about the population. The BLP is something we wish to learn. Fortunately, the BLP (and BLA and LP) is precisely what OLS estimates.

**Discussion Question 7.2 (BLP).** Let $Y$ be income (thousands of dollars per year) and $X$ be number of siblings. When $X = 0$, the
mean $Y$ is 60 and $50 \leq Y \leq 70$. When $X = 1$, the mean $Y$ is 40 and $30 \leq Y \leq 50$. When $X = 2$, it’s the same as when $X = 1$: the mean $Y$ is 40 and $30 \leq Y \leq 50$. In a population with mostly $X = 1$ and $X = 2$, the BLP is $\text{LP}(Y \mid 1, X) = 43 - 2X$.

a) What $Y$ does the BLP predict when $X = 0$?

b) Is the prediction from (a) good? Why/not?

### 7.6 Causality Under Misspecification

Some things can be said about causality under misspecification, but none as pleasing as the BLP for prediction or BLA for description. For example, if the structural error $U$ satisfies the CEF error property $E(U \mid X) = 0$, then the structural function is the CEF, so the linear projection is also the best linear approximation of the structural function. Alternatively, if the structural model is linear, $Y = \beta_0 + \beta_1 X + U$, and if $\text{Cov}(X, U) = 0$, then $\beta_1$ equals the linear projection slope coefficient. However, the linear structural model may be misspecified, too. This is one motivation for “nonparametric” CEF estimation (Section 8.3).

### 7.7 OLS Estimation and Inference

OLS estimation was initially discussed in Section 6.7, along with important terms like fitted values and residuals. Here, additional insights, statistical properties, and code are provided.

#### 7.7.1 OLS Estimator Insights

For the OLS estimator, it is most important to know the statistical properties and R functions, but I can’t resist a couple comments on the nature of the estimator itself.

First, the “least squares” formulation of the OLS estimator from (6.46) mirrors the BLP definition in (7.15). That is, replacing the population mean ($E$) in (7.15) with the sample mean ($\bar{\text{E}}$) yields (6.46). This reinforces that OLS fundamentally estimates the BLP, or equivalently the LP or BLA, not the CEF. The CEF equals the LP only in the very special case of a linear CEF.

Second, the OLS estimator can be written parallel to the population linear projection coefficients in (7.8). Again, replacing population mean with sample mean, and replacing population variance and covariance with sample variance and covariance, (7.8) turns into

$$
\hat{\beta}_1 = \frac{\hat{\text{Cov}}(Y, X)}{\hat{\text{Var}}(X)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}),
$$

$$
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \bar{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i, \quad \bar{X} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i.
$$

(7.16)
This matches the formulas from solving the minimization problem in (6.46) directly. This may seem surprising at first, but recall that the population formula in (7.8) came from solving the population minimization problem. For this reason, \( \hat{\beta}_1 \) may be called the sample analog of \( \beta_1 \), and similarly for \( \hat{\beta}_0 \) and \( \beta_0 \), just as the sample mean \( \bar{E}(Y) = \frac{1}{n} \sum_{i=1}^{n} Y_i \) is the sample analog of the population mean \( E(Y) \) (Sections 3.3 and 3.4).

Third, the OLS estimator essentially performs orthogonal projection in the linear algebra sense. The actual math is beyond our scope, but to get the fitted values \( \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \), the vector of \( Y_i \) values is projected onto a certain subspace defined by the \( X_i \) values.

### 7.7.2 Statistical Properties

The following statistical properties consider OLS as an estimator of the linear projection coefficients (LPCs). These properties hold true under very general assumptions. If the CEF is linear, then it equals the linear projection, so these properties would equally apply to CEF estimation. If the CEF is linear and additional assumptions hold such that the CEF slope identifies the ASE of \( X \) on \( Y \) (Section 6.5.2), then the following properties apply to ASE estimation.

However, as before, the measures of statistical uncertainty (like confidence intervals) say nothing about uncertainty in the identifying assumptions. The statistical uncertainty only captures uncertainty about the LPCs.

#### Assumptions

The following assumptions combined are sufficient for Theorems 7.1–7.3 but not necessary (using logical terms from Section 6.1).

**Assumption A7.1** (iid sampling). Sampling of \((Y_i, X_i)\) is iid.

**Assumption A7.2** (non-constant regressor). The regressor \( X \) is not a constant, i.e., there is no single value \( x \) such that \( P(X = x) = 1 \).

**Assumption A7.3** (finite variances). The variances of \( Y \) and \( X \) are finite: \( \text{Var}(Y) < \infty \), \( \text{Var}(X) < \infty \). Or, equivalently, the expected values of \( Y^2 \) and \( X^2 \) (i.e., second moments) are finite: \( E(Y^2) < \infty \), \( E(X^2) < \infty \).

**Assumption A7.4** (finite fourth moments). The expected values of \( Y^4 \) and \( X^4 \) (i.e., fourth moments) are finite: \( E(Y^4) < \infty \), \( E(X^4) < \infty \).

Assumption A7.1 was discussed in Section 3.2 for \( Y_i \) by itself. If we let vector \( W_i \equiv (Y_i, X_i) \) be what’s observed about individual \( i \), and vector \( W_k \equiv (Y_k, X_k) \) be the observation for individual \( k \), then the iid assumption is essentially the same as before: \( W_i \perp \perp W_k \) for \( i \neq k \) (“independent”), and \( W_i \) and \( W_k \) have the same distribution (“identically distributed”). More specifically, “independent” means \((Y_i, X_i) \perp \perp (Y_k, X_k)\) for \( i \neq k \), which implies \( Y_i \perp \perp Y_k, X_i \perp \perp X_k \),
\[ Y_i \perp X_k, \text{ and } X_i \perp Y_k, \] but implies nothing about (in)dependence between \( X_i \) and \( Y_i \) (or \( X_k \) and \( Y_k \)). "Identically distributed" says \((Y_i, X_i)\) and \((Y_k, X_k)\) have the same joint distribution, which implies the conditional and marginal distributions (and their features) are also identical. For example, \( E(Y_i) = E(Y_k) \), \( \text{Var}(X_i) = \text{Var}(X_k) \), \( E(Y_i | X_i = x) = E(Y_k | X_k = x) \), \( P(Y_i \leq 0 | X_i = x) = P(Y_k \leq 0 | X_k = x) \), etc. All this readily generalizes to multiple regressors, just redefining \( W_i \equiv (Y_i, X_{i1}, X_{i2}, \ldots) \).

There can be dependence among the elements of \( W_i \). Specifically, the outcome \( Y_i \) and regressor \( X_i \) may be correlated or otherwise dependent. The iid assumption does not restrict the relationship between \( Y_i \) and \( X_i \) at all.

Assumptions A7.3 and A7.4 are similar, but A7.4 is stronger. That is, \( A7.4 \implies A7.3 \), i.e., \( E(Y^4) < \infty \implies E(Y^2) < \infty \) and similarly for \( X \).

Assumptions A7.3 and A7.4 are usually true with economic data, but there are some exceptions. They are true for any variable whose absolute value is bounded, like \( |Y| \leq b < \infty \), because then \( E(Y^4) \leq E(b^4) = b^4 < \infty \). For example, if \( X \) is age or education, then \( |X| \leq 200 \), so \( E(X^4) \leq (200)^4 < \infty \).

Nonetheless, some economic variables may violate A7.4 or even A7.3. (Or, there are variables best modeled by distributions that violate these assumptions.) One example is stock returns, or other asset returns. Whether to model such financial returns with finite or infinite variance is a matter of ongoing debate; e.g., see Grabchak and Samorodnitsky (2010) and references therein.

Assumption A7.2 is qualitatively similar to the overlap assumption (A6.6). They both say we must see different values of \( X \) in order to learn about a relationship involving \( X \). They both seem obvious with a single \( X \).

Conveniently, if A7.2 seems false in the data, then your statistical software will report an error or warning. So, don’t worry about A7.2, unless you get such a warning.

**Theoretical Results**

**Theorem 7.1** (OLS consistency, 1 regressor). If A7.1–A7.3 are true, then the OLS intercept and slope estimators are consistent for the population linear projection intercept and slope.

Theorem 7.1 says that with enough data, the OLS coefficient estimators should be close to the true linear projection coefficients with high probability. Whether the linear projection is a CEF or whether the slope has a causal interpretation are questions of identification, not estimation. OLS estimates the linear projection, and leaves further interpretation up to us.

Logically, Theorem 7.1 does not say that OLS is a bad estimator if sampling is not iid (the “inverse”), as discussed in Section 6.1. In fact, the iid assumption can be relaxed in certain ways: even with some
dependence” (instead of independence), or survey weights, OLS can still consistently estimate the population LPCs.

**Theorem 7.2** (OLS approximate normality, 1 regressor). If A7.1, A7.2, and A7.4 are true, then the OLS intercept and slope estimators are asymptotically normal, i.e., with large n, approximately \( \hat{\beta}_0 \sim N(\beta_0, \text{SE}_0^2) \) and \( \hat{\beta}_1 \sim N(\beta_1, \text{SE}_1^2) \), where the true standard errors \( \text{SE}_0 \) and \( \text{SE}_1 \) are unknown but can be estimated and are proportional to \( 1/\sqrt{n} \).

Theorem 7.2 is practically useful for constructing confidence intervals, whose properties are in Theorem 7.3.

**Theorem 7.3** (coverage probability, 1 regressor). If A7.1, A7.2, and A7.4 are true, then the heteroskedasticity-robust confidence intervals in Section 7.7.3 are asymptotically correct. That is, with large enough n, the coverage probability is approximately equal to the desired confidence level.

### 7.7.3 Code

The following code is based on the example from Section 7.1. Each row in the final output shows the estimate \( \hat{\beta}_1 \) (in the column titled Estimate) along with the heteroskedasticity-robust standard error (column Std. Error), t-statistic for \( H_0: \beta_1 = 0 \) (column t value) and p-value for \( H_0: \beta_1 = 0 \) (column Pr(>|t|)), and 95% CI for \( \beta_1 \) (lower endpoint in column 2.5 %, upper endpoint in column 97.5 %).

There are three randomly simulated datasets for which these quantities are estimated. All have \( X \in \{0,1,2\} \). The first dataset comes from a linear CEF with \( m(x) = 60 - 20x \), where \( P(X = j) = 1/3 \) for \( j = 0,1,2 \). The next two datasets have nonlinear CEF \( m(0) = 60, m(1) = m(2) = 40 \), but different distributions of \( X \). The first distribution has \( P(X = j) = (3 - j)/6 \), while the second has \( P(X = j) = (j + 1)/6 \), for \( j = 0,1,2 \). As seen, the distribution of \( X \) affects the linear projection slope when the CEF is nonlinear, as discussed in Section 7.4.

Finally, dummy variables are used to estimate a properly specified nonlinear CEF, as in (7.4). Only the estimated coefficients are displayed below, using the coefficients() function. Specifically, the number under (Intercept) is the estimated intercept, the number under D1 is the estimated coefficient on D1, and the number under D2 is the estimated coefficient on D2.

```r
library(lmtest); library(sandwich)
set.seed(112358)
n <- 500  # sample size
m012 <- c(60,40,20)  # m(0),m(1),m(2) (linear CEF)
df <- data.frame(X=sample(x=0:2, size=n, prob=c(1,1,1)/3, replace=TRUE),
                 U=rnorm(n))
df$Y <- rnorm(n=n, mean=m012[1+df$X]) + df$U
```
```r
ret <- lm(formula=Y~X, data=df)
retVC1 <- vcovHC(ret, type="HC1")
CEF <- c(coeftest(ret, vcov. = retVC1)['X'],
         coefci(ret, vcov. = retVC1)['X'],
        
# Now: nonlinear CEF; LPC depends on X dist
set.seed(112358)
ret <- lm(formula=Y~X, data=df)
retVC1 <- vcovHC(ret, type="HC1")
ret <- lm(formula=Y~X, data=df)
retVC1 <- vcovHC(ret, type="HC1")
```

```r
## Estimate Std. Error t value Pr(>|t|) 2.5 % 97.5 %
## CEF -19.8 0.077 -257.1 0  19.98 19.67
## LP1 -12.3 0.310 -39.6 0  12.91 11.69
## LP2 -7.7 0.310 -24.8 0  8.31  7.09
```

```r
# Use dummies to estimate nonlinear CEF
df$D <- (df$X==1)
df$D1 <- as.integer(df$X==1)  # D1=1 iff X=1
df$D2 <- as.integer(df$X==2)  # D2=1 iff X=1
ret <- lm(formula=Y~D1+D2, data=df)
```

```r
coefficients(ret)
```

## (Intercept) D1 D2
## 59.8 -19.8 -19.8

### 7.8 Simple Linear Regression

⇒ Kaplan video: OLS in R

The prior results are essentially the same when X has more than
three possible values, too. There could even be an infinite number of possible values; e.g., if there is no upper bound for \(X\), or if \(X\) could be any real (decimal) number between 0 and 1. Misspecification is likely. The linear projection, best linear approximation, and best linear predictor interpretations all still apply. OLS estimation and heteroskedasticity-robust standard errors and confidence intervals are computed the same way.

The main difference is that it is harder to use dummy variables to properly model a nonlinear CEF. If \(X\) has only four values, then it is not too difficult. But if \(X\) has hundreds or thousands of values, or an infinite number, then the dummy variable approach may fail. Chapter 8 addresses alternative ways to model a CEF that is not linear in \(X\).

**Practice 7.2** (linear fit). For each scatterplot in Figure 7.3, guess what the OLS estimated regression line looks like, i.e., the line \(\hat{\beta}_0 + \hat{\beta}_1 X\). (Hint: remember OLS minimizes the sum of the squares of the vertical distances from each point to the fit line.) You can also make your own puzzles in R: first make a scatterplot like

\[
Y \leftarrow c(1,2,3,4,13); \ X \leftarrow c(1,2,3,4,5); \ \text{plot}(X,Y)
\]

and then (after guessing) plot the OLS fit with

\[
\text{abline(lm(Y~X))}
\]

![Figure 7.3: Scatterplots for Practice 7.2.](attachment:image.png)

**Practice 7.3** (regression units). Consider a regression of wage \(Y\) \(($/hr)\) on “distance to nearest university” \(X\). Let \(\gamma_1\) be the estimated slope when \(X\) is measured in miles, and let \(\delta_1\) be the estimated slope when \(X\) is measured in kilometers, where \(1 \text{ mi} = 1.6 \text{ km}\).

a) What are the units of \(\gamma_1\)? \(\delta_1\)?
b) Do you think $\gamma_1 = \delta_1$, $\gamma_1 > \delta_1$, or $\gamma_1 < \delta_1$?

c) Can you come up with a formula relating $\gamma_1$ and $\delta_1$? (Hint: what change in $Y$ is associated with a 1.6 km increase in $X$, in terms of $\gamma_1$? In terms of $\delta_1$?)

**Discussion Question 7.3** (student-teacher ratio simple regression).

Let $Y$ be the average math standardized test score (in units of points) for a school’s 5th-grade students. Let $X$ be the 5th-grade student-teacher ratio (total number of 5th-grade students divided by total number of 5th-grade teachers; like the average class size), generally around $15 \leq X \leq 25$. For schools $i = 1, \ldots, n$, the values $(Y_i, X_i)$ are recorded. A linear regression is run to estimate $\beta_0$ and $\beta_1$ in the CEF model $Y = \beta_0 + \beta_1 X + V$, $E(V \mid X) = 0$.

a) What are the units of $\beta_0$ and $\beta_1$?

b) What’s the interpretation of $\beta_0$? What is it useful for?

c) Consider the estimate $\hat{\beta}_1 = -2.28$. What does this imply about the average score difference between 15-student classes and 25-student classes? Is it economically significant (Section 3.9.7)? (Hint: make additional assumptions about the scoring system/scale if you need to.)

d) Consider further that $\hat{\beta}_1$ has heteroskedasticity-robust standard error 0.8, so the $p$-value for $H_0: \beta_1 = 0$ is 0.004. Discuss the statistical significance (Section 3.8.4) of $\hat{\beta}_1$.

e) Describe one reason you doubt $\hat{\beta}_1$ has a causal interpretation.

f) Describe one reason you think the linear CEF model is misspecified.
Empirical Exercises

Empirical Exercise EE7.1. You will analyze data on colleges’ athletic success and number of applications. The data were collected by Patrick Tulloch for an economics term project, from various college and sports data records. As the R description says, “The ‘athletic success’ variables are for the year prior to the enrollment and academic data.”

a. Load the data (assuming you’ve already installed the R package or Stata command).

R: library(wooldridge)

Stata: bcuse athlet1 , nodesc clear

b. Keep only data from 1993.

R: dat <- athlet1[athlet1$year==1993 , ] Stata: keep if year==1993

c. Create a new variable equal to the sum of bowl (football bowl game) and finfour (men’s basketball Final Four).

R: dat$bowl4 <- dat$bowl + dat$finfour

Stata: generate bowl4 = bowl + finfour

d. Display the number of observations with each possible value of bowl4 (0, 1, or 2).

R: table(dat$bowl4)

Stata: tabulate bowl4

e. Regress the number of applications (for admission) on the prior year’s athletic success.

R: ret <- lm(apps~bowl4, data=dat)

Stata: regress apps bowl4 , vce(robust)

f. R only: save the fitted OLS values of \( \hat{Y} \) for the three possible values of \( X \) (bowl4) with fit012 <- predict(ret, newdata =data.frame(bowl4=0:2)) and optionally add helpful labels with names(fit012) <- c(‘X=0’,’X=1’,’X=2’) 

R: mean(dat$apps[dat$bowl4==0]) to estimate \( m(0) \), and replace 0 with 1 to estimate \( m(1) \) and with 2 to estimate \( m(2) \); store these into a vector named m012 with m012 <- c( m0 , m1 , m2 ) where m0 is your code for estimating \( m(0) \) and similarly for m1 and m2.

Stata: bysort bowl4 : egen CEF = mean(apps) to compute the sample mean of apps within each group of observations with the same value of bowl4, storing it into a new variable named CEF
h. Plot the fitted OLS line against the estimated CEF points.
   R: `plot(0:2, m012)` (to plot estimated CEF points) followed by `abline(ret)` (to plot the OLS fit line)
   Stata: `twoway scatter CEF bowl4 || lfit apps bowl4`

i. Make the same plot, but adjust the line color and style, the title, the axis labels, and whatever else you’d like to adjust.
   R: inside the `plot()` command, add argument `main='...'` to set the title and similarly for `xlab='...'` and `ylab='...'` to set the x-axis and y-axis labels (where you replace all the ... with whatever names you want); inside the `abline()` function, add arguments `col=2` to change the line’s color, `lty=2` to change the line style, and `lwd=3` to change the line width; again, you can set whatever values you like.
   Stata: `twoway scatter CEF bowl4 || lfit apps bowl4`, XXX but replace the XXX with options to change the graph’s appearance (all separated by spaces, not any more commas), like `title("...") xtitle("...") ytitle("...")` for the title and axis labels, and `lcolor(red) lpattern(dash)` for the line color and style; use whatever values you’d like.

j. Display the numerical values of the OLS fit and the estimated CEF.
   R: `rbind(m012, fit012)`
   Stata: `collapse (mean) meanapps=apps`, by(bowl4) followed by `predict OLSfit`, `xb` and `list`
Chapter 8

Nonlinear and Nonparametric Regression

⇒ Kaplan video: Chapter Introduction

Depends on: Chapter 7 (which depends on Chapters 2–4 and 6)

Unit learning objectives for this chapter

8.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]

8.2. Interpret the coefficients in various nonlinear regression models [TLOs 3 and 5]

8.3. Judge which model seems most appropriate, using both economic reasoning and statistical insights [TLO 6]

8.4. In R (or Stata): estimate nonlinear and nonparametric regression models, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

Optional resources for this chapter

- Functional form misspecification (Lambert video)
- Log-log example (Lambert video)
- Overfitting (Lambert video)
- Sections 2.4 ("Nonlinearities," including log models), 6.1.3 ("Logarithms"), and 6.1.4 ("Quadratics and Polynomials") in Heiss (2016)
- Section 8.2 ("Nonlinear Functions of a Single Independent Variable") in Hanck et al. (2018)
- Nonparametric regression: Chapter 7 ("Moving Beyond Linearity") in James et al. (2013), including §7.5 ("Smoothing Splines"); and Chapter 5 ("Basis Expansions and Regularization") in Hastie, Tibshirani, and Friedman (2009), including §5.4 ("Smoothing Splines")
Model selection: Chapter 7 ("Model Assessment and Selection") in Hastie, Tibshirani, and Friedman (2009)

Bias–variance tradeoff: James et al. (2013, §2.2.2), Hastie, Tibshirani, and Friedman (2009, §§2.9,5.5.2,7.2,7.3)

Part V ("Nonparametric Regression") in Kaplan (2020)

R package splines

Having mastered regression with a linear functional form, we now consider nonlinear functions. First nonlinear functions of $X$ are allowed, and then nonparametric estimation and machine learning are introduced.

8.1 Log Transformation

Sometimes a simple regression model improves greatly by transforming $Y$ or $X$ or both. The most common transformation in economics is the natural logarithm function, which economists just call “log.”

Three different log models are discussed below. A model with the familiar form $Y = \beta_0 + \beta_1 X + U$ could be called a “linear-linear” model (although it’s just called a linear model), meaning both $Y$ and $X$ are in their original units, i.e., in levels. If $Y$ is replaced by its log, $\ln(Y)$, it’s called a log-linear model; if instead we have $Y$ and $\ln(X)$, then it’s linear-log; and if both are in logs, then log-log.

Here in Section 8.1, the distinction among causal, CEF, and linear projection models is unimportant. The interpretation of $U$ is left ambiguous intentionally. Instead, emphasis is on the interpretation of $\beta_1$ in terms of units of measure.

8.1.1 Properties of the Natural Log Function

Basic Shape and Properties

The natural log function is peculiar, especially if you haven’t taken calculus. It is written $\ln(\cdot)$, although often people will simply say “log” (without “natural”) and write $\log(\cdot)$, since the natural log is the only one commonly used in economics; in R, the function is $\text{log()}$.

The log function is the inverse of the exponential function: $\ln(\exp(x)) = x$, where $\exp(x)$ is the same as $e^x$. Consequently, if $e^x = M$, then $\ln(M) = \ln(e^x) = x$.

Figure 8.1 shows the log function, giving a general idea of its shape. However, two important features are unclear. First, as $x$ gets closer and closer to 0, $\ln(x)$ decreases toward $-\infty$. Second, $\ln(x)$ keeps increasing to $\infty$ as $x$ increases to $\infty$.

The log function has many properties, including the following.

1. $\ln(x)$ is only defined for $x > 0$
2. $\ln(x)$ is strictly increasing; for any $x_2 > x_1 > 0$, $\ln(x_2) > \ln(x_1)$
3. $\ln(x)$ increases more slowly with larger $x$; it is very steep for $x$ near zero, but less and less steep (i.e., flatter) as $x$ increases
4. For any \( x > 0 \) and any \( b \), \( \ln(x^b) = b \ln(x) \)

5. For any \( x_1 > 0 \) and \( x_2 > 0 \), \( \ln(x_1/x_2) = \ln(x_1) - \ln(x_2) \) and \( \ln(x_1x_2) = \ln(x_1) + \ln(x_2) \)

6. \( \lim_{x \downarrow 0} \ln(x) = -\infty \) and \( \lim_{x \to \infty} \ln(x) = \infty \)

**Percentage Approximation**

Near \( x = 1 \), \( \ln(x) \) is approximately the same as the linear function \( f(x) = x - 1 \), i.e., \( \ln(x) \approx x - 1 \). Equivalently, letting \( w \equiv x - 1 \), if \( w \) is near zero, then \( \ln(1 + w) \approx w \). For example, with \( w = 0.01 \), \( \ln(1 + 0.01) = 0.00995 \). Negative \( w < 0 \) is fine, too: \( \ln(1 - 0.01) = -0.01005 \). Even with \( w = 0.1 \), \( \ln(1.1) = 0.0953 \), not far from 0.1. The approximation is perfect at \( w = 0 \) since \( \ln(1) = 0 \) exactly, and it gets worse as \( w \) increases; \( \ln(1.5) = 0.405 \), not good.

The log function can approximate small percent changes. Consider \( v_2 > v_1 \); how much bigger is \( v_2 \)? In percentage (of \( v_1 \)) terms, \( v_2 \) is

\[
100\left(\frac{v_2 - v_1}{v_1}\right)\% = 100\left(\frac{v_2}{v_1} - 1\right)\%
\]

larger than \( v_1 \). For example, if \( v_1 = 100 \) and \( v_2 = 102 \), then \( v_2/v_1 - 1 = 1.02 - 1 = 0.02 \), so we’d say \( v_2 \) is 100(0.02)% = 2% larger than \( v_1 \). In other words, the increase (in level) of \( v_2 - v_1 = 2 \) is 2% of \( v_1 \). This 2% can be approximated by the log increase, \( \ln(v_2) - \ln(v_1) \). Let \( p \equiv v_2/v_1 - 1 \), like \( p = 0.02 \) in the example, so \( v_2 = v_1(1 + p) \).

Combining two properties above, if \( p \) is near zero, then

\[
\ln(v_2) = \ln(v_1(1+p)) = \ln(v_1)+\ln(1+p) \approx \ln(v_1)+p \implies p \approx \ln(v_2) - \ln(v_1)
\]

Put differently, a log difference of \( p = \ln(v_2) - \ln(v_1) \) is approximately a 100p% change in level. In fact, the above math is identical when \( v_2 < v_1 \), so \( p \) can be positive or negative (increase or decrease). However, as before, the approximation is poor if \( p \) is larger, like 0.5.
8.1.2 The Log-Linear Model

Interpretation

A log-linear model specifies

$$ \ln(Y) = \beta_0 + \beta_1 X + U. $$  \hspace{1cm} (8.2)

Since $X$ is in levels, the coefficient $\beta_1$ tells us about a one unit increase in $X$. Specifically, a one unit increase in $X$ is associated with a $\beta_1$ change in $\ln(Y)$ (increase if $\beta_1 > 0$, decrease if $\beta_1 < 0$). Sometimes, people call this a $\beta_1$ change in $Y$ in log units.

If $\beta_1$ is close to zero, then (8.1) offers another interpretation: a one unit increase in $X$ is associated with approximately a $100\beta_1\%$ change in $Y$. For example, if $\beta_1 = 0.02$, then a one unit increase in $X$ is associated with approximately a 2% increase in $Y$.

Recall the difference between a percentage change and a percentage point change. For example, a 1% increase in $Y$ means increasing to $1.01Y$. “Percentage point” only applies when the units are already percentages; e.g., a 1 percentage point increase is changing from 10% to 11%, or from 67% to 68%.

However, even if $\beta_1$ is near zero, the approximation in (8.1) may be poor if we consider large changes in $X$. For example, if again $\beta_1 = 0.02$, but we consider a 50-unit increase in $X$, the increase in $Y$ is poorly approximated by $100(\beta_1)(50)\% = 100\%$. A 100% increase would be from value $v_1$ to value $v_2 = 2v_1$. But, if $\ln(v_2) - \ln(v_1) = 1$ (a change of 1 “log unit”), then $\ln(v_2/v_1) = 1$, meaning $v_2/v_1 = e \approx 2.72$, not $v_2/v_1 = 2$.

When to Use It

When does a log-linear model make sense? Sometimes, scatterplots of the raw $Y$ and $X$ data suggest it. For example, maybe the relationship between $Y$ and $X$ looks nonlinear, but the relationship between $\ln(Y)$ and $X$ looks approximately linear.

Sometimes, even before looking at data, the log-linear model makes more sense economically or intuitively. For example, with $Y$ variables like income, it may seem more natural to model effects as (approximate) percentage changes in $Y$, like a 1% higher income instead of a $500/yr higher income. Further, the log-linear form derives from economic models of human capital where there is a multiplicative effect on wage. The most famous of these is the “Mincer equation” for earnings as a function of education (schooling) and experience, named after the log-linear model in Mincer (1974, Ch. 5, p. 84).

Issue with Prediction

Unfortunately, the log-linear model is not optimal for predicting $Y$, even if $E(U \mid X) = 0$. From (8.2), the CEF is

$$ E(Y \mid X = x) = e^{\beta_0 + \beta_1 x} E(e^U \mid X = x). $$
It is easy to plug in \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), but difficult to estimate \( E(e^U \mid X = x) \). We could simply ignore the difficult term, but \( e^{\hat{\beta}_0 + \hat{\beta}_1 x} \) is generally not the best predictor of \( Y \) given \( X = x \). There are alternatives, but they are beyond our scope.

### 8.1.3 The Linear-Log Model

#### Interpretation

A linear-log model specifies

\[
Y = \beta_0 + \beta_1 \ln(X) + U. \tag{8.3}
\]

When \( X \) increases by one log unit, the corresponding change in \( Y \) is \( \beta_1 \); but one log unit is a very big change (more than doubling). To use the percentage approximation, a smaller change in \( X \) must be used. Specifically, an increase of \( X \) by 1% is associated with a change in \( Y \) of \( \beta_1 / 100 \) units. A 1% increase is a change from \( X \) to \( 0.01X \), which is different than a 1 percentage point change in \( X \).

Mathematically, the interpretation of \( \beta_1 \) can be seen in two steps. First, let \( Z = \ln(X) \), and imagine a linear model with \( Y \) and \( Z \):

\[
Y = \beta_0 + \beta_1 Z + U.
\]

Then, an increase in \( Z \) by 0.01 units corresponds to a change in \( Y \) of \( (\beta_1)(0.01) = \beta_1 / 100 \) units of \( Y \). Second, from (8.1), an increase in \( Z = \ln(X) \) by 0.01 is approximately a \( (100)(0.01)\% = 1\% \) increase in \( X \).

For larger changes, instead of using the percentage approximation, just plug in two values of \( X \). For example, consider increasing \( X = 40 \) to \( X = 60 \) (a 50% increase). The associated change in \( Y \) is

\[
\frac{X=60}{X=40} \left( \beta_0 + \beta_1 \ln(60) - [\beta_0 + \beta_1 \ln(40)] \right) = (\beta_0 - \beta_0) + \beta_1 [\ln(60) - \ln(40)] = \beta_1 \ln(1.5) = 0.41\beta_1.
\]

The same 0.41\( \beta_1 \) change results for any 50% increase in \( X \), regardless of starting value, because \( \ln(1.5X) - \ln(X) = \ln(1.5X/X) = \ln(1.5) = 0.41 \) log units. More generally, a change from \( X = x_1 \) to \( X = x_2 \) is associated with a change in \( Y \) of \( \beta_1 \ln(x_2/x_1) \).

#### When to Use It

When does a linear-log model make sense? Sometimes, the scatterplot of \( Y \) and \( X \) reveals a shape that looks like a log function: increasing steeply at first, then getting less and less steep, but without ever decreasing. (Or: switch “increasing” and “decreasing,” if \( \beta_1 < 0 \).) That is, the relationship between \( Y \) and \( X \) looks nonlinear, but maybe plotting \( Y \) against \( \ln(X) \) looks closer to linear. The log function’s shape also helps model diminishing marginal benefits: the first unit of \( X \) helps increase \( Y \) a lot, but each additional unit of \( X \) helps less and less.
8.1.4 The Log-Log Model

Interpretation

A log-log model specifies
\[ \ln(Y) = \beta_0 + \beta_1 \ln(X) + U. \] (8.4)

A 1\% increase in \( X \) is associated with an approximate \( \beta_1 \% \) change in \( Y \). This percentage interpretation is particularly nice: \( \beta_1 \) represents an elasticity of \( Y \) with respect to \( X \). But, if the percentages are too large, then the approximation is poor.

When to Use It

When does a log-log model make sense? First, it’s a simple way to get an elasticity interpretation. Second, a scatterplot of \( \ln(Y) \) against \( \ln(X) \) may look roughly linear. Third, if you suspect a power law type of relationship between \( Y \) and \( X \), exponentiating both sides of (8.4) yields

\[ \exp\{\ln(Y)\} = \exp\{\beta_0 + \beta_1 \ln(X) + U\}, \implies Y = e^{\beta_0} \exp\{\ln(X^{\beta_1})\} e^U \]

Issue with Prediction

As with the log-linear model, \( e^{\beta_0} X^{\beta_1} \) is generally not the CEF because \( E(e^U \mid X) = 1 \) is not implied by \( E(U \mid X) = 0 \). Consequently, predicting \( Y \) as \( e^{\beta_0} X^{\beta_1} \) is generally not optimal.

<table>
<thead>
<tr>
<th>In Sum: Regression Models with Log Transformations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-linear: 1-unit ↑ ( X ) associated with approximate 100( \beta_1 % ) change in ( Y )</td>
</tr>
<tr>
<td>Linear-log: 1% ↑ ( X ) associated with approximate ( \beta_1/100 )-unit change in ( Y ); more precisely, change from ( x_1 ) to ( x_2 ) associated with ( \beta_1 \ln(x_2/x_1) )-unit change in ( Y )</td>
</tr>
<tr>
<td>Log-log: 1% ↑ ( X ) associated with approximate ( \beta_1 % ) change in ( Y ) (elasticity)</td>
</tr>
</tbody>
</table>

Discussion Question 8.1 (pollution and house price). Consider the relationship between the price of a house and the concentration of air pollution. Explain which type of model (linear, log-linear, linear-log, or log-log) you think would best fit, and why. (Hint: think especially about changes in levels vs. in logs.)

8.1.5 Warning: Model-Driven Results

\[ \implies \text{Kaplan video: Warnings About Model-Driven Results} \]

When choosing a model, beware self-fulfilling prophecy. Empirical results are driven by data, but also by your model’s structure. For example, the function \( \beta_0 + \beta_1 X \) specifies a constant (\( \beta_1 \)) change
for every unit increase in $X$; different datasets can lead to different estimated slopes ($\hat{\beta}_1$), but the slope will always be constant, regardless of the data. The log-linear model may seem more flexible than a linear model, but it is not: it still only has two parameters. It is just different, not more flexible. Consequently, the fitted log-linear model always shows a diminishing effect of $X$ on $Y$ as $X$ increases. This pattern does not come from the data, but from the model itself, regardless of the data.

Figure 8.2, based on the comic at https://xkcd.com/2048, illustrates such self-fulfilling prophecy. Each graph shows the same scatterplot from the same data (the dots), but with a very different fitted model in each (the line). Clearly, the differences do not come from the data since it’s the exact same data. All differences are entirely due to the model. The top-left shows the linear model, which by construction imposes a constant slope $\beta_1$. Below that is a log-linear model; the constant percentage increase of $Y$ with each unit of $X$ leads to exponential growth (hence the “exponential” label in the comic). The top-right shows the “tapering off” of the linear-log model. Although mostly beyond our scope, some comments on “model selection” are in Sections 8.3 and 15.2.

### 8.1.6 Code

![Figure 8.2: Same data, different models.](https://xkcd.com/2048)
results illustrate the concerns of Section 8.1.5.

```r
par(family='serif', mar=c(3,3,1,1), mgp=c(2.1,0.8,0), mfrow=c(2,2))
set.seed(112358)
n <- 31
X <- sort(runif(n=n, min=1, max=9))
Y <- 1 + pnorm(q=X, mean=5, sd=1.5) +
    2*( rbeta(n=n, shape1=10-X, shape2=X) - (10-X)/10 )
df <- data.frame(X=X, Y=Y)
ret.linlin <- lm(Y~X, data=df)
ret.loglin <- lm(log(Y)~X, data=df)
ret.linlog <- lm(Y~log(X), data=df)
ret.loglog <- lm(log(Y)~log(X), data=df)

# XL <- ''; YL <- ''
plot(x=df$X, y=df$Y, type='p', pch=16, main='', xlab=XL, ylab=YL)
lines(predict(ret.linlin)~df$X, col=2)
title("Linear", line=-1, adj=0.1)

# plot(x=df$X, y=df$Y, type='p', pch=16, main='', xlab=XL, ylab=YL)
lines(predict(ret.linlog)~df$X, col=2)
title("Linear-Log", line=-1, adj=0.1)

# plot(x=df$X, y=df$Y, type='p', pch=16, main='', xlab=XL, ylab=YL)
lines(exp(predict(ret.loglin))~df$X, col=2)
title("Log-Linear", line=-1, adj=0.1)

# plot(x=df$X, y=df$Y, type='p', pch=16, main='', xlab=XL, ylab=YL)
lines(exp(predict(ret.loglog))~df$X, col=2)
title("Log-Log", line=-1, adj=0.1)
```

### 8.2 Nonlinear-in-Variables Regression

**Discussion Question 8.2** (nonlinear OVB). Imagine a structural model $Y = \beta_0 + \beta_1 X + \beta_2 X^2$, with no error term: $X$ completely determines $Y$. To be more concrete, imagine $Y = 1 + X^2$ (i.e., $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$), with $0 \leq X \leq 5$. You run a linear-in-variables regression; OLS estimates the function $\hat{\gamma}_0 + \hat{\gamma}_1 X$.

a) Approximately what value would you expect $\hat{\gamma}_1$ to be? (Hint: recall Sections 7.3–7.5.)

b) What does $\hat{\gamma}_0 + \hat{\gamma}_1 X$ suggest about the relationship between $X$ and $Y$? What features are similar or different compared to the true $1 + X^2$? (Hint: draw a picture.)

Beyond replacing $X$ with a single transformation of $X$ like $\ln(X)$, we can replace $X$ with a more complicated nonlinear function involving multiple terms and multiple parameters. OLS can still be used
There are two types of (non)linearity. They are often confused. Further, people often say “linear model” or “nonlinear model” without clarifying which type they mean.

8.2.1 Linearity

The root of “linearity” is linear combination. A linear combination is like a weighted sum. For example, a linear combination of \( A \) and \( B \) is anything with the form

\[
w_1 A + w_2 B,
\]

where \( w_1 \) and \( w_2 \) are weights that may take any value, including zero or even negative numbers. Linear combinations may involve more than two terms, like \( w_1 A + w_2 B + w_3 C + w_4 D \). In some cases, instead of \( A, B, C, \) and \( D \), we have something like \( Y_1, Y_2, Y_3, \) and \( Y_4 \), in which case the linear combination may be written in summation notation:

\[
w_1 Y_1 + w_2 Y_2 + w_3 Y_3 + w_4 Y_4 = \sum_{i=1}^{4} w_i Y_i.
\]

For example, the expected value formula for discrete random variables in (2.16) is a special case of a linear combination, where the linear combination weights are the probabilities of the different possible values. Also, the sample mean is a linear combination of observed \( Y_i \) values, with weights \( w_i = 1/n \).

A function is linear-in-parameters if it is a linear combination of the parameters. For example, \( \beta_0 + \beta_1 X \) is linear-in-parameters because it is a linear combination of the parameters \( \beta_0 \) and \( \beta_1 \) with weights \( w_1 = 1 \) and \( w_2 = X \):

\[
w_1 \beta_0 + w_2 \beta_1 = (1)(\beta_0) + (X)(\beta_1) = \beta_0 + \beta_1 X.
\]

The function \( \beta_0 + \beta_1 X \) is also linear-in-variables. This is trickier to see since it is not actually a linear combination of \( X \) alone. (The function \( \beta_0 + \beta_1 X \) is called an affine function of \( X \), which means a linear function of \( X \) plus a constant.) Secretly, we have actually had a second regressor all along: \( X_0 = 1 \). Since this second regressor \( X_0 \) is just always 1, it has not been treated like a true regressor, but mathematically it is. Seen this way, the linear combination of \( X_0 \) and \( X \) has weights \( w_1 = \beta_0 \) and \( w_2 = \beta_1 \):

\[
(w_1)(X_0) + (w_2)(X) = (\beta_0)(1) + (\beta_1)(X) = \beta_0 + \beta_1 X.
\]

For this reason, in economics, people often call \( \beta_0 + \beta_1 X \) “linear in \( X \)” even though technically it is “affine in \( X \)” and “linear in \( X_0 \) and \( X \).”

These two types of linearity can apply specifically to CEFs or linear projections. For example, if the CEF is \( \text{E}(Y \mid X = x) = \)
\( \beta_0 + \beta_1 x \), then the CEF is linear-in-parameters and linear-in-variables. Regardless of the CEF, the linear projection of \( Y \) onto \((1, X)\) is \( \text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X \), which is always linear-in-parameters and linear-in-variables by definition.

Confusingly, even if the models are written in error form, people still refer to them as “linear.” For example, consider the CEF model \( Y = \beta_0 + \beta_1 X + U \) with \( \text{E}(U \mid X) = 0 \). Despite the \(+U\) at the end, sometimes people say this is linear-in-variables and linear-in-parameters, presumably because \( \text{E}(Y \mid X) = \beta_0 + \beta_1 X \) indeed satisfies both types of linearity. Similarly, consider the linear projection model in error form: \( Y = \beta_0 + \beta_1 X + U \) with \( \text{E}(U) = \text{E}(XU) = 0 \). Again, despite the \(+U\) at the end, sometimes people say this is linear-in-variables and linear-in-parameters, presumably because \( \text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X \) indeed satisfies both types of linearity.

Even more confusingly, sometimes even a structural model of the form \( Y = \beta_0 + \beta_1 X + U \) is called linear-in-parameters and linear-in-variables. In that case, there is no CEF or LP that is implicitly the linear function. Regardless, it is helpful to be aware of conventional terminology even if it’s not the best, so you can understand others when they mention a “linear structural model.”

### 8.2.2 Nonlinearity

Often a quadratic term is added to a model to increase flexibility. Specifically,

\[
Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U \tag{8.7}
\]

is called a **quadratic model** since the right-hand side is a quadratic function of \( X \) (plus an error term). This is now **nonlinear-in-variables** because of the \( X^2 \) term. That is, \( \beta_0 + \beta_1 X + \beta_2 X^2 \) cannot be written as a linear combination of \( X_0 = 1 \) and \( X \), so it is not linear-in-variables. However, (8.7) is still linear-in-parameters: with linear combination weights, \( 1 \), \( X \), and \( X^2 \),

\[
(1)(\beta_0) + (X)(\beta_1) + (X^2)(\beta_2) = \beta_0 + \beta_1 X + \beta_2 X^2.
\]

There are (infinitely) many other examples of functions that are linear-in-parameters but nonlinear-in-variables. For example,

\[
\begin{align*}
\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4, \\
\beta_0 + \beta_1 \sin(X) + \beta_2 \cos(X), \\
\beta_0 + \beta_1 \ln(X) + \beta_2 \sqrt{X} + \beta_3 X^{1/3}.
\end{align*}
\]

Each can be written in terms of functions \( f_j(\cdot) \) in the form

\[
\sum_{j=0}^{J} \beta_j f_j(X). \tag{8.8}
\]

For example, the polynomial example has \( f_j(X) = X^j \) and \( J = 4 \).
A **nonlinear-in-parameters** model cannot be written as a linear combination of the parameters. For example, in the power law model

\[ Y = \beta_0 X^{\beta_1} + U, \quad (8.9) \]

the term \( \beta_0 X^{\beta_1} \) cannot be written as a linear combination of \( \beta_0 \) and \( \beta_1 \). Nonlinear-in-parameters models are not discussed further.

### 8.2.3 Estimation and Inference

OLS can estimate nonlinear-in-variables models as long as they are linear-in-parameters. As always, the OLS estimates are the parameter values that minimize the sum of squared residuals, solving the empirical analog of the optimal prediction problem (minimizing mean quadratic loss).

Inference on parameters is also the same. For example, the same R code to compute a confidence interval for \( \beta_1 \) earlier still works, and a confidence interval for \( \beta_2 \) can be computed the same way. The underlying code/math is very similar, too. However, confidence intervals for predicted values or predicted differences now involve multiple coefficients.

### 8.2.4 Parameter Interpretation

Unlike estimation and inference, which remain similar, interpretation of parameters changes greatly with nonlinear-in-variables models.

**Insufficiency of Linear Coefficient**

In (8.7), \( \beta_1 \) is no longer the change in \( Y \) associated with a unit increase in \( X \). That is, when \( X \) increases, so does \( X^2 \), so both \( \beta_1 \) and \( \beta_2 \) are needed. Not only does the change in \( Y \) associated with a unit increase in \( X \) now depend on the initial value of \( X \), but even the sign of the change (i.e., increase or decrease) may depend on \( X \).

For example, consider the function \( 5X - X^2 \), i.e., \( \beta_0 = 0 \), \( \beta_1 = 5 \), and \( \beta_2 = -1 \). Going from \( X = 0 \) to \( X = 1 \), the change is

\[
[(5)(1) - 1^2] - [(5)(0) - 0^2] = 4 - 0 = 4.
\]

Going from \( X = 1 \) to \( X = 2 \), the change is

\[
[(5)(2) - 2^2] - [(5)(1) - 1^2] = 6 - 4 = 2,
\]

still positive, but smaller. From \( X = 2 \) to \( X = 3 \),

\[
[(5)(3) - 3^2] - [(5)(2) - 2^2] = 6 - 6 = 0,
\]

no change at all. And from \( X = 3 \) to \( X = 4 \),

\[
[(5)(4) - 4^2] - [(5)(3) - 3^2] = 4 - 6 = -2,
\]

a negative change, i.e., a decrease. Even though \( \beta_1 = 5 \) is positive, sometimes an increase in \( X \) is associated with a decrease in \( Y \). Not even the sign of \( \beta_1 \) (positive, negative) tells us anything.
Summarizing Nonlinear Functions

With only one $X$, the best summary is to plot the function (along with a scatterplot of data), like in Figure 8.2. As the saying goes, “A picture is worth a thousand words [or numbers].” However, if there are many different regressors (as in later chapters), pictures get confusing (trying to show slices of many-dimensional manifolds...).

Another approach is to plug in changes of $X$ that are relevant to policy or a particular economic question. For example, if $Y$ is income, $X$ is education, and we want to understand the value of the 12th year of education, then comparing $X = 12$ to $X = 11$ is relevant. With a quadratic model, the associated change in $Y$ is

$$[\beta_0 + \beta_1(12) + \beta_2(12)^2] - [\beta_0 + \beta_1(11) + \beta_2(11)^2] = \beta_1(12-11) + \beta_2(12^2 - 11^2)$$

Generally, we could write

$$Y = f(X) + U,$$

where in the quadratic model $f(X) = \beta_0 + \beta_1X + \beta_2X^2$. Plugging in OLS parameter estimates yields $\hat{f}(X)$, like $\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1X + \hat{\beta}_2X^2$ for the quadratic. We can graph this estimated function by evaluating it at many $X$ values and drawing a line through them. Similarly, for a change from $X = x_1$ to $X = x_2$, the associated change in $Y$ is $\hat{f}(x_2) - \hat{f}(x_1)$.

**In Sum: Interpreting and Summarizing Nonlinear Models**

The $\beta_1X$ term alone has no meaning.

Given (8.10), a change from $X = x_1$ to $X = x_2$ is associated with a change in $Y$ of $\hat{f}(x_2) - \hat{f}(x_1)$.

**Practice 8.1** (quadratic example). You regress $Y$ on $X$ and $X^2$ and get the fitted function $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1X + \hat{\beta}_2X^2$ with $\hat{\beta}_0 = 2$, $\hat{\beta}_1 = 4$, and $\hat{\beta}_2 = -2$.

- a) What’s the predicted value of $Y$ when $X = 0$? $X = 1$? $X = 2$?
- b) What’s the predicted change in $Y$ when $X$ changes from 0 to 1? from 1 to 2?

**Discussion Question 8.3** (nonlinear wage model interpretation). Let $Y$ be wage ($$/hr$$) and $X$ years of education. Given a sample of data, you estimate $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1X + \hat{\beta}_2X^2$ with $\hat{\beta}_0 = 14.4$, $\hat{\beta}_1 = -1.6$, and $\hat{\beta}_2 = 0.1$.

- a) Does $\hat{\beta}_1 < 0$ mean that more education is associated with lower wage? Why/not?
- b) What does this estimated function suggest about the (descriptive) relationship between wage and education? (Hint: try plugging in salient values like $X = 12$ [high school] or $X = 16$ [college], or graph the whole function.)
8.2.5 Description, Prediction, and Causality

The interpretation of a nonlinear-in-variables model as causal, CEF, or linear projection is similar to linear-in-variables models. The main difference is that we may wish to clarify the word “linear” in linear projection, best linear approximation, and best linear predictor.

Description and Prediction

Consider the quadratic model from (8.7) when the true CEF is not quadratic. Then, the “linear” projection of $Y$ onto $X_0 = 1$, $X$, and $X^2$ is defined the same way as in (7.6) before:

$$\text{LP}(Y | 1, X, X^2) = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$= \arg \min_{a,b,c} d(Y, a + bX + cX^2)$$

$$= \arg \min_{a,b,c} \sqrt{E[(Y - a - bX - cX^2)^2]}.$$  \hspace{1cm} (8.11)

These linear projection coefficients are what OLS estimates. This same function of $X$ is again a “best” CEF approximation and “best” predictor of $Y$. Specifically, mirroring (7.14) and (7.15),

$$\text{LP}(Y | 1, X, X^2) = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$= \text{BLA} \left\{ \arg \min_{a,b,c} E\{ E(Y | X) - (a, b, c) \} \right\}$$

$$= \text{BLP} \left\{ \arg \min_{a,b,c} E\{ Y - (a + bX + cX^2) \} \right\}. \hspace{1cm} (8.12)$$

As before, if the true CEF actually is quadratic, the these all equal the true CEF.

Structural Identification

The ASE is identified if the CEF is properly specified and independence (Assumption A6.7) holds. Then, the ASE on $Y$ of changing $X$ from $x_1$ to $x_2$ equals the difference in the CEF at those two points:

$$\text{ASE}(x_1 \rightarrow x_2) = E(Y | X = x_2) - E(Y | X = x_1) \equiv m(x_2) - m(x_1). \hspace{1cm} (8.13)$$

If the CEF is correctly specified, then OLS can consistently estimate this CEF difference.

For example, if the true CEF is actually quadratic,

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

then regressing (with OLS) $Y$ on $1$, $X$, and $X^2$ yields consistent estimators of $\beta_0$, $\beta_1$, and $\beta_2$, under certain finite-moment and sampling
assumptions (e.g., iid sampling and finite fourth moments of \( Y \) and \( X \)). Then, \( \hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \), so a consistent ASE estimator is

\[
\hat{\text{ASE}}(x_1 \to x_2) = \hat{m}(x_2) - \hat{m}(x_1)
= \hat{\beta}_0 + \hat{\beta}_1 x_2 + \hat{\beta}_2 x_2^2 - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2)
= \hat{\beta}_1 (x_2 - x_1) + \hat{\beta}_2 (x_2^2 - x_1^2),
\]

(8.14)

Alternatively, if the true structural model is \( Y = h(X) + U \) and the structural error \( U \) satisfies \( E(U \mid X) = 0 \), then the structural function \( h(\cdot) \) is also the CEF \( m(\cdot) \). Thus, if \( h(\cdot) \) is linear-in-parameters, then it can be estimated by OLS.

### Code

Figure 8.3 is generated by the following code that fits the same data with four models: linear, quadratic, and cubic polynomials, and a trigonometric model with a sine and cosine term. Figure 8.3 shows four identical scatterplots with the four different fitted models. Note how the four fitted lines have very different qualitative features, even though they use the same data. This illustrates the same concerns about model-driven results and “self-fulfilling prophecy” as in Section 8.1.5 and Figure 8.2.

```r
par(family=’serif’, mar=c(3,3,1,1), mgp=c(2.1, /zero.alt3.8, /zero.alt3), mfrow=c(2,2))
set.seed(112358)
n <- 31
X <- sort(3*rbeta(n=n,shape1=1,shape2=1))
```
8.3 Nonparametric Regression

In nonparametric regression, the functional form of the CEF \( m(\cdot) \) is unknown. This is more general than nonlinear-in-variables regression, where \( m(\cdot) \) is nonlinear but has a known functional form, like a cubic polynomial or log-linear model, in which only the coefficient values are unknown.

In principle, this allows a very flexible model for \( m(\cdot) \), although in practice the (hopefully) optimal level of flexibility must be chosen somehow. There is no universal quantitative definition of “flexible,” but the qualitative meaning is the same as the physical flexibility of a hose or cable: can it bend around sharply in many places to take whatever shape you wish (flexible), or can it only take on particular shapes? The number of parameters (terms) in a model is a general guide to how flexible the model is. For example, a model with 20 parameters is more flexible than a model with only 2 parameters.

Many machine learning methods are nonparametric CEF estimators. In machine learning, often prediction is emphasized over
description and causality, but recall that the CEF is the best predictor of \( Y \) given \( X \) (under quadratic loss).

One view of nonparametric regression is that it is like nonlinear regression, but choosing the model with a formal statistical procedure instead of guessing. The steps are basically:

1. Choose a group of possible regression models.
2. Choose a way to evaluate models.
3. Evaluate the quality of each model, given the data.
4. Select the best (least bad) model.
5. Use the estimates from the selected model.

Steps 1–4 describe **model selection**, i.e., choosing which model to use for estimation. This is unavoidable. Sometimes model selection is informal; e.g., somebody just feels like using a quadratic model today. With nonparametric regression, usually Step 1 is done informally (but thoughtfully). For Step 2, there are many formal, statistical evaluation procedures to choose from; this choice (of procedure) is also done informally but thoughtfully. Steps 3 and 4 are done by the chosen statistical procedure using the data.

In R, usually Steps 1 and 2 require you to pick a particular R function (and certain arguments), and then the function computes Steps 3 and 4 (and Step 5) for you. Depending on the chosen model, Step 5 may be identical to Section 8.2.

Some intuitive ways to evaluate models are really bad. First, maximizing \( R^2 \) is bad. Whenever you add a term to your model, \( R^2 \) always increases, even if the model is worse (i.e., yields worse CEF estimates and predictions). Adjusted \( R^2 \) is better but still not designed for optimal model selection. Second, hypothesis testing is bad. Different significance levels yield different chosen models, and the answer to “which model is best?” never starts with “I controlled the type I error rate....”

The first difficulty in selecting a good CEF model is that \( m(\cdot) \) could be very nonlinear. Imagine \( Y = m(X) \) exactly. Even without any error term, we could get a bad estimate if we specify \( m(x) = \beta_0 + \beta_1 x \) when really \( m(\cdot) \) is not linear-in-variables. So, our model must be flexible enough to approximate the true \( m(\cdot) \) well.

The second difficulty is distinguishing \( m(X_i) \) from the CEF error \( V_i \equiv Y_i - m(X_i) \) in the data. If we knew \( Y = m(X_i) \), then we could learn \( m(x) \) perfectly for all \( x = X_i \). But in reality, we observe \( Y_i = m(X_i) + V_i \). If \( Y_i \) is big, we don’t know if \( m(X_i) \) is big or \( V_i \) is big. You can think of \( m(X_i) \) as the “signal” and \( V_i \) as the “noise”; we want to distinguish the signal from the noise. If our model is too flexible, we risk **overfitting**, mistaking noise for signal. For example, perhaps the true \( m(\cdot) \) is linear, but we estimate a very nonlinear function.

In practice, the key is balancing the two difficulties described above. If the model is too simple, it may fail to approximate the true CEF. If the model is too complex, it may lead to overfitting. The CEF estimate is bad in either case.
In more complex models, optimal model selection for prediction may not be optimal for causality. Historically, model selection has focused on prediction. Model selection for causal estimation is a cutting edge area of econometrics research.

The following code shows a particular example of nonparametric regression. Specifically, it uses something called a smoothing spline estimator, implemented in function `smooth.spline()` in R. The different estimates shown (thick red lines) correspond to different levels of flexibility of the model. The plots labeled “GCV” and “LOOCV” refer to formal model selection procedures, provided through the `smooth.spline()` function automatically. The others show intentionally bad fits: one model is “Too flexible,” the other is “Not flexible enough.” Note that the same data is used for each estimate, as seen in the scatter plots. The thin black line is the true CEF.

Figure 8.4 shows the results from the following code.

```r
par(family=’serif’, mar=c(3,3,1,1), mgp=c(2.1,0.7,3.8), mfrow=c(2,2))
set.seed(112358)
n <- 48; CEF <- function(x) { 1 + pnorm(12*(x-1/2)) }
df <- data.frame(X=sort(runif(n)))
df$Y <- CEF(df$X) + rbeta(n=n,shape1=2,shape2=2)*2-1
rets <- list()
titles <- c(‘GCV’,’LOOCV’,’Too flexible’, ’Not flexible enough’)
rets[[1]] <- smooth.spline(x=df$X, y=df$Y, cv=FALSE) #GCV
rets[[2]] <- smooth.spline(x=df$X, y=df$Y, cv=TRUE) #LOOCV
rets[[3]] <- smooth.spline(x=df$X, y=df$Y, df=n)
```

Figure 8.4: Smoothing spline estimates: same data, different amounts of flexibility.
rets[[4]] <- smooth.spline(x=df$X, y=df$Y, df=2)
xx <- seq(from=0, to=1, by=0.005)
for (ifig in 1:4) {
  plot(x=df$X, y=df$Y, type='p', pch=16, xlab='', ylab='', main='', xlim=0:1, ylim=0:1*3.04)
  lines(x=xx, y=CEF(xx), col=1)
  lines(predict(rets[[ifig]], x=xx), col=2)
  title(main=titles[ifig], line=-1, adj=0.1)
}

**Discussion Question 8.4** (model evaluation). In practice, why don’t we just make graphs like in Figure 8.4 and see which fitted function looks best? (Hint: can we make such graphs in practice? If so, how can we agree on which “looks best”? What does “best” mean?)
Empirical Exercise EE8.1. You will analyze data on law schools and their student outcomes, originally collected by Kelly Barnett for an economics term project. The idea is to compare median starting salaries of graduates from each law school with the school’s cost. Of course, these are not causal estimates: does a Harvard Law graduate make a lot of money because Harvard is expensive, or because she's very skilled (enough to get into Harvard)? Since school cost is essentially a continuous variable, you will explore possible nonlinearity in the (statistical) relationship between cost and salary.

a. Load the data (assuming you’ve already installed that R package or Stata command).

R: library(wooldridge)

Stata: bcuse lawsch85, nodesc clear

b. Stata only: make a graph with a local linear nonparametric CEF estimate (of salary given cost), a linear fit, and a quadratic fit, with command

lpoly salary cost, degree(1) n(100) addplot(lfit salary cost || qfit salary cost) where

n(100) simply specifies the number of CEF values to estimate and plot, and lfit and qfit stand for linear fit and quadratic fit, and model selection is done with a “rule-of-thumb” formula that attempts to optimally balance variance and squared bias.

c. R only: make a data frame named df with only salary and cost variables, and only when both are observed, with

df <- data.frame(Y=lawsch85$salary, X=lawsch85$cost)
df <- df[!(is.na(df$Y) | is.na(df$X)),]

where is.na() is TRUE if the entry is missing and FALSE if not.

d. R only: compute and store linear and quadratic (in variables) regressions with

reutm <- lm(Y~X, data=df) and retnl <- lm(Y~X+I(X^2), data=df)

e. R only: compute and store a nonparametric smoothing spline CEF estimate with GCV model selection with command

retss <- smooth.spline(x=df$X, y=df$Y, cv=FALSE)

f. R only: specify a sequence of X values and compute CEF estimates at each value from each of the three models (linear, quadratic, nonparametric). Store the sequence as xx with

xx <- seq(from=min(df$X), to=max(df$X), length.out=100) and then compute the estimates as

fitlm <- predict(reutm, newdata=data.frame(X=xx))
fitnl <- predict(retnl, newdata=data.frame(X=xx))
fitss <- predict(retss, newdata=data.frame(X=xx))

g. R only: make a scatterplot of raw data with

plot(x=df$X, y=df$Y, xlab='Cost', ylab='Starting Salary')
h. R only: plot the three estimated CEFs as lines over the scatter-plot with
   \begin{align*}
   &\text{lines}(x=xx, y=fitlm, \text{col}=1, \text{lty}=1) \\
   &\text{lines}(x=xx, y=fitnl, \text{col}=2, \text{lty}=5) \\
   &\text{lines}(fitss, \text{col}=4, \text{lty}=3)
   \end{align*}

i. Repeat your analysis but with the school’s rank (variable rank) instead of cost.

j. Repeat again but with log salary and log rank. Log salary is already in the dataset as variable lsalary (that’s a lowercase L before salary).

R: \text{df} \leftarrow \text{data.frame}(Y=lsalary, X=\text{log(lawsch85$rank}))

Stata: \text{generate lrank = log(rank)} then use lrank and lsalary

**Empirical Exercise EE8.2.** You will analyze data on sleep and wages, originally from Biddle and Hamermesh (1990). Specifically, you’ll estimate the CEF of daily hours of sleep conditional on hourly wage. For now, just drop missing values without worry, and focus on the linear, quadratic, and nonparametric estimation.

a. Load the data (assuming you’ve already installed that R package or Stata command).

R: \text{library(wooldridge)}

Stata: \text{bcuse sleep75 , nodesc clear}

b. R only: follow the same steps (identical code) as EE8.1 after setting up the data frame named df. Specifically, replace part (c) with
   \begin{align*}
   &\text{df} \leftarrow \text{data.frame}(Y=\text{sleep75$slpnaps/7/6/zero.alt3}, X=\text{sleep75$hrwage}) \\
   &\text{df} \leftarrow \text{df}[!(\text{is.na(df$Y) | is.na(df$X)) }, ]
   \end{align*}
   and then use the same code for all subsequent steps

c. Stata only: generate a new variable that translates the total weekly minutes of sleep into average daily hours of sleep with \text{generate sleephrsdaily = slpnaps/7/60}

d. Stata only: graph linear, quadratic, and nonparametric (local linear) CEF estimates similar to EE8.1(b), with command \text{lpoly sleephrsdaily hrwage , degree(1) n (100) addplot(lfit sleephrsdaily hrwage || qfit sleephrsdaily hrwage )}

e. Repeat your analysis, but instead of hrwage use totwrk as the conditioning variable (regressor); this is total minutes of work per week. (You could also adjust it to be average daily hours of work, to make it more comparable to the sleep variable you use.)
Empirical Exercise EE8.3. You will analyze data from the 1994–1995 men’s college basketball season scores and Las Vegas betting “spreads,” originally collected by Scott Resnick. Before each game, people can bet on whether the score difference will be “over” or “under” the spread set by bookmakers in Las Vegas. (In the data, the “difference” is the favored team’s score minus the other team’s score; so the variable spread is always positive, but the actual score difference scrdiff can be negative if the favored team loses.) Basically, the bookmaker adjusts the spread so that half the bets are “over” and half “under,” so regardless of the actual score outcome, half win and half lose (and the bookmaker always profits): the losers pay the winners, and the bookmaker keeps the transaction fees. (It’s a little complicated since bets can be placed at different times, and the spread can change over time, but we can imagine a simplified version where everyone bets at once and the spread is set so that half bet “over” and half “under.”) See the Wikipedia entry at https://en.wikipedia.org/wiki/Spread_betting for more on spread betting. Consequently, the spread does not reflect the bookmaker’s belief, but rather the aggregate beliefs of everybody betting on the game. The accuracy of such aggregate wisdom has spurred the creation of “prediction markets” for events beyond sports, like presidential elections, although there have been notable failures (e.g., 2016 U.S. presidential election).¹ You will check whether the Las Vegas spread is indeed a good predictor of the actual score difference.

Technically, the above arguments suggest that given the spread, the median score difference should equal the spread, not the mean. But, such an investigation would require “median regression” (a type of “quantile regression”), which is beyond our scope. Instead, you will investigate whether the spread is still a good predictor of the actual score difference with quadratic loss. Specifically, you can check if the OLS fit has intercept close to 0 and slope close to 1 (and whether those values are in the respective confidence intervals).

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

```
library(wooldridge); library(sandwich); library(lmtest)
?pntsprd
```

b. Stata only: load the data with `bcuse pntsprd , nodesc`
clear (assuming `bcuse` already installed)

c. For each observation (each game), compute whether the actual score difference was over, under, or equal to the spread. In math and in the code below, the “sign” function (not to be confused with “sine”) equals +1 for strictly positive values, −1 for strictly negative values, and 0 for zero

R: `overunder <- sign(pntsprd$scrdiff-pntsprd$spread)`

Stata: `generate overunder = sign(scrdiff - spread)`

¹See https://en.wikipedia.org/wiki/Prediction_market for more.
d. Display the frequency of over, under, and equal.
   R: `table(overunder, useNA='ifany')`
   Stata: `tabulate overunder, missing`

e. Regress the score difference on the spread.
   R: `ret <- lm(scrdiff~spread, data=pntsprd)`
   Stata: `regress scrdiff spread, vce(robust)`

f. R only (since already reported by Stata): display the point estimates and heteroskedasticity-robust 95% confidence intervals for the intercept and slope with
   ```r
   cbind(coefTest(ret, vcov.=vcovHC(ret, type='HC1')
        [,1:2],
        coefCI(ret, vcov.=vcovHC(ret, type='HC1')
   )
   ```

  g. Plot nonparametric CEF fitted values against the line $Y = X$ (intercept zero, slope one).
   R: `plot(smooth.spline(x=pntsprd$spread, y=pntsprd$scrdiff))` then `abline(a=0, b=1, col=2)`
   Stata: `lpoly scrdiff spread, degree(1) addplot(function y=x, range(spread)) noscatter`

h. Repeat your analysis in parts (e)–(g) but with the reverse regression: regress the spread on the score difference. (Is the slope still close to 1? Are you surprised? Consider games with the biggest possible score difference; should the spread be even bigger half the time?)
Chapter 9

Regression with Two Binary Regressors

⇒ Kaplan video: Chapter Introduction

Depends on: Chapters 6 and 7 (which depend on Chapters 2–4)

Unit learning objectives for this chapter

9.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]

9.2. Assess whether there is bias from omitting a variable in a real-world example, including the direction of bias [TLOs 5 and 6]

9.3. Interpret (appropriately) the coefficients of a regression with two binary variables, mathematically and intuitively, for description, prediction, and causality [TLO 3]

9.4. Assess whether comparing changes in two groups over time can be interpreted causally, and interpret such differences appropriately [TLOs 2, 3, and 6]

9.5. In R (or Stata): estimate regression models with two binary variables, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

Optional resources for this chapter

- ATT (Masten video)
- Potential outcomes and CATE (Masten video)
- OVB/confounders (Masten video)
- CIA/unconfoundedness (Masten video)
- ATE/CIA example (Masten video)
- Difference-in-differences (Masten video)
• Parallel trends (Masten video)
• Diff-in-diff example: immigration and unemployment (Masten videos)
• Parallel trends example: immigration and unemployment (Masten videos)
• Diff-in-diff example: minimum wage (Masten video)
• Diff-in-diff example: posting calorie counts (Masten video)
• OVB example: test score and class size (Lambert video)
• OVB example: wages and education (Lambert video)
• Sections 3.3 ("Ceteris Paribus Interpretation and Omitted Variable Bias") and 6.1.5 ("Interaction Terms") in Heiss (2016)
• Section 13.2 ("Difference-in-Differences") in Heiss (2016)
• Collider bias examples: https://doi.org/10.1093/ije/dyp334
• Collider bias review (very detailed): https://doi.org/10.1146/annurev-soc-071913-043455
• Sections 6.1 ("Omitted Variable Bias") and 8.3 ("Interactions Between Independent Variables") in Hanck et al. (2018)

Perhaps surprisingly, there is a lot to think about with even just two binary regressors. Topics include (mis)specification of a CEF model, interaction between regressors as a type of nonlinearity, interpretation of regression coefficients, causality, estimation, and more.

9.1 Omitted Variable Bias

⇒ Kaplan video: Omitted Variable Bias

For causality, omitted variable bias (OVB) is a common problem in economics. More broadly, it is a common problem in any field that uses observational (non-experimental) data and has many variables interact in complex ways. Generally, OVB arises because a variable outside our model is moving with $X$ and causing $Y$ to change, but our model assumes these changes are entirely from $X$.

9.1.1 An Allegory

Imagine a ghost ($Q$) that often accompanies a child ($X$), i.e., the ghost and child are often in the same place at the same time. The ghost always makes a huge mess ($Y$): spilling flour, knocking over chairs, drawing on walls, etc. The child’s parents only observe the child and the mess; they do not observe the ghost. The parents note that when the child is in the kitchen, then there is often a mess in the kitchen, and when the child is in the bathroom, then there is often a mess in.
the bathroom, etc. Thus, they infer that the child \((X)\) causes the mess \((Y)\). However, we know that it only appears that way because

HOST.1 the ghost \((Q)\) often accompanies the child \((X)\) and

HOST.2 the ghost \((Q)\) causes a mess \((Y)\).

The child is the regressor. The ghost is the omitted variable. The parents are economists who over-estimate how much mess the child causes. This phenomenon is OVB.

### 9.1.2 Formal Conditions

The ghost of OVB can be formalized as follows. Consider the structural model

\[
Y = \beta_0 + \beta_1 X + \beta_2 Q + V, \tag{9.1}
\]

where \(\text{Cov}(X, V) = 0\). If we don’t observe \(Q\), then instead we have the structural model

\[
Y = \beta_0 + \beta_1 X + U, \quad U \equiv \beta_2 Q + V. \tag{9.2}
\]

Here, \(X\) is sometimes called the included regressor (included in the model; not omitted). If \(X\) is binary, then for OLS to estimate \(\beta_1\) requires \(E(U | X = 0) = E(U | X = 1)\): the average effect of the structural error term \(U\) must be the same for both \(X\) groups. For simplicity, imagine \(Q\) is also binary.

Condition GHOST.1, “the ghost follows the child,” means that we usually see \(Q = 1\) when \(X = 1\), and \(Q = 0\) when \(X = 0\). More generally, it means \(Q\) is correlated with \(X\). This correlation does not need to have a causal interpretation. It does not matter why the ghost follows the child: maybe the ghost likes the child’s company (or vice-versa), or maybe they just get hungry at the same time. It only matters that they tend to be in the same place: \(Q\) and \(X\) tend to have the same value. OVB can also occur if there is a negative correlation, e.g., if usually \(Q = 1\) when \(X = 0\), and \(Q = 0\) when \(X = 1\).

Condition GHOST.2, “the ghost causes a mess,” means that \(Q\) is a causal determinant of \(Y\). In (9.1), this means \(\beta_2 \neq 0\). Although in the example \(\beta_2 > 0\) (more mess), OVB can occur with \(\beta_2 < 0\), too. For example, maybe the child is really messy, but the ghost cleans everything up; then the parents would incorrectly think the child is not messy.

To summarize: for variable \(Q\) that is not included as a regressor (it is omitted from the model), it will cause OVB if both of the following conditions hold.

**OVB.1** \(\text{Corr}(Q, X) \neq 0\): the omitted variable is correlated with the included regressor.

**OVB.2** The omitted variable \(Q\) is a causal determinant of \(Y\) (not only through \(X\)).

The variable \(Q\) may be called an omitted variable or a confounder.
Assessing OVB Conditions Empirically

If \( Q \) is observed in the data, then you can compare \( \hat{\beta}_1 \) (the estimated coefficient on \( X \)) when \( Q \) is included as a regressor to \( \hat{\beta}_1 \) when \( Q \) is omitted. If the estimates are meaningfully different (economically), then it may be best to include \( Q \) to reduce OVB. However, there are other types of variables that would also lead to a different \( \hat{\beta}_1 \) but are actually worse to include, so careful thought is required; see Section 9.6.

If \( Q \) is not observed in the data, then even \( \text{Corr}(Q, X) \) in OVB.1 cannot be assessed empirically (i.e., using data).

Beware of “omitted variable” tests that are not concerned with this type of OVB. For example, Stata’s `ovtest` implements the Ramsey test (RESET). Although the `ov` in `ovtest` indeed stands for “omitted variables,” the Ramsey test only looks for (certain types of) nonlinearity, to see whether a polynomial model might be better than a linear model. That is, it is about nonlinearity in \( X \) (Section 8.2), not about a separate \( Q \) variable. Besides, as you learned in Section 8.3, hypothesis testing is a bad way to do model selection.

Example

For example, imagine we want to learn the effect of kindergarten classroom size on earnings as an adult. (This is inspired by Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011), who actually have randomized experimental data to answer this question.)

Let \( Y \) denote the annual earnings of the individual at age 30. Let \( X = 1 \) if (as a child) the individual was in a kindergarten classroom with more than 24 students and \( X = 0 \) otherwise. Imagine \( X \) is not randomized. We are curious whether we can just regress \( Y \) on \( X \), or if there is OVB. Consider the following possible omitted variables.

First, consider \( Q \) to be somebody’s first grade class size. (First grade is the year after kindergarten in the U.S.) As with \( X \), \( Q = 1 \) if it is above 24 students and \( Q = 0 \) otherwise. Since it seems like kindergarten class size has an effect on adult earnings (\( Y \)) according to Chetty et al. (2011), probably first grade class size does, too, satisfying OVB.2. If all students in the population are completely randomly assigned to classes each year, \( \text{Corr}(X, Q) = 0 \); then, OVB.1 does not hold, so this \( Q \) would not cause OVB. However, students tend to stay in the same school, and some schools tend to have smaller class sizes than others, so OVB.1 probably does hold. Since both OVB.1 and OVB.2 are true, there is OVB.

Second, consider \( Q \) as the number of cubbies (places to put clothes, backpacks, etc.) in somebody’s kindergarten classroom. Presumably larger classes (\( X = 1 \)) require more cubbies since there are more students, so \( \text{Corr}(Q, X) > 0 \), satisfying OVB.1. However, I’d guess the number of cubbies does not have a causal effect on future earnings \( Y \). That is, if we simply went into classrooms and added a few cubbies (without adding students), I don’t think it would affect students’ future earnings. Thus, OVB.2 does not hold, and this \( Q \) does not
cause OVB.

Third, consider \( Q = 1 \) if the kindergarten is in a high-income area and \( Q = 0 \) otherwise. Areas with higher income are more likely to be able to afford more teachers to keep class sizes small. That is, it’s more likely to see \( Q = 1 \) and \( X = 0 \), or \( Q = 0 \) and \( X = 1 \), so \( \text{Corr}(Q, X) < 0 \), satisfying OVB.1. Also, Chetty, Hendren, and Katz (2016) provide evidence that growing up in a higher-income area has a positive causal effect on earnings as an adult (not only because of smaller kindergarten classes), meaning \( Q \) is a causal determinant of \( Y \), satisfying OVB.2. Thus, omitting this \( Q \) causes OVB.

**In Sum: Possible Omitted Variables (Q) in Kindergarten Example**

First grade class size: affects earnings (OVB.2), and probably correlated with kindergarten class size (OVB.1) if population includes multiple schools \( \implies \) OVB

Cubbies: more if more students (OVB.1), but no causal effect on earnings (no OVB.2) \( \implies \) no OVB

Neighborhood income: smaller classes if higher income (OVB.1), and affects earnings (OVB.2) \( \implies \) OVB

**Discussion Question 9.1** (assessing OVB). Among public elementary schools (students mostly 5–11 years old) in California, let \( Y \) be the average standardized math test score among a school’s 5th-graders, and let \( X \) be the school’s student-teacher ratio for 5th-graders (like average number of students per class). Consider a simple regression of \( Y \) on \( X \). For any two of the following variables, assess each OVB condition separately, and then decide whether you think it’s a source of OVB.

a) School’s parking lot area per student. (Remember 5–11-year-olds don’t have cars to park.)
b) Time of day of the test.
c) School’s total spending per student (including books, facilities, etc.).
d) Percentage of English learners (non-native speakers) among a school’s 5th-grade students.

**9.1.3 Consequences**

The practical problem of OVB is that we systematically over-estimate or under-estimate the true structural parameter. This consequence is quantified below.

**Formulas**

The following results are much more general than OVB with binary regressors. Beyond OVB, they quantify the consequences of any source of endogeneity that causes correlation between the regressor \( X \) and structural error term \( U \). Other sources of endogeneity are discussed
Given structural model \( Y = \beta_0 + \beta_1 X + U \), the OLS estimator of \( \beta_1 \) has the property
\[
\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)},
\] (9.3)
or equivalently (in different notation)
\[
\text{plim}_{n \to \infty} \hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)}.
\] (9.4)
That is, for large samples (large \( n \)), the estimator \( \hat{\beta}_1 \) is close to the right-hand side expression in most randomly sampled datasets. (To review \( \xrightarrow{p} \) and consistency, see Section 3.7.3.)

Equations (9.3) and (9.4) show OVB is not solved by having lots of data. Unless \( \text{Corr}(X, U) = 0 \), the OLS estimator is not consistent for the structural \( \beta_1 \).

Rearranging (9.4), the asymptotic bias (as in (3.32)) is
\[
\text{plim}_{n \to \infty} \hat{\beta}_1 - \beta_1 = \frac{\text{Cov}(X, U)}{\text{Var}(X)} = \text{slope coefficient in } \text{LP}(U \mid 1, X).
\] (9.5)
The characterization as a linear projection slope coefficient comes from replacing \( Y \) with \( U \) in (7.8). This can’t be computed from data since \( U \) is unobserved, but it is helpful for thinking about the direction and magnitude of asymptotic bias.

Although technically this is “asymptotic bias” rather than “bias” (Section 3.7.1), the practical implication is the same. Although very different mathematically, we won’t worry about such technicalities.

**Direction of Asymptotic Bias**

(Recall terms and definitions from Sections 3.7.1 and 3.7.3.)

The direction (+ or −) of the asymptotic bias in (9.5) depends on the sign (+ or −) of the slope in \( \text{LP}(U \mid 1, X) \). The sign of this slope is equivalent to the sign of \( \text{Corr}(X, U) \).

If \( \text{Corr}(X, U) > 0 \), then \( \text{plim}_{n \to \infty} \hat{\beta}_1 - \beta_1 > 0 \). This is positive (upward) asymptotic bias, meaning we systematically estimate a value “above” the true \( \beta_1 \). “Above” does not mean “bigger in magnitude”: it could be that \( \beta_1 = -9 \) and positive asymptotic bias causes \( \text{plim}_{n \to \infty} \hat{\beta}_1 = 0 \). This is “positive” since \( 0 - (-9) > 0 \) (positive), but we might also say that we’re estimating a “smaller” effect (in fact zero effect) in the sense that \( |0| < |-9| \). This can be confusing.

If \( \text{Corr}(X, U) < 0 \), then \( \text{plim}_{n \to \infty} \hat{\beta}_1 - \beta_1 < 0 \), meaning negative (downward) asymptotic bias. Again confusing, negative asymptotic bias can actually make effects look bigger, e.g., if \( \beta_1 = 0 \) and \( \text{plim}_{n \to \infty} \hat{\beta}_1 = -9 \): the true effect is zero, but the negative asymptotic bias makes it appear like there is an effect.
9.1. OMITTED VARIABLE BIAS

Results in Terms of \( Q \)

For OVB specifically, the general results in terms of \( U \) can be translated to \( Q \). As in (9.2), let \( U = \beta_2 Q + V \), with \( \text{Cov}(X, V) = 0 \). Then, using a linearity property of covariance,

\[
\text{Cov}(X, U) = \text{Cov}(X, \beta_2 Q + V) = \beta_2 \text{Cov}(X, Q) + \text{Cov}(X, V).
\]

Plugging this into the first expression in (9.3),

\[
\hat{\beta}_1 \overset{p}{\to} \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)} = \beta_1 + \beta_2 \frac{\text{Cov}(X, Q)}{\text{Var}(X)} = \beta_1 + \beta_2 \text{Corr}(X, Q) \sqrt{\frac{\text{Var}(Q)}{\text{Var}(X)}}.
\]

Equation (9.7) shows why both Conditions OVB.1 and OVB.2 are required for OVB. Condition OVB.1 says \( \text{Corr}(X, Q) \neq 0 \), while OVB.2 says \( \beta_2 \neq 0 \). If either \( \beta_2 = 0 \) or \( \text{Corr}(X, Q) = 0 \) in (9.7), then \( \beta_2 \text{Corr}(X, Q) = 0 \), and the asymptotic bias disappears: \( \hat{\beta}_1 \overset{p}{\to} \beta_1 \).

The direction of asymptotic bias can also be interpreted in terms of \( Q \). Using (9.7), the sign of the asymptotic bias is the sign of \( \beta_2 \text{Corr}(X, Q) \). That is, if \( \beta_2 \text{Corr}(X, Q) > 0 \), then there is positive (upward) asymptotic bias; if \( \beta_2 \text{Corr}(X, Q) < 0 \), then there is negative (downward) asymptotic bias.

Example

Consider the asymptotic bias direction in the example where \( X = 1 \) if the kindergarten class size is large and \( Q = 1 \) if the neighborhood income is high. Earlier, we thought probably \( \text{Corr}(X, Q) < 0 \) and \( \beta_2 > 0 \). Thus, there is negative OVB since \( \beta_2 \text{Corr}(X, Q) < 0 \). That is, if the true effect of class size on earnings is \( \beta_1 \), then we systematically estimate something below \( \beta_1 \).

Does this make the effect size (absolute value) appear bigger or smaller? Since smaller classes are better, average earnings (\( Y \)) are higher when \( X = 0 \) than when \( X = 1 \). This means a negative slope: \( \beta_1 < 0 \). That is, the effect of changing from a smaller class (\( X = 0 \)) to a larger class (\( X = 1 \)) is lower future earnings (\( \beta_1 < 0 \)). Negative asymptotic bias means we estimate something even more negative: \( \lim_{n \to \infty} \hat{\beta}_1 < \beta_1 < 0 \). This makes the size of the effect appear larger than it really is: we estimate something farther away from zero.

Intuitively, this OVB direction makes sense. Individuals who had a small kindergarten class tend to have grown up in wealthier areas with lots of other advantages that also cause higher earnings. If we ascribe the entire mean earnings difference to kindergarten, then it falsely appears that kindergarten alone cause the big difference, when
in reality many different forces were all working together in the same direction.

### In Sum: OVB Assessment

1. Think of a specific variable \( Q \)
2. Assess OVB.1: correlated with \( X \)?
3. Assess OVB.2: causal effect on \( Y \)? (separate from \( X \) effect)
4. If both OVB.1 and OVB.2 \( \implies \) OVB
5. OVB direction: positive bias if \( \text{Corr}(X, Q) \) and effect of \( Q \) on \( Y \) are either both + or both −; otherwise, negative bias
6. OVB magnitude: all else equal, larger (in absolute value) if i) larger effect of \( Q \) on \( Y \), ii) larger \( \text{Corr}(X, Q) \), iii) larger \( \text{Var}(Q)/\text{Var}(X) \).

### Practice 9.1 (OVB: kindergarten)

Consider the OVB example with earnings as an adult (\( Y \)), kindergarten classroom size (\( X \)), and childhood neighborhood income (\( Q \)). But, reverse the definition of \( X \): let \( X = 1 \) for smaller classrooms (24 or fewer students) and \( X = 0 \) for larger classrooms. Say whether you think each of the following is positive or negative, and explain why: a) \( \beta_1 \); b) \( \text{Corr}(X, Q) \); c) \( \beta_2 \); and d) OVB. Also discuss: e) will our estimated effect \( \hat{\beta}_1 \) tend to be larger or smaller than the true effect \( \beta_1 \), and why?

### Discussion Question 9.2 (OVB: ES habits)

Recall from DQ 6.3 the example with \( Y \) as a student’s final semester score \((0 \leq Y \leq 100)\) and \( X = 1 \) if a student starts the exercise sets well ahead of the deadline (and \( X = 0 \) otherwise).

a) What’s one variable that might cause OVB? Explain why you think both OVB conditions are satisfied.

b) Which direction of asymptotic bias would your omitted variable cause? Explain.

### 9.1.4 OVB in Linear Projection

For linear projection (without causal interpretation), the OVB formula is actually the same as (9.7), just with \( \beta_1 \) and \( \beta_2 \) interpreted as linear projection coefficients rather than structural coefficients. Similar results for larger linear projection models are in Hansen (2020, §2.24), for example.

However, if we are interested in prediction, we don’t care whether our \( \hat{\beta}_1 \) estimates a particular linear projection coefficient; we only care whether we can predict \( Y \) well. Of course, we don’t want to omit \( Q \) if it’s helpful for prediction, but we don’t care about OVB itself. That is, OVB is only a problem for causality, not prediction.
9.2 Linear-in-Variables Model

The simplest CEF model with two binary variables is linear-in-variables (Section 8.2.1),

\[ E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \]  

(9.8)

Misspecification

Unfortunately, (9.8) may be misspecified. Recall from Section 7.1 that misspecification arose when \( X \) had three values but the CEF model \( \beta_0 + \beta_1 X \) had only two parameters. The case here is similar: (9.8) has only 3 parameters, but there are 4 possible values of \((X_1, X_2)\). Specifically, \((X_1, X_2)\) could equal \((0, 0)\), \((0, 1)\), \((1, 0)\), or \((1, 1)\). Consequently, there are four CEF values:

\[
\begin{align*}
m(0, 0) &= E(Y \mid X_1 = 0, X_2 = 0), \quad m(0, 1) = E(Y \mid X_1 = 0, X_2 = 1), \\
m(1, 0) &= E(Y \mid X_1 = 1, X_2 = 0), \quad m(1, 1) = E(Y \mid X_1 = 1, X_2 = 1).
\end{align*}
\]  

(9.9)

To see the possible misspecification, we can write the \( \beta_j \) regression coefficients in terms of the CEF values \( m(x_1, x_2) \). If (9.8) were true, then

\[
\begin{align*}
m(0, 0) &= \beta_0 + (\beta_1)(0) + (\beta_2)(0) = \beta_0, \quad (9.10) \\
m(0, 1) &= \beta_0 + (\beta_1)(0) + (\beta_2)(1) = \beta_0 + \beta_2, \quad (9.11) \\
m(1, 0) &= \beta_0 + (\beta_1)(1) + (\beta_2)(0) = \beta_0 + \beta_1, \quad (9.12) \\
m(1, 1) &= \beta_0 + (\beta_1)(1) + (\beta_2)(1) = \beta_0 + \beta_1 + \beta_2. \quad (9.13)
\end{align*}
\]

Consequently, \( \beta_1 \) has two interpretations. It equals either (9.13) minus (9.11), or (9.12) minus (9.10):

\[
\begin{align*}
m(1, 1) - m(0, 1) &= (\beta_0 + \beta_1 + \beta_2) - (\beta_0 + \beta_2) = \beta_1, \\
m(1, 0) - m(0, 0) &= (\beta_0 + \beta_1) - \beta_0 = \beta_1.
\end{align*}
\]

Thus, the model implicitly assumes \( m(1, 1) - m(0, 1) = m(1, 0) - m(0, 0) \), which may not be true of the real CEF. For example,

\[
m(0, 0) = 0, m(1, 0) = 1, m(0, 1) = 2, m(1, 1) = 4 \implies m(1, 1) - m(0, 1) \\

\]

Because \( m(1, 1) - m(0, 1) \neq m(1, 0) - m(0, 0) \), the CEF model in (9.8) is misspecified (wrong). That is, there are no possible \((\beta_0, \beta_1, \beta_2)\) such that \( m(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \).

As discussed in Chapter 7, if the CEF model is wrong, then OLS estimates the linear projection. Here, OLS estimates \( \text{LP}(Y \mid 1, X_1, X_2) \). However, this is not useful for causality, and the misspecification is easily fixed.
Before we fix the misspecification, consider more carefully why (9.8) is usually misspecified. To be concrete, imagine $Y$ is wage, $X_1 = 1$ if an individual has a college degree (and $X_1 = 0$ if not), and $X_2 = 1$ if an individual has at least 10 years of work experience (and $X_2 = 0$ if not). For simplicity, we’ll call $X_1$ “education” and $X_2$ “experience.” The quantity $m(1,1) - m(0,1)$ compares the mean wage in the high-education, high-experience group (subpopulation) with the mean wage in the low-education, high-experience group. That is, within the high-experience subpopulation, it compares the mean wage of the high-education and low-education sub-sub-populations. The quantity $m(1,0) - m(0,0)$ also compares mean wages across high and low education, but within the low-experience subpopulation. Thus, assuming $m(1,1) - m(0,1) = m(1,0) - m(0,0)$ can be interpreted as assuming that the mean wage difference between high-education and low-education groups is identical within the high-experience subpopulation and within the low-experience subpopulation. This is a strong assumption that is probably not true in this example (or in most examples).

### 9.3 Fully Saturated Model

Misspecification is avoided by adding the interaction term $X_1 X_2$:

$$E(Y | X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$  \hspace{1cm} (9.14)

Mathematically, interaction terms often involve the product of two regressors, like $X_1 X_2$ here. Economically, the interaction term allows the mean $Y$ difference associated with $X_1$ to depend on the value of $X_2$. Similarly, it allows the mean $Y$ difference associated with $X_2$ to depend on the value of $X_1$. For example, the mean wage difference associated with education can depend on the value of experience. More generally, interaction terms allow the change in $Y$ associated with a unit increase in one regressor to depend on the value of another regressor.

The CEF model in (9.14) is also called **fully saturated** (Section 7.2.2) since it is flexible enough to allow a different CEF value for each value of $(X_1, X_2)$. Logically, having the same number (four) of possible values of $(X_1, X_2)$ as $\beta_j$ parameters is necessary but not sufficient for the model to be fully saturated.

Interpretation of the coefficients requires writing them in terms of different CEF values. First, similar to (9.10)–(9.13), each CEF value
can be written in terms of the $\beta_j$:

$$m(x_1, x_2) = \beta_0 + (\beta_1)(x_1) + (\beta_2)(x_2) + (\beta_3)(x_1)(x_2),$$

$$m(0, 0) = \beta_0 + (\beta_1)(0) + (\beta_2)(0) + (\beta_3)(0)(0) = \beta_0, \quad (9.15)$$

$$m(0, 1) = \beta_0 + (\beta_1)(0) + (\beta_2)(1) + (\beta_3)(0)(1) = \beta_0 + \beta_2, \quad (9.16)$$

$$m(1, 0) = \beta_0 + (\beta_1)(1) + (\beta_2)(0) + (\beta_3)(1)(0) = \beta_0 + \beta_1, \quad (9.17)$$

$$m(1, 1) = \beta_0 + (\beta_1)(1) + (\beta_2)(1) + (\beta_3)(1)(1) = \beta_0 + \beta_1 + \beta_2 + \beta_3. \quad (9.18)$$

From (9.15)–(9.18) and their differences,

$$\beta_0 = m(0, 0), \quad (9.19)$$

$$\beta_1 = (\beta_0 + \beta_1) - \beta_0 = m(1, 0) - m(0, 0), \quad (9.20)$$

$$\beta_2 = (\beta_0 + \beta_2) - \beta_0 = m(0, 1) - m(0, 0), \quad (9.21)$$

$$\beta_3 = [\beta_2 + \beta_3] - [\beta_2] = [(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1)] - [(\beta_0 + \beta_2) - \beta_0]. \quad (9.22)$$

Because of the difference-in-differences structure seen in (9.22) and (9.23), this model is sometimes called a difference-in-differences model, particularly when $X_2$ represents time and $X_1$ represents a “treatment” (see Section 9.7).

Using (9.19)–(9.23), the four $\beta_j$ in (9.14) have the following interpretations, both in terms of the wage example ($Y$ wage, $X_1$ education, $X_2$ experience) and more generally.

- $\beta_0 = m(0, 0)$ is the mean wage among low-education, low-experience individuals.

More generally, $\beta_0$ is the mean $Y$ in the subpopulation with $X_1 = 0$ and $X_2 = 0$.

Caution: generally $\beta_0 \neq E(Y)$.

- $\beta_1 = m(1, 0) - m(0, 0)$ is the mean wage difference between high-education and low-education individuals within the low-experience subpopulation.

More generally, $\beta_1$ is the mean $Y$ difference between $X_1 = 1$ and $X_1 = 0$ individuals within the $X_2 = 0$ subpopulation.

Caution: generally $\beta_1 \neq E(Y \mid X_1 = 1) - E(Y \mid X_1 = 0)$; it additionally conditions on $X_2 = 0$. 

• \( \beta_2 = m(0, 1) - m(0, 0) \) is the mean wage difference between high-experience and low-experience individuals within the low-education subpopulation.

More generally, \( \beta_2 \) is the mean \( Y \) difference between \( X_2 = 1 \) and \( X_2 = 0 \) individuals within the \( X_1 = 0 \) subpopulation.

Caution: generally \( \beta_2 \neq E(Y \mid X_2 = 1) - E(Y \mid X_2 = 0) \); it additionally conditions on \( X_1 = 0 \).

• \( \beta_3 = [m(1, 1) - m(1, 0)] - [m(0, 1) - m(0, 0)] \) is the mean wage difference associated with experience in the high-education subpopulation minus the mean wage difference associated with experience in the low-education subpopulation.

More generally, \( \beta_3 \) is the mean \( Y \) difference associated with \( X_2 \) in the \( X_1 = 1 \) subpopulation minus the mean \( Y \) difference associated with \( X_2 \) in the \( X_1 = 0 \) subpopulation.

• \( \beta_3 = [m(1, 1) - m(0, 1)] - [m(1, 0) - m(0, 0)] \) is also the mean wage difference associated with education in the high-experience subpopulation minus the mean wage difference associated with education in the low-experience subpopulation.

More generally, \( \beta_3 \) is the mean \( Y \) difference associated with \( X_1 \) in the \( X_2 = 1 \) subpopulation minus the mean \( Y \) difference associated with \( X_1 \) in the \( X_2 = 0 \) subpopulation.

The \( \beta_j \) interpretations can also be seen by considering the regression of \( Y \) on \( X_1 \) when \( X_2 = 0 \) and separately when \( X_2 = 1 \). That is, plugging in \( x_2 = 0 \) first and then \( x_2 = 1 \) second,

\[
m(x_1, 0) = \beta_0 + \beta_1 x_1 + (\beta_2)(0) + (\beta_3)(x_1)(0) = \beta_0 + \beta_1 x_1, \quad (9.24)
m(x_1, 1) = \beta_0 + \beta_1 x_1 + (\beta_2)(1) + (\beta_3)(x_1)(1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1. \quad (9.25)
\]

That is, when changing from \( X_2 = 0 \) to \( X_2 = 1 \), the intercept changes by \( \beta_2 \) and the slope changes by \( \beta_3 \). These changes could be positive or negative, or zero. The interaction coefficient \( \beta_3 \) describes how the slope with respect to \( X_1 \) differs when \( X_2 = 1 \) versus \( X_2 = 0 \).

Equivalently, we could switch all the \( X_1 \) and \( X_2 \) and interpret \( \beta_3 \) as the difference between the slope with respect to \( X_2 \) when \( X_1 = 1 \) versus when \( X_1 = 0 \):

\[
m(0, x_2) = \beta_0 + (\beta_1)(0) + \beta_2 x_2 + (\beta_3)(0)(x_2) = \beta_0 + \beta_2 x_2, \quad (9.26)
m(1, x_2) = \beta_0 + (\beta_1)(1) + \beta_2 x_2 + (\beta_3)(1)(x_2) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x_2. \quad (9.27)
\]

**Practice 9.2** (binary interaction). Let \( Y \) be wage ($/hr), \( D_1 = 1 \) if an individual has a college degree \((D_1 = 0 \text{ if not})\), and \( D_2 = 1 \) if an individual has more than 15 years of experience \((D_2 = 0 \text{ if not})\).

You have a sample of data and run OLS on the fully saturated model, yielding \( \hat{Y} = 10 + 5D_1 + D_2 + 2D_1 D_2 \).
a) For the college-educated subpopulation, what is the estimated change in mean wage associated with changing from low to high experience?
b) Within the low-experience subpopulation, what’s the estimated difference in mean wage between the college and no-college subpopulations?
c) How do you interpret the $2$ (the coefficient on $D_1D_2$)?

### 9.4 Structural Identification by Exogeneity

Imagine $Y$ is determined by the structural model

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + U. \quad (9.28)$$

The qualitative condition for identification is the same as in Section 6.6.1. Specifically, if $U$ (which contains other causal determinants of $Y$) is unrelated to the regressors, then the structural parameters are identified. Recall that a regressor unrelated to $U$ is called exogenous; otherwise, it’s endogenous.

Mathematically, one sufficient definition of “unrelated” here is “uncorrelated.” If

$$\text{Cov}(U, X_1) = \text{Cov}(U, X_2) = \text{Cov}(U, X_1X_2) = 0, \quad (9.29)$$

then $\beta_1$, $\beta_2$, and $\beta_3$ are the linear projection slope coefficients from $\text{LP}(Y \mid 1, X_1, X_2, X_1X_2)$. Other mathematical definitions of “unrelated” imply (9.29) and are thus sufficient for identification. For example, $U \perp (X_1, X_2)$ logically implies (9.29), as does mean independence $\text{E}(U \mid X_1, X_2) = \text{E}(U)$.

If the structural $\beta_1$, $\beta_2$, and $\beta_3$ are also linear projection coefficients, then they can be estimated by OLS. That is, we can interpret the OLS-estimated slope coefficients as the structural parameters in (9.28).

### 9.5 Identification by Conditional Independence

By extending the independence assumption ($A6.1$ and $A6.5$), variants of the ASE and ATE can be identified. (Note: more details and examples are in the Spring 2020 edition.)

Consider the subpopulation with $X_2 = 1$, and whether the mean difference $\text{E}(Y \mid X_1 = 1, X_2 = 1) - \text{E}(Y \mid X_1 = 0, X_2 = 1)$ has a causal interpretation. This is equivalent to redefining the population as everybody with $X_2 = 1$ and asking if the mean difference $\text{E}(Y \mid X_1 = 1) - \text{E}(Y \mid X_1 = 0)$ has a causal interpretation. This question was studied in Section 6.6, for both structural and potential outcomes models.

The key identifying assumption from Section 6.6 was independence. In the structural model, this meant independence between the regressor $X_1$ and the unobserved determinants of $Y$. In the potential
CHAPTER 9. REGRESSION WITH TWO BINARY REGRESSORS

Outcomes model, this meant independence between the treatment and the pair of potential outcomes.

Extending independence is conditional independence, which essentially assumes independence within each subpopulation ($X_2 = 1$ and $X_2 = 0$). Conditional independence has other names like unconfoundedness, selection on observables, and ignorability; see Imbens and Wooldridge (2007, p. 6) and references therein. Mathematically, both structural and potential outcomes versions of conditional independence are stated in Assumption A9.1.

Assumption A9.1 (conditional independence). Let $Y = h(X_1, X_2, U)$ be the structural model. Conditional on the control variable $X_2$, the regressor of interest $X_1$ is independent of the vector of unobserved causal determinants $U$: $U \perp \perp X_1 \mid X_2$. Alternatively, in potential outcomes notation, where $Y^T$ and $Y^C$ are the treated and untreated potential outcomes, binary treatment $X_1$ is independent of the potential outcomes, conditional on the control variable $X_2$: $(Y^0, Y^1) \perp \perp X_1 \mid X_2$. More generally, in either case, $X_2$ may be replaced by multiple control variables, $X_2, X_3, X_4, \ldots$, and $X_1$ can be discrete (including binary) or continuous in the structural model.

Consequently, the ASE or ATE within subpopulation $X_2 = 1$ is identified and equal to the conditional mean difference $E(Y \mid X_1 = 1, X_2 = 1) - E(Y \mid X_1 = 0, X_2 = 1)$. Similarly, the ASE or ATE within subpopulation $X_2 = 0$ is identified and equal to the conditional mean difference $E(Y \mid X_1 = 1, X_2 = 0) - E(Y \mid X_1 = 0, X_2 = 0)$. Because these are causal effects within a subpopulation (not full population), i.e., conditional on $X_2$, the ATE is sometimes called a conditional ATE, or the ASE a conditional ASE.

The (unconditional) ATE can be computed from the conditional ATEs. Specifically, the ATE is the mean conditional ATE. Writing the conditional ATE for $X_2 = 1$ as $\text{CATE}(1)$, and similarly $\text{CATE}(0)$ for $X_2 = 0$, then $\text{CATE}(X_2)$ is a random variable. Specifically, $P(\text{CATE}(X_2) = \text{CATE}(1)) = P(X_2 = 1)$ and $P(\text{CATE}(X_2) = \text{CATE}(0)) = P(X_2 = 0)$. Thus, $\text{CATE}(X_2)$ has a mean. Ultimately,

$$\text{ATE} = E[Y^T - Y^C] = E[E(Y^T - Y^C \mid X_2)] = E[\text{CATE}(X_2)] = P(X_2 = 1) \text{CATE}(1) + P(X_2 = 0) \text{CATE}(0).$$

Similar arguments apply to the conditional ASE.

9.6 Collider Bias

Although OVB shows the risk of omitting certain types of variables (confounders), other types of variables actually should be omitted, otherwise there is a different type of (asymptotic) bias.

A collider or common outcome is a variable on which both $X$ and $Y$ have a causal effect. (Whereas a confounder has a causal effect on both $X$ and $Y$.) For example, imagine you want to learn the effect of a firm’s ownership structure (say $X = 1$ for family-owned, $X = 0$
otherwise) on its research and development expenditure $Y$. Both $X$ and $Y$ affect the firm’s performance $Z$, so $Z$ is a collider.

Including a collider as a regressor causes **collider bias** when estimating a causal relationship. This is not as intuitive as OVB, but consider the following example.\(^1\)

Imagine you’re interested in the causal effect of eating falafel or salad on having the flu (which is zero effect), and you have a sample of 200 individuals. You randomly assigned 100 people to eat falafel for lunch, and 100 salad; a few hours later, you test each for flu (assume there is no testing error). Let $Y = 1$ if somebody has the flu (otherwise $Y = 0$), and $X = 1$ if somebody ate falafel for lunch ($X = 0$ if salad). Let $Z = 1$ if the individual has a fever (otherwise $Z = 0$). Sadly, the salad had some romaine contaminated with E. coli, so 40% of those who ate salad got a fever from the E. coli, unrelated to whether or not they had the flu. Among individuals with flu, 90% have a fever, but 10% don’t.

Table 9.1: Counts in falafel/salad/flu example.

<table>
<thead>
<tr>
<th></th>
<th>Flu</th>
<th>No flu</th>
<th>Fever</th>
<th>No fever</th>
</tr>
</thead>
<tbody>
<tr>
<td>Falafel</td>
<td>50</td>
<td>50</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>Salad</td>
<td>50</td>
<td>50</td>
<td>47</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 9.1 shows the number of individuals in different categories. Overall, there is no relationship between lunch and flu, so the flu rate is the same in the falafel and salad groups. To make the numbers easier, the overall flu rate is 50% (100/200 overall, 50/100 in each group). Since nobody who ate falafel got E. coli, the only reason for fever is the flu, which has a 90% fever rate. Thus, among the 50 with flu who at falafel, $(50)(0.9) = 45$ have a fever and 5 do not. This entirely explains the Falafel row. In the salad row, given the statistical independence of flu (probability 0.5) and E. coli (probability 0.4), the probability of having neither is

$$P(\text{not flu and not E. coli}) = P(\text{not flu})P(\text{not E. coli}) = \left[1 - P(\text{flu})\right]\left[1 - P(\text{E. coli})\right] = (0.5)(0.6) = 0.3,$$

hence $(100)(0.3) = 30$ salad-eaters who have neither flu nor E. coli, and thus no fever. This explains the No fever / No flu entry of 30 in the Salad row. Similarly,

$$P(\text{flu, not E. coli}) = (0.5)(0.6) = 0.3 \quad (30 \text{ people}),$$
$$P(\text{flu, E. coli}) = (0.5)(0.4) = 0.2 \quad (20 \text{ people}),$$
$$P(\text{not flu, E. coli}) = (0.5)(0.4) = 0.2 \quad (20 \text{ people}).$$

\(^1\)Modified from [https://doi.org/10.1093/ije/dyp334](https://doi.org/10.1093/ije/dyp334)
The “not flu and E. coli” are the 20 individuals who have a fever (from the E. coli) but not flu. The 20 with both flu and E. coli all have a fever, due to E. coli. Among the 30 with flu but not E. coli, 90% have a fever, i.e., (30)(0.9) = 27 have a fever, so 3 do not. This 3 is the No fever / Flu entry in the Salad row. The 27 combine with the 20 who had both illnesses to make 47 who have both flu and a fever in the Salad row.

If we regress \( Y \) (flu) on \( X \) (food), then we correctly estimate zero effect, but if we also use \( Z \) (fever), then we incorrectly estimate a non-zero effect. If we only look at the “no fever” group, then there is (appropriately) zero difference: the flu rate for the falafel eaters is \( 5/55 = 1/11 \), identical to the \( 3/33 = 1/11 \) for the salad eaters. Mathematically, these “rates” are estimates of the conditional mean of the binary \( Y \) flu variable; e.g., \( 5/55 = \hat{E}(Y \mid \text{falafel, no fever}) \), recalling \( E(Y) = P(Y = 1) \) for binary \( Y \). However, if we also look at the “fever” group, the flu rate is much higher in the falafel group. In fact, the falafel group’s flu rate is \( 45/45 = 100\% \), whereas the salad group’s flu rate is only \( 47/(47+20) = 70\% \), substantially lower. Mathematically, \[
\begin{align*}
\frac{5}{55} - \frac{3}{33} &= 0, \\
\frac{45}{45} - \frac{47}{67} &= 0.30.
\end{align*}
\] (9.30)
This suggests eating falafel causes flu, but this incorrect conclusion is entirely collider bias.

9.7 Causal Identification: Difference-in-Differences

If \( X_1 \) is a treatment indicator and \( X_2 \) is a time period indicator, then the fully saturated model with two binary regressors is called a difference-in-differences (diff-in-diff) model. This is a special case of (9.14), whose coefficients were interpreted in Section 9.3.

Below, the parameter \( \beta_3 \) from (9.14) is shown to have a certain causal interpretation under certain conditions.

The general setup is that some individuals (or firms, or cities, etc.) were exposed to some “treatment,” like a training program or law or other policy. The treatment wasn’t randomized, but there’s a group of untreated individuals whose outcomes can be used to form a counterfactual: what’s the mean outcome of treated individuals in the parallel universe where they weren’t treated?

Such setups are sometimes called natural experiments or quasi-experiments (see also Section 4.3.2). Since they weren’t fully randomized experiments, it’s invalid to simply compare treated and untreated outcomes, as seen in Section 9.7.1. However, there is enough randomness that a valid comparison can be found, with some additional work (like diff-in-diff).
For example, maybe $Y$ is annual labor income, and we are interested in the effect of minimum wage. Imagine our city recently implemented a large minimum wage increase. The goal is to learn the effect of this particular minimum wage increase on $Y$ (income), for individuals in our city. Notationally, $X_1 = 1$ if the individual lives in our city (and $X_1 = 0$ otherwise), and $X_2 = 1$ if the observation is from the year after the minimum wage increase (and $X_2 = 0$ if before the increase).

Notationally, $X_1 = 1$ is the “treated group” and $X_1 = 0$ the “untreated group”; $X_2 = 0$ is the time period “before” treatment and $X_2 = 1$ is “after.”

### 9.7.1 Bad Approaches

**Discussion Question 9.3** (bad panel approach #1, for Mariel boatlift).
Consider the basic setup from Card (1990). Due to a seemingly random/exogenous political decision, Cubans were temporarily permitted to immigrate to the U.S. for a few months in 1980. About half settled in Miami, FL, while the other half went to live in other cities around the U.S. We could compare wages of native-born workers in Miami in 1979 (before boatlift) and 1981 (after). Explain why this change in average wage would not be a good estimate of the average treatment effect of the Mariel boatlift on native worker wage. (Hint: are 1979 Miami and 1981 Miami the same except for how many Cubans live there, or might something else have changed?)

**Discussion Question 9.4** (bad panel approach #2, for Mariel boatlift).
Consider the same setup as in DQ 9.3. But now compare 1981 wages of native workers in Miami and Houston, TX, a city that did not receive a large influx of Cuban immigrants in 1980. Explain why this difference (Miami minus Houston) in average wage would not be a good estimate of the average treatment effect of the Mariel boatlift on native worker wage. (Hint: are 1981 Miami and Houston the same except for how many Cubans live there, or might there be other differences between the cities that might cause omitted variable bias?)

**Discussion Question 9.5** (bad panel approach #1, for fracking).
Discussion Questions 9.5 and 9.6 are based loosely on the setting of Street (2018), who uses much better approaches. For counties in North Dakota, let $Y$ denote crime rate. Consider the average crime rate in counties that started fracking activity, before and after the fracking started. (Fracking was a new technology that allowed extraction of certain underground oil and natural gas reserves that were previously infeasible or unprofitable to extract.) Explain why this change in average crime rate would not be a good estimate of the average treatment effect of the fracking activity on crime rate.

**Discussion Question 9.6** (bad panel approach #2, for fracking).
Consider the same setup as in DQ 9.5, but now compare the “after” crime rates in North Dakota counties with fracking to those without fracking. Explain why this difference (fracking minus non-fracking)
in average crime rate would not be a good estimate of the average treatment effect of fracking on crime rate.

Continuing the minimum wage example, one bad approach is to use only data from our city, before and after the minimum wage increase. That is, we could try to estimate \( E(Y \mid X_2 = 1, X_1 = 1) - E(Y \mid X_2 = 0, X_1 = 1) \). However, coincidentally, there may have been a national (or global) recession right after the minimum wage law was passed. This may make everybody’s income lower in the year after. It would look like the minimum wage hurt incomes, but really it was the recession. Alternatively, there may have been great national (macroeconomic) conditions that made incomes go up, which would make us incorrectly conclude that the law increased incomes greatly. There is almost always OVB with such before vs. after comparisons, which invalidates causal interpretation.

Another bad approach is to compare incomes in our city and another city in the year after our law passed. By using the other city as a sort of control group, we avoid the problem of misinterpreting macroeconomic changes as treatment effects. However, it’s hard to know which other city to pick. We could pick one that has the same population, for example, but our city may still have much higher (or lower) income for reasons other than our minimum wage. For example, San Francisco and Columbus, OH have very similar populations, but they have (and have for a while had) very different incomes.

9.7.2 Counterfactuals and Parallel Trends

The difference-in-differences idea is to combine the before vs. after comparison with the treated vs. untreated comparison.

Conceptually, the goal is to construct a counterfactual (link to pronunciation), like what our city’s mean income would have been if there were not a minimum wage increase. Thinking of the potential outcomes framework, the counterfactual is the parallel universe where the treatment never happened.

The key identifying assumption is called parallel trends. Conceptually, in the running example, parallel trends says that without the minimum wage law, our city’s mean income would have increased by exactly the same amount as the other city’s mean income. Mathematically, with \( m(x_1, x_2) \equiv E(Y \mid X_1 = x_1, X_2 = x_2) \), the other city’s mean income increase (i.e., “after” minus “before”) is

\[
m(0, 1) - m(0, 0) = E(Y \mid X_1 = 0, X_2 = 1) - E(Y \mid X_1 = 0, X_2 = 0). \tag{9.31}
\]

Parallel trends assumes that adding this increase to the “before” mean income in our city, \( m(1, 0) = E(Y \mid X_1 = 1, X_2 = 0) \), gives us the counterfactual income for our city in the “after” time period.
Given parallel trends, we can learn about causality by comparing

$$\text{actual (our city, after)} \quad E(Y \mid X_1 = 1, X_2 = 1) \quad \text{vs.} \quad \text{counterfactual}$$

$$E(Y \mid X_1 = 1, X_2 = 0) + E(Y \mid X_1 = 0, X_2 = 1) - E(Y \mid X_1 = 0, X_2 = 0)$$

$$\text{our city, before} \quad \text{increase in other city over time} \quad (9.32)$$

$$\begin{align*}
\text{actual} & : m(1,1) - \{m(1,0) + [m(0,1) - m(0,0)]\} = [m(1,1) - m(1,0)] - [m(0,1) - m(0,0)] \\
\text{counterfactual} & : m(1,0) + [m(0,1) - m(0,0)] \\
\beta_3 & : \text{in (9.14)}
\end{align*}$$

Figure 9.1 visualizes this effect. We can think of constructing the counterfactual outcome, and then subtracting it from the actual outcome $m(1,1)$, or we can think of taking the before/after difference for our city, $m(1,1) - m(1,0)$, and subtracting off the before/after difference in the other city, $m(0,1) - m(0,0)$.

**Discussion Question 9.7** (parallel trends skepticism). Consider U.S. state traffic fatality (i.e., car accident death) rates ($Y$), where the year 1980 is “before” ($X_2 = 0$) and 1990 is “after” ($X_2 = 1$). Consider states that adopt a 0.08 blood alcohol content (BAC) limit law sometime between 1980 and 1990 ($X_1 = 1$) and states that never have such a law ($X_1 = 0$). Explain why you might doubt the parallel trends assumption. Hint #1: is a BAC law the only way states try to reduce fatal accidents? Hint #2: this is more difficult than simply thinking of an omitted variable that would cause OVB in a cross-sectional regression, because parallel trends allows certain types of such omitted variables.

**9.7.3 Identification**

**Population Object of Interest: ATT**

Most fundamentally, the difference-in-differences approach only learns the average treatment effect for the group that was actually treated
This is called the **average treatment effect on the treated** (ATT) (or sometimes ATTE or ATET). Mathematically, ATE meant \(E(Y^1 - Y^0),\) where \(Y^1\) and \(Y^0\) are the treated and untreated potential outcomes, respectively (previously \(Y^T\) and \(Y^C\)). ATT is the same, but for the subpopulation who was actually treated in our universe. Since \(X_1 = 1\) if somebody is actually treated, the ATT is

\[
\text{ATT} \equiv E(Y^1 - Y^0 \mid X_1 = 1).
\]  

(9.33)

It’s possible but uncommon that \(\text{ATT} = \text{ATE}\). For example, maybe there are different demographics in our city than the comparison city, or different levels of unionization, or different other labor laws, or different industry mix, so the minimum wage effect is different in our city \((X_1 = 1)\) than elsewhere. (This is essentially a question of external validity; see Chapter 12.)

**Identification of ATT**

Parallel trends is sufficient to identify the counterfactual. In potential outcomes notation, “parallel trends” is

\[
E(Y^0 \mid X_1 = 1, X_2 = 1) - E(Y^0 \mid X_1 = 1, X_2 = 0)
\]

\[
= E(Y^0 \mid X_1 = 0, X_2 = 1) - E(Y^0 \mid X_1 = 0, X_2 = 0).
\]  

(9.34)

That is, the mean untreated potential outcome changes over time \((X_2 = 0\) to \(X_2 = 1)\) by the same amount in the treated \((X_1 = 1)\) and untreated \((X_1 = 0)\) groups. The term \(E(Y^0 \mid X_1 = 1, X_2 = 1)\) is the counterfactual, like our city’s mean wage in the “after” period in the parallel universe where minimum wage never increased. In the other three terms, \(Y^0 = Y,\) i.e., the untreated \(Y^0\) is the observed \(Y.\) Only when \(X_1 = X_2 = 1\) is the treated \(Y^1\) observed, \(Y = Y^1.\) Thus, the counterfactual can be written uniquely in terms of the joint distribution of \((Y, X_1, X_2):\)

\[
E(Y^0 \mid X_1 = 1, X_2 = 1)
\]

\[
= E(Y^0 \mid X_1 = 1, X_2 = 0) + [E(Y^0 \mid X_1 = 0, X_2 = 1) - E(Y^0 \mid X_1 = 0, X_2 = 0)]
\]

\[
= E(Y \mid X_1 = 1, X_2 = 0) + [E(Y \mid X_1 = 0, X_2 = 1) - E(Y \mid X_1 = 0, X_2 = 0)]
\]

\[
= m(1,0) + [m(0,1) - m(0,0)].
\]  

(9.35)

Because the counterfactual is identified, so is the ATT. Specifically, the ATT equals \(\beta_3\) in the fully saturated CEF model (9.14),

\[
\text{ATT} = E(Y^1 - Y^0 \mid X_1 = 1, X_2 = 1)
\]

\[
\stackrel{Y^1=Y \text{ since } X_1=1, X_2=1}{=} E(Y^1 \mid X_1 = 1, X_2 = 1) - E(Y^0 \mid X_1 = 1, X_2 = 1)
\]

\[
\stackrel{\text{counterfactual}}{=} E(Y \mid X_1 = 1, X_2 = 1) - \{m(1,0) + [m(0,1) - m(0,0)]\}
\]

\[
= m(1,1) - \{m(1,0) + [m(0,1) - m(0,0)]\}
\]

\[
= [m(1,1) - m(1,0)] - [m(0,1) - m(0,0)]
\]

\[
= \beta_3.
\]
9.8. ESTIMATION AND INference

Skepticism About Parallel Trends

In practice, the parallel trends condition may not hold for various reasons. For example, maybe our city was experiencing fast wage growth, whereas the comparison city was declining (maybe due to reliance on different industries). Maybe our city passed the minimum wage law partly because everybody’s wages were increasing anyway. In that case, we can’t tell whether our city’s wages grew more than the other city’s wages because of the minimum wage, or because of other factors (our industries were growing, theirs were declining, etc.).

Parallel trends is also a bit fragile since nonlinear functions of $Y$ change whether it’s true or not. For example, if there are parallel trends when $Y$ is wage, then there are not parallel trends for log-wage $\ln(Y)$. Similarly, if there are parallel log-wage trends, then the wage trends cannot be parallel.

In the data, you can try to see if parallel trends seems plausible, but it is not directly testable. Specifically, “pre-trend analysis” compares trends for a few periods before the treatment takes place. But even if the trends were parallel before, it does not mean for sure that the trends would have remained parallel after the treatment year. We can never know because the “trend” refers to the treated group’s untreated potential outcomes, which by definition are not observed. So, there is no empirical test that can replace careful critical thought.

9.7.4 Extensions

There are many interesting extensions of the basic diff-in-diff idea, although all are beyond our scope. For example, there are related models that allow additional regressors, or more time periods, or quantile treatment effects.

9.8 Estimation and Inference

Since (9.14) is just a special case of a regression model, standard regression techniques and R functions can be used. For estimation, OLS consistently estimates each $\beta_j$ under fairly general conditions; remember to use sample/survey weights if they are available in the data. The same heteroskedasticity-robust methods from earlier (like Section 7.7.3) can be used to compute confidence intervals if sampling is iid.

The following code shows different R syntax to get the same coefficient estimates, with simulated data. The notation $X1:X2$ is the interaction term (or in the output, its coefficient). Heteroskedasticity-robust CIs are also reported.

```r
library(sandwich); library(lmtest)
n <- 4*8
set.seed(112358)
```
m00 <- 10; m10 <- 15; m01 <- 16; m11 <- 25
df <- data.frame(X1=rep(0,n/2),rep(1,n/2),
                 X2=rep(0:1,each=n/4, times=2))
df$Y <- c(rep(m00,n/4),rep(m01,n/4),
          rep(m10,n/4),rep(m11,n/4) ) + rnorm(n)

# Three equivalent estimates
ret1 <- lm(Y~X1*X2, data=df)
ret2 <- lm(Y~X1+X2+X1:X2, data=df)
df$Xint <- df$X1*df$X2
ret3 <- lm(Y~X1+X2+Xint, data=df)
TrueBetas <- c(m00,m10-m00,m01-m00,(m11-m01)-(m10-m00))
retmat <- rbind(coef(ret1),coef(ret2),coef(ret3),TrueBetas)
rownames(retmat) <- c('est1','est2','est3','true')
print(round(retmat, digits=2))

# (Intercept) X1 X2 X1:X2
# est1 10 5.17 6.12 3.73
# est2 10 5.17 6.12 3.73
# est3 10 5.17 6.12 3.73
# true 10 5.00 6.00 4.00

round(coefci(ret1, vcov.=vcovHC(ret1,type='HC1')),digits=2)

# (Intercept) 9.30 10.77
# X1 4.31 6.04
# X2 4.97 7.27
# X1:X2 2.20 5.27
Empirical Exercise EE9.1. You will analyze data on driving laws and fatal accident rates, originally from Freeman (2007). In particular, you’ll compare weekend driving fatality (death) rates for states that adopted a 0.08 blood alcohol content (BAC) law and states that didn’t, comparing rates before and after the law adoption. Standard errors can be smaller if the full dataset is used, but such methods are beyond our scope. Either way, the difference-in-differences approach is probably not identifying a treatment effect: probably states that adopted such laws also adopted other ways to discourage drunk driving, whether official laws or just changing cultural norms. This violates the parallel trends assumption.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

```r
library(wooldridge); library(sandwich); library(lmtest)
?driving
```

b. Stata only: load the data with

```stata
use http://faculty.missouri.edu/kaplandm/intro_text/driving, clear
```

c. Keep only years 1980 and 1990.

R: `df <- driving[driving$year==1980 | driving$year ==1990 , ]`

Stata: `keep if year==1980 | year==1990`

d. Create a dummy variable for the “after” period (year 1990).

R: `df$after <- (df$year==1990)`

Stata: `generate after = (year==1990)`

e. Create variable `bac` equal to 1 (or TRUE) if there’s any BAC law that year.

R: `df$bac <- (df$bac08+df$bac10>=1)`

Stata: `generate bac = (bac08 + bac10 >= 1)`

f. Drop states that already had a BAC law in the “before” period (1980), leaving only states that never had the law or adopted it between 1980 and 1990.

R: `dropst <- unique(df$state[!df$after & df$bac])` to get a list of the states to drop, and then remove them with `df <- df[!df$state %in% dropst , ]`

Stata: `generate dropflag = (!after & bac)`

`bysort state : egen dropst = max(dropflag)`

`drop if dropst`
g. Create a treatment dummy equal to 1 for states that adopted a BAC law by 1990.

R: `treatst <- unique(df$state[df$bac])` followed by `df$treat <- df$state %in% treatst`

Stata: `bysort state : egen treat = max(bac)`

h. Run a difference-in-difference regression with the intercept, "after" dummy, treatment dummy, and interaction term. Below, the * in R and the ## in Stata automatically generate the desired interaction term.

R: `ret <- lm(wkndfatrte~treat*after, data=df)`

Stata: `regress wkndfatrte treat##after , vce(robust)`

i. To see how the OLS coefficient estimates relate to the conditional means (CEF estimates), compute the sample mean weekend driving fatality rate within each of the four groups defined by the time period and “treatment” status.

R: `agg <- aggregate(wkndfatrte~treat*after, data=df, FUN=mean)`

Stata: `tabulate treat after , summarize(wkndfatrte) means missing`

j. Display the CEF-based replication of the OLS estimates.

R: `c(agg[1,3], agg[2,3]-agg[1,3], agg[3,3]-agg[1,3])` for the first three coefficient estimates and `c((agg[4,3]-agg[3,3])-(agg[2,3]-agg[1,3]), (agg[4,3]-agg[2,3])-(agg[3,3]-agg[1,3]))` to show both (equivalent) ways to compute the interaction coefficient estimate.

Stata: `collapse (mean) wkndfatrte, by(treat after)`

k. Repeat part (h) but with a different outcome variable to replace `wkndfatrte`, like the weekend fatalities per 100 million miles driven (instead of population), or the total fatality rate (not just weekends), etc.

l. Repeat parts (e)–(h) but replacing your `bac` treatment variable created in part (e) with a treatment dummy equal to 1 if `perse` (a different driving law) equals 1 (and equal to 0 otherwise).

Empirical Exercise EE9.2. You will analyze wage data for different types of individuals from the 1976 Current Population Survey
EMPIRICAL EXERCISES (conducted by the U.S. Census Bureau). Specifically, you’ll look at dummy variables for **nonwhite** (race) and **female**, as well as their interaction. The results are clearly not causal, but the interaction term shows (descriptively) the difference in the white/nonwhite wage gap for females compared to non-females, or (equivalently) the difference in the female/non-female wage gap for nonwhites compared to whites.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

```r
library(wooldridge); library(sandwich); library(lmtest)
?wagel
```

b. Stata only: load the data with `bcuse wage1, nodesc clear` (assuming `bcuse` is already installed)

c. Display the group mean wage for the four groups defined by the **nonwhite** and **female** dummy variables.

R:

```r
(agg <- aggregate(wage~nonwhite*female, data=wage1, FUN=mean))
```

Stata:

```
tabulate female nonwhite, summarize(wage)
means missing
```

d. Run a “difference-in-differences” type of regression with the intercept, non-white dummy, female dummy, and interaction term.

R:

```r
ret <- lm(wage~nonwhite*female, data=wage1)
coeftest(ret, vcov.=vcovHC(ret, type="HC1"))
coefci( ret, vcov.=vcovHC(ret, type="HC1"))
```

Stata:

```
regress wage female ##nonwhite, vce(robust)
```

e. Compute the OLS coefficient estimates manually from the four conditional means.

R: store the conditional means with `m00 <- agg$wage[1]; m10 <- agg$wage[2]; m01 <- agg$wage[3]; m11 <- agg$wage[4]` and show that you can replicate the OLS estimates with `rbind(coef(ret), c(m00, m10-m00, m01-m00, (m11-m01)-(m10-m00)) )` and also note that `c( (m11-m01) - (m10-m00) , (m11-m10) - (m01-m00) )` shows the equivalence of the two interpretations of the interaction term coefficient.

Stata: collapse the dataset to just the four conditional means with `collapse (mean) wage , by(female nonwhite)` and then display the manually calculated coefficient estimates with

```
display wage[1]
display wage[3]-wage[1]
display wage[2]-wage[1]
display (wage[4]-wage[3])-(wage[2]-wage[1])
display (wage[4]-wage[2])-(wage[3]-wage[1])
```
f. Repeat part (d) but using **south** instead of **female**

g. Repeat part (d) again with any two dummy variables of your choice; you may use one from a previous analysis as long as it is combined with a different dummy. The dataset comes with many dummy variables already, like **nonwhite, female, south** (and other regions), **servocc** (and other occupational fields and industries), and **married**, or you can create your own. For example, you can generate a “more than high school education” dummy with R code `wage1$gtHS <- (wage1$educ>12)` or Stata command `generate gtHS = (educ>12)`
Chapter 10

Regression with Multiple Regressors

⇒ Kaplan video: Chapter Introduction

Depends on: Chapters 8 and 9 (which depend on Chapters 2–4, 6, and 7)

Unit learning objectives for this chapter

10.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]

10.2. Assess in a real-world example whether there is bias from omitted variables and whether a linear model seems realistic [TLOs 2 and 6]

10.3. Describe and interpret models with multiple regressors, including those in which two variables interact [TLO 3]

10.4. Judge which assumptions seem true and which interpretation seems most appropriate for real-world regressions [TLOs 2 and 6]

10.5. In R (or Stata): estimate a regression with multiple variables, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

Optional resources for this chapter

• James et al. (2013, §3.2)
• Hastie, Tibshirani, and Friedman (2009, §§2.3.1,2.4,3.1–3.2)
• Linear projection (theory): Hansen (2020, §7)
• Average structural effects and their identification: Hansen (2020, §2.30)
• Regression example (Masten video)
• Perfect multicollinearity (Lambert video)
• Imperfect multicollinearity example (Lambert video)
• Dummy coefficients (Lambert video)
• Dummy interactions (Lambert video)
• Continuous interactions (Lambert video)
• Sections 3.1 (“Multiple Regression in Practice”) and 6.1.5 (“Interaction Terms”) in Heiss (2016)
• Section 4.4 (“Reporting Regression Results”) in Heiss (2016)
• Section 8.3 (“Interactions Between Independent Variables”) in Hanck et al. (2018)

Allowing multiple regressors opens a multitude of combinations, especially when combined with nonlinear functions like in Chapter 8. Most of Chapter 10 focuses on the different functional forms themselves, with the different types of flexibility they do (and don’t) allow. These discussions apply equally to descriptive, predictive, and causal models.

10.1 Omitted Variable Bias

One motivation for this chapter is that omitted variable bias (OVB, Section 9.1) can still be a problem even if we include two regressors. We may need to include three, or even 10 or 100 regressors to avoid OVB. But even with 100 regressors, OVB can still be a big problem.

Consider OVB with the linear structural model

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + V. \] (10.1)

For OLS to consistently estimate \( \beta_j \) for \( j = 1, 2, 3 \) (the slope coefficients) requires \( \text{Cov}(X_j, V) = 0 \) for \( j = 1, 2, 3 \). Imagine this is true, but \( X_3 \) is omitted, so

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U, \quad U \equiv \beta_3 X_3 + V. \] (10.2)

In (10.2), OLS consistency for \( \beta_1 \) and \( \beta_2 \) requires \( \text{Cov}(X_1, U) = \text{Cov}(X_2, U) = 0 \). Since

\[ \text{Cov}(X_j, U) = \beta_3 \text{Cov}(X_j, X_3) + \text{Cov}(X_j, V), \] (10.3)

this requires either \( \beta_3 = 0 \) (i.e., \( X_3 \) is not a causal determinant of \( Y \)) or else \( \text{Cov}(X_1, X_3) = \text{Cov}(X_2, X_3) = 0 \).

There are other mathematical formulations, but they all make the point that even including 100 regressors is not sufficient to avoid OVB if there is still an important omitted variable. That is, even if (10.2) becomes

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + U, \quad U \equiv \gamma Q + V, \] (10.4)

then we still have OVB if \( \gamma \neq 0 \) and any \( \text{Cov}(X_j, Q) \neq 0 \).

That is, there is OVB if both of the following conditions hold.
OVB.1' The omitted variable $Q$ is a causal determinant of $Y$; in (10.4), $\gamma \neq 0$.

OVB.2' The omitted variable is correlated with an included regressor; in (10.4), $\text{Corr}(X_i, Q) \neq 0$ for some $j$.

**Discussion Question 10.1** (OVB with multiple regressors). Consider the example of California schools where $Y$ is a school’s average standardized math test score for 5th-graders, $X_1$ is the 5th-grade student-teacher ratio, and $X_2$ is the percentage of 5th-graders who are English learners (non-native speakers). Judge whether a school’s total expenditures per student satisfies each of Conditions OVB.1' and OVB.2' for OVB.

### 10.2 Linear-in-Variables Model

#### 10.2.1 Model and Coefficient Interpretation

The linear-in-variables model and discussion from Section 9.2 naturally generalize to non-binary and/or more than two regressors. With $J$ regressors $X_1, X_2, \ldots, X_J$,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J + U = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + U \equiv g(X_1, \ldots, X_J) + U. \quad (10.5)$$

If $U$ is a CEF error, then $g(\cdot)$ represents the CEF. However, the following discussion is essentially the same if $U$ is a linear projection error and $g(\cdot)$ is the linear projection, or if the $\beta_j$ have a causal interpretation.

Regardless of interpretation, the coefficient $\beta_j$ shows how the function $g(\cdot)$ changes when $X_j$ increases by one unit. This is true whether $X_j$ is binary, discrete, or continuous. For example, $X_1$ only appears in the $\beta_1 X_1$ term, so if we change from $X_1 = x_1$ to $X_1 = x_1 + 1$ (unit increase), that term changes from $\beta_1 x_1$ to $\beta_1(x_1 + 1) = \beta_1 x_1 + \beta_1$, a change of $\beta_1$. That is, for any starting values $X_1 = x_1, X_2 = x_2, \text{etc.}$, a unit increase in $X_1$ changes the function by

$$g(x_1 + 1, x_2, \ldots, x_J) - g(x_1, x_2, \ldots, x_J) = [\beta_0 + \beta_1(x_1 + 1) + \sum_{j=2}^{J} \beta_j x_j] - [\beta_0 + \beta_1 x_1 + \sum_{j=2}^{J} \beta_j x_j] = \beta_1(x_1 + 1 - x_1) = \beta_1. \quad (10.6)$$

For example, if $Y$ is wage in $\$/hr, and $X_1$ is years of education, and $\beta_1 = (\$5/hr)/yr$, then each additional year of education is associated with a $(\$5/hr)/yr change, regardless of the initial education level or other variables like experience.
More generally, if $X_1$ changes by $\Delta_1$ units, then the function’s value changes by $\beta_1 \Delta_1$. Regardless of the starting values, if $X_1$ changes from $x_1$ to $x_1 + \Delta_1$, then similar to (10.6),

$$g(x_1 + \Delta_1, x_2, \ldots, x_J) - g(x_1, x_2, \ldots, x_J) = [\beta_0 + \beta_1(x_1 + \Delta_1) + \sum_{j=2}^{J} \beta_j x_j] - [\beta_0 + \beta_1 x_1 + \sum_{j=2}^{J} \beta_j x_j] = \beta_1 (x_1 + \Delta_1) - \beta_1 x_1 = \beta_1 \Delta_1.$$  

10.2.2 Limitations

While pleasingly simple, these formulas may not be realistic. That is, the change in $Y$ may depend on not only $\Delta_1$, but the starting value $x_1$, or other $x_j$. For example, let $Y$ be wage, $X_1$ years of experience, and $X_2$ years of education. Due to diminishing marginal benefits, perhaps the first years of experience are associated with bigger increases in mean wage than later years of experience. The wage increase associated with the change from $X_1 = 0$ to $X_1 = 1$ is probably larger than the increase from $X_1 = 40$ to $X_1 = 41$, even though $\Delta_1 = 1$ in both cases. Further, the change from $X_1 = 0$ to $X_1 = 1$ may be associated with a larger wage increase for highly educated individuals (large $X_2$) than for less-educated individuals. Mathematically, the change depending on the starting value of $X_1$ implies some nonlinearity in $X_1$, and the dependence on the value of $X_2$ implies some sort of interaction term(s).

Nonlinear and nonparametric functions of a single variable are discussed in Sections 8.2 and 8.3; interactions are discussed in Sections 9.3 and 10.3. Nonparametric models with multiple regressors are beyond our scope.

10.2.3 Code

The following code shows a simple linear-in-variables regression with simulated data. In the output, the row labeled X1 shows results for the corresponding slope coefficient $\beta_1$. Specifically, the output shows the OLS estimate $\hat{\beta}_1$ (Estimate), the heteroskedasticity-robust standard error estimate (Std. Error), and a 95% confidence interval (lower endpoint under 2.5 %, upper endpoint under 97.5 %). Similarly for the other regressors and results.

```r
library(sandwich); library(lmtest)
set.seed(112358)
n < 50
CEF <- function(x1,x2,x3) { 1*x1+2*x2+3*x3 }
df <- data.frame(X1=runif(n), X2=runif(n), X3=runif(n))
dfSy <- CEF(df$X1, df$X2, df$X3) + rnorm(n)
ret <- lm(Y~X1+X2+X3, data=df)
retVC1 <- vcovHC(ret, type="HC1")
round(cbind(coefret, vcov. = retVC1)[,1:2],
```
### 10.3 Interaction Terms

To start, imagine there are two regressors, one of which is binary. To help us remember which is which, let $D$ (for “dummy”) be the binary regressor ($D = 1$ or $D = 0$) and $X$ the other regressor. Assume $X$ is the regressor of interest.

#### 10.3.1 Limitation of Linear-in-Variables Model

With a linear-in-variables model,

$$ Y = g(X, D) + U, \quad g(X, D) = \beta_0 + \beta_1 X + \beta_2 D. \quad (10.8) $$

A unit increase in $X$ always changes the function $g(X, D)$ by $\beta_1$ units, regardless of the starting value of $X$ or the value of $D$. As discussed in Section 10.2, this is often unrealistic.

Since $D$ has only two possible values, we can plug them each into $g(X, D)$:

$$ g(X, 0) = \beta_0 + \beta_1 X, \quad (10.9) $$

$$ g(X, 1) = \beta_0 + \beta_1 X + (\beta_2)(1) = \beta_0 + \beta_2 + \beta_1 X. \quad (10.10) $$

These are two functions of $X$: one when $D = 0$, one when $D = 1$. They have the same slope ($\beta_1$) but different intercepts ($\beta_0$ and $\beta_0 + \beta_2$).

#### 10.3.2 Interpretation of Interaction Term

To allow both the intercept and slope to differ between $g(X, 0)$ and $g(X, 1)$, an interaction term can be used, specifically the product $DX$. Mathematically, adding this term to (10.8),

$$ g(X, D) = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 DX. \quad (10.11) $$
The function in (10.11) is more general because setting $\beta_3 = 0$ yields (10.8). Given (10.11), instead of (10.9) and (10.10),

\[
g(X, 0) = \beta_0 + \beta_1 X + (\beta_2)(0) + (\beta_3)(0)(X) = \beta_0 + \beta_1 X, \quad (10.12)
\]

\[
g(X, 1) = \beta_0 + \beta_1 X + (\beta_2)(1) + (\beta_3)(1)(X) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X. \quad (10.13)
\]

Now, the slope differs (by $\beta_3$), too. Just as $\beta_2 > 0$, $\beta_2 < 0$, and $\beta_2 = 0$ are all possible, so are $\beta_3 > 0$, $\beta_3 < 0$, and $\beta_3 = 0$.

![Figure 10.1: Visualization of $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 DX$.](image)

Figure 10.1 illustrates the interpretation of the function from (10.11). In the figure’s example, $\beta_2 > 0$ and $\beta_3 > 0$. Omitting the interaction term is equivalent to assuming $\beta_3 = 0$, in which case the two lines would be parallel (same slope).

If you’re interested in $D$, don’t only look at $\beta_2$. Rearranging (10.11),

\[
g(X, D) = (\beta_0 + \beta_1) + D(\beta_2 + \beta_3 X), \quad (10.14)
\]

so the slope coefficient on $D$ is $\beta_2 + \beta_3 X$. For example, even if $\beta_2 = -2$, the slope $\beta_2 + \beta_3 X$ is positive if $\beta_3 X > 2$. The opposite is also possible, e.g., if $\beta_2 = 5$, $\beta_3 = -1$, and $X > 5$: then $\beta_2 > 0$, but the slope is negative, $\beta_2 + \beta_3 X < 0$.

**Discussion Question 10.2** (sleep and interactions). Let $Y$ be a person’s hours of sleep per night, $X$ the person’s age, and $D = 1$ if the person lives in the same house as children under 8 years old (and $D = 0$ if not). Consider the model from (10.11).

a) What would you guess for the signs (+, −, or zero) of $\beta_2$ and $\beta_3$?

b) Given the same set of regressors $(X, D)$, describe another non-linear term that would improve the CEF estimate, and why you think that term would help.

**Discussion Question 10.3** (wage and interactions). Repeat DQ 10.2, but let $Y$ be wage.
10.3.3 Non-Binary Interaction

Discussion Question 10.4 (linear-in-variables?). Let \( Y \) be log wage, \( X_1 \) years of education, and \( X_2 \) years of experience. Consider possible linear-in-variables CEF model

\[
E(Y \mid X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.
\]

a) Explain one reason you think this CEF model is misspecified (wrong).

b) How do you think the CEF slope with respect to experience might differ for different values of education? (Hint: draw a graph with different lines like \( E(Y \mid X_1 = 12, X_2 = x_2) \), where you fix the \( X_1 \) value and then graph the CEF as a function of only \( x_2 \).)

Even if neither regressor were binary, an interaction term allows the slope to depend on the other regressor’s value. Replacing \( X = X_1 \) and \( D = X_2 \) in (10.11),

\[
g(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \tag{10.15}
\]

Consider the slope of \( g(X_1, X_2) \) with respect to \( X_1 \), at different values of \( X_2 \). Generally, rearranging (10.15) as a function of \( X_1 \),

\[
g(X_1, X_2) = \underbrace{(\beta_0 + \beta_2 X_2)}_{\text{intercept}} + \underbrace{(\beta_1 + X_2 \beta_3)}_{\text{slope}} X_1. \tag{10.16}
\]

Plugging values \( X_2 = a \) and \( X_2 = b \) into (10.16), similar to (10.12) and (10.13),

\[
g(X_1, a) = \underbrace{(\beta_0 + a \beta_2)}_{\text{intercept}} + \underbrace{(\beta_1 + a \beta_3)}_{\text{slope}} X_1, \tag{10.17}
\]

\[
g(X_1, b) = \underbrace{(\beta_0 + b \beta_2)}_{\text{intercept}} + \underbrace{(\beta_1 + b \beta_3)}_{\text{slope}} X_1. \tag{10.18}
\]

Changing \( X_2 \) from \( a \) to \( b \) changes the intercept from \( \beta_0 + a \beta_2 \) to \( \beta_0 + b \beta_2 \), and it changes the slope from \( \beta_1 + a \beta_3 \) to \( \beta_1 + b \beta_3 \). Alternatively, we could plug in \( X_1 = a \) and \( X_1 = b \) and consider \( g(a, X_2) \) and \( g(b, X_2) \) as functions of \( X_2 \), where again both the intercept and slope may change.

Don’t get fooled by looking at \( \beta_1 \) alone (a common mistake). For example, imagine \( X_1 \) is experience and \( X_2 \) is years of education, and \( Y \) is wage (\$/hr). Imagine

\[
\hat{Y} = \hat{g}(X_1, X_2) = 5 - 15X_1 + 2X_2 + 2X_1 X_2, \tag{10.19}
\]

i.e., \( \hat{\beta}_0 = 5, \hat{\beta}_1 = -15, \hat{\beta}_2 = 2, \) and \( \hat{\beta}_3 = 2 \). Superficially, \( \hat{\beta}_1 = -15 \) seems like a negative relationship between experience (\( X_1 \)) and wage: it looks like more experience is associated with much lower wage. However, the interaction term affects the slope with respect to \( X_1 \). Using (10.16), that slope is \( \beta_1 + \beta_3 X_2 = 2X_2 - 15 \). If everyone in the data has at least 10 years of education, then \( X_2 \geq 10 \), so
$2X_2 - 15 \geq (2)(10) - 15 = 5$: the slope with respect to $X_1$ is always positive. Even though $\hat{\beta}_1 < 0$, $\hat{g}(X_1, X_2)$ is always increasing in $X_1$, for any possible $X_2 \geq 10$.

This interaction model is more general than the linear-in-variables model, but not fully general. For example, imagine $Y$ is wage, $X_1$ is education, and $X_2$ is experience. Maybe the slope with respect to $X_2$ should be increasing a lot with $X_1$ when $X_1$ is around 12 or 16, but less so around $X_1 = 20$ (or maybe more so?). This type of nonlinearity in the interaction is not allowed by simply including $X_1X_2$. There are nonlinear and nonparametric models to address such situations, but details are beyond our scope.

### 10.3.4 Code

The following code illustrates estimation, heteroskedasticity-robust inference, and prediction with a model including an interaction term. In R formula syntax, the term $D:X$ is the same as including the interaction term $DX$ like in (10.11). Alternatively, $D*X$ includes both linear and interaction terms, i.e., it is equivalent to $D+X+D:X$. So, both estimation models below are identical to (10.11).

```r
library(sandwich); library(lmtest)
set.seed(112358)
n <- 50
CEF <- function(d, x) { 2+3*d+4*x+5*x*d }
df <- data.frame(X=runif(n), D=sample(x=0:1,size=n,replace=TRUE))
df$Y <- CEF(df$D, df$X) + rnorm(n)
# Equivalent estimates
ret <- lm(Y~D*X, data=df)
ret2 <- lm(Y~D+X+D:X, data=df)
retVC <- vcovHC(ret, type="HC1")
round(cbind(coeftest(ret, vcov. = retVC)[,1:2],
          coefci(ret, vcov. = retVC)), digits=2)
## Estimate Std. Error 2.5 % 97.5 %
## (Intercept) 2.00 0.39 1.21 2.80
## D 2.69 0.51 1.67 3.71
## X 3.86 0.65 2.54 5.17
## D:X 5.66 0.85 3.96 7.36
predict(ret, newdata=data.frame(X=c(12,8),D=1:0))
## 1 2
## 118.9 32.8
```

### 10.4 Other Examples

⇒ Kaplan video: Wage Regression Example (again again)
Models can get very complex with multiple regressors. We could have more than 2 regressors; we could have many nonlinear functions of each regressor by itself; and we could have many interactions. For example, even if we only have 5 regressors, there are 10 pairs of regressors (like $X_1$ and $X_4$, $X_2$ and $X_3$, etc.), and each pair may have multiple interaction terms (i.e., not just $X_1X_4$, but also $X_1X_4^2$ or something). With each regressor by itself, we may have multiple nonlinear terms. There could be 40 or 50 terms in our regression just from 5 original regressors. Even if all 5 are binary, the fully saturated model requires $2^5 = 32$ parameters.

With such complicated models, it is better to look at predicted changes using the full model instead of looking at individual coefficients. This is done in R with the `predict()` function.

### 10.5 Assumptions for Linear Projection

Below are formal assumptions sufficient for consistency and asymptotic normality of the OLS estimator of the linear projection coefficients. Asymptotic normality in turn justifies confidence intervals (and $p$-values), which should be approximately correct in large samples (large $n$). These are relatively weak assumptions. However, stronger assumptions are required to interpret the linear projection as a CEF or structural model.

The assumptions are basically the same as in Section 7.7.2, with one exception (perfect multicollinearity). Like before, iid sampling is sufficient but not necessary; OLS consistently estimates the linear projection coefficients with various types of dependent data and complex sampling designs, and the estimators remain asymptotically normal, although the standard errors are different.

#### 10.5.1 Multicollinearity (Two Types)

The one new assumption is that there cannot be perfect multicollinearity. This essentially says redundant regressors are not allowed. For example, if $X_3 = X_1 + X_2$, then $X_3$ is a linear function of other regressors, so we cannot include all of $X_1$, $X_2$, and $X_3$. Remember that the intercept term can be seen as the coefficient on regressor $X_0 = 1$. So if we had $X_1 = 1$ for females and $X_2 = 1$ for non-females, then $X_1 + X_2 = 1 = X_0$, which means perfect multicollinearity.

Something nice about perfect multicollinearity is that computers can check it for us. If you try to run a regression with perfect multicollinearity, R will simply report `NA` for coefficients of the “redundant” regressors (without warning or error). Other statistical packages may give you a warning or error.

For prediction, redundant variables don’t help, so dropping them is fine. For causality, we are unable to distinguish the separate effects among redundant variables. But if they are merely “control variables,” then we do not care.
A related concept is **imperfect multicollinearity**. This refers to regressors being strongly correlated, but not perfectly correlated (i.e., not completely redundant).

This makes it more difficult to learn about the slope coefficients on the highly correlated regressors, but it does not invalidate any results on identification, estimation, or inference. “More difficult” means confidence intervals can be large. This makes sense: if regressors $X_1$ and $X_2$ are highly correlated, and we observe that $Y$ is high when $X_1$ and $X_2$ are high, it’s unclear whether $Y$ is high because $X_1$ is high or because $X_2$ is high. Since they are highly correlated, there are few observations where only $X_1$ or $X_2$ (not both) is high to help distinguish the effect of $X_1$ from that of $X_2$. This is similar to the logic behind omitted variable bias, except we can see the ghost. With prediction, it may be best to include only $X_1$ or $X_2$ (not both), but standard model selection procedures can handle this without any special consideration. (But: if you have a job interview and sense that your interviewer thinks imperfect multicollinearity is really important for some reason, just go with it.)

## 10.5.2 Formal Assumptions and Results

The assumptions and results refer to the linear projection model

$$\text{LP}(Y \mid X_1, \ldots, X_J) = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J. \quad (10.20)$$

The $X_j$ may include nonlinear functions of an original set of regressors. For example, if $X$ and $D$ are observed regressors, then the model could have $X_1 = D$, $X_2 = X$, and $X_3 = DX$. It could also include $X_4 = X^2$, etc.

**Assumption A10.1.** Sampling of $(Y_i, X_{1i}, \ldots, X_{Ji})$ is iid from the population joint distribution of $(Y, X_1, \ldots, X_J)$.

**Assumption A10.2.** There is no perfect multicollinearity. That is, no $X_j$ is a linear combination of other regressors (including the intercept). Equivalently, the only constants $c_j$ ($j = 0, \ldots, J$) that make $0 = c_0 + \sum_{j=1}^{J} c_j X_j$ are $c_0 = c_1 = \cdots = c_J = 0$.

**Assumption A10.3.** The variances of $Y$ and all $X_j$ are finite: $\text{Var}(Y) < \infty$, $\text{Var}(X_j) < \infty$ for $j = 1, \ldots, J$. Or, equivalently, the expected values of $Y^2$ and $X_j^2$ (i.e., second moments) are finite: $E(Y^2) < \infty$, $E(X_j^2) < \infty$ for $j = 1, \ldots, J$.

**Assumption A10.4.** The expected values of $Y^4$ and $X_j^4$ (i.e., fourth moments) are finite: $E(Y^4) < \infty$, $E(X_j^4) < \infty$ for $j = 1, \ldots, J$.

**Theorem 10.1** (OLS consistency). *Given the linear projection model in (10.20), if A10.1–A10.3 are true, then $\hat{\beta}_j \overset{p}{\to} \beta_j$ for $j = 0, 1, \ldots, J$.**

**Theorem 10.2** (OLS approximate normality). *Given the linear projection model in (10.20), if A10.1, A10.2, and A10.4 are true, then the OLS coefficient estimators are asymptotically normal, i.e., with large $n$, approximately $\hat{\beta}_j \sim N(\beta_j, \text{SE}_j^2)$, where the true standard error $\text{SE}_j$ can be estimated and is proportional to $1/\sqrt{n}$.***
In more formal mathematical econometrics, Theorem 10.2 is written in terms of the distribution of $\sqrt{n}(\hat{\beta}_j - \beta_j)$, instead of $\hat{\beta}_j$. This can be confusing. (It is helpful mathematically, but for reasons beyond our scope.)

**Theorem 10.3** (coverage probability, multiple regressors). If $A10.1$, $A10.2$, and $A10.4$ are true, then heteroskedasticity-robust confidence intervals are asymptotically correct. That is, with large enough $n$, the coverage probability is approximately equal to the desired confidence level.

### 10.6 Structural Identification

There are many identification results in which there is a causal interpretation for something OLS can estimate. Here are a few.

#### 10.6.1 Linear Structural Model

Imagine the structural model is

$$Y = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + U.$$  \hspace{1cm} (10.21)

It is possible that some $X_j$ are nonlinear functions of regressors, including interaction terms. If $\text{Cov}(U, X_j) = 0$ for all $j = 1, \ldots, J$, then the structural $\beta_j$ are also linear projection coefficients, which OLS can estimate. That is, if all terms $X_j$ in the regression are “exogenous” in the sense of uncorrelated with $U$, then OLS can consistently estimate all the structural slope coefficients.

#### 10.6.2 General Structural Model

As alluded to in Section 9.5, a conditional average structural effect can be identified under conditional independence ($A9.1$). Consider the structural model

$$Y = h(X_1, X_2, \ldots, X_J, U),$$  \hspace{1cm} (10.22)

where $U = (U_1, U_2, \ldots)$ is a vector containing all the unobserved determinants of $Y$ besides the $X_j$. The structural effect of a one-unit increase in $X_1$ (hence 1 subscript on $s_1$) may depend on the initial value of $X_1$ as well as values of other variables and $U$:

$$s_1(X_1, X_2, \ldots, X_J, U) \equiv h(X_1+1, X_2, \ldots, X_J, U) - h(X_1, X_2, \ldots, X_J, U).$$  \hspace{1cm} (10.23)

Fixing the $X_j$ but averaging over $U$, the **conditional average structural effect** (CASE) of $X_1$ (hence 1 subscript on CASE$_1$) is

$$\text{CASE}_1(x_1, \ldots, x_J) \equiv E[s_1(X_1, \ldots, X_J, U) \mid X_1 = x_1, \ldots, X_J = x_J].$$  \hspace{1cm} (10.24)
Under the conditional independence assumption (CIA, Assumption A9.1), this structural object is identified:

\[
\text{CASE}_1(x_1, \ldots, x_J) = E[Y \mid X_1 = x_1 + 1, X_2 = x_2, \ldots, X_J = x_J] - E[Y \mid X_1 = x_1, \ldots, X_J = x_J] \tag{10.25}
\]

Thus, if we can guess the correct CEF functional form, then we can estimate CASEs by OLS. First, OLS can consistently estimate the LP. Second, given the correct CEF specification, we can interpret the LP as the CEF. Third, we can estimate CEF differences using the estimated LP coefficients; e.g., we can use the `predict()` function in R. Fourth, given the identification results, we can interpret these CEF differences as causal effects (CASEs).

However, if we do not guess the CEF’s functional form correctly (i.e., our model is misspecified), then the second and third steps fail. That is, even if the CASEs are identified, we may fail to estimate them if we don’t estimate the CEF correctly. One response to this disappointing fact is to use nonparametric CEF models, but nonparametric regression with multiple regressors is beyond our scope.

### 10.6.3 Conditional ATE

As alluded to in Section 9.5, a **conditional average treatment effect** (CATE) can be identified under conditional independence (A9.1). Here, \(X_1\) is the binary treatment variable. Given Assumption A9.1 (and SUTVA and overlap), the CATE equals a CEF difference:

\[
E[Y^T - Y^U \mid X_2 = x_2, \ldots, X_J = x_J] = E[Y \mid X_1 = 1, X_2 = x_2, \ldots, X_J = x_J] - E[Y \mid X_1 = 0, X_2 = x_2, \ldots, X_J = x_J] \tag{10.26}
\]

Intuitively, the CIA says that within the subpopulation defined by \((X_2, \ldots, X_J)\), treatment is “as good as random,” so comparing mean treated and untreated observed outcomes (within the subpopulation) has a causal interpretation.

For estimation, as with the CASE, we either need to know the CEF’s functional form (and use OLS) or use nonparametric estimation techniques.
Empirical Experiments

Empirical Exercise EE10.1. You will analyze data collected from Botswana’s 1988 Demographic and Health Survey by James Heakins for an economics term project. In particular, you’ll see how the number of living children a woman (in Botswana) has relates to various other variables, with particular interest in the woman’s years of education. You’ll start with a simple regression of \textit{children} on \textit{educ} that shows an economically significant negative coefficient. Then, you’ll see how this coefficient changes (generally moving toward zero) as you add other regressors as control variables, like the husband’s education (\textit{heduc}) and the woman’s age (\textit{age}). These changes in the estimated coefficient suggest omitted variable bias in the original simple regression. But, even with a large number of control variable regressors, there is probably still omitted variable bias.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

\begin{verbatim}
library(wooldridge); library(sandwich); library(lmtest)
?fertil2
\end{verbatim}

b. Stata only: load the data with \texttt{bcuse fertil2, nodesc clear} (assuming \texttt{bcuse} is already installed)

c. Run a simple regression of \textit{children} on \textit{educ}.

\begin{verbatim}
R: ret1 <- lm(children~educ, data=fertil2)
Stata: regress children educ, vce(robust)
\end{verbatim}

d. Repeat but adding \textit{heduc} as a control variable regressor.

\begin{verbatim}
R: ret2 <- lm(children~educ+heduc, data=fertil2)
Stata: regress children educ heduc, vce(robust)
\end{verbatim}

e. Repeat but adding yet another regressor (woman’s age).

\begin{verbatim}
R: ret3 <- lm(children~educ+heduc+age, data=fertil2)
Stata: regress children educ heduc age, vce(robust)
\end{verbatim}

f. Repeat but add even more regressors (in addition to \textit{educ}, \textit{heduc}, and \textit{age}): \textit{agesq}, \textit{knowmeth}, \textit{usemeth}, \textit{electric}, \textit{urban}, and \textit{catholic}, as well as interactions between \textit{age} and \textit{knowmeth} and between \textit{age} and \textit{usemeth}.

\begin{verbatim}
R: store the result as \texttt{ret4}, and you can simply write \texttt{knowmeth:age} and \texttt{usemeth:age} in the regression formula to generate the interactions.
Stata: first create the two interaction variables like with \texttt{generate know_age = knowmeth*age} and then add those new variables in your list of regressors.
\end{verbatim}

g. R only (since already displayed by Stata): output the four sets of estimated regression coefficients with
coef(ret1)
coef(ret2)
coef(ret3)
coef(ret4)

h. Repeat one more time, with whichever regressors (in addition to educ) you think appropriate; feel free to create additional interaction terms and/or nonlinear terms (like age^3, etc.).

Empirical Exercise EE10.2. You will analyze data originally from Harrison and Rubinfeld (1978), including housing prices and pollution measures. The data are not for individual houses, but instead small areas (census tracts, I’d guess), within which the median housing price is computed along with other characteristics that may affect housing prices, including pollution. You’ll start with a simple regression of log price on log nox (the pollution measure). The coefficient is around −1, meaning a 1% increase in pollution is associated with (approximately) a 1% decrease in price. Then, you’ll add other regressors to try to reduce omitted variable bias. By adding just a couple variables, the pollution coefficient estimate’s magnitude is cut in half, suggesting that there was indeed much OVB. However, even with a large number of regressors, serious OVB may remain.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:
   library(wooldridge); library(sandwich); library(lmtest)
   ?hprice2
b. Stata only: load the data with bcuse hprice2, nodesc clear (assuming bcuse is already installed)
c. Run a simple log-log regression of price on nox.
   R: ret1 <- lm(log(price)~log(nox), data=hprice2)
   Stata: regress lprice lnox, vce(robust)
d. Repeat but adding rooms as a control variable regressor.
   R: ret2 <- lm(log(price)~log(nox)+rooms, data=hprice2)
   Stata: regress lprice lnox rooms, vce(robust)
e. Repeat but adding yet another regressor (crime rate per capita).
   R: ret3 <- lm(log(price)~log(nox)+rooms+crime, data=hprice2)
   Stata: regress lprice lnox rooms crime, vce(robust)
f. Repeat but add even more regressors: dist, radial, stratio, and lowstat. Store the result as ret4 in R.
g. R only (since already displayed by Stata): output the four sets of estimated regression coefficients with
coef(ret1)
coef(ret2)
coef(ret3)
coef(ret4)

h. Repeat one more time, with whichever regressors you think appropriate; try to use interaction terms and/or nonlinear terms (like rooms^2, etc.).
Chapter 11

Midterm Exam #2

⇒ Kaplan video: Chapter Introduction

When I teach this class, the second midterm exam is this week. This “chapter” makes the chapter numbers match the week of the semester. This midterm covers all chapters between the first midterm and now. It does not explicitly include questions about the material before the first midterm exam, but of course that materials was foundational for the material covered on the new exam, so it may (or may not) still help to review it.
Chapter 12

Internal and External Validity

⇒ Kaplan video: Chapter Introduction

Depends on: Chapter 7 (which depends on Chapters 2–4 and 6); for deeper understanding, also Chapters 8 and 10

Unit learning objectives for this chapter

12.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]

12.2. Assess possible problems with regression results and their application to real-world questions (of description, prediction, and causality), and the likely direction of bias [TLOs 5 and 6]

12.3. In R (or Stata): check datasets for possible issues like missing data [TLO 7]

Optional resources for this chapter

- Sample selection from survey non-response (Masten video)
- External validity (Masten video)
- Missing data approaches (Masten video)
- Reverse causality and simultaneity (Masten video)
- Reverse causality example: violence (Lambert video)
- Reverse causality example: HDI (Lambert video)
- Greater external validity for “structural” results (Masten video)
- Sections 9.2 (“Measurement Error”) and 9.3 (“Missing Data and Nonrandom Samples”) in Heiss (2016)
- Chapter 22 (“Missing Data”) in Kaplan (2020)
This chapter discusses many reasons to worry about the validity of econometric results and their application to decisions. Like statistical and economic significance, “validity” is better thought of as a continuum rather than a yes/no property. To tweak Box’s aphorism, “All results are invalid, but some are useful.”

12.1 Terminology

An econometric study has **internal validity** if the methods are appropriate for the study’s setting and sample, i.e., if all the identifying assumptions and other assumptions hold.

An econometric study has **external validity** for a different setting if the results can be used to learn something about the new setting. This does not mean the values or overall effects are identical, but rather that the prediction model or structural model is the same. Ideally, the estimated model can be applied successfully to new variable values and distributions.

The **population studied** refers to the population from which the data was sampled, whereas the **population of interest** is the one that you (as the researcher, policy maker, or decision maker) want to learn about.

For example, you may see an econometric study of the causal effect of a minimum wage change from $4.25/hr to $5.05/hr in New Jersey in 1992, but your job is to advise Missouri about a possible minimum wage increase next year. The study is internally valid if it properly estimates the causal effect of the New Jersey minimum wage increase on people in New Jersey in 1992, i.e., for the population studied. It is externally valid if the estimates can be used to learn about the (potential) policy effects in Missouri, your population of interest. Again, this doesn’t mean the effect next year in Missouri must be identical to the effect in 1992 New Jersey, but that the model estimated with the 1992 New Jersey data can be applied to current Missouri data to learn the potential policy effect.

This chapter briefly discusses many **threats to validity**, i.e., reasons an analysis may not be internally or externally valid.

12.2 Threats to External Validity

Threats to external validity are generally more obvious than threats to internal validity, but they harm evidence-based decisions just as much. For example, consider the descriptive task of estimating the median house price in Missouri. Obviously, a sample of house prices from California (which is much more expensive) does not help. Even
with the price of every house in California, we learn little about Mis-
souri. We can try to learn about relationships between price and
house features (size, land area, etc.) in California, but probably even
such relationships themselves differ in Missouri. This problem is a
lack of external validity.

The Lucas critique (Lucas, 1976, also Section 4.3.3) can also be in-
terpreted in terms of external validity. When macroeconomic policy
changes, that fundamentally changes the setting. Even if our esti-
mates from historical data have internal validity, they might not be
accurate in the new setting under the new policy, i.e., they might not
have external validity.

A few common threats to external validity are now discussed.
Although these don’t automatically imply lack of validity, they are
reasons for skepticism.

12.2.1 Different Place

Different places have different legal, political, cultural, and economic
settings. The house price example highlights just one of many im-
portant differences between California and Missouri. Ideally, you can
always find an empirical study from the same place you’re interested
in. If not, you have to decide whether you think the other place is
similar enough to still help you make a good decision.

For example, imagine you need to quantify costs and benefits of
expanding public bus systems in Missouri. The neighboring states of
Oklahoma and Illinois recently collected data during their (hypothet-
ical) bus system expansions, with different results. Although both
states are very close geographically to Missouri, other characteristics
matter, too. Illinois has almost double the state gas tax of Missouri,
and its urban population share is over 15 percentage points greater
than Missouri’s; both of these may be important for both people’s
decision to ride the bus (versus drive) and the cost of bus operation.
In contrast, Oklahoma’s gas tax and urban population share are very
similar to Missouri’s, so there is probably greater external validity.
Still, there may be other important differences between Missouri and
Oklahoma, some of which may be difficult to measure accurately or
quantify, like cultural attitudes.

12.2.2 Different Time

Even in the same place, time changes the legal, political, cultural, and
economic setting.

For example, consider again the median house price in Missouri,
which is also the (unconditional) best prediction under absolute loss.
Having learned not to use California data, we get Missouri data—from
the year 1975. This is also bad since house prices were much lower
in 1975 than today. Adjusting for inflation would help some, but the
housing market supply and demand have both changed substantially,
even basics like how many people live in Missouri and houses’ size,
age, and quality. We could try to use all these variables in a model,
but some may not have data available, and the model itself may have changed since 1975.

What if we had Missouri data from two years ago; is that close enough? One year? One month? It depends how quickly things are changing and on the decision you need to make. In normal conditions, the median house price does not change more than a few thousand dollars each month. That difference may not matter much for some decisions, like trying to find another state that’s similar to Missouri. But that difference may be too big for other decisions, like a high-frequency investment strategy. Or if instead of “normal conditions” there was just a financial crisis, or a new law passed, the month-old data may be off by more than just a few thousand dollars.

Alternatively, sometimes we can model how variables change over time, in order to predict how they have changed since the data sample was collected; see Chapters 14 and 15.

12.2.3 Different Population

Even in the same place, at the same time, the population studied may differ from your population of interest.

For example, to guide tax incentives for first-time homebuyers in Missouri this year, you want to estimate the median first-year mortgage payment for first-time homebuyers. If you find a study estimating the median mortgage payment among all home owners in Missouri this year, then your number will be much too big because the studied population (all owners) differs greatly from the population of interest (first-time owners), even though they’re in the same place (Missouri) at the same time (this year).

As another example, imagine you’re estimating the benefits of expanding government subsidies for college in the U.S. You (amazingly) find an internally valid estimate of the mean wage increase from college, in the same place (U.S.), from just a few months ago. However, the estimate is for the whole U.S. population (the population studied), including individuals who already got college degrees even without the additional subsidy. Instead, your population of interest is individuals who currently do not (or cannot) choose to graduate from college, but who would with the additional subsidy. Such individuals may not have the same causal effect of college on their wages.

Discussion Question 12.1 (external validity: minimum wage). You’re deciding whether to vote for a minimum wage increase in your state or country (yes, you! wherever you live or vote right now), from $10/hr to $15/hr (or an equivalent increase in your country’s currency). You find a study (Card and Krueger, 1994) of effects of a minimum wage increase from $4.25/hr to $5.05/hr in New Jersey in 1992. Explain your specific concerns about external validity. (Note: this is only a question about external validity; arguments about whether minimum wage should be lower or higher are completely irrelevant.)
12.3 Threats to Internal Validity

For description and prediction, see Items 1–5 in the list below. For causality, the following common threats to internal validity are described below.

1. Functional form misspecification (Section 12.3.1)
2. Measurement error (Sections 12.3.2 and 12.3.3)
3. Non-iid sampling and weights (Section 12.3.4)
4. Missing data (Section 12.3.5)
5. Sample selection (Section 12.3.6)
6. Omitted variables (Section 12.3.7)
7. Simultaneity and reverse causality (Section 12.3.8)

Additionally, violation of SUTVA (as discussed earlier) is another threat to internal validity for treatment effect analysis.

12.3.1 Functional Form Misspecification

Misspecifying the functional form leads to inconsistent estimates of the CEF. This is bad for description, prediction, and causality alike. Details are in Chapters 7–10, including reasons for misspecification, ways to address it, and interpretations of what OLS estimates when it’s not a CEF.

12.3.2 Measurement Error in the Outcome Variable

Without good data, it’s hard to get valid econometric results. As they say, “Garbage in, garbage out.” But, it is not as simple as “good” and “bad” data. Certain data problems can safely be ignored; others can’t be ignored, but they can be fixed; and yet other problems cannot be fixed by any amount of econometrics magic.

Sometimes the true value of a variable is not the value seen in the data. This is especially true in survey data, where individuals (or firms, schools, etc.) report their own information (“self-reported”), and with macroeconomic variables that are difficult to measure accurately. With survey data, people may simply forget the exact value, or they may intentionally lie in some cases.

Notation and Terminology

To define some notation and terms, consider the example of exercise. In a survey, people are asked how many minutes of exercise they did last week, and their responses $Y$ are recorded in the data. Let $Y^*$ be how much exercise somebody truly did last week. This $Y^*$ is the latent (unobserved) true value. In contrast, the observed value
is \( Y = Y^* + M \), where \( M \) is the measurement error. That is, \( M = Y - Y^* \) is the difference between the observed and true values.

All of \((Y^*, Y, M)\) are uppercase to show they’re random variables. For example, one individual could have true \( Y^* = 98.52 \) and report \( Y = 100 \), so \( M = 100 - 98.52 = 1.48 \), whereas another individual could have \( Y^* = 271 \), report \( Y = 250 \), and have \( M = 250 - 271 = -21 \), where all values are in units of “exercise minutes per week.” There are different values of \((Y^*, Y, M)\) for different individuals; the population distribution describes the probabilities of these different possible values.

**Discussion Question 12.2** (exercise error). Consider the example where \( Y^* \) is true exercise minutes last week and \( Y \) is the value somebody reports. Explain one reason (each) why an individual could have a) \( M = 0 \), b) \( M < 0 \), or c) \( M > 0 \). Overall, would you guess \( E(M) = 0 \), \( E(M) < 0 \), or \( E(M) > 0 \)? Why? (Hint: you’ll probably need to make additional assumptions and definitions; e.g., what does “exercise” mean?)

**Regression**

Imagine the true linear projection in error form is

\[
Y^* = \beta_0 + \beta_1 X + V, \quad E(V) = \text{Cov}(X, V) = 0. \tag{12.1}
\]

We want to learn \( \beta_1 \). Substituting in \( Y^* = Y - M \),

\[
Y - M = \beta_0 + \beta_1 X + V,
\]

\[
Y = \beta_0 + \beta_1 X + (V + M) = \beta_0 + \beta_1 X + U. \tag{12.2}
\]

The OLS estimator \( \hat{\beta}_1 \) may be asymptotically biased if \( X \) and \( M \) are related because then \( X \) and \( U \) are related. From (9.5), the asymptotic bias is

\[
\text{plim}_{n \to \infty} \hat{\beta}_1 - \beta_1 = \frac{\text{Cov}(X, U)}{\text{Var}(X)} = \frac{\text{Cov}(X, V + M)}{\text{Var}(X)} = \frac{-\text{Cov}(X, V) + \text{Cov}(X, M)}{\text{Var}(X)}. \tag{12.3}
\]

This is the slope coefficient in \( \text{LP}(M \mid 1, X) \), the linear projection of the measurement error onto the regressor (and an intercept). That is, writing \( \text{LP}(M \mid 1, X) = \gamma_0 + \gamma_1 X \), then the asymptotic bias is \( \text{plim}_{n \to \infty} \hat{\beta}_1 - \beta_1 = \gamma_1 \).

With binary \( X \), \( \gamma_1 \) is a mean difference as in (6.29), so (12.3) is equivalent to

\[
\text{plim}_{n \to \infty} \hat{\beta}_1 - \beta_1 = E(M \mid X = 1) - E(M \mid X = 0). \tag{12.4}
\]

This helps us think about both the sign (direction) and magnitude of asymptotic bias. For example, if there tends to be more positive measurement error when \( X = 1 \) than when \( X = 0 \), then \( \gamma_1 > 0 \), so the OLS estimator \( \hat{\beta}_1 \) has positive asymptotic bias, \( \text{plim}_{n \to \infty} \hat{\beta}_1 > \beta_1 \).
In Sum: Measurement Error in the Outcome

\[ Y_{\text{observed}}; Y^*_{\text{latent/true}}; M_{\text{measurement error}} \]

\[ Y = Y^* + M \iff M = Y - Y^* \]

Binary \( X \): \( \hat{\beta}_1 \) asymptotic bias is \( E(M | X = 1) - E(M | X = 0) \); see (12.4)

General \( X \): \( \hat{\beta}_1 \) asymptotic bias is \( \gamma_1 \) in \( LP(M | 1, X) = \gamma_0 + \gamma_1 X \); see (12.3).

Example

Continuing with \( Y^* \) as weekly exercise, let \( X = 1 \) if somebody has a gym membership and \( X = 0 \) otherwise. The goal is to learn \( \beta_1 \) in \( LP(Y^* | 1, X) = \beta_0 + \beta_1 X \). Since \( X \) is binary, \( \beta_1 \) is also the mean difference \( E(Y^* | X = 1) - E(Y^* | X = 0) \). Also due to binary \( X \), the slope in \( LP(M | 1, X) = \gamma_0 + \gamma_1 X \) is the same as the mean difference, \( \gamma_1 = E(M | X = 1) - E(M | X = 0) \).

There’s no asymptotic bias in a few cases. Obviously, if \( M = 0 \) for everybody, then \( Y = Y^* \), so regressing \( Y \) on \( X \) is identical to regressing \( Y^* \) on \( X \). Even if everyone overreports (\( E(M) > 0 \)) or underreports (\( E(M) < 0 \)), as long as it’s the same for both gym members and non-members, then \( \gamma_1 = 0 \), so there is no asymptotic bias. It’s also fine if \( E(M | X = 0) = E(M | X = 1) = 0 \) but \( \text{Var}(M | X = 1) < \text{Var}(M | X = 0) \), i.e., the gym members report more accurately (smaller variance of \( M \); in the extreme, even \( M = 0 \)), but both groups are accurate on average.

However, there is asymptotic bias if there’s systematic overreporting by only gym members. Maybe gym members are more likely to feel guilty about not exercising and not using their membership, which may cause them to report going to the gym and exercising more than they actually do. Or, conversely, perhaps individuals who think they exercise more than they do (and thus have large \( M \)) are more likely to become gym members because they think it’ll be worth it. Either way, more positive \( M \) (overreporting) is associated with \( X = 1 \) compared to \( X = 0 \), i.e., \( \gamma_1 > 0 \). This leads to positive (upward) asymptotic bias of \( \hat{\beta}_1 \).

Figure 12.1 illustrates the upward bias of \( \hat{\beta}_1 \) in the gym/exercise example. The \( X = 0 \) group does not report perfectly, but there is no systematic reporting bias. The \( X = 1 \) group systematically overreports exercise. Consequently, the red line’s slope (using observed \( Y \)) is much larger than the black line’s slope (using true but unobserved \( Y^* \)). That is, if we could observe \( Y^* \), we would estimate the black line; but we can’t, and using the observed \( Y \) yields a very different (biased) estimate of the slope \( \beta_1 \).

Alternatively, maybe non-gym members tend to have larger \( M \). Maybe gym members only report gym time, whereas non-members include walking the dog, lifting groceries, etc. In that case, \( E(M | X = 0) > E(M | X = 1) \), so \( \gamma_1 < 0 \) and there’s negative asymptotic
**Figure 12.1**: Bias from measurement error in $Y$.

Discussion Question 12.3 (measurement error: scrap rate). Imagine the government wants to help increase the efficiency of chalk manufacturing firms. Specifically, $Y^*$ is a firm’s “scrap rate”: what proportion of their output has to be “scrapped” (trashed/not sold) due to manufacturing defects? For example, $Y^* = 0.04$ means 4% scrap rate. The government randomly assigns firms to a control group and treatment group, to run an experiment. On January 1, the treated firms receive grant money, which they are supposed to use to improve efficiency. All firms self-report their scrap rates on December 31; this is $Y$.

a) Describe a reason why treated firms might systematically over-report ($M > 0$) or underreport ($M < 0$) their scrap rates.

b) In that case, and assuming untreated firms report accurately ($M = 0$), would we overestimate or underestimate the treatment effect of a grant? Why?

c) If the government uses these incorrect estimates to decide whether or not to continue the program, what incorrect decision might they make? Why?

Methods to Address Measurement Error

In some cases, there are methods to reduce or eliminate the bias from measurement error. However, such methods often have additional requirements, like a second measurement of the same variable, and they are beyond our scope.

12.3.3 Measurement Error in the Regressors

There are similarities between measurement error in $X$ and measurement error in $Y$. Much of the math is similar. The causes of measurement error are the same, since a variable may be the $Y$ variable in one model but the $X$ variable in another.

To see how measurement error might cause asymptotic bias, equations like (12.1) and (12.2) can be derived. The true LP with latent
$X^*$ is

$$Y = \beta_0 + \beta_1 X^* + R, \quad E(R) = \text{Cov}(X^*, R) = 0.$$  \hspace{1cm} (12.5)

Since the observed $X$ is $X = X^* + M$, substituting in $X^* = X - M$,

$$Y = \beta_0 + \beta_1 (X - M) + R = \beta_0 + \beta_1 X + (R - \beta_1 M).$$  \hspace{1cm} (12.6)

Like (12.3), the asymptotic bias is

$$\text{plim}_{n \to \infty} \hat{\beta}_1 - \beta_1 = \frac{\text{Cov}(X, R - \beta_1 M)}{\text{Var}(X)},$$

so the asymptotic bias is zero if and only if $\text{Cov}(X, R - \beta_1 M) = 0$, i.e., if the observed $X$ is uncorrelated with the unobserved “error term” $R - \beta_1 M$. Using (12.5) and linearity,

$$\text{Cov}(X, R - \beta_1 M) = \text{Cov}(X, R) - \beta_1 \text{Cov}(X, M)$$

$$= \text{Cov}(X^* + M, R) - \beta_1 \text{Cov}(X, M)$$

$$= \text{Cov}(X^*, R) + \text{Cov}(M, R) - \beta_1 \text{Cov}(X, M).$$

If $M$ is uncorrelated with the LP error $R = Y - \beta_0 - \beta_1 X^*$, and if $\beta_1 = 0$ (which means $Y$ and the true $X^*$ are not correlated), then this is zero. Otherwise, there is almost certainly asymptotic bias, in particular when $\text{Cov}(X, M) \neq 0$.

**Attenuation Bias: Assumptions and Result**

Unfortunately, $\text{Cov}(X, M) = 0$ is very unlikely. Consider what seems to be the best-case scenario: $M$ is just random noise unrelated to the true value $X^*$, so $\text{Cov}(X^*, M) = 0$. Unfortunately, using $\text{Cov}(X^*, M) = 0$,

$$\text{Cov}(X, M) = \text{Cov}(X^* + M, M) = \text{Cov}(X^*, M) + \text{Cov}(M, M) = \text{Var}(M).$$  \hspace{1cm} (12.7)

Assuming not everybody has $M = 0$, then $\text{Var}(M) > 0$, so $\text{Cov}(X, M) > 0$. Thus, even if $\text{Cov}(M, R) = 0$, the asymptotic bias is not zero, $-\beta_1 \text{Cov}(X, M) \neq 0$.

In this case with $\text{Cov}(X, M) > 0$ and $\text{Cov}(M, R) = 0$, the resulting bias is called **attenuation bias**. This means that the estimates $\hat{\beta}_1$ tend to be in between 0 and $\beta_1$: $0 < \text{plim} \hat{\beta}_1 / \beta_1 < 1$, implying $|\text{plim} \hat{\beta}_1| < |\beta_1|$. That is, the estimates are systematically pushed closer to zero by the measurement error. This is different than positive (upward) bias, which tends to make $\hat{\beta}_1 > \beta_1$, or negative (downward) bias, which tends to make $\hat{\beta}_1 < \beta_1$. With attenuation bias, if $\beta_1 > 0$, then generally $0 < \hat{\beta}_1 < \beta_1$, whereas if $\beta_1 < 0$, then generally $0 > \hat{\beta}_1 > \beta_1$.

Even if we cannot fix the attenuation bias, it is helpful to know the direction of the bias. For example, if we estimated $\hat{\beta}_1 = 7$, and we suspect attenuation bias, then we may think $\beta_1$ might be even larger, but probably not smaller.
Figure 12.2 illustrates attenuation bias. It shows a simple example where \( P(X^* = 1) = P(X^* = 2) = 0.5 \), and \( Y = X^* \) (no error term). The linear projection is just the line through \((X^*, Y) = (1, 1)\) and \((2, 2)\), which has \( \beta_0 = 0 \) and \( \beta_1 = 1 \) (intercept zero, slope one). Then, imagine adding error: \( P(M = -1) = P(M = 1) = 0.5 \), regardless of \( X^* \) or \( Y \). Then the \( X^* = 1 \) values become \( X = X^* + M \): either \( X = 1 - 1 = 0 \) or \( X = 1 + 1 = 2 \). Similarly, the \( X^* = 2 \) values become either \( X = 2 - 1 = 1 \) or \( X = 2 + 1 = 3 \). Now we have four possible values of \((X, Y)\), each with equal 0.25 probability: \((0, 1)\), \((2, 1)\), \((1, 2)\), and \((3, 2)\), forming a parallelogram. The result is \( \text{LP}(Y \mid 1, X) = 1 + X/3 \) (slope is 1/3), very different than \( \text{LP}(Y \mid 1, X^*) = X^* \) (slope is 1). That is, when we add horizontal noise (errors in \( X \)), the slope of the linear projection \( \text{LP}(Y \mid 1, X) \) is flatter (closer to zero) than the slope of \( \text{LP}(Y \mid 1, X^*) \).

**General Bias**

Unfortunately, outside this very special case, the type of bias may differ. It is not necessarily attenuation bias.

In particular, if \( \text{Cov}(M, R) \neq 0 \) and \( |\text{Cov}(M, R)| > |\beta_1 \text{Cov}(X, M)| \), then the sign of the bias is the sign of \( \text{Cov}(M, R) \), i.e., positive bias if \( \text{Cov}(M, R) > 0 \) or negative bias if \( \text{Cov}(M, R) < 0 \). So, generally, any type of asymptotic bias is possible, depending how the measurement error is related to other variables.

There are methods that address measurement error in \( X \), but these are beyond our scope.
12.3. THREATS TO INTERNAL VALIDITY

you don’t actually need to use weights, but it’s safer to just always use them.

Sampling may be non-iid for reasons other than weights. Clustered and/or stratified sampling can cause non-iid sampling, as discussed in Section 3.5. Time series data also usually lack iid sampling; see Part III.

Generally, these types of non-iid sampling do not affect consistency of estimators, but they often cause incorrect standard errors and confidence intervals. That is, constructing an asymptotic 95% CI based on iid sampling may produce an interval with only 90% coverage probability, or even 80% or 50% or lower. Similarly, \( p \)-values may tend to be too small, or hypothesis test type I error rates may be much larger than desired.

Thankfully, valid (consistent) standard error estimators exist in almost all these cases. However, it can get complicated. For now, just be aware of when sampling is non-iid.

The following code shows unweighted and weighted results using simulated data. Without worrying about the details, some patterns are clear. First, the weighted and unweighted estimates are significantly different; it is important to use weights in estimation. Second, the unweighted and weighted SEs also differ significantly. Third, although the weighted estimates are identical, the three weighted standard errors are different. Judging which is most appropriate requires understanding different types of weights and is beyond our scope.

```r
library(survey); library(sandwich); library(lmtest)
set.seed(112358)
n <- 20
dat <- data.frame(X=rnorm(n))
dat$Y <- 1 + 2*dat$X + 3*dat$X^2 + rnorm(n)
dat$wgt <- 1/zero.alt3/zero.alt3*runif(n)+1/zero.alt3
dsgn <- svydesign(ids=~1, weights=~wgt, data=dat)
c(mean(dat$Y), svymean(dat$Y, dsgn), sum(dat$Y*dat$wgt)/sum(dat$wgt))
## [1] 4.25 4.28 4.28
ret.un <- lm(Y~X, data=dat, weights=NULL)
ret.w <- lm(Y~X, data=dat, weights=wgt)
m1 <- c('lm/unwgtd', 'lm/wgtd', 'svyglm/wgtd', 'lm/unwgtd'),
m2 <- c('lm/un', 'lm/w', 'svyglm/w', 'coeftest/un', 'coeftest/w', 'svyglm/w',
out <- data.frame(est.method=m1, SE.method=m2, est=NA, SE=NA)
vcu1 <- vcovHC(ret.un, type="HC1")
vcw1 <- vcovHC(ret.w, type="HC1")
out[4,3:4] <- coeftest(ret.un, vcov.=vcu1)[X',1:2]
out[5,3:4] <- coeftest(ret.w, vcov.=vcw1)[X',1:2]
print(out)
```

### 12.3.5 Missing Data

Like with measurement error in $Y$ (Section 12.3.2), the reason why there is missing data determines whether or not it’s a problem. As we saw with measurement error in $Y$, if the error is completely random (independent of $X$), then it will not bias linear projection slope estimates. Similarly, if data is missing completely at random (like, a cat walked across your computer keyboard or something), then it’s fine to just drop observations with missing data and proceed as usual. This is called **complete case analysis**, where a **complete case** is an observation in which no values are missing (i.e., all values are observed). For example, if the dataset is $(Y_i, X_i)$ for $i = 1, \ldots, n$, then the complete cases are the $i$ for which both $Y_i$ and $X_i$ are observed (not missing).

In other cases, we can’t ignore the missing data problem, but there are methods that can fix the problem and avoid asymptotic bias.

In yet other cases, it is very difficult to address the missing data problem. In particular, when the value of $Y$ affects whether or not data are missing, it is very difficult. For example, if $Y$ is income and people with high (or low) income tend not to report their income on a survey, then regression estimates will be biased.

Figure 12.3 shows an example of missingness related to $Y$. Here, $Y$ is income and $X = 1$ if an individual has a college degree, $X = 0$ if not. In the example, the highest-income individuals do not report $Y_i$ but everyone else does. This mostly affects $X_i = 1$ individuals, but also the very highest $Y_i$ in the no-college group. If we just run OLS on observations with both $Y_i$ and $X_i$ observed, then both the OLS slope and sample mean are biased downward. The OLS intercept is very slightly downward biased, too, since the top $Y_i$ when $X_i = 0$ are missing.

**Discussion Question 12.4** (program attrition). Consider a job training program like the federally funded Job Training Partnership Act (JTPA) of 1982. Each eligible individual was randomly assigned to either take the job training or not. You want to estimate the average treatment effect on annual income ($Y$) of being assigned to the training (the “intent to treat” effect). However, some individuals’ data is missing because they moved to a different state to take a high-paying job. Explain why this could be a threat to internal validity, and in which direction you think the resulting bias might be.
Discussion Question 12.5 (missing salary data). You get data on a sample of professors from research universities in the U.S., which is the population of interest. However, you only find salary data for public universities, not private.

a) How/does this bias your estimate of the population mean salary? Why?

b) How/does this bias your regression of salary on a dummy for being a professor in a STEM field? Why? (Hint: consider the intercept and slope separately.)

c) Discuss your previous answers in terms of external validity.

12.3.6 Sample Selection

Whereas missing data means some values are missing in the dataset, \textbf{sample selection} means entire individuals (observations) are missing. Whereas missing values are indicated by \texttt{NA} in R, there may be no indication that entire individuals are missing. The number of missing individuals may be unknown.

As with missing data, the reason behind the sample selection is crucial for whether it results in sample selection bias. For example, if individuals are “selected” into the sample at random (unrelated to their $Y_i$ or $X_i$), then it’s just like we’re taking a random sample of a random sample, so we can just proceed as normal. However, if individuals are selected into the sample based on their $Y_i$, then OLS (and other estimators) can be very biased.

For example, similar to Figure 12.3, imagine $Y$ is wage, and individuals with high wage are less likely to take a survey at all. If our dataset only shows individuals who did take the survey, then...
sample selection bias is likely. The picture is basically the same as Figure 12.3, just that the “missing” data points are now entirely unobserved (those \( i \) are not even in our sample). This particular example describes 

**non-response bias**, a common problem for surveys; i.e., people who actually answer the survey are not representative of the population of interest, differing in important ways compared to people who do not answer the survey.

An important economic example of sample selection is that wages are only observed for currently employed individuals. We may want to learn what determines the wage an individual is offered by a firm. However, if the wage a firm is willing to pay is below the individual’s reservation wage or a legal minimum wage, then the individual won’t or can’t take the offer. But if they don’t work, then we can’t observe the hypothetical wage. This was the motivation for the famous approach to correct for sample selection due to Heckman (1979).

Methods to address sample selection bias are beyond our scope, but you can at least try to think critically about whether sample selection bias might be an issue in real-world examples.

### 12.3.7 Omitted Variable Bias and Collider Bias

Omitted variable bias is discussed in Sections 9.1 and 10.1. It is very common with observational economic data: many variables are (cor)related in economics, and many important ones are difficult to measure (human capital, technology, marginal cost, etc.). If they are actually observed in the data, then they can just be included, although recall that including colliders actually makes bias worse (Section 9.6). If not, then other methods can be used under certain specific conditions. For example, difference-in-differences (Section 9.7) allows certain types of omitted variables. Other estimators with panel data (observations for the same unit \( i \) over multiple time periods) also allow certain types of omitted variables, like those that do not change over time. However, these and yet other estimators that address omitted variable bias are beyond our scope.

### 12.3.8 Simultaneity and Reverse Causality

When we regress \( Y \) on \( X \), we often (perhaps subconsciously) assume that \( X \) may have a causal effect on \( Y \), but that \( Y \) does not have an effect on \( X \). However, sometimes in reality \( Y \) affects \( X \), too. This is called reverse causality or simultaneous causality.

The issue of **simultaneity** is basically the same (and often synonymous), but emphasizes that it is not necessarily a direct causal effect of \( Y \) on \( X \), just that \( X \) and \( Y \) are determined by the same system at the same time (simultaneously). Economic systems are often complex, where conditions “determine” the values of multiple variables at the same time. For example, supply and demand curves simultaneously determine the equilibrium market price and quantity. Rather than trying to say price affects quantity and quantity affects price (simultaneous causality), it’s more precise to say that price and
Because economists often study systems with complex interactions among many variables, and with observational data, simultaneity and reverse causality are common.

For example, one question economists have studied is the effect of police officers per capita $X$ on crime rate $Y$ in a city. (Note: as with other examples like minimum wage and right-to-work laws, this has nothing to do with “good” or “bad,” but only how simplistic econometric analysis can fail to have a causal interpretation.) Of course, it is possible that the density of police has a causal effect on crime rate. But it is also possible that crime rate $Y$ has a causal effect on $X$, through policy decisions. That is, imagine you are in charge of the city’s decision of how many police officers to have. Aside from budget constraints, one of the biggest factors in your decision is probably the city’s crime rate. If the city has a very low crime rate, then you would probably not decide to spend more to hire more police officers; in fact, you may decide to have fewer and spend the savings on other city needs. However, if the city’s crime rate is very high, then you would seriously consider hiring more police officers. That is, your decision about $X$ is determined partly by $Y$.

With simultaneity or reverse causality, OLS regression of $Y$ on $X$ does not consistently estimate structural or treatment effects. In the police example, even if there were zero effect of $X$ on $Y$, the response of $X$ to $Y$ would cause positive correlation between $X$ and $Y$ (cities with more crime would have more police), i.e., OLS estimates a positive slope that falsely suggests a positive effect.

There are methods like instrumental variables that can (sometimes) solve the problem of simultaneity or reverse causality, but they are beyond our scope. For now, you can just try to think critically about whether or not simultaneity or reverse causality is a problem in real-world examples.

**Discussion Question 12.6** (health and medical expenditure). You want to learn the causal effect of how much an individual spends on medical insurance and care ($X$, dollars per year) on health ($Y$, higher value means healthier).

a) Explain why a regression of $Y$ on $X$ would not estimate this causal effect.

b) Would the regression slope be higher or lower than the causal effect? Why?
Empirical Exercise EE12.1. You will analyze data from Rouse (1998) on a “school voucher” program in Milwaukee, Wisconsin. As Rouse (1998) explains, “In 1990 Wisconsin began providing vouchers to a small number of low-income students to attend nonsectarian private schools.” Wooldridge notes that many observations with missing data have already been dropped, so there is sample selection. He also notes you can use variable mnce90 to try to control for this, but mnce90 is missing for 2/3 students, so then there’s a missing data problem, too. If everything were perfect, the estimated ATE of eligibility (binary variable select) shouldn’t depend too much on the control variables or the subsample of individuals; but clearly it does.

a. Load and see a description of the data.
   R: library(wooldridge) and ?voucher
   Stata: use http://faculty.missouri.edu/kaplandm/intro_text/voucher, clear describe
b. R only: copy the dataset into data frame df with df <- voucher
c. Display the total number of observations (rows) in the dataset.
   R: nrow(df)
   Stata: count
d. Display summary statistics of mnce90 and mnce, including the number of missing observations.
   R: summary(df[,c('mnce','mnce90')])
   Stata: count if missing(mnce90) and summarize mnce mnce90
e. Run a simple regression of mnce (the 1994 math test score) on select (the dummy variable for whether a child was ever allowed to use a voucher).
   R: (ret1 <- lm(mnce~select, data=df))
   Stata: regress mnce select , vce(robust)
f. Repeat but adding the 1990 math test score mnce90 as a regressor. Also, compare the number of observations used in the regression to the total number of observations in the dataset.
   R: (ret2 <- lm(mnce~select+mnce90, data=df)) and then length(ret2$residuals) or summary(ret2) to see the number of observations actually used.
   Stata: regress mnce select mnce90 , vce(robust) noting that observations with missing mnce90 are automatically (and silently) omitted from the regression, but the output shows the number of observations actually used, which you can compare to the number in the full dataset.
g. To try to see how much of the estimate's change is due to controlling for \texttt{mnce90} versus sample selection bias, re-run your first simple regression but with only the observations used in the second regression, i.e., only observations with non-missing \texttt{mnce90}.

R: \texttt{(ret2b <- lm(mnce~select, data=df[!is.na(df$mnce90),])})

Stata: \texttt{regress mnce select if !missing(mnce90), vce(robust)}

h. Repeat the above three regressions but with \texttt{selectyrs} (number of years eligible for voucher program) instead of the binary \texttt{select}

i. Repeat the first three regressions but with additional regressors like \texttt{female} to see if they further change the coefficient on \texttt{select}

\textbf{Empirical Exercise EE12.2.} You will analyze data from Card (1995), first seen in EE3.1, with individual-level observations of wages, years of education, and other variables. You’ll focus on the relationship between wage and education. The variable \texttt{IQ} seems like a helpful control variable, but it is not observed for all individuals, which may cause bias depending on why it is missing. You’ll estimate the coefficient on education with different sets of regressors and different subsets of data. You’ll also look at the difference it makes using the sampling weights (as you should).

a. Load and see a description of the data.

R: \texttt{library(wooldridge) and ?card}

Stata: \texttt{bcuse card, clear}

b. R only: copy the dataset into data frame \texttt{df} with \texttt{df <- card}

c. Display the total number of observations (rows) in the dataset.

R: \texttt{nrow(df)}

Stata: \texttt{count}

d. Show how many observations are missing \texttt{IQ}.

R: \texttt{table(is.na(df$IQ))}

Stata: \texttt{count if missing(IQ)}

e. Run a simple regression of log wage on years of education.

R: \texttt{(ret1u <- lm(log(wage)~educ, data=df))}

Stata: \texttt{regress lwage educ, vce(robust)}

f. Run the same regression but with the provided weights.

R: \texttt{(ret1w <- lm(log(wage)~educ, data=df, weights=weight))}

Stata: \texttt{regress lwage educ [pweight=weight], vce(robust)}
g. Run the same simple weighted regression but with the subset of observations for which IQ is observed.

R: replace \texttt{df} with \texttt{df[!is.na(df$IQ),]}

Stata: add \texttt{if !missing(IQ)} after \texttt{educ} (with a space on either side)

h. Regress log wage on education and IQ (which automatically uses only observations where IQ is non-missing).

R: \texttt{(ret2w <- lm(log(wage)~educ+IQ, data=df, weights=weight))}

Stata: \texttt{regress lwage educ IQ [pweight=weight] , vce(robust)}

i. Repeat parts (f)–(h) but with additional regressors of your choice.
Part III

Time Series
Introduction

Time series data and models are considered in Part III. The focus is on forecasting, i.e., prediction of future values or events. Foundational concepts like stationarity, autocorrelation, and appropriately adjusted standard errors are introduced.

Related (free) material is from Diebold (2018b) and Hanck et al. (2018, Ch. 14). Chapter 1 in the DataCamp intro time series course is also free.
Chapter 13

Time Series: One Variable

⇒ Kaplan video: Chapter Introduction

Depends on: Chapters 2 and 3

Unit learning objectives for this chapter

13.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]
13.2. Identify and describe different components and properties of a time series [TLOs 2 and 3]
13.3. Interpret transformed and decomposed time series [TLOs 2 and 3]
13.4. In R (or Stata): estimate basic descriptions of a time series [TLO 7]
13.5. In R (or Stata): decompose a time series into different components [TLO 7]

Optional resources for this chapter

- Deterministic and stochastic trends (Lambert video)
- Chapter 14 (“Time Series”) in Hansen (2020)
- Transformations: Section 3.2 (“Transformations and adjustments”) in Hyndman and Athanasopoulos (2019)
- Seasonality and holidays: Section 5.4 (“Some useful predictors”) in Hyndman and Athanasopoulos (2019)
Chapter 13 extends Chapter 2 to the time series setting. New concepts like stationarity and autocorrelation are introduced. There are even new complications just with estimating a variable’s mean and computing a standard error.

13.1 Terms and Notation

A time series of a single variable is written as $Y_t$ for time periods $t = 1, \ldots, T$. For example, $Y$ could be annual GDP of the U.S., with $t = 1$ indicating the year 2001 and $T = 10$ indicating a total of ten years of data (here 2001, 2002, \ldots, 2010). Or, $Y$ could be quarterly GDP from 2001Q1 (year 2001, quarter 1) through 2010Q4, a total of $T = 40$ periods where $t = 1$ is 2001Q1, $t = 2$ is 2001Q2, $t = 9$ is 2003Q1, etc. Or, $Y$ could be the weekly return on a certain stock observed over a single calendar year, $t = 1, \ldots, 52$.

In practice, there are many possible complications with timing and measurement, although details are beyond our scope. First, instead of “discrete time” periods $t = 1, \ldots, T$, “continuous time” models let $t$ be any real (decimal) number, not just integers. Second, even with discrete time, the periods may be of different lengths. Third, even with equal discrete periods, it is important to know precisely when and how the “time $t$” observation is measured. For example, imagine annual data, where $t$ represents an entire year. Is $Y_t$ measured on January 1 of year $t$? Or December 31? Or is $Y_t$ the average value across the entire year? Such timing is particularly important when analyzing multiple time series. For example, if $Y_t$ is measured on January 1 of year $t$, but $X_t$ is measured on December 31, then $X_t$ is measured 364 days after $Y_t$ but only 1 day before $Y_{t+1}$.

Similar to physics, the sampling frequency is the inverse of the length of each time period. For example, if each period is one year,
then there is one observation per year, so the sampling frequency is yearly (or “annual”). If each period is one quarter, then the sampling frequency is quarterly. Similarly, time series can be monthly, weekly, daily, or even hourly or higher frequency (like for stock prices, website traffic, energy use, etc.).

The following terms describe relationships among observations. Relative to $Y_t$, the **first lag** (or first lagged value) is $Y_{t-1}$, i.e., the value from the immediately prior period. Similarly, the second lag is $Y_{t-2}$, and generally the $j$th lag is $Y_{t-j}$. The **first difference** is

$$\Delta Y_t \equiv Y_t - Y_{t-1}. \tag{13.1}$$

(But “second difference” does not refer to $Y_t - Y_{t-2}$.) Looking to the future, $Y_{t+1}$ is the **first lead** (of $Y_t$), and $Y_{t+j}$ is the $j$th lead. In many cases, modeling the relationship between $Y_{t+1}$ and $Y_t$ is equivalent to modeling $Y_t$ and $Y_{t-1}$, for example. If we use observations $Y_1, \ldots, Y_T$ for estimation, then anything in the period $t = 1, \ldots, T$ is called **in-sample**, as opposed to $t = T+1, T+2, \ldots$, which is **out-of-sample**. Sometimes, fewer than $T$ observations are used for estimation, and the definitions are adjusted accordingly (Section 15.2).

### 13.2 Populations, Randomness, and Sampling

We continue the perspective of $Y_t$ as a random variable, just as $Y_i$ was earlier (Sections 2.1 and 2.3). Earlier, $Y_i$ was “random” since we could have sampled a different value from the population. But, what is the “population” for a time series?

One view is like the superpopulation from Section 2.2. That is, we can imagine many (infinite) possible universes. In each, there are the same mechanisms underlying how the time series values are generated, but the actual numerical values differ across universes. Like before, $E(Y_t)$ is the average of the $Y_t$ values across all the different universes. Similarly, $\text{Var}(Y_t)$ is the variance across universes. Measures like $\text{Corr}(Y_t, Y_{t+1})$ show whether $Y_t$ and $Y_{t+1}$ tend to both be high (or low), or opposite, or unrelated. For example, maybe GDP growth is high in both 2018 and 2019 in many universes, and low in both in other universes, but very few universes have high growth in 2018 and low in 2019, or low and then high. Then, in the (super)population, $\text{Corr}(Y_{2018}, Y_{2019}) > 0$.

Another view is that we observe a sequence of $T$ values within an infinitely long sequence of $Y_t$. We could think about, e.g., what the sample average would be if we had a very long sequence, or other “asymptotic” properties.

All that said, there are only a few chapters left, so in order to focus on practical descriptions and predictions (forecasts), these “deeper” issues are not explored further.
13.3 Stationarity

Will the future be like the past? This question arose in Section 12.2, on external validity. Here, “be like” is formalized in terms of probability distributions.

A time series \( Y_t \) is \textbf{stationary} if its future is like its past, probabilistically. A necessary (but not sufficient) aspect of this is \( E(Y_t) = E(Y_s) \) for any time periods \( t \) and \( s \): the mean never changes. Likewise, the median never changes, nor the standard deviation; the entire (marginal) distribution of \( Y_t \) is identical to that of \( Y_s \). Further, the relationship between this time period and next period must be stable over time, i.e., the joint distribution of \( (Y_t, Y_{t+1}) \) is identical for all \( t \). Similarly, the joint distribution of the previous, current, and next periods’ values, \( (Y_{t-1}, Y_t, Y_{t+1}) \), never changes. In full, stationarity is defined as the joint distribution of \( (Y_{t-J}, \ldots, Y_t, Y_{t+1}) \) not depending on \( t \), for any \( J \).

The foregoing describes \textbf{strict stationarity} (also called \textbf{strong stationarity}); a “weaker” concept called \textbf{covariance stationarity} (also called \textbf{wide-sense stationarity} or \textbf{weak-sense stationarity}) requires only the means and autocovariances (Section 13.4) to be the same at all \( t \), not the full joint distributions. Technically, it is not “weaker” in the logical sense (Section 6.1.1) because of weird distributions whose mean is undefined (e.g., Cauchy), but if you assume \( Y_t \) has finite variance, then strict (strong) stationarity implies covariance (weak) stationarity. That is, given finite variance, all strictly stationary series are also covariance stationary, but some covariance stationary series are not strictly stationary.

With either type of stationarity, an estimate of \( E(Y_t) \) from historical data can be interpreted as an estimate of the future \( E(Y_{T+1}) \), which is the (unconditional) best prediction of \( Y_{T+1} \) under quadratic loss. Stationarity essentially assumes external validity over time, allowing us to extrapolate the past into the future. In Chapters 14 and 15, we’ll improve upon the unconditional forecast by incorporating other information, but stationarity (and its variations) remain important considerations for external validity.

In practice, you should not blindly assume stationarity, but examine it empirically and economically. That is, you can look at the data to see if it appears stationary, and you can also think about what is happening in the world now that may change the future behavior. A previously stationary time series may no longer be stationary if there is a sudden law change or other event with permanent effect.

Section 13.6 contains more on data that’s nonstationary, i.e., not stationary.

13.4 Autocovariance and Autocorrelation

⇒ Kaplan video: \textbf{Autocorrelation}
An important feature of a time series is the correlation between this period’s value and last period’s value, i.e., between \( Y_t \) and \( Y_{t-1} \). This correlation is called the first **autocorrelation** or **serial correlation**.

The first autocorrelation can be positive, negative, or zero. For example, if today’s price change is not systematically related to yesterday’s price change, then the time series of price changes has zero autocorrelation. If high quarterly GDP growth follows high growth, and low follows low, rather than jumping around randomly each quarter, then GDP growth has a positive autocorrelation. Conversely, negative first autocorrelation implies high values are followed by low values, and low by high, more often than high following high or low following low. In economics, positive autocorrelation is most common.

The sampling frequency affects the first autocorrelation. Generally, first autocorrelations are closer to positive one with high frequency and closer to zero with low frequency. For example, today’s U.S. unemployment rate will be extremely close to yesterday’s rate, so the first autocorrelation is near one with daily data. However, with yearly data (lower frequency), the first autocorrelation is lower. If each period is one decade (even lower frequency), then the first autocorrelation may be near zero.

Generally, for a stationary series, the \( j \)th autocorrelation (or \( j \)th **autocorrelation coefficient** \( \rho_j \)) describes the relationship between \( Y_t \) and \( Y_{t-j} \), as does the related \( j \)th **autocovariance** \( \gamma_j \). Stationarity implies these values do not vary with \( t \), only \( j \) (the lag). Consequently, it is the same (statistically) if we look \( j \) periods in the past or \( j \) periods in the future, since period \( t-j \) is \( j \) periods before \( t \) just as \( t \) is \( j \) periods before \( t+j \), and \( \text{Cov}(W,Z) = \text{Cov}(Z,W) \).

Mathematically,

\[
\gamma_j \equiv \text{Cov}(Y_t, Y_{t-j}) = \text{Cov}(Y_{t+j}, Y_t) = \gamma_{-j},
\]

\[
\rho_j \equiv \text{Corr}(Y_t, Y_{t-j}) = \text{Corr}(Y_{t+j}, Y_t) = \rho_{-j},
\]

\[
\gamma_0 \equiv \text{Cov}(Y_t, Y_t) = \text{Var}(Y_t) = \sigma^2_Y, \quad \rho_0 = \text{Corr}(Y_t, Y_t) = 1,
\]

\[
\rho_j \equiv \text{Corr}(Y_t, Y_{t-j}) = \frac{\text{Cov}(Y_t, Y_{t-j})}{\sqrt{\text{Var}(Y_t)\text{Var}(Y_{t-j})}} = \frac{\gamma_j}{\sigma^2_Y} = \frac{\gamma_j}{\gamma_0}.
\]

In (13.5), the denominator simplifies because stationarity implies \( \text{Var}(Y_{t-j}) = \sigma^2_Y \), and \( \sigma^2_Y = \gamma_0 \) from (13.4):

\[
\sqrt{\text{Var}(Y_t)\text{Var}(Y_{t-j})} = \sqrt{\sigma^2_Y\sigma^2_Y} = \sigma^2_Y = \gamma_0.
\]

Although sometimes autocovariances are more convenient mathematically, autocorrelations are easier to interpret. The units of autocovariance are the square of the units of \( Y_t \) (like “squared dollars”), which is difficult to interpret. The autocorrelation does not depend on the units of \( Y_t \) and has the same interpretation as a correlation, where possible values are between \(-1\) (perfect negative linear correlation) and \(+1\) (perfect positive linear correlation). The usual caveats about interpreting correlation (nonlinearity, causality, magnitude of change, etc.) apply equally to autocorrelation.\(^1\)

\(^1\)E.g., https://en.wikipedia.org/wiki/Correlation_and_dependence
Discussion Question 13.1 (autocorrelation). Do you think $\rho_1 > 0$, $\rho_1 \approx 0$, or $\rho_1 < 0$ for the following?
   a) An individual’s employment status ($Y_t = 1$ if employed at time $t$, otherwise $Y_t = 0$), observed weekly.
   b) GDP growth, quarterly.
   c) GDP growth, annual.
   d) Stock market (S&P 500 index) price change, annual.

13.5 Estimation

13.5.1 Mean

With stationarity, the mean is the same $\mu = E(Y_t)$ for all $t$, so intuition suggests the sample mean may still be a good estimator. Indeed, it often is, although consistency requires additional technical conditions. For example, if the joint distributions are Gaussian and the autocorrelations are zero at very distant lags (or $\rho_j \to 0$ as $j \to \infty$), then the sample mean is consistent,

$$\frac{1}{T} \sum_{t=1}^{T} Y_t \overset{p}{\to} \mu.$$  \hspace{1cm} (13.6)

See DasGupta (2008, p. 40) for more, and Proposition 7.5 in Hamilton (1994) for slightly different sufficient conditions.

13.5.2 Autocovariances and Autocorrelations

With stationarity, similarly, autocovariances and autocorrelations do not depend on $t$, so the sample autocovariances and sample autocorrelations are reasonable estimators.

To estimate the $j$th autocovariance $\text{Cov}(Y_t, Y_{t-j})$, the sample $j$th autocovariance is

$$\hat{\gamma}_j = \hat{E}[(Y_t - \hat{E}(Y))(Y_{t-j} - \hat{E}(Y))] = \frac{1}{T} \sum_{t=1+j}^{T} (Y_t - \bar{Y})(Y_{t-j} - \bar{Y}).$$  \hspace{1cm} (13.7)

This estimator is often consistent, meaning $\hat{\gamma}_j \overset{p}{\to} \gamma_j$ for a given $j$ as $T \to \infty$. For example, see equation [7.2.15] in Hamilton (1994). Using $j = 0$ estimates the variance, $\gamma_0 = \sigma_Y^2$.

You can check that R’s `acf()` function uses the formula in (13.7).

The data are $Y_1 = -1$, $Y_2 = -1$, $Y_3 = 1$, and $Y_4 = 1$, so $T = 4$. The sample average is zero: $\bar{Y} = ((-1) + (-1) + 1 + 1)/4 = 0$. Thus, (13.7) simplifies to

$$\hat{\gamma}_j = (1/4) \sum_{t=1+j}^{4} Y_t Y_{t-j},$$

$$\hat{\gamma}_0 = (1/4)(Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2) = (1/4)(4) = 1,$n
$$\hat{\gamma}_1 = (1/4)(Y_2 Y_1 + Y_3 Y_2 + Y_4 Y_3) = (1/4)(1 + (1) + 1) = 1/4 = 0.25,$n
$$\hat{\gamma}_2 = (1/4)(Y_3 Y_1 + Y_4 Y_2) = (1/4)((-1) + (1)) = -2/4 = -0.5,$n
$$\hat{\gamma}_3 = (1/4)(Y_4 Y_1) = -1/4 = -0.25.
These match the following output.

\[
c(acf(x=c(-1,-1,1,1), type='covariance', plot=FALSE)$acf)
\]

## [1] 1.00 0.25 -0.50 -0.25

Since \( \rho_j = \gamma_j / \gamma_0 \) as in (13.5), the estimated autocorrelations are

\[
\hat{\rho}_j = \hat{\gamma}_j / \hat{\gamma}_0.
\]  

(13.8)

**Difficulties**

There are limits to what we can learn from data. In (13.7), there are only \( T - j \) terms being averaged. In the extreme, if \( j = T - 1 \), then there is only a single term in the average, like in the above example of \( \hat{\gamma}_3 \) with \( T = 4 \). Intuitively, an average of a single number is a bad estimator; arguments about consistency of averages assume a large number of values being averaged. If \( T \) is large, then we can learn about \( \gamma_1 \) very well, but we cannot learn about \( \gamma_j \) for large \( j \) (near \( T \)).

Even in more complex models, estimation is difficult if the \( t = 1 \) and \( t = T \) variables are strongly correlated.

**Code**

In R, \texttt{acf()} estimates autocovariances and autocorrelations, as seen in the following code using monthly international airline passenger data. The result \( \hat{\rho}_{12} > \hat{\rho}_6 \) seems surprising at first: \( Y_t \) is more strongly correlated with \( Y_{t-12} \) than \( Y_{t-6} \), even though \( Y_{t-6} \) is closer in time. However, these are monthly data with strong seasonality (Section 13.6.2), so the fact that \( t-12 \) is the same calendar month as \( t \) causes stronger correlation than with \( t-6 \), which is a very different season (e.g., \( t-6 \) is summer if \( t \) is winter). To see a graph, run the code yourself with \texttt{plot=TRUE} instead of \texttt{FALSE}.

\[
\text{retcorr <- acf(AirPassengers, lag.max=12, type='correlation', plot=FALSE, ci.type='ma')}
\]

\[
\text{retcov <- acf(AirPassengers, lag.max=12, type='covariance', plot=FALSE, ci.type='ma')}
\]

\[
\text{print(data.frame(lagmonth=0:12, rho.j=round(retcorr$acf,digits=2), gamma.j=round(retcov$acf,digits=0))), row.names=F)
\]

## lagmonth rho.j gamma.j
## 0 1.00 14292
## 1 0.95 13549
## 2 0.88 12514
## 3 0.81 11529
## 4 0.75 10757
## 5 0.71 10201
## 6 0.68 9743
## 7 0.66 9474
13.6 Nonstationarity

⇒ Kaplan video: Nonstationarity

This section describes the most common reasons a time series is nonstationary, i.e., not stationary.

In Sum: Reasons for Nonstationarity

- Stochastic trend (unit root; e.g., random walk): variance increasing over time
- Deterministic trend: mean changing over time
- Seasonality: mean changing over time (repeating up-and-down pattern)
- Cycles: up-and-down patterns without fixed frequency
- Breaks: permanent changes

13.6.1 Trends

Stochastic Trends

A random walk as in (13.9) generates nonstationary $Y_t$. This is a special case of a more general unit root process, which all share qualitatively similar properties (including nonstationarity). It is also sometimes called a stochastic trend. Let $Y_0$ be the initial value. Let

$$Y_t = Y_{t-1} + \epsilon_t,$$  \hspace{1cm} (13.9)

where the increments $\epsilon_t$ are iid, mean zero, and independent of all past values $Y_s$ for $s \leq t - 1$; i.e., the $\epsilon_t$ are independent white noise (Section 14.1).

One way to see the nonstationarity is that

$$\text{Var}(Y_t) = \text{Var}(Y_{t-1} + \epsilon_t) = \text{Var}(Y_{t-1}) + \text{Var}(\epsilon_t) + 2 \text{Cov}(Y_{t-1}, \epsilon_t) > \text{Var}(Y_{t-1})$$  \hspace{1cm} (13.10)

violating the property of stationarity that the variance is the same at all $t$. Logically, stationarity implies same variance at all $t$, so by the contrapositive, different variance at difference $t$ implies nonstationarity.

For prediction, given (13.9), the “best” guess (under quadratic loss) of next period’s $Y_{t+1}$ is the current period’s $Y_t$. Because $\epsilon_{t+1}$ is
mean zero and independent of \( Y_t \),

\[
E(Y_{t+1} \mid Y_t = y_t) = E(Y_t + \epsilon_{t+1} \mid Y_t = y_t) = E(Y_t \mid Y_t = y_t) + E(\epsilon_{t+1} \mid Y_t = y_t) = y_t + E(\epsilon_{t+1}) = y_t + 0 = y_t.
\] (13.11)

From (6.30), the conditional mean is the best predictor under quadratic loss, so the current value \( y_t \) is the best predictor of next period’s value \( Y_{t+1} \). This remains the best predictor even if instead of only \( y_t \) we know the entire history \( y_t, y_{t-1}, y_{t-2}, \ldots \) (mathematically, if we condition on these realizations).

Another interpretation of (13.9) is that \( Y_t \) contains all the relevant historical information about the future \( Y_{t+1} \). Additionally knowing \( Y_{t-1} \) or other past values does not help. Thus, the random walk has the Markov property (or is a Markov chain): all “information” about future values is contained in the current value, and additionally knowing past values adds no new information.

Although nonstationary, the random walk can be transformed into a stationary process by taking a first difference (Section 13.1). Subtracting \( Y_{t-1} \) from both sides of (13.9),

\[
Y_t - Y_{t-1} = Y_{t-1} + \epsilon_t - Y_{t-1} = \epsilon_t,
\] (13.12)

and \( \epsilon_t \) is iid, which is a special case of stationarity. Generally, when a first difference of a time series produces a stationary series, the original time series is called difference stationary.

Deterministic Trends

With a deterministic trend, the time series goes up (or down, or up and down) in a non-random pattern. For example, imagine \( Y_t = t + \epsilon_t \), where \( \epsilon_t \) are mean-zero iid variables. Then,

\[
E(Y_t) = E(t + \epsilon_t) = t + E(\epsilon_t) = t + 0 = t,
\]

which changes with \( t \), violating stationarity. Analogous to difference stationarity, a time series is trend stationary if removing its deterministic trend produces a stationary series. In the above example, \( Y_t \) is trend stationary because \( Y_t - t = \epsilon_t \), which is stationary.

Distinguishing Trend Types

Despite their seeming so different, in practice it can be difficult to distinguish a stochastic trend from a deterministic trend. For example, in climate econometrics,\(^2\) there is ongoing debate about whether the earth’s temperature currently has a stochastic trend or a deterministic trend that changed at some point in the past; e.g., see Kaufmann,

\[^2\]Although not exactly climate econometrics, half the 2018 Nobel Prize was awarded to William Nordhaus “for integrating climate change into long-run macroeconomic analysis”; see https://www.nobelprize.org/prizes/economic-sciences/2018/press-release/
However difficult, it is important to distinguish stochastic and deterministic trends because they affect forecasts. Roughly, a trend stationary time series is expected to return to its deterministic trend line relatively quickly, whereas the stochastic trend makes deviations more persistent. For example, if we knew $Y_t = t + \epsilon_t$ with mean-zero iid $\epsilon_t$, and we observed $Y_t$ values 1.09, 1.98, 3.05, 4.6 for $t = 1, 2, 3, 4$, the best forecast of $Y_5$ is $\hat{Y}_5 = 5$, even though $Y_4 = 4 + 0.6$ was well above the trend line. In contrast, with a stochastic trend, we’d expect the effect of $\epsilon_4 = 0.6$ to persist in $t = 5$, so our forecast would be higher than 5. Specifically, if $Y_t$ is difference stationary with $Y_t - Y_{t-1} = 1 + \epsilon_t$ (with mean-zero iid $\epsilon_t$), then

$$
E(Y_5 | Y_4 = 4.6) = E(Y_4 + 1 + \epsilon_5 | Y_4 = 4.6) = 4.6 + 1 = 5.6. \quad (13.13)
$$

Diebold (2018c, §8.1) shows a similar example with U.S. gross national product.

### 13.6.2 Seasonality

A time series with seasonality tends to have higher values in some seasons of the year than in others. For example, retail sales are highest near the Christmas holiday season, and some agricultural crops are only harvested in one season. Residential energy use is also seasonal, with the most heating in the winter and most cooling in the summer, and lowest energy use in the fall and spring. The pattern of seasonality may vary by location, too: energy use may be highest in winter in colder places like Montana, but highest in summer in warmer places like Louisiana. Many other variables show seasonality, too, either due to human-imposed seasons (holidays, school schedules, elections, etc.) or natural seasons (weather, crops, sunlight, etc.).

Seasonality is more general than seasons within a calendar year. For example, restaurant dinner sales are higher on Friday and Saturday than other days of the week. Crime rates fluctuate with the day of the week and even the hour of the day, as do things like electricity usage. There can also be “seasons” like Congressional elections that occur only every two years (or longer).

The presence of seasonality depends on the length of time period, too. For example, if $Y_t$ is retail sales in year $t$, then seasonality won’t matter because all seasons are lumped into a single $t$. However, if $t$ is quarterly, then seasonality appears; e.g., $Y_t$ always jumps up during the fourth quarter (October, November, December). If $t$ is divided into even shorter periods, then seasonality is still seen: with monthly data, $Y_t$ jumps up in December, or with weekly data, $Y_t$ jumps up in the weeks leading up to Christmas.

Some “seasons” are not actually seasons with a fixed frequency, so they must be handled differently. For example, the calendar date of Easter differs from year to year. For forecasting regression models,
13.6. NONSTATIONARITY

You can add dummy variables for such events. For Easter specifically, the function `easter()` in the `forecast` package is helpful.

Figure 13.1 illustrates how seasonality can be seen in plots of $Y_t$ over $t$ that show an up-and-down pattern that repeats every year (or other period). The left graph is from `plot(AirPassengers)` and shows monthly numbers of international airline passengers (in thousands). There is a clear up-and-down seasonal pattern that repeats every year. You can also try using `seasonplot(AirPassengers)`, a function in the `forecast` package (Hyndman et al., 2020; Hyndman and Khandakar, 2008).

![AirPassengers graph](image)

![log(AirPassengers) graph](image)

Figure 13.1: Seasonality in international air travel.

The right graph of Figure 13.1 is from `plot(log(AirPassengers))` and shows $\ln(Y_t)$ against $t$. Although both show seasonality, the peak-to-trough magnitude (height) of the seasonal variation is more constant every year for $\ln(Y_t)$; see Section 13.7.

13.6.3 Cycles

What about up-and-down patterns caused by macroeconomic business cycles, or El Niño–Southern Oscillation cycles? Cycles are often important but more difficult to understand. One added difficulty is the unknown and changing length of cycles; e.g., El Niño does not come precisely every five years, nor is there a recession every five years. Here, like in Hyndman and Athanasopoulos (2019, §6), the “trend” is actually a **trend–cycle component** that includes cycles,
13.6.4 Structural Breaks

Sometimes there are big, permanent changes in the world, and the properties of a time series also change permanently. This is often called a structural break. For example, in the U.S., many macroeconomic time series look very different before and after 1985; in particular, the reduction in volatility led to the term “Great Moderation.”

Dealing with breaks is beyond our scope, but they are important to be aware of; see also Section 14.5.

13.7 Decomposition

The observed time series $Y_t$ can be written in terms of unobserved components of “trend” (really trend–cycle), seasonality, and a remainder (Diebold, 2018b, §2.10). The remainder, also called the random or irregular or residual or noise component, is what remains of $Y_t$ after removing the trend and seasonality.

Notationally, following Hyndman and Athanasopoulos (2019, §6), let $T_t$ denote trend, $S_t$ seasonality, and $R_t$ remainder. Then,

$$R_t \equiv Y_t - T_t - S_t \implies Y_t = T_t + S_t + R_t.$$  

(13.14)

This is an additive decomposition: $Y_t$ is “decomposed” into additive trend, seasonality, and remainder components, which all have the same units as $Y_t$.

Alternatively, a multiplicative decomposition is

$$Y_t = T_t \times S_t \times R_t.$$  

(13.15)

Now, $T_t$ still has the same units as $Y_t$, but $S_t$ and $R_t$ represent percentage deviations from the trend. For example, $S_t = 1.05$ means 5% higher, or $R_t = 0.85$ means 15% lower. (Often a percentage seasonal component makes more sense.) Actually, taking the log of both sides of (13.15) yields an additive model:

$$\ln(Y_t) = \ln(T_t) + \ln(S_t) + \ln(R_t).$$  

(13.16)

Finally, sometimes the decomposition is a mix: $Y_t = T_t \times S_t + R_t$.

There are R functions to decompose time series into trend, seasonal, and remainder components. To choose the right method, you must decide whether the seasonality is additive or multiplicative. For example, compared to sales on July 1, are sales on December 1 usually higher by $500 (additive), or by 30% (multiplicative)? In other words, is (13.14) or (13.15) more sensible?

---

3See https://en.wikipedia.org/wiki/Great_Moderation
For intuition, the following roughly describes a **classical additive decomposition** (Hyndman and Athanasopoulos, 2019, §6.3). First, the trend is estimated, usually by some nonparametric smoother, yielding the estimated trend $\hat{T}_t$. Second, the “seasonal” averages of $Y_t - \hat{T}_t$ (the detrended data) are computed. For example, with monthly data, all January values of $Y_t - \hat{T}_t$ are averaged to estimate $\hat{S}_t$ when $t$ is in January, and then all February values are averaged to get $\hat{S}_t$ for February $t$, etc. Third, $\hat{R}_t = Y_t - \hat{T}_t - \hat{S}_t$. There are many variations, with different estimators of $\hat{T}_t$, or allowing $\hat{S}_t$ to change over time. For **multiplicative decomposition**, either apply the above to $\ln(Y_t)$, or replace subtraction with division: use $Y_t/\hat{T}_t$ in the second step, and $Y_t/(\hat{T}_t\hat{S}_t)$ in the third step.

Figure 13.2 shows an additive decomposition produced by the following R code that uses `decompose()` (in the built-in `stats` package).

```r
par(family='serif', mgp=c(2.1,0.8,0))
ret <- decompose(co2, type='additive')
plot(ret)
```

Figure 13.3 shows a multiplicative decomposition generated by the following R code. When seasonality is multiplicative instead of
additive, specify `type='multiplicative'` as below.

```r
par(family='serif', mgp=c(2.1,0.8,0))
ret <- decompose(AirPassengers, type='multiplicative')
plot(ret)
```

Other R decomposition functions to try (or Google) include `stl()`, `HoltWinters()`, and the `forecast` package’s `mstl()` (multiple seasonal).

**Discussion Question 13.2** (nonstationarity). Explain specifically why you doubt the strict stationarity of each of the following time series: a) GDP, annual; b) stock market index, annual; c) world population, annual; and d) residential water usage, monthly.

### 13.8 Transformations

To improve interpretation or statistical properties, it may help to transform a time series before analyzing it. Three common transformations are now briefly discussed.
First, the first difference looks at changes in $Y_t$, defined in (13.1) as $\Delta Y_t \equiv Y_t - Y_{t-1}$. One motivation is Section 13.6: some nonstationary $Y_t$ are difference stationary, so $\Delta Y_t$ is stationary. For example, if $Y_t = Y_{t-1} + U_t$, where $U_t$ is iid, then $Y_t$ is a random walk and thus nonstationary. However, $\Delta Y_t = U_t$ is iid, which is stationary. Methods that only work with stationary data could be applied to $\Delta Y_t$ but not $Y_t$.

Second, log transformations sometimes help, like in (13.16) where a multiplicative model becomes additive. That is, instead of $Y_t$, we analyze $Z_t = \ln(Y_t)$.

Third, taking a log difference $\ln(Y_t) - \ln(Y_{t-1})$ yields the compound growth rate. This is the first difference of the log-transformed series: letting $Z_t = \ln(Y_t)$, then $\Delta Z_t = Z_t - Z_{t-1} = \ln(Y_t) - \ln(Y_{t-1}) = \ln(Y_t/Y_{t-1})$. For example, the formula for the final level $A$ after continuously compounded growth at effective annual rate $r$ for $t$ years, starting at initial level $P$, is $A = Pe^{rt}$, the “Pert” formula you may have learned in high-school for computing compound interest rates. For a single year ($t = 1$ in the formula), the rate $r$ is then solved by $A = Pe^r$ implying $e^r = A/P$ and thus $r = \ln(A/P) = \ln(A) - \ln(P)$, using a log property (from Section 8.1.1) for the last equality. Thus, with annual data, the log difference $\ln(Y_t) - \ln(Y_{t-1})$ represents the effective annual rate.
Empirical Exercises

Empirical Exercise EE13.1. You will analyze monthly U.S. unemployment data. You’ll notice that the unemployment rate is not very seasonal (by month), but it is very persistent (positively autocorrelated). Note that \texttt{urate} is in percent units, so 5.2 means 5.2\%, etc.

a. Load and see a description of the data.
   
   R: \texttt{library(wooldridge) and ?beveridge}
   
   Stata: \texttt{bcuse beveridge, clear}

b. Tell your software that you have monthly time series data.
   
   R: \texttt{tsdat <- ts(data=beveridge$urate, frequency=12, start=c(2/zero.alt3/zero.alt3/zero.alt3,12))} creates a time series variable named \texttt{tsdat} that’s a time series (ts) with the unemployment rate data (urate) starting in year 2000 month 12 (the first value of \texttt{beveridge$month}). Argument \texttt{frequency=12} says there are 12 “seasons” before getting back to the first one; in this case, 12 different months per year. (Daily data could use \texttt{frequency=7} to allow day-of-week “seasonality.”)
   
   Stata: \texttt{tsset ym, monthly}

c. R only: decompose (additively) the unemployment rate time series into trend, seasonal, and remainder components with \texttt{tsdec <- decompose(tsdat)} to compute and \texttt{plot(tsdec)} to plot. You can also see that the magnitude of the seasonal component is relatively small with \texttt{max(abs(tsdec$seasonal))}

Stata only: to additively decompose the time series, first estimate the trend component with a nonparametric “moving average smoother” with command
   
   \texttt{tssmooth ma furate=urate, weights(1 2 2 2 2 2 <2> 2 2 2 2 2 1)}
   
   and plot this smoothed trend against the raw time series with \texttt{tsline urate furate, name(furate) ylabel(#3)}

d. Stata only: compute the seasonal effects by averaging the difference between the data and the trend within each month (e.g., average among all January values, then separately among all February values, etc.). Generate the month variable with \texttt{generate month = month(dofm(ym))} and compute the within-month averages with \texttt{bysort month : egen seasadd = mean(urate-furate)}

Stata only: normalize the seasonal effects to average to zero. Compute the average of the raw seasonal effects, and then subtract that value from the seasonal effects (to make them average to zero) with commands (note: the broken-up “line” \texttt{scalar normadd ... should all be on the same line of code})
sort ym
replace seasadd = seasadd - normadd
g. Stata only: see how big (or small) the seasonal effects are with commands
list month seasadd if year(dofm(ym))==year(dofm(ym [1]))+1
summarize seas , detail
h. Stata only: generate the remainder term as the raw data minus trend minus seasonality, with command generate remadd = urate - furate - seasadd
i. Stata only: plot the seasonal and remainder series, and then make a combined graph with everything (similar to what R shows):
tsline seasadd , name(seasadd) ylabel(#3)
tsline remadd , name(remadd) ylabel(#3)
graph combine furate seasadd remadd , cols(1) name( decompurateadd)
j. Plot the autocorrelation function (ACF) up to 48 months lag.
R: acf(tsdat, lag.max=48, ci=0)
Stata: ac urate , level(95) lags(48)
k. Display the autocorrelation values up to 24 months.
R: acf(tsdat, lag.max=24, type='correlation', plot=FALSE)
Stata: corrgram urate , lags(24) noplot
l. Repeat the decomposition plot and ACF plot for the vacancy rate variable vrate

Empirical Exercise EE13.2. You will analyze monthly data on industrial cement production from Shea (1993). If you’re curious, you can view and download more recent cement data from the Federal Reserve Bank of St. Louis. You’ll notice that seasonality is very important. You’ll also notice that the autocorrelations of the raw data reflect the up-and-down seasonality, whereas the autocorrelations of the seasonally-adjusted data show more consistently positive autocorrelation (up to two years lag or so).

a. Load and see a description of the data.
R: library(wooldridge) and ?cement
Stata: bcuse cement , clear

\footnote{https://fred.stlouisfed.org/series/IPG3273N}
b. Tell your software that you have monthly time series data.

R: use

```r
tmdat <- ts(data=cement$ipcem, frequency=12,  
             start=c(cement$year[1],cement$month[1]))
```

to create a time series variable named `tsdat` that’s a time series (`ts`) with the industrial cement production index data (`ipcem`).

Stata:

```stata
generate yrmo = ym(year, month)  
format yrmo %tm  
tsset yrmo
```

c. R only: compute, store, and plot a multiplicative decomposition, to see how important seasonality is for industrial cement production:

```r
tsdec <- decompose(tmdat, type='mult')  
plot(tsdec)  
window(tsdec$seasonal, start=c(1964,1), end=c(1964,12))
```

The last line above prints the numerical values for the seasonality plot (which are the same for each year; e.g., 1964 could be replaced by 1971).

d. Stata only: estimate the trend and plot it against the raw data:

```stata
tssmooth ma fipcem1=ipcem, weights(1 2 2 2 2 2 2 2 2 2 2 2 1)  
tsline ipcem fipcem1, name(fipcem1) ylabel(#3)
```

e. Stata only: compute multiplicative seasonal effects with

```stata
bysort month : egen seasmult = mean(ipcem/fipcem1)  
(but don’t worry about normalizing these to average to 1 like is sometimes done)
```

f. Stata only: compute the multiplicative remainder as the observed value divided by the trend value, divided yet again by the seasonal effect:

```stata
generate rem1mult = ipcem/fipcem1/seasmult
```

g. Stata only: plot the seasonal and remainder series, and then all series together (similar to the R plot):

```stata
tsline seasmult, name(seasmult) ylabel(#3)  
tsline rem1mult, name(rem1mult) ylabel(#3)  
graph combine fipcem1 seasmult rem1mult, cols(1)  
    name(decompmult)
```

h. Plot the autocorrelation function (ACF) of the raw data up to 48 months lag.

R: `acf(tmdat, lag.max=48, ci=0, na.action=na.omit)`

Stata: `ac ipcem , level(95) lags(48)`
i. Plot the ACF of the seasonally-adjusted data.

R: `acf(tsdat/tsdec$seasonal, lag.max=48, ci=0, na.action=na.omit)`

Stata:

`generate saipcem = ipcem / seasmult
ac saipcem, level(95) lags(48)`

j. Repeat the decomposition plot and ACF plots for a different variable in the dataset.
CHAPTER 13. TIME SERIES: ONE VARIABLE
Chapter 14

First-Order Autoregression

⇒ Kaplan video: Chapter Introduction

Depends on: Chapters 8 and 13 (which depend on Chapters 2–4, 6, and 7)

Unit learning objectives for this chapter

14.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]
14.2. Describe the first-order autoregressive model and its features, including interpretation for description and prediction [TLOs 2 and 3]
14.3. Interpret and evaluate forecasts, including multi-step and interval forecasts [TLOs 2 and 3]
14.4. In R (or Stata): estimate the parameters of a first-order autoregression [TLO 7]
14.5. In R (or Stata): generate interval and multi-step forecasts [TLO 7]

Optional resources for this chapter

- AR(1) (Lambert video)
- AR(1) series with different autocorrelations (Lambert video)
- Chapter 12 (“Serial Correlation”) in Diebold (2018a)
- AR(1) model and properties: Hamilton (1994, §3.4)
- Asymptotic theory: Hamilton (1994, §§8.2,17.4)
Decent forecasts are often achieved by simply regressing $Y_t$ on $Y_{t-1}$. Chapter 14 explores this model, which is also useful for description (if not causal inference). Some extensions are discussed, with additional extensions in Chapter 15.

Although you don’t need to be able to reproduce (or even understand) the mathematical derivations in this chapter, they are provided in case you’re interested.

14.1 Model

The first-order autoregressive model, or AR(1) model, is essentially a simple linear regression in which the regressor is the first lag of the outcome variable:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t,$$

where $\phi_0$ and $\phi_1$ are constant coefficients, with $\phi_1$ called the autoregressive parameter (or autoregressive coefficient), and $\epsilon_t$ is something called white noise. A special case called independent white noise is if the $\epsilon_t$ are iid, with mean zero and finite variance, and independent of all past $Y_s$ values for $s < t$:

$$\epsilon_t \sim \text{iid}, \quad E(\epsilon_t) = 0, \quad \sigma^2_t \equiv \text{Var}(\epsilon_t) < \infty, \quad \epsilon_t \perp \perp Y_{t-1}, Y_{t-2}, \ldots, \text{for all } t.$$  

Diebold (2018a, §13.6) and Diebold (2018b, §6.2) have many more details on white noise that are beyond our scope.

Given (14.1) and (14.2), stationarity (either type) of $Y_t$ depends on the parameter values. Specifically,

$$Y_t \text{ is stationary } \iff |\phi_1| < 1.$$  

If instead $|\phi_1|$, then $Y_t$ has a unit root (Section 13.6.1). For example, with $\phi_0 = 0$ and $\phi_1 = 1$, (14.1) becomes the random walk in (13.9).

“Explosive processes” with $|\phi_1| > 1$ are sometimes considered to model stock market bubbles (or other bubbles) but are beyond our scope.

Assuming stationarity (either type), the mean of $Y_t$ can be solved for in terms of parameters $\phi_0$ and $\phi_1$. Let $\mu \equiv E(Y_t)$, which is the same for all $t$ if $Y_t$ is stationary. Using (14.1),

$$\mu = E(Y_t) = E(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t) = \phi_0 + \phi_1 E(Y_{t-1}) + E(\epsilon_t) = \phi_0 + \phi_1 \mu.$$  

(14.4)
Solving $\mu = \phi_0 + \phi_1 \mu$ for $\phi_0$ and then $\mu$,

$$\phi_0 = \mu (1 - \phi_1), \quad \mu = \frac{\phi_0}{1 - \phi_1}. \quad (14.5)$$

The AR(1) model in (14.1) can be written equivalently in terms of demeaned values. Generally, a **demeaned** random variable has had its mean subtracted (like a “deboned” fish has had its bones removed), so it has mean zero, like the population mean model’s error term defined in (6.4). Here, $Y_t - \mu$ is demeaned since $E(Y_t) = \mu$, so

$$E(Y_t - \mu) = E(Y_t) - \mu = 0.$$ 

Similarly, since $\mu = E(Y_{t-1})$, $E(Y_{t-1} - \mu) = E(Y_{t-1}) - \mu = 0$, and similarly $E(Y_{t-j} - \mu) = 0$ for all $j$ because all $E(Y_{t-j}) = \mu$ due to stationarity.

The demeaned AR(1) model is

$$Y_t - \mu = \phi_1 (Y_{t-1} - \mu) + \epsilon_t. \quad (14.6)$$

This is equivalent to (14.1). After adding $\mu$ to both sides of (14.6),

$$Y_t = \mu + \phi_1 (Y_{t-1} - \mu) + \epsilon_t = \mu + \phi_1 Y_{t-1} - \phi_1 \mu + \epsilon_t = \mu (1 - \phi_1) + \phi_1 Y_{t-1} + \epsilon_t \quad (14.7)$$

### 14.2 Description

Certain properties of $Y_t$ are implied by (14.1) and (14.2). Here, we look at the mean, variance, autocovariances, and autocorrelations of $Y_t$ in terms of the model parameters.

The mean is $\mu = \phi_0 / (1 - \phi_1)$ given covariance stationarity, as shown in (14.5).

The variance is derived by taking the variance of each side of (14.1). Assuming covariance stationarity, let $\sigma_Y^2 \equiv \text{Var}(Y_t) = \text{Var}(Y_{t-1})$. Using variance identities and $\text{Cov}(Y_{t-1}, \epsilon_t) = 0$ (since $\epsilon_t \perp Y_{t-1}$),

$$\text{Var}(Y_t) = \text{Var}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t) = \text{Var}(\phi_1 Y_{t-1} + \epsilon_t) = \varphi_1^2 \text{Var}(Y_{t-1}) + \sigma_\epsilon^2 + 2 \phi_1 \text{Cov}(Y_{t-1}, \epsilon_t) = \varphi_1^2 \sigma_Y^2 + \sigma_\epsilon^2.$$ 

Rearranging to solve for $\sigma_Y^2$,

$$\sigma_Y^2 = \varphi_1^2 \sigma_Y^2 + \sigma_\epsilon^2 \implies \sigma_Y^2 (1 - \varphi_1^2) = \sigma_\epsilon^2 \implies \sigma_Y^2 = \frac{\sigma_\epsilon^2}{1 - \varphi_1^2}. \quad (14.8)$$
The autocovariances can also be calculated given covariance stationarity. Substituting for $Y_t$ using (14.1), and using the same properties from above,

\[
\gamma_1 \equiv \text{Cov}(Y_t, Y_{t-1}) = \text{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-1})
\]

= 0 since $\phi_0 = \text{const}$

= Cov($\phi_0, Y_{t-1}$) + Cov($\phi_1 Y_{t-1}, Y_{t-1}$) + Cov($\epsilon_t, Y_{t-1}$) = $\phi_1$ Cov($Y_{t-1}$)

= $\phi_1 \sigma_Y^2$.

(14.9)

Using (14.9) recursively,

\[
\gamma_2 \equiv \text{Cov}(Y_t, Y_{t-2}) = \text{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-2})
\]

= 0 since $\phi_0 = \text{const}$

= Cov($\phi_0, Y_{t-2}$) + $\phi_1$ Cov($Y_{t-1}, Y_{t-2}$) + Cov($\epsilon_t, Y_{t-2}$)

= $\phi_1 \gamma_1 = \phi_1 \phi_1 \sigma_Y^2$

= $\phi_1^2 \sigma_Y^2$.

(14.10)

More generally, by induction, if $\gamma_{j-1} = \phi_1^{j-1} \sigma_Y^2$, then

\[
\gamma_j \equiv \text{Cov}(Y_t, Y_{t-j}) = \text{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-j})
\]

= 0 since $\phi_0 = \text{const}$

= Cov($\phi_0, Y_{t-j}$) + $\phi_1$ Cov($Y_{t-1}, Y_{t-j}$) + Cov($\epsilon_t, Y_{t-j}$)

= $\phi_1 \phi_1^{j-1} \sigma_Y^2$

= $\phi_1^j \sigma_Y^2$.

(14.11)

which holds for all $j \geq 0$.

The autocorrelations combine (14.11) with (13.5):

\[
\rho_j \equiv \text{Corr}(Y_t, Y_{t-j}) = \gamma_j / \sigma_Y^2 = (\phi_1 \sigma_Y^2) / \sigma_Y^2 = \phi_1^j.
\]

(14.12)

With $j = 1$, the first autocorrelation is $\rho_1 = \phi_1$, the autoregressive coefficient in (14.1).

**In Sum: AR(1) for Description**

Given (14.1) and (14.2) with $|\phi_1| < 1$ (\Rightarrow stationary),

- mean: $\mu = \text{E}(Y_t) = \phi_0/(1 - \phi_1)$
- variance: $\sigma_Y^2 = \text{Var}(Y_t) = \sigma_\epsilon^2/(1 - \phi_1^2)$
- $j$th autocovariance: $\gamma_j = \phi_1^j \sigma_Y^2$
- $j$th autocorrelation: $\rho_j = \phi_1^j$

14.3 Prediction (Forecasting): Optimality

For time series, “prediction” usually means **forecasting** future values of $Y_t$ given the current and past values. As in Section 2.5, given a loss function, the optimal forecast (prediction) minimizes mean loss.
This optimal forecast is defined in the population (without data) and can be estimated with data. In practice, in many cases, given the observed $Y_t$ for $t = 1, \ldots, T$, the goal is to forecast $Y_{T+1}$.

As in Part II, the focus here is on the CEF, the best forecast given quadratic loss. The white noise property of $\epsilon_t$ implies (14.1) is a CEF model:

$$E(Y_t | Y_{t-1}) = \hat{E}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t | Y_{t-1})$$

$$= \hat{E}(\phi_0 | Y_{t-1}) + \hat{E}(\phi_1 Y_{t-1} | Y_{t-1}) + \hat{E}(\epsilon_t | Y_{t-1})$$

$$= \phi_0 + \phi_1 E(Y_{t-1} | Y_{t-1}) + \hat{E}(\epsilon_t)$$

$$= \phi_0 + \phi_1 Y_{t-1}.$$  (14.13)

Consequently, given sample $Y_1, \ldots, Y_T$ and corresponding OLS estimates $\hat{\phi}_0$ and $\hat{\phi}_1$, a reasonable forecast of $Y_{T+1}$ is

$$\hat{Y}_{T+1} = \hat{\phi}_0 + \hat{\phi}_1 Y_T.$$  (14.14)

Even if the AR(1) model is wrong, if the time series is covariance stationary, then OLS estimates the best linear predictor (Section 7.5) of $Y_t$ given $Y_{t-1}$. But “best” does not mean “good” (Section 7.4.2); forecast accuracy may be improved by using additional lags, other variables, and/or nonlinearity (Chapter 15).

Under absolute loss, or with asymmetric loss, the CEF is not optimal, but rather the conditional median (or other percentiles). In that case, a quantile autoregression model and estimator could be used, but details are beyond our scope.

**Discussion Question 14.1** (forecast and reality). Given sample $Y_1, \ldots, Y_T$, you construct the forecast $\hat{Y}_{T+1} = \hat{\phi}_0 + \hat{\phi}_1 Y_T$. Then you wait one period and observe the actual $Y_{T+1}$.

a) Will you be surprised if $Y_{T+1} > \hat{Y}_{T+1}$? Or if $Y_{T+1} < \hat{Y}_{T+1}$? Why/not?

b) How often do you expect to see $Y_{T+1} = \hat{Y}_{T+1}$? Why?

c) Is it usually true that $Y_{T+1} = \phi_0 + \phi_1 Y_T$? Why/not? Hint: for any random variable $W$, how often does $W = E(W)$, i.e., what’s $P(W = E(W))$?

### 14.4 Estimation

Before discussing AR(1) estimation by OLS, the results of Section 13.5 can be interpreted as consistency of OLS in an intercept-only regression. Section 13.5 gave conditions in which a good estimator of the population mean is the sample mean. As in (3.10) and (3.11), the sample mean is the OLS estimator for the intercept-only model

$$Y_t = \mu + \epsilon_t.$$
In the AR(1), skipping technical details, the OLS estimators \( \hat{\phi}_0 \) and \( \hat{\phi}_1 \) are consistent in many cases. If \( |\phi_1| < 1 \), then OLS is consistent, meaning \( \hat{\phi}_0 \xrightarrow{p} \phi_0 \) and \( \hat{\phi}_1 \xrightarrow{p} \phi_1 \). Technical details for a more general version of this result may be found in Case 4 on pages 215–217 of Hamilton (1994, §8.2). In fact, \( \hat{\phi}_1 \xrightarrow{p} \phi_1 \) even if \( \phi_1 = 1 \) (Hamilton, 1994, §17.4).

There are other consistent estimators, too, and some research has tried to compare the small-sample properties of these, but such comparison is beyond our scope.

Using the estimated coefficients, a **point forecast** (our single, best guess) \( \hat{Y}_{T+1} \) is computed as in (14.14). For the demeaned model in (14.6),

\[
\hat{Y}_{T+1} = \hat{\mu} + \hat{\phi}_1 (Y_T - \hat{\mu}).
\] (14.15)

### 14.4.1 Code

The following code shows an example. The data \( Y \) are simulated from an AR(1) model with \( \phi_0 = 0 \) (so \( \mu = 0 \)) and \( \phi_1 = 0.25 \), using `arima.sim()`. The argument `n.ahead` tells `predict()` how many time periods past the end of the sample to make predictions for. In R, `ar()` by default estimates the demeaned model. In the code, \( \hat{\mu} \) is `ret$x.mean`, \( \hat{\phi}_1 \) is `ret$ar`, and \( \hat{\phi}_0 \) is \( \hat{\mu} \times (1 - \hat{\phi}_1) \). The predicted value in `pr$pred[1]` is shown to be equivalent to (14.14) and (14.15). (Alternatively, with argument `method='ols'`, you can estimate \( \phi_0 \) and \( \phi_1 \) directly, by OLS, as in the final, commented-out lines of code below.) Finally, note that the “standard errors” in `pr$se` below are for the predicted values, not just AR coefficient uncertainty, which is shown separately with `sqrt(ret$asy.var.coef)`.

```r
set.seed(112358)
RHO <- 0.25; n <- 100
Y <- arima.sim(n=n, model=list(ar=RHO), sd=1)
ret <- ar(x=Y, aic=FALSE, order.max=1)
cat(sprintf("PhiHat=\%5.3f, PhiHat1=\%5.3f\n",
        ret$x.mean*(1-ret$ar), ret$ar))
## PhiHat=0.024, PhiHat1=0.143
cat(sprintf("SE(PhiHat1)=\%5.3f\n", sqrt(ret$asy.var.coef)))
## SE(PhiHat1)=0.100
pr <- predict(ret, n.ahead=1)
# output point forecast and prediction SE
c(round(pr$pred, digits=3), round(as.numeric(pr$se), digits=3))
## [1] 0.196 0.966
# check prediction against formulas
c(pr$pred[1], ret$x.mean + ret$ar*(Y[n]-ret$x.mean),
    ret$x.mean*(1-ret$ar) + Y[n]*ret$ar)
```
## 14.5 Parameter Stability

Parameter stability pertains to external validity, as in Section 12.2: is \( \phi_1 \) truly a constant, or has it changed over time, and might it change in the future? This is also related to structural breaks (Section 13.6.4). With enough data, we could form multiple historical datasets and see if the estimates \( \hat{\phi}_1 \) change much over time. But either way, this does not tell us what will happen in the future. Historical data cannot predict a future black swan, something new not seen in the past. As usual, purely statistical analysis may fall short; a combination of your statistical and economic expertise (and critical thinking) yields better results.

This issue of parameter stability is related to the Lucas critique (Lucas, 1976). For example, a partial equilibrium model’s parameters may change when there is a new macroeconomic policy with general equilibrium effects (Section 4.3.3). The policy’s effects might change optimal forecasts and time series descriptions.

AR(1) models (and more complex models) allowing time-varying coefficients have been developed but are beyond our scope. There are also methods for identifying when in time a certain parameter changed, also beyond our scope.

**Discussion Question 14.2** (recession-affected coefficient). Name a variable you think might have different \( \phi_1 \) during a recession (than not during a recession), and explain why. Hint: this is not simply asking which variables are higher or lower in a recession, because that’s not what \( \phi_1 \) describes; e.g., the time series \( Z_t = Y_t + 10 \) would have the exact same \( \phi_1 \) as \( Y_t \), just a different \( \phi_0 \) or \( \mu \).

## 14.6 Multi-Step Forecast

Instead of forecasting \( Y_{t+1} \) given \( Y_t \), you may need to forecast \( Y_{t+h} \) given \( Y_t \) for a particular \( h > 1 \). This is called the **h-step-ahead forecast**.

For example, if you must make a decision that affects your business or government policy for the next year, and you have monthly data, you might like to predict \( Y_{t+12} \) given \( Y_t \), i.e., predict the value 12 months in the future. In fact, you may want to predict \( Y_{t+h} \) for all \( h = 1, \ldots, 12 \), i.e., predict each of the next 12 months. (Ideally,
instead of always using quadratic loss, your forecast would use a loss function more appropriate for your decision, but often practice is not ideal.)

14.6.1 Intuition: Mean Zero (Special Case)

For intuition, imagine the AR(1) \( Z_t = \phi_1 Z_t + \epsilon_t \), so \( E(Z_t) = 0 \), and

\[
E(Z_{T+1} \mid Z_T) = E(\phi_1 Z_T + \epsilon_{T+1} \mid Z_T) = \phi_1 E(Z_T \mid Z_T) + E(\epsilon_{T+1} \mid Z_T)
\]

\[
= \phi_1 Z_T.
\]

(14.16)

Iterating once, again using linearity of \( E(\cdot) \),

\[
E(Z_{T+2} \mid Z_T) = E(\phi_1 Z_{T+1} + \epsilon_{T+2} \mid Z_T) = \phi_1 E(Z_{T+1} \mid Z_T) + E(\epsilon_{T+2} \mid Z_T)
\]

\[
= \phi_1^2 Z_T.
\]

This pattern continues: given quadratic loss, the optimal forecast of \( Z_{T+h} \) given \( Z_T \) is

\[
E(Z_{T+h} \mid Z_T) = \phi_1^h Z_T.
\]

(14.17)

With data, the forecast is

\[
\hat{Z}_{T+h} = \hat{\phi}_1^h Z_T.
\]

(14.18)

With stationary \( Z_t \), \( |\phi_1| < 1 \), so \( \phi_1^h \) gets closer to zero as \( h \) gets bigger. For example, with \( \phi_1 = 0.5 \) and \( h = 10 \), \( (0.5)^{10} \approx 0.001 \). Thus, the farther into the future we forecast, the closer our (optimal) forecast gets to \( 0 = E(Z_t) = E(Z_{T+1}) \).

If instead \( \phi_1 = 1 \), then \( \phi_1^h = 1 \) for any \( h \), so the optimal forecast is \( E(Z_{T+h} \mid Z_T) = Z_T \). Even with very large \( h \), the optimal forecast is \( Z_T \), not \( 0 = E(Z_t) = E(Z_{T+h}) \). However, even if the true \( \phi_1 = 1 \), estimating \( |\hat{\phi}_1| < 1 \) leads to \( \hat{Z}_{T+h} \approx 0 \) for large \( h \). This is one reason there are many methods to try to distinguish unit root (here, \( \phi_1 = 1 \)) from stationary (here, \( |\phi_1| < 1 \)) processes, though they are beyond our scope.

14.6.2 General Results with AR Parameters

To extend the results, let \( Z_t \equiv Y_t - \mu \), with \( \mu \equiv E(Y_t) \), continuing to assume stationarity. Then, (14.17) and (14.18) become

\[
E(Y_{T+h} - \mu \mid Y_T) = \phi_1^h (Y_T - \mu) \implies E(Y_{T+h} \mid Y_T) = \mu + \phi_1^h (Y_T - \mu),
\]

(14.19)

\[
\hat{Y}_{T+h} = \hat{\mu} + \hat{\phi}_1^h (Y_T - \hat{\mu}) = \hat{\mu}(1 - \hat{\phi}_1^h) + \hat{\phi}_1^h Y_T. \hat{Y}_{T+h} = \frac{\hat{\phi}_0}{1 - \hat{\phi}_1}
\]

(14.20)

Given \( Y_T \), \( \hat{\phi}_0 \), and \( \hat{\phi}_1 \), this can forecast \( Y_{t+h} \) by plugging in \( h \). With \( h = 1 \), matching (14.14),

\[
\hat{Y}_{T+1} = \frac{\hat{\phi}_0}{1 - \hat{\phi}_1}(1 - \hat{\phi}_1) + \hat{\phi}_1 Y_T = \hat{\phi}_0 + \hat{\phi}_1 Y_T.
\]
14.6.3 Direct Approach

Alternatively, to forecast $Y_{T+h}$ given $Y_T$, simply regress $Y_{t+h}$ on $Y_t$ (and an intercept). This makes sense: such a regression estimates the best linear predictor of $Y_{t+h}$ given $Y_t$. This regression estimates the parameters in

$$Y_{t+h} = \phi_0 + \phi_1 Y_t + \epsilon_{t+h}. \quad (14.21)$$

The forecast of $Y_{T+h}$ is then

$$\hat{Y}_{T+h} = \hat{\phi}_0 + \hat{\phi}_1 Y_T. \quad (14.22)$$

For example, if $Y_t$ is quarterly GDP growth, and we want to predict GDP growth four quarters (i.e., one year) in the future, then $h = 4$. We regress $Y_{t+4}$ on an intercept and $Y_t$ in our quarterly data, predicting $\hat{Y}_{T+4} = \phi_0 + \phi_1 Y_T$. The forecast in (14.14) showed the special case with $h = 1$.

14.6.4 Code

There are functions in R that do multi-step forecasts automatically. In particular, the `forecast` function in the `forecast` package (Hyndman et al., 2020; Hyndman and Khandakar, 2008), which also does multi-step interval forecasts; see Section 14.8.

Discussion Question 14.3 (long-horizon AR forecast). Let $\hat{\phi}_1 = 0.5$.

a) What’s $\hat{\phi}_1^h$ when $h = 10$? $h = 20$?

b) Given (14.20), what does this imply about the forecast $\hat{Y}_{T+h}$ when $h$ is large?

c) Name one variable for which such a long-term forecast doesn’t make sense, and explain why not.

d) Does $\hat{Y}_{T+20} \approx \hat{Y}_{T+21} \approx \cdots$ imply we’d be surprised if the actual $Y_{T+20}, Y_{T+21}, \ldots$ go up and down (versus being all roughly the same value)? Why/not?

14.7 Interval Forecasts

The idea of a prediction interval from Section 2.5.5 extends to time series forecasting. Like a confidence interval, an interval forecast (or forecast interval) incorporates uncertainty. Instead of giving a single number point forecast like (14.14), an interval of numbers can help show how much uncertainty there is around the point forecast.

14.7.1 Goal and Sources of Uncertainty

The goal is to construct intervals that contain the true future value with some specified probability, like 95%. This is similar to the goal of a confidence interval, to contain the true parameter value 95% of the time. For intuition, imagine your job is to create 95% interval forecasts, and you make one every day for 1000 days. That is, on
each day \( t \), you make an interval forecast for the next day’s value \( Y_{t+1} \). Thus, the next day you can check whether or not the true value was inside your interval or not. If you’re doing your job well, then you should find that (approximately) 950 days out of 1000 your interval contained the true value, and the other 50 days it didn’t.

There are two sources of uncertainty in forecasting. The first source of uncertainty is the same as in a confidence interval: parameter uncertainty. That is, we only have estimated parameter values \( {\hat{\phi}_0} \) and \( {\hat{\phi}_1} \); we do not know the true population parameters \( \phi_0 \) and \( \phi_1 \). As before, the standard error helps quantify this type of uncertainty.

The second (and usually larger) source of uncertainty is the error term. Knowing the true parameters alone is not sufficient to perfectly forecast \( Y_{T+1} \) with complete certainty. For example, consider a simple case where \( \phi_0 = \phi_1 = 0 \), so \( Y_{T+1} = \epsilon_{T+1} \). Even if somehow we knew \( \phi_0 = \phi_1 = 0 \) and thus knew that \( Y_{T+1} = \epsilon_{T+1} \), we still wouldn’t know \( \epsilon_{T+1} \) at time \( T \), when we are trying to predict \( Y_{T+1} \). That is, in addition to uncertainty about parameters, there is uncertainty about \( \epsilon_{T+1} \).

### 14.7.2 Intervals Assuming Normality

Continuing the example with \( \phi_0 = \phi_1 = 0 \), what would a 95% interval forecast be? Since \( Y_{T+1} = \epsilon_{T+1} \), the interval should contain \( \epsilon_{T+1} \) with 95% probability. So, the interval depends on the distribution of the random variable \( \epsilon_{T+1} \). In the extremely unlikely case that \( \epsilon_{T+1} \sim N(0,1) \), then we know the interval \([-1.96,1.96]\) would work since \( P(-1.96 \leq \epsilon_{T+1} \leq 1.96) = 95\% \) is a property of the standard normal distribution. If instead \( \epsilon_{T+1} \sim N(0,\sigma^2_\epsilon) \), then the interval \([-1.96\sigma_\epsilon,1.96\sigma_\epsilon]\) would work; we’d probably have to estimate \( \sigma_\epsilon \) from the data and use the estimated \( \hat{\sigma}_\epsilon \) instead. This is the foundation for many interval forecasts; e.g., see Diebold (2018b, §7.3.3).

Continuing to assume normally distributed \( \epsilon_{T+1} \), the 95% interval forecast can be stated more generally. This covers the AR(1) as well as any other forecast. The 95% forecast interval becomes

\[
[\hat{Y}_{T+1} - 1.96\hat{\sigma}_\epsilon, \hat{Y}_{T+1} + 1.96\hat{\sigma}_\epsilon],
\]

where \( \hat{Y}_{T+1} \) is the point forecast. This ignores parameter uncertainty, which is usually much smaller than the uncertainty from \( \epsilon_{T+1} \). To get a 90% interval, simply replace 1.96 with 1.64; or for a 100(1 − 2\( \alpha \))% interval, use the 100(1 − \( \alpha \))th percentile of the standard normal distribution. (Most built-in statistical functions only ask you for the desired percentage and compute the critical value automatically.)

But what if \( \epsilon_{T+1} \) isn’t normally distributed? Then the forecast interval in (14.23) is not valid. It may be “too wide,” containing the true value with more than 95% probability, or it may be “too short,” containing the true value with less than 95% probability. For example, if \( \epsilon_{T+1} \sim \text{Unif}(−\sqrt{3},\sqrt{3}) \), uniformly distributed over all real (decimal) numbers between \(-\sqrt{3}\) and \(\sqrt{3}\), then \( \sigma_\epsilon = 1 \). But \( \sqrt{3} \approx 1.73 \), so \( \epsilon_{T+1} \) is always between \(-1.73\) and \(1.73\), which means
it is always in the interval $[-1.96, 1.96]$, 100% of the time (not 95%). The correct 95% interval is shorter: $P(-1.65 \leq \epsilon_{T+1} \leq 1.65) = 95%$. Further, if the distribution of $\epsilon_{T+1}$ is not symmetric, then the best forecast interval may not be symmetric either; i.e., the point forecast may not be exactly in the middle of the interval.

It’s possible to estimate the distribution of $\epsilon_{T+1}$ (assuming it’s strictly stationary) and use the estimated percentiles to construct forecast intervals, but that approach is beyond our scope.

### 14.7.3 Code

The following code shows basic interval forecasts using the `forecast` package (which also shows the point forecasts). The argument `h=12` specifies forecasting values for the next 12 time periods, so the results include multi-step interval forecasts. Argument `level=c(80,95)` specifies both 80% and 95% prediction intervals. Although the code is easy to run, an AR(1) is not always appropriate, so critical thought is required; e.g., see DQ 14.4.

```r
library(forecast)
ret <- ar(AirPassengers, aic=FALSE, order.max=1)
forecast(ret, h=12)
```

<table>
<thead>
<tr>
<th>Period</th>
<th>Point Forecast</th>
<th>Lo.80</th>
<th>Hi.80</th>
<th>Lo.95</th>
<th>Hi.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-1961</td>
<td>424</td>
<td>375</td>
<td>473</td>
<td>349</td>
<td>499</td>
</tr>
<tr>
<td>Feb-1961</td>
<td>417</td>
<td>349</td>
<td>484</td>
<td>313</td>
<td>520</td>
</tr>
<tr>
<td>Mar-1961</td>
<td>410</td>
<td>329</td>
<td>490</td>
<td>286</td>
<td>533</td>
</tr>
<tr>
<td>Apr-1961</td>
<td>403</td>
<td>312</td>
<td>494</td>
<td>264</td>
<td>542</td>
</tr>
<tr>
<td>May-1961</td>
<td>396</td>
<td>297</td>
<td>496</td>
<td>245</td>
<td>548</td>
</tr>
<tr>
<td>Jun-1961</td>
<td>390</td>
<td>284</td>
<td>496</td>
<td>228</td>
<td>553</td>
</tr>
<tr>
<td>Jul-1961</td>
<td>385</td>
<td>273</td>
<td>497</td>
<td>214</td>
<td>556</td>
</tr>
<tr>
<td>Aug-1961</td>
<td>379</td>
<td>262</td>
<td>496</td>
<td>200</td>
<td>558</td>
</tr>
<tr>
<td>Sep-1961</td>
<td>374</td>
<td>253</td>
<td>495</td>
<td>189</td>
<td>560</td>
</tr>
<tr>
<td>Oct-1961</td>
<td>369</td>
<td>244</td>
<td>494</td>
<td>178</td>
<td>560</td>
</tr>
<tr>
<td>Nov-1961</td>
<td>365</td>
<td>236</td>
<td>493</td>
<td>168</td>
<td>561</td>
</tr>
<tr>
<td>Dec-1961</td>
<td>360</td>
<td>229</td>
<td>491</td>
<td>160</td>
<td>561</td>
</tr>
</tbody>
</table>

**Discussion Question 14.4** (forecast sanity check). Do the point forecasts shown above pass a sanity check? That is, they show steadily decreasing values from January to December 1961; does this seem reasonable given Figure 13.1? Why/not?

### 14.8 More R Examples

#### 14.8.1 AR(1) Multi-Step Forecast Intervals

The following code simulates data from an AR(1) model, and then computes (and outputs and plots) various estimates and forecasts.
Note that $T = 100$ ($n \leftarrow 100$), $\phi_1 = 0.8$ ($\text{RHO}$), $\mu = \text{E}(Y_t) = 5$, and $\sigma_\epsilon = 1$ (from the sd=1 option). The estimated $\hat{\phi}_1$ is not particularly good, although the true value is within two standard errors (there is just a lot of uncertainty).

![Figure 14.1: Point and interval forecasts.](image)

```r
set.seed(112358)
RHO <- 0.80; n <- 100
Y <- 5 + arima.sim(n=n, model=list(ar=RHO), sd=1)
ret <- ar(X=Y, aic=FALSE, order.max=1)
cat(sprintf("PhiHat1=%5.3f, SE(PhiHat1)=%5.3f
",
ret$ar, sqrt(ret$asy.var.coef)))
## PhiHat1=0.685, SE(PhiHat1)=0.074

(fc <- forecast(ret, h=15, level=c(80,95)))
plot(fc)
```

<table>
<thead>
<tr>
<th>Period</th>
<th>Point Forecast</th>
<th>Lo.80</th>
<th>Hi.80</th>
<th>Lo.95</th>
<th>Hi.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>3.10</td>
<td>1.83</td>
<td>4.37</td>
<td>1.15</td>
<td>5.04</td>
</tr>
<tr>
<td>102</td>
<td>3.59</td>
<td>2.05</td>
<td>5.13</td>
<td>1.23</td>
<td>5.95</td>
</tr>
<tr>
<td>103</td>
<td>3.93</td>
<td>2.27</td>
<td>5.58</td>
<td>1.40</td>
<td>6.46</td>
</tr>
<tr>
<td>104</td>
<td>4.16</td>
<td>2.45</td>
<td>5.86</td>
<td>1.55</td>
<td>6.76</td>
</tr>
<tr>
<td>105</td>
<td>4.32</td>
<td>2.59</td>
<td>6.04</td>
<td>1.68</td>
<td>6.96</td>
</tr>
<tr>
<td>106</td>
<td>4.42</td>
<td>2.69</td>
<td>6.16</td>
<td>1.77</td>
<td>7.08</td>
</tr>
<tr>
<td>107</td>
<td>4.50</td>
<td>2.76</td>
<td>6.24</td>
<td>1.83</td>
<td>7.16</td>
</tr>
<tr>
<td>108</td>
<td>4.55</td>
<td>2.81</td>
<td>6.29</td>
<td>1.88</td>
<td>7.22</td>
</tr>
<tr>
<td>109</td>
<td>4.58</td>
<td>2.84</td>
<td>6.33</td>
<td>1.92</td>
<td>7.25</td>
</tr>
<tr>
<td>110</td>
<td>4.61</td>
<td>2.86</td>
<td>6.35</td>
<td>1.94</td>
<td>7.28</td>
</tr>
<tr>
<td>111</td>
<td>4.63</td>
<td>2.88</td>
<td>6.37</td>
<td>1.95</td>
<td>7.30</td>
</tr>
<tr>
<td>112</td>
<td>4.64</td>
<td>2.89</td>
<td>6.38</td>
<td>1.97</td>
<td>7.31</td>
</tr>
<tr>
<td>113</td>
<td>4.64</td>
<td>2.90</td>
<td>6.39</td>
<td>1.97</td>
<td>7.32</td>
</tr>
<tr>
<td>114</td>
<td>4.65</td>
<td>2.90</td>
<td>6.40</td>
<td>1.98</td>
<td>7.32</td>
</tr>
<tr>
<td>115</td>
<td>4.65</td>
<td>2.91</td>
<td>6.40</td>
<td>1.98</td>
<td>7.32</td>
</tr>
</tbody>
</table>
Figure 14.1 was generated by the above code and shows some patterns. The graph essentially plots the table of results (point and interval forecasts) after plotting the original time series. First, in the data itself, we can see some persistence (high values tend to be followed by high values, and low by low), but the values never get too far from the mean $E(Y_t) = 5$. Second, the point forecasts $\hat{Y}_{t+h}$ get closer and closer to the sample average $\bar{Y} = \frac{1}{T} \sum_{t=1}^{T} Y_t$ as $h$ increases. This is because we chose an AR(1) forecasting model; even if the data were not generated by an AR(1), the forecasts would show the same pattern. Third, the forecast intervals get wider and wider as $h$ increases. This makes intuitive sense (although we skipped the math): the farther in the future, the less certainty we have.

14.8.2 General R Forecast Allowing Seasonality and Trend

Figure 14.2 uses the `stlf()` and `auto.arima()` functions from the `forecast` package. They do much better than the earlier forecast in Section 14.7.3 that ignored seasonality and trend. This general application of `stlf()` or `auto.arima()` can sometimes be improved by more carefully considering the type of trend, the properties of the remainder, the type of seasonality, etc., but clearly for series where
the trend and/or seasonality is important, it is much better to use these functions that incorporate trend and seasonality than a model that does not allow for trend and seasonality, like the basic AR(1). But, AR models are still very useful: they (or more general ARIMA models) are used by \texttt{stlf()} and \texttt{auto.arima()} to fit the detrended, seasonally-adjusted data.

Either way, it is always good to “sanity check” your forecasts visually. In this case, the point forecasts in Figure 14.2 look reasonable, unlike the earlier basic AR(1). It is also reasonable that the interval forecasts get longer (taller) farther into the future, appropriately reflecting greater uncertainty. However, the \texttt{stlf()} interval forecasts seem too narrow; even multiple years in the future, the interval is relatively short. Reading the \texttt{?stlf} help file suggests one reason why: it says, “Note that the prediction intervals ignore the uncertainty associated with the seasonal component.” That is, it assumes the estimated seasonality is actually the true seasonality, with no uncertainty. Even the \texttt{auto.arima()} intervals may be “too short” since (as usual) they do not account for uncertainty about the true model itself changing over time (i.e., structural breaks), only uncertainty about the parameter values.

Figure 14.2 was generated by the following code.

```r
library(forecast)
par(family='serif', mar=c(1.8,1.8,0.3,0.6), mgp=c(2.1,0.8,0))
ret1 <- stlf(y=log(AirPassengers), h=48)
plot(ret1)
ret2 <- auto.arima(y=log(AirPassengers))
plot(forecast(ret2, h=48))
```
Empirical Exercises

Empirical Exercise EE14.1. You will analyze the New York Stock Exchange (NYSE) value-weighted price index, specifically the weekly close prices every Wednesday. (Unfortunately, the dataset does not note the dates or data source.) You’ll consider forecasting price as well as the price change, using an AR(1) model, with both point and interval forecasts. In practice, if you could reliably predict the price change, then you could make a lot of money; so you should be (very) skeptical that you can forecast the price change. (This is related to the “efficient market hypothesis.”)

Mathematically, assume the price change $U_t = Y_t - Y_{t-1}$ is indeed unrelated to $Y_t$ and $Y_{t-1}$ (and other past values), and let $\phi_0 = E(U_t)$ and $V_t = U_t - E(U_t)$, so $E(V_t) = 0$. Then $Y_t = Y_{t-1} + U_t = \phi_0 + Y_{t-1} + V_t$ is an AR(1) with $\phi_1 = 1$, in which case $Y_T + \phi_0$ is the best forecast of $Y_{T+1}$. You will check if $\hat{\phi}_1 \approx 1$ and estimate the value of $\phi_0$, among other computations.

a. R only: load the needed packages (and install them if necessary) and look at a description of the dataset:
   ```r
   library(wooldridge); library(forecast)
   ?nyse
   ```
b. Stata only: load the data with `bcuse nyse , nodesc clear` (assuming `bcuse` is already installed)
c. Tell your software that you have weekly time series data.
   ```r
   R: tsdat <- ts(data=nyse$price, frequency=52.18)
   Stata: tsset t, weekly
   ```
d. Define a variable `holdout` for how many time periods at the end of the sample to “hold out” when fitting your model.
   ```r
   R: holdout <- 2
   Stata: scalar holdout = 20
   ```
e. R only: using `holdout`, define the time period at the end of the “training” data (just before the “testing” data) as `midpt <- length(tsdat) - holdout` and use it to define the training and testing data respectively:
   ```r
   tsdattrain <- subset(tsdat, start=1, end=midpt)
   tsdattest <- subset(tsdat, start=midpt+1, end=length(tsdat))
   ```
f. Estimate an AR(1) model to produce “dynamic” forecasts, i.e., what would be forecast if we were living at the end of the training data.
   ```r
   R: ret <- ar(x=tsdattrain, aic=FALSE, order.max=1, method='yw')
   Stata: arima price if _n<=_N-holdout , arima(1,0,0)
   ```
g. Pretend you travel back in time to the very end of the training data, and produce dynamic forecasts for the next 20 periods (weeks).

R: (fc <- forecast(ret, h=holdout, level=c(80,95)))
Stata: predict fmulti, y dyn(t[_N-holdout+1]) where fmulti is the name for a newly created variable and dyn tells it to make dynamic forecasts

h. Plot the forecasts against the actual historical data.

R: plot(fc) and lines(window(tsdattest), col=1)
Stata: 
\texttt{twoway tsline price || tsline fmulti if _n>_N-holdout, lcolor(red)}

i. Repeat your analysis, but with an AR(1) model of the first-differenced price ($\Delta Y_t = Y_t - Y_{t-1}$), which is already in the dataset as the variable cprice (“c” for “change”).

R: when you create tsdat, use data=nyse$cprice[-1] to exclude the first value of cprice (which is missing); otherwise the code should be the same; you may also like to draw a line with \texttt{lines(x=c(-1,1)*999,y=c(0,0))} at the very end for reference.

Stata: just use cprice and make sure to name a different new variable in your predict command, which you’ll reference in your graphing command. Note also that instead of \texttt{arima cprice}, you could use OLS estimation with \texttt{regress cprice cprice_1}, or equivalently \texttt{regress D.price L.D.price} where D.price means “take the first difference of the variable price” and L.D.price means “lag of first difference of price”
Chapter 15

Higher-Order Autoregression and Autoregressive Distributed Lag Regression

⇒ Kaplan video: Chapter Introduction

Depends on: Chapter 14 (which depend on Chapters 2–4, 6–8, and 13)

Unit learning objectives for this chapter

15.1. Define new vocabulary words (in bold), both mathematically and intuitively [TLO 1]

15.2. Explain the problem of choosing the best model both mathematically and intuitively, along with possible solutions [TLO 2]

15.3. Implement and compare different ways to select the best forecasting model [TLOs 2 and 6]

15.4. In R (or Stata): estimate more general time series regression models for the purpose of forecasting future values [TLO 7]

Optional resources for this chapter

- AIC and BIC: Hanck et al. (2018, §14.6)
- Forecast model evaluation and selection: Hyndman and Athanasopoulos (2019, §§3.4,5.5) and function forecast::CV()
- Autoregression: Hyndman and Athanasopoulos (2019, §8.3)
- Lagged predictors: Hyndman and Athanasopoulos (2019, §9.6)
- Example data: fpp2 package in R (Hyndman, 2018)
Sometimes, accuracy improves by forecasting $Y_{t+1}$ using not only $Y_t$ but also $Y_{t-1}$. And why stop at $Y_{t-1}$? Maybe $Y_{t-2}$ contains additional information not found in $Y_t$ and $Y_{t-1}$; or maybe $Y_{t-3}$ does, or even longer lags of $Y_t$. Additionally, other variables and possibly their lags may further improve forecasting accuracy.

### 15.1 The AR($p$) Model

The **AR($p$) model** generalizes the AR(1) model in (14.1):

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t = \phi_0 + \sum_{j=1}^{p} \phi_j Y_{t-j} + \epsilon_t. \quad (15.1)$$

Again $\epsilon_t$ is white noise, with properties as in (14.2). Coefficient $\phi_j$ is called the $j$th **partial autocorrelation**, for $j = 1, \ldots, p$. (This can be confusing since $\phi_j \neq \text{Corr}(Y_t, Y_{t-j})$, the $j$th autocorrelation.)

Theoretical details and properties are mostly omitted here, but there are concepts similar to the AR(1). For example, there is the concept of a unit root, which generates nonstationary $Y_t$, but its mathematical characterization is more complicated than just $\phi_1 = 1$. The autocovariances and autocorrelations can be derived from the coefficients and properties of $\epsilon_t$, but the derivations and formulas are again more complicated.

Instead, the next sections focus on good forecasts.

### 15.2 Model Selection: How Many Lags?

[⇒ Kaplan video: Model Selection for Forecasting](#)

In practice, which $p$ should we use? This is a question of model selection (see also Section 8.3). Choosing $p$ is equivalent to setting $\hat{\phi}_j = 0$ for $j > p$, instead of estimating those $\phi_j$ from data.

#### 15.2.1 Difficulties and Intuition

Recall the intuition from Section 8.3. If $p$ is too small, then the model is not flexible enough; implicitly, this sets $\hat{\phi}_j = 0$ for some important $\phi_j \neq 0$. Even if the $\phi_j$ are estimated perfectly for $j = 0, 1, \ldots, p$, the estimated model may not forecast very well because $\hat{\phi}_j = 0 \neq \phi_j$ for some $j > p$. However, if $p$ is too big, then the model can be too flexible, overfitting the data. This also causes poor forecasts. We want the “just right” $p$ that balances these two sources of error.

Only looking at in-sample fit leads to overfitting (Section 8.3). For example, minimizing the sum of squared residuals (SSR), or equivalently maximizing the $R^2$, always picks the largest possible $p$, regardless of the dataset and which model is actually best. The “adjusted $R^2$” is better, but still not designed for picking the best forecasting model. Similarly, hypothesis testing is not designed to pick the best forecasting model.
With time series, large \( p \) additionally limits the amount of usable data. For example, if we observe \( Y_t \) for \( t = 1, \ldots, T \), and we regress \( Y_t \) on lags up to \( Y_{t-50} \) \( (p = 50) \), then we can only use \( t \) for which both \( Y_t \) and \( Y_{t-50} \) are observed. If \( t > T \), then \( Y_t \) isn’t observed; if \( t \leq p \), then \( Y_{t-p} \) isn’t observed. If \( T = 51 \), then there is only one usable data point: regressing \( Y_{51} \) on \( Y_{50}, Y_{49}, \ldots, Y_1 \). Since it’s impossible to estimate 51 parameters from 1 data point, \( p \) must be (much) smaller. Even with \( p = 25 \), there are \( p + 1 = 26 \) parameters and \( T - p = 26 \) usable data points; estimates could be computed but certainly suffer from overfitting. With \( T \) total observations, you can only estimate an AR\((p)\) with \( p < T/2 \), and \( p \) must be even smaller for reliable estimation.

The most common model selection methods for AR\((p)\) models use information criteria. Basically, an information criterion tries to quantify how bad a model is for prediction, so lower values are better (less bad). The two most common are the Akaike information criterion (AIC), proposed by Akaike (1974), and the Bayesian information criterion (BIC) (or sometimes SIC, SBC, or SBIC) of Schwarz (1978). There is also a “corrected” AICc; e.g., see Hyndman and Athanasopoulos (2019, §8.6). As seen below, both AIC and BIC try to avoid overfitting by adding a penalty to the in-sample fit. The penalty is larger when the model is larger (more flexible). AIC and BIC can also be used for model selection with other types of models beyond autoregression.

Instead of picking a single “best” forecasting model, averaging multiple forecasts (“forecast averaging,” or more generally “model averaging”) often performs even better but is beyond our scope.

### In Sum: Model Selection for Forecasting

After you think critically about which variables and lags might help forecast future values, AIC (and AICc) and BIC can help you pick which model produces the best forecasts.

#### 15.2.2 AIC and BIC Formulas

There are many different but equivalent formulas for AIC and BIC. This is because the selected model is the one whose value is lower than any other model’s value, so only the relative values matter, not the numeric values themselves. Thus, we could add 5 to all values, or multiply by \( T \), or take the log, etc., because this would not change which value of \( p \) (number of lags) minimizes the AIC or BIC.

The AIC can be written in terms of the sum of squared residuals (SSR) and a penalty based on \( p \). Specifically,

\[
AIC(p) = \hat{T} \ln(\text{SSR}) + 2(p + 1).
\]

Intuitively, we’d like our models to fit the data well (small SSR), but given the same fit we prefer less flexible models (small penalty). The
penalty prevents overfitting, where a model fits the data sample “too well” because it fits all the noise, which in turn makes its out-of-sample forecasts poor.

The BIC also involves the SSR and a penalty. Specifically,

$$\text{BIC}(p) = T \ln(\text{SSR}) + (p + 1) \ln(T).$$  \tag{15.3}$$

When comparing models with different lag lengths, due to the different number of usable data points, some care is required to ensure a fair comparison. For now, you can try to use built-in functions and hope that they were implemented carefully; e.g., in the forecast package, \texttt{auto.arima()} does automatic model selection using the AICc (which you can change to AIC or BIC with the \texttt{ic} argument).

### 15.2.3 Comparison of AIC and BIC

Compared to the AIC, the BIC has a larger penalty for large models since \(\ln(T) > 2\) if \(T > 7\). (And if \(T \leq 7\), you should collect more data.) That is, the BIC is more likely to pick smaller \(p\), i.e., shorter lag lengths (smaller models).

Related to this difference, the BIC is better than AIC if the true model is small but worse if the true model is large (Shao, 1997, p. 235). For example, if the true model is an AR(1), and you’re selecting among AR\((p)\) models for \(p = 0, 1, \ldots, 24\), then BIC is more likely to pick the true model than AIC. However, if the true model is AR(100) and \(T = 50\) (in which case picking the true model is impossible), then AIC is more likely than BIC to pick the best feasible model. Similarly, AIC is generally better for generating accurate forecasts if you only consider lag length up to \(p\), but the true lag length is even larger. Thus, whether AIC or BIC is best depends on what you think about the true model.

For example, imagine choosing from either one or two lags. The AR(2) model always fits the data better (lower SSR) than the AR(1) model. To be concrete, imagine \(T \ln(\text{SSR}) = 11\) with \(p = 1\), and \(T \ln(\text{SSR}) = 8\) with \(p = 2\). With AIC, the penalty term equals 4 for \(p = 1\) and equals 6 for \(p = 2\); the AIC penalty depends only on \(p\), not the data or even \(T\). For BIC, the penalty terms for \(p = 1\) and \(p = 2\) are \(2 \ln(T)\) and \(3 \ln(T)\), respectively; e.g., if \(T = 50\), then these are approximately 7.8 and 11.7. Thus, plugging these values into (15.2) and (15.3),

\[
\begin{align*}
\text{AIC}(1) &= T \ln(\text{SSR}) + 2(p + 1) = 15, & \text{BIC}(1) &= T \ln(\text{SSR}) + (p + 1) \ln(T) = 11.7 \\
\text{AIC}(2) &= T \ln(\text{SSR}) + 2(p + 1) = 14, & \text{BIC}(2) &= T \ln(\text{SSR}) + (p + 1) \ln(T) = 11.7.
\end{align*}
\]

Since \(\text{AIC}(2) < \text{AIC}(1)\), \(p = 2\) is better according to AIC. However, \(\text{BIC}(1) < \text{BIC}(2)\), so \(p = 1\) is better according to BIC. If we use AIC, we then fit an AR(2) model and use its estimates to forecast \(Y_{T+1}\). If
instead we had used BIC for model selection, we’d estimate an AR(1) model and use it to forecast $Y_{T+1}$.

**Discussion Question 15.1** (lag choice for forecasting). Imagine $Y_t = 50 + 0.5Y_{t-1} + 0.00001Y_{t-2} + \epsilon_t$, where the $\epsilon_t$ are independent of past values $Y_{t-1}, Y_{t-2}, \ldots$ and are iid and mean-zero. Do you think an estimated AR(0), AR(1), AR(2), or AR(3) would produce the best forecasts? Explain why you think your estimated model would produce better forecasts than each of the other three estimated models. Hint #1: if you need to make assumptions about things like the value of $T$, please feel free as long as you say so explicitly. Hint #2: thinking about extreme situations is sometimes helpful; e.g., what if $\epsilon_t = 0$ for all $t$, or what if $T = 8$, etc.

### 15.2.4 Code

The following code uses the AIC to choose $p$, then makes a forecast of $Y_{T+1}$ using an AR($p$) model. The AIC-chosen $p$ is shown along with the $p$ used to generate the data. Finally, the BIC is computed for the AIC-chosen $p$ and that $p-1$; the BIC is lower for the latter value, so it prefers a smaller model (smaller $p$) than AIC in this case.

```r
set.seed(112358)
MAXP <- 15 #max lag length for AR(p)
ARCOEFFS <- c(0.6, -0.4, 0.4, 0.1)
TRUEP <- length(ARCOEFFS) #p in true AR(p) DGP
# simulate data
Y <- arima.sim(n=60, model=list(ar=ARCOEFFS), sd=1)
# fit AR(p), using AIC to choose best p
ret <- ar(x=Y, aic=TRUE, order.max=MAXP)
# output optimal p
cat(sprintf("true p=%d; AIC-chosen p=%d\n", TRUEP, ret$order))
## true p=4; AIC-chosen p=7
pr <- predict(ret, n.ahead=1) #compute point forecast
c(round(pr$pred, digits=3)) #output
## [1] -0.434
# check BIC for AIC-chosen p and one smaller
# probably BIC prefers smaller (ret2)
ret1 <- arima(Y, order=c(ret$order,0,0))
ret2 <- arima(Y, order=c(ret$order-1,0,0))
c(BIC(ret1),BIC(ret2))
## [1] 186 185
```
15.3 Autoregressive Distributed Lag Regression

The autoregressive distributed lag (ADL) model (or “dynamic distributed lag” model) adds other variables and their lags to the AR\((p)\) model. That is, instead of forecasting \(Y_{t+1}\) using only \(Y_t, Y_{t-1}\), and other lags of \(Y\), we could also use \(X_t, X_{t-1}\), etc. Since \(X_{t+1}\) is not available at time \(t\), it should not be included as an explanatory variable if we are interested in forecasting. Equivalently, if we regress \(Y_t\) on \(Y_{t-1}, Y_{t-2}\), and other lags, we could add \(X_t-1, X_t-2\), etc., but not \(X_t\). If the goal is not forecasting but rather understanding the economic relationship between \(Y_t\) and \(X_t\), then this comment does not apply.

The same ideas from before apply to the ADL model. For example, it could be used for multi-step forecasting by replacing \(Y_{t+1}\) with \(Y_{t+h}\), or used for interval forecasts, and forecasts may be evaluated and compared as in Section 15.2.

To handle seasonality, decomposition or seasonal dummies can be used. The first option is to “seasonally adjust” your data by removing the seasonal component, and then fit the ADL model (and add back the seasonality into the forecast \(\hat{Y}_{T+1}\)). The second option is to use the raw data but replace the intercept term with dummies for each possible season. For example, with quarterly data, let \(D_{1t} = 1\) if time period \(t\) is in quarter 1 of some year (and \(D_{1t} = 0\) otherwise), and similarly \(D_{2t} = 1\) if \(t\) is in quarter 2, \(D_{3t} = 1\) for quarter 3, and \(D_{4t} = 1\) for quarter 4. All four dummies can be included as regressors if the intercept is removed; alternatively, you can keep the intercept and just add \(D_{2t}, D_{3t}, \text{ and } D_{4t}\) as regressors. Or, for monthly data, you can include the intercept along with \(D_{2t}, \ldots, D_{12t}\), where \(D_{2t}\) is the dummy for February, \(D_{3t}\) for March, up to \(D_{12t}\) for December, or else remove the intercept and include all \(D_1t, \ldots, D_{12t}\).

The following code uses ADL models to forecast quarterly GDP growth. First, quarterly GDP \(G_t\) is transformed to \(Y_t = \ln(G_t) - \ln(G_{t-1})\) and stored in variable \(\text{GDPgr}\) (“gr” for “growth”). Second, lags of T-bill rates are generated. Third, various ADL models are fit and their AIC (actually AICc) calculated. Fourth, the best ADL model is used to forecast \(Y_{T+1}\); the output at the end shows the point forecast along with forecast intervals. Note that \texttt{auto.arima()} automatically chooses the best lag length for \(Y_t\), but the best T-bill lag is determined “manually,” by calling \texttt{auto.arima()} once for each possible T-bill lag.

```r
library(AER); library(forecast); data('USMacroSWQ')
GDPgr <- diff(x=log(USMacroSWQ[, 'gdp'])) # GDP growth
Tlags <- cbind(Tlag1=lag(USMacroSWQ[, 'tbill'], -1),
               Tlag2=lag(USMacroSWQ[, 'tbill'], -2),
               Tlag3=lag(USMacroSWQ[, 'tbill'], -3),
               Tlag4=lag(USMacroSWQ[, 'tbill'], -4))
Tlags <- subset(Tlags, end=NROW(GDPgr))
```
```r
fit1 <- auto.arima(y=subset(GDPgr,start=4),
                   xreg=subset(Tblags[,1:1],start=4))
fit2 <- auto.arima(y=subset(GDPgr,start=4),
                   xreg=subset(Tblags[,1:2],start=4))
fit3 <- auto.arima(y=subset(GDPgr,start=4),
                   xreg=subset(Tblags[,1:3],start=4))
fit4 <- auto.arima(y=subset(GDPgr,start=4),
                   xreg=subset(Tblags[,1:4],start=4))
AICcs <- c(fit1["aicc"],fit2["aicc"], fit3["aicc"],fit4["aicc"])
best <- which.min(AICcs)
# fit3 has lowest AIC/AICc; fit1 lowest BIC
# Now fit w/ all available data
fit <- auto.arima(y=GDPgr, xreg=Tblags[,1:best])
tnow <- ptqy(USMacroSWQ)
xr <- cbind(Tblag1=USMacroSWQ[tnow-/zero.alt3,'tbill'],
            Tblag2=USMacroSWQ[tnow-1,'tbill'],
            Tblag3=USMacroSWQ[tnow-2,'tbill'],
            Tblag4=USMacroSWQ[tnow-3,'tbill'])
xr <- matrix(xr[,1:best], nrow=1)
(fc <- forecast(fit, h=1, xreg=xr))
```
Empirical Exercise EE15.1. You will analyze annual U.S. unemployment and inflation data from the 2004 Economic Report of the President, Tables B-42 and B-64. The goal is to forecast the unemployment rate. We’ll use the first $T-1$ observations to build a forecast, then compare our forecast to the actual observation in time $T$.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

```r
library(wooldridge); library(forecast)
?phillips
```

b. Stata only: load the data with `bcuse phillips , nodesc` (assuming `bcuse` is already installed)

c. R only: define `thisyr <- 1995` since the Stata dataset only has through year 1996, so that we can get comparable results. Also define `yr1 <- min(phillips$year)`

d. Tell your software that you have annual (yearly) time series data.

R: `tsdat <- ts(phillips[phillips$year<=thisyr, ], frequency=1, start=yr1)`

Stata: `tsset year , yearly`

e. Stata only: define `scalar holdout = 1` and `scalar endyr = year[_N]`

f. Plot the unemployment and inflation time series.

R: `plot(tsdat[,c('unem','inf')])`

Stata: `tsline unem inf`

g. Considering AR($p$) models with $p = 0, 1, 2, 3, 4$, use the AIC to choose the best model, and estimate such a model.

R: `ret <- ar(tsdat[,,'unem'], aic=TRUE, order.max=4)`

Stata: `varsoc unem , maxlag(4)` and then `arima unem if year<=endyr-holdout , arima(p,0,0)` but replacing the $p$ in `arima(p,0,0)` with whatever lag length the previous `varsoc` command said is optimal. (It’s possible to do this programmatically, but it gets complicated.)

h. R only (since Stata displayed this already): compute the BIC values for $p = 0, 1, 2, 3, 4$ with `ret$aic+(log(ret$n.used)-2) *1:length(ret$aic)` which adjusts the AIC values to reflect the BIC’s different penalty

i. Using the estimates based on data years up to 1995, compute (dynamic) forecasts for the next ten years, 1996–2005, and plot them.

R: `(fcARp <- forecast(ret, h=10)) and plot(forecast(ret, h=10))`
EMPIRICAL EXERCISES

Stata:

tsucc , add(9)
predict fcur if year>endyr-holdout , y
order year unem fcur
list year unem fcARp if year>=endyr-holdout
twoway tsline unem || tsline fcARp

j. Stata only: delete the previously added rows with drop if year>endyr

k. Now consider autoregressive distributed lag (ADL) models with up to 2 lags of unemployment and up to 2 lags of inflation. Compute all the AIC values.

R:

unem <- ts(phillips[, 'unem'], frequency=1, start=yr1)
inf <- ts(phillips[, 'inf'], frequency=1, start=yr1)
dat <- cbind(Y=unem, L1Y=lag(unem,-1),
L2Y=lag(unem,-2), L1X=lag(inf,-1), L2X=lag(inf,-2))
dat1 <- window(dat, start=yr1+2, end=thisyr)
ret <- lm(Y~L1X, data=dat1)
ret1 <- lm(Y~L1Y, data=dat1)
ret11 <- lm(Y~L1Y+L1X, data=dat1)
ret12 <- lm(Y~L1Y+L1X+L2X, data=dat1)
ret2 <- lm(Y~L1Y+L2Y, data=dat1)
ret21 <- lm(Y~L1Y+L2Y+L1X, data=dat1)
ret22 <- lm(Y~L1Y+L2Y+L1X+L2X, data=dat1)
AICs <- data.frame(L/zero.alt3.inf=c(AIC(ret), AIC(ret1), AIC(ret11), AIC(ret12), AIC(ret2), AIC(ret21), AIC(ret22))
rownames(AICs) <- c("L/zero.alt3.unem","L1.unem","L2.unem")
print(AICs, digits=4)

Stata:

varsoc unem, maxlag(2) exog()
varsoc unem, maxlag(2) exog(L.inf)
varsoc unem, maxlag(2) exog(L.inf L2.inf)

l. Estimate the ADL model with the smallest AIC. For example, if the AIC is smallest with one lag of each variable, then use R command (ret <- lm(Y~L1Y+L1X, data=window(dat,end =thisyr))) or Stata command arima unem L.inf if year<=endyr-holdout, arima(1,0,0)
m. Compute the ADL forecast for unemployment rate in 1996 and compare it with the AR\(p\) forecast and actual 1996 value.

**R:**

```r
newdat <- window(dat, start=thisyr+1, end=thisyr+1)
fcADL <- predict(ret, newdata=newdat)
res <- rbind(fcARp$mean[1], fcADL,
             window(unem,start=thisyr+1,end=thisyr+1))
rownames(res) <- c('AR(p)', 'ADL', 'Actual')
colnames(res) <- thisyr+1
print(res)
```

**Stata:**

```stata
predict fcADL if year>endyr-holdout, y
order year unem fcARp fcADL
list year unem fcARp fcADL if year>=endyr-holdout
```
When I teach this class, Week 16 is final exams week. There is no new material this week (since there are no classes). My final exam is cumulative: questions may be about any material from any time during the semester. The exception is that there are no questions about coding in R, although there may be some questions showing statistical results in R.
Bibliography


Index

AR(1) model, 282
AR(p) model, 298

ACE, see average causal effect
ADL, see autoregressive distributed lag
affine function, 181
after sampling, 17
AIC, see Akaike information criterion
Akaike information criterion, 299
alternative hypothesis, 80
analogy principle, 62
ASE, see average structural effect
associated with, 138, 161
asymptotic approximation, 71
asymptotic bias, 75
ATE, see average treatment effect
ATT, see average treatment effect on the treated
attenuation bias, 247
attrition, 111
autocorrelation, 265
partial, 298
autocorrelation coefficient, 265
autocovariance, 265
autoregressive coefficient, 282
autoregressive distributed lag, 302
autoregressive parameter, 282
average causal effect, 106
average structural effect, 142
average treatment effect, 106
average treatment effect on the treated, 214
base category, 158
Bayesian, 54
Bayesian information criterion, 299
before sampling, 17
Bernoulli random variable, 23
best linear approximation, 162
best linear predictor, 163
bias, 72
attenuation, 73
downward, 73
negative, 73
positive, 73
toward zero, 73
upward, 73
BIC, see Bayesian information criterion
binary variable, 23
BLA, see best linear approximation
black swan, 287
BLP, see best linear predictor
CASE, see conditional average structural effect
CATE, see conditional average treatment effect
categorical variable, 31
nominal, 31
ordinal, 31
causal inference, 99
CDF, see cumulative distribution function
CEF, see conditional expectation function
CEF error term, 136
central limit theorem, 71
CI, see confidence interval
classical, 54
CLT, see central limit theorem
clustered sampling, 60
coefficients, 137
collider, 208
collider bias, 209
common outcome, 208
complete case, 250
complete case analysis, 250
conditional average structural effect, 231
conditional average treatment effect, 232
conditional distribution, 130
conditional expectation function, 135
conditional independence, 208
conditional mean, 132
conditional probability, 130
conditioning variable, 134
confidence interval, 77
   one-sided, 78
   two-sided, 78
confidence level, 77
confounder, 97, 197
consistency, 75
OLS, 166
continuous variable, 34
contrapositive, 126
converse, 125
correctly specified, 157
counterfactual, 264
covariate, 134
covariance stationarity, 264
covariate, 134
covariance stationary, 264
covariate, 134
covariance, 264
credit probability, 77
gen variate, 55
critical value, 78, 82
 cumulative distribution function, 22
data, 54
data-generating process, 18
decomposition, 272
classical additive, 273
classical multiplicative, 273
remainder, 272
demeaned, 283
dependence, 134
dependent variable, 134
deterministic trend, 269
detrended, 273
DGP, see data-generating process
diff-in-diff, see difference-in-differences
difference stationary, 269
difference-in-differences, 205, 210
discrete variable, 26
distribution
   Bernoulli, 24
dummy variable, 23, 158
ECDF, see empirical CDF
economic significance, 88
empirical CDF, 62
empirical distribution, 61
endogenous, 144
error form, 128
error term, 128
CEF, 136
structural, 141
estimand, 63
exogenous, 144
expectation, see expected value
expected value, 25
external validity, 240
false positive, 81
finite-sample, 72
first difference, 263
first lag, 263
first lead, 263
first-order autoregressive model, 282
fitted value, 64
fitted values, 150
flexible, 187
forecast
   h-step-ahead, 287
   multi-step, 287
forecast interval, 289
forecasting, 259, 284
fourth moments, 165, 230
frequency weights, 64
frequentist, 54
fully saturated, 159, 204
functional form, 157

Gaussian distribution, 36
GE, see general equilibrium
genral equilibrium, 100
general equilibrium effects, 100, 105

hat, 62
heterogeneity, 103
heteroskedasticity, 151
heteroskedasticity-robust, 151
homoskedasticity, 151
hypothesis testing, 80

identically distributed, 58
identification, 109
identified, 109
identifying assumptions, 109, 143

if, 125
if and only if, 125
ignorability, 147, 208
iid, see independent and identically distributed
implied by, 125
implies, 124
in-sample, 263
included regressor, 197
independence, 147
independent, 134
independent and identically distributed, 57
independent but not identically distributed, 60
independent variable, 134
indicator function, 24
indicator variable, 23
inference, 76
information criterion, 299
inid, see independent but not identically distributed
intention-to-treat, 111
interaction term, 204, 225
internal validity, 240
interquartile range, 23
interval estimation, 63
interval forecast, 289
interval prediction, 50
invariant, 141

inverse, 125

joint distribution, 129
joint probability, 129

latent, 243
least squares, 64
left-hand side variable, 134
level, 81
linear combination, 181
linear projection, 160
linear projection coefficient, 161
linear-in-parameters, 181
linear-in-variables, 181
linearity, 29
log function, 174
log units, 176
loss
expected, 39
loss function, 37, 42–50
0–1, 42
$L_2$, 43
quadratic, 43
squared, 43
squared error, 43

LP, see linear projection
LPC, see linear projection coefficient

machine learning, 187
marginal distribution, 129
marginal probability, 129
Markov chain, 269
Markov property, 269
mean, 23, 25
mean independence, 134
mean squared error, 74
measurement error, 244
median, 23

misspecification

functional form, 157

misspecified, 156
mode, 23, 33, 50
model, 134

linear, 174
linear-log, 177
log-linear, 176
log-log, 178

model selection, 188
MSE, see mean squared error
multicollinearity
imperfect, 230
perfect, 229
multiple comparisons problem, 86
multiple testing problem, 86
natural experiments, 99, 210
necessary, 125
Neyman–Rubin causal model, 101
no interference, 104, 146
nominal coverage probability, 77
nominal variable, 31
non-compliance, 111
non-interference, 146
non-random attrition, 111
non-response bias, 252
nonlinear-in-parameters, 183
nonlinear-in-variables, 182
nonparametric regression, 187
nonstationary, 268
normal distribution, 36
standard, 37
null hypothesis, 80
object of interest, 63
OLS, see ordinary least squares
omitted variable bias, 196
only if, 125
ordinal variable, 31
ordinary least squares, 149
out-of-sample, 263
outcome variable, 134
OVB, see omitted variable bias
overfitting, 188
oversampled, 65
p-value, 79
parallel trends, 212
parameters, 137
partial equilibrium, 100
PDF, see probability density function
PE, see partial equilibrium
percentile, 23
percentiles, 30
plug-in principle, 62
PMF, see probability mass function
point estimation, 63
point forecast, 286
point prediction, 50
population, 18
finite, 18
infinite, 19
super-, 19
population of interest, 240
population studied, 240
posterior, 55
potential outcome
treated, 102
untreated, 102
power, 81
predicted value, 64
prediction, 41
prediction interval, 50
predictor, 134
prior, 55
probability density function, 35
probability mass function, 22
program evaluation, 99
projection, 159
properly specified, 157
publication bias, 87
quadratic model, 182
quantiles, 30
quasi-experiments, 210
random coefficient, 140
random draw, 16
random sample, 16, 57
random variable, 16
random walk, 268
randomized, 99
randomized controlled trial, 110
randomized experiment, 110
RCT, see randomized controlled trial
realization, 16
realized value, 16
reduced form, 99
regressand, 134
regression, 124
regressor, 134
rejection probability, 81
repeated sampling, 56
residual, 64, 150
response variable, 134
reverse causality, 252
right-hand side variable, 134
risk, 39
robust to heteroskedasticity, 151
see also Neyman–Rubin causal model

sample, 54
sample analog, 62, 165
sample average, 63
sample mean, 63
sample selection, 251
sample size, 57
sample values, 54
sampling
  independent, 58
sampling bias, 58
sampling distribution, 65
sampling frequency, 262
sampling weights, 64
script, 8
SE, see standard error
seasonal dummies, 302
seasonality, 270
second moments, 165, 230
selection on observables, 208
self-selection, 111
serial correlation, 265
simultaneity, 252
simultaneous causality, 252
size (of a hypothesis test), 81
slope coefficient, 137
spillover effects, 105
SSR, see sum of squared residuals
stable unit treatment value assumption, 104
standard deviation, 23, 26
standard error, 76
standard error of the mean, 76
stationarity
  strict, 264
  strong, 264
weak-end, 264
wide-sense, 264
stationary, 264
statistic, 62
statistical significance, 79
statistically significant, 79
stratified sampling, 59
strong ignorability, 147
stronger, 125
structural approach, 99
structural break, 272
structural effect, 142
structural model, 141
structural parameter, 141
sufficient, 125
sum of squared residuals, 64
survey weights, 64
SUTVA, see stable unit treatment value assumption
threats to validity, 240
time series, 262
time-varying coefficients, 287
treatment, 101
treatment dummy, 140
treatment effect, 101, 103
treatment indicator, 140
trend
  stochastic, 268
trend stationary, 269
trend–cycle component, 271
type I error, 81
type I error rate, 81
type II error, 81
type II error rate, 81
unbiased, 73
unconditional probability, 129
unconfoundedness, 147, 208
unit root, 268
units, 57
units of measure, 34

variance, 26
weaker, 125
white noise, 282
  independent, 268, 282