Introductory Econometrics
Description, Prediction, and Causality

First 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 43

4 One Variable, Two Populations 83

5 Midterm Exam #1 105

II Regression 107

Introduction 109
6 Comparing Two Distributions by Regression
111

7 Comparing Three (or More) Distributions
143

8 Nonlinear and Nonparametric Regression
163

9 Regression with Two Binary Regressors
189

10 Regression with Multiple Regressors
217

11 Midterm Exam #2
233

12 Internal and External Validity
235

III Time Series
255

Introduction
257

13 Time Series: One Variable
259

14 First-Order Autoregression
277

15 AR(p) and ADL Models
295

16 Final Exam
305

Bibliography
307

Index
313
Contents

List of Discussion Questions xiv
Preface xvii
Textbook Learning Objectives xix
Notation xxi

1 Getting Started with R (or Stata) 1
  1.1 Comparison of R and Stata 1
  1.2 R 2
    1.2.1 Accessing the Software 2
    1.2.2 Installing Packages 5
  1.3 Stata 5
    1.3.1 Accessing the Software 5
    1.3.2 Installing Additional Commands 7
  1.4 Optional Resources 7
    1.4.1 R Tutorials 7
    1.4.2 R Quick References 8
    1.4.3 Running Code in This Textbook 8
    1.4.4 Stata Resources 8
    Empirical Exercises 9

1 Analysis of One Variable 11

  Introduction 13
2 One Variable: Population

2.1 The World is Random

2.1.1 Before and After: Two Perspectives

2.1.2 Outcomes and Mechanisms

2.1.3 Probability Distributions

2.1.4 Sums and Averages are Random

2.1.5 Random Terminology

2.1.6 Employment is Random

2.2 The Population

2.2.1 Population Types

2.2.2 Types of Population Distributions

2.3 Description of a Population Distribution

2.3.1 Units of Measure

2.3.2 Complete Description: Distributions

2.3.3 Population Summary: Mean

2.3.4 Population Summary: Standard Deviation

2.3.5 Continuous Distributions

2.4 Prediction with a Known Distribution

2.4.1 Example: Rain

2.4.2 Loss Functions

2.4.3 Optimal Prediction: Generic Examples

2.4.4 Optimal Prediction: Specific Examples

2.4.5 Mean and Mode as Optimal Predictions

2.4.6 Interval Prediction

3 One Variable: Sample

3.1 Bayesian and Frequentist Perspectives

3.2 Types of Sampling

3.3 Estimation

3.3.1 The Sample Mean and Least Squares Estimation

3.3.2 Non-iid Sampling: Survey Weights

3.3.3 Brief Bayesian Comparison

3.4 Estimator Properties

3.4.1 Bias

3.4.2 Mean Squared Error

3.4.3 Asymptotic Theory

3.4.4 Consistency

3.5 Quantifying Uncertainty

3.5.1 Frequentist and Bayesian Views of Uncertainty

3.5.2 Sampling Distribution
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5.3 Standard Errors</td>
<td>65</td>
</tr>
<tr>
<td>3.5.4 Confidence Intervals</td>
<td>66</td>
</tr>
<tr>
<td>3.5.5 ( p )-values</td>
<td>69</td>
</tr>
<tr>
<td>3.5.6 Statistical and Economic Significance</td>
<td>72</td>
</tr>
<tr>
<td>3.5.7 Hypothesis Testing</td>
<td>75</td>
</tr>
<tr>
<td>3.5.8 Mental Math for Statistical Uncertainty</td>
<td>78</td>
</tr>
<tr>
<td>3.5.9 Statistical Decision Theory</td>
<td>79</td>
</tr>
<tr>
<td>Empirical Exercises</td>
<td>80</td>
</tr>
<tr>
<td>4 One Variable, Two Populations</td>
<td>83</td>
</tr>
<tr>
<td>4.1 Description</td>
<td>85</td>
</tr>
<tr>
<td>4.1.1 Population Mean Difference</td>
<td>85</td>
</tr>
<tr>
<td>4.1.2 Estimation</td>
<td>86</td>
</tr>
<tr>
<td>4.1.3 Quantifying Uncertainty</td>
<td>86</td>
</tr>
<tr>
<td>4.2 Prediction</td>
<td>87</td>
</tr>
<tr>
<td>4.3 Causality</td>
<td>88</td>
</tr>
<tr>
<td>4.3.1 Some Dichotomies</td>
<td>88</td>
</tr>
<tr>
<td>4.3.2 Differences with Description and Prediction</td>
<td>89</td>
</tr>
<tr>
<td>4.3.3 Potential Outcomes Framework</td>
<td>91</td>
</tr>
<tr>
<td>4.3.4 Average Treatment Effect</td>
<td>95</td>
</tr>
<tr>
<td>4.3.5 Identification: ATE</td>
<td>98</td>
</tr>
<tr>
<td>4.3.6 Estimation and Inference</td>
<td>101</td>
</tr>
<tr>
<td>Empirical Exercises</td>
<td>103</td>
</tr>
<tr>
<td>5 Midterm Exam #1</td>
<td>105</td>
</tr>
<tr>
<td>II Regression</td>
<td>107</td>
</tr>
<tr>
<td>Introduction</td>
<td>109</td>
</tr>
<tr>
<td>6 Comparing Two Distributions by Regression</td>
<td>111</td>
</tr>
<tr>
<td>6.1 Logic</td>
<td>112</td>
</tr>
<tr>
<td>6.1.1 Terminology</td>
<td>113</td>
</tr>
<tr>
<td>6.1.2 Assumptions</td>
<td>114</td>
</tr>
<tr>
<td>6.1.3 Theorems</td>
<td>115</td>
</tr>
<tr>
<td>6.2 Preliminaries</td>
<td>116</td>
</tr>
<tr>
<td>6.2.1 Population Mean Model in Error Form</td>
<td>116</td>
</tr>
<tr>
<td>6.2.2 Joint and Marginal Distributions</td>
<td>116</td>
</tr>
<tr>
<td>6.2.3 Conditional Distributions</td>
<td>117</td>
</tr>
</tbody>
</table>
### 6.2.4 Conditional Mean
- Page 119

### 6.2.5 Comparison of Joint, Marginal, and Conditional Distributions
- Page 120

### 6.2.6 Independence and Dependence
- Page 120

### 6.3 Population Model: Conditional Expectation Function
- Page 121
  - 6.3.1 Conditional Expectation Function
    - Page 122
  - 6.3.2 CEF Error Term
    - Page 122
  - 6.3.3 CEF Model in Error Form
    - Page 123
  - 6.3.4 Linear CEF
    - Page 123
  - 6.3.5 Interpretation, Description, and Prediction
    - Page 124
  - 6.3.6 Interpretation with Values Besides 0 and 1
    - Page 124

### 6.4 Population Model: Potential Outcomes
- Page 126

### 6.5 Population Model: Structural
- Page 127
  - 6.5.1 Linear Model
    - Page 127
  - 6.5.2 General Model
    - Page 128
  - 6.5.3 ASE and CEF
    - Page 129

### 6.6 Identification
- Page 130
  - 6.6.1 Structural Model
    - Page 130
  - 6.6.2 Average Treatment Effect
    - Page 133
  - 6.6.3 Average Structural Effect
    - Page 135

### 6.7 Estimation: OLS
- Page 135

### 6.8 Quantifying Uncertainty
- Page 137
  - 6.8.1 Types of Uncertainty
    - Page 138
  - 6.8.2 Heteroskedasticity
    - Page 138
  - 6.8.3 Code
    - Page 139

### 6.6.4 Empirical Exercises
- Page 142

### 7 Comparing Three (or More) Distributions
- Page 143

### 7.1 Misspecification
- Page 144

### 7.2 Coping with Misspecification
- Page 146

### 7.3 Linear Projection
- Page 148

### 7.4 “Best” Linear Approximation
- Page 150

### 7.5 “Best” Linear Predictor
- Page 151

### 7.6 Causality Under Misspecification
- Page 152

### 7.7 OLS Estimation and Inference
- Page 152
  - 7.7.1 OLS Estimator Insights
    - Page 152
  - 7.7.2 Statistical Properties
    - Page 153
  - 7.7.3 Code
    - Page 155

### 7.8 Simple Linear Regression
- Page 157

### 8 Nonlinear and Nonparametric Regression
- Page 163
# Contents

8.1 Log Transformation ........................................... 164
   8.1.1 Properties of the Natural Log Function ................. 165
   8.1.2 The Log-Linear Model .................................. 166
   8.1.3 The Linear-Log Model .................................. 168
   8.1.4 The Log-Log Model .................................... 169
   8.1.5 Warning: Model-Driven Results ........................ 170
   8.1.6 Code ................................................. 170

8.2 Nonlinear-in-Variables Regression ......................... 171
   8.2.1 Linearity ........................................... 173
   8.2.2 Nonlinearity ......................................... 174
   8.2.3 Estimation and Inference .............................. 175
   8.2.4 Parameter Interpretation .............................. 175
   8.2.5 Description, Prediction, and Causality ............... 176
   8.2.6 Other Nonlinear Models ............................... 178
   8.2.7 Examples ........................................... 178

8.3 Nonparametric Regression .................................. 181

Empirical Exercises ............................................ 184

9 Regression with Two Binary Regressors ..................... 189
   9.1 Omitted Variable Bias .................................... 191
      9.1.1 An Allegory ....................................... 191
      9.1.2 Formal Conditions .................................. 191
      9.1.3 Consequences ...................................... 193
      9.1.4 OVB in Linear Projection ............................ 195
   9.2 Linear-in-Variables Model ............................... 196
   9.3 Fully Saturated Model ..................................... 197
   9.4 CEF Coefficient Interpretation ........................... 198
   9.5 Structural Identification by Exogeneity ............... 200
   9.6 Identification by Conditional Independence ............ 200
      9.6.1 Setup ............................................ 201
      9.6.2 Formal Results ..................................... 202
      9.6.3 Potential Outcomes and CIA ........................ 202
   9.7 Causal Identification: Difference-in-Differences ... 205
      9.7.1 Bad Approaches .................................... 205
      9.7.2 Counterfactuals and Parallel Trends ............... 207
      9.7.3 Identification ...................................... 209
      9.7.4 Extensions .......................................... 211
   9.8 Estimation and Inference ................................ 211

Empirical Exercises ............................................ 213

10 Regression with Multiple Regressors .................... 217

13.2 Stationarity .................................................. 261
13.3 Autocovariance and Autocorrelation ......................... 262
13.4 Estimation ..................................................... 263
   13.4.1 Mean ................................................... 263
   13.4.2 Autocovariances and Autocorrelations ................. 264
13.5 Nonstationarity ............................................ 265
   13.5.1 Trends ................................................ 265
   13.5.2 Seasonality ......................................... 267
   13.5.3 Cycles ............................................... 268
   13.5.4 Decomposition ................................. 269
13.6 Transformations .......................................... 273
Empirical Exercises ........................................... 274

14 First-Order Autoregression ................................... 277
14.1 Model ....................................................... 278
14.2 Description ................................................. 279
14.3 Prediction (Forecasting) ................................... 281
14.4 Estimation .................................................. 282
14.5 Parameter Stability ...................................... 284
14.6 Multi-Step Forecast ...................................... 285
14.7 Interval Forecasts ........................................ 287
14.8 More R Examples ........................................ 289
Empirical Exercises ........................................... 292

15 AR(p) and ADL Models ........................................ 295
15.1 The AR(p) Model ........................................... 296
15.2 Model Selection: How Many Lags? ......................... 296
   15.2.1 Problems and Intuition ............................ 296
   15.2.2 AIC and BIC Formulas ............................ 297
   15.2.3 Comparison of AIC and BIC ..................... 298
   15.2.4 Code ............................................... 299
15.3 Autoregressive Distributed Lag Regression ................ 300
Empirical Exercises ........................................... 302

16 Final Exam .................................................. 305

Bibliography .................................................... 307

Index ............................................................ 313
# List of Discussion Questions

<table>
<thead>
<tr>
<th>Section</th>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>web traffic</td>
<td>20</td>
</tr>
<tr>
<td>2.2</td>
<td>student data</td>
<td>22</td>
</tr>
<tr>
<td>2.3</td>
<td>banana loss function</td>
<td>40</td>
</tr>
<tr>
<td>2.4</td>
<td>optimal banana prediction</td>
<td>41</td>
</tr>
<tr>
<td>3.1</td>
<td>rural household sampling</td>
<td>47</td>
</tr>
<tr>
<td>3.2</td>
<td>probability of positive mean</td>
<td>50</td>
</tr>
<tr>
<td>3.3</td>
<td>estimator MSE</td>
<td>55</td>
</tr>
<tr>
<td>3.4</td>
<td>equal p-values, equal belief?</td>
<td>71</td>
</tr>
<tr>
<td>3.5</td>
<td>significance: distance and education</td>
<td>74</td>
</tr>
<tr>
<td>3.6</td>
<td>jellybean solution?</td>
<td>77</td>
</tr>
<tr>
<td>3.7</td>
<td>nova</td>
<td>77</td>
</tr>
<tr>
<td>3.8</td>
<td>Ebola drug</td>
<td>79</td>
</tr>
<tr>
<td>4.1</td>
<td>DPC with two populations</td>
<td>84</td>
</tr>
<tr>
<td>4.2</td>
<td>description, prediction, causality</td>
<td>91</td>
</tr>
<tr>
<td>4.3</td>
<td>cash transfer spillovers</td>
<td>95</td>
</tr>
<tr>
<td>4.4</td>
<td>unrepresentative ATE</td>
<td>98</td>
</tr>
<tr>
<td>4.5</td>
<td>breakfast effect?</td>
<td>101</td>
</tr>
<tr>
<td>6.1</td>
<td>median theorem logic</td>
<td>115</td>
</tr>
<tr>
<td>6.2</td>
<td>mean theorem logic</td>
<td>115</td>
</tr>
<tr>
<td>6.3</td>
<td>logic with feathers</td>
<td>115</td>
</tr>
<tr>
<td>6.4</td>
<td>joint distribution and causality</td>
<td>117</td>
</tr>
<tr>
<td>6.5</td>
<td>regression parameter units</td>
<td>124</td>
</tr>
<tr>
<td>6.6</td>
<td>ES habits and final scores</td>
<td>130</td>
</tr>
<tr>
<td>6.7</td>
<td>marriage and salary</td>
<td>132</td>
</tr>
<tr>
<td>6.8</td>
<td>heteroskedasticity</td>
<td>138</td>
</tr>
<tr>
<td>6.9</td>
<td>regression significance</td>
<td>141</td>
</tr>
<tr>
<td>7.1</td>
<td>misspecification</td>
<td>145</td>
</tr>
<tr>
<td>7.2</td>
<td>Facebook</td>
<td>147</td>
</tr>
<tr>
<td>7.3</td>
<td>BLP</td>
<td>151</td>
</tr>
</tbody>
</table>
LIST OF DISCUSSION QUESTIONS

7.4 income dummies .................................................. 158
7.5 linear fit ......................................................... 158
7.6 regression units .................................................. 158
7.7 regression significance ........................................... 160
8.1 pollution and house price ........................................ 170
8.2 nonlinear OVB .................................................... 171
8.3 quadratic example ................................................ 176
8.4 nonlinear wage model interpretation ............................ 176
8.5 model evaluation ................................................ 183
9.1 assessing OVB .................................................... 193
9.2 OVB: kindergarten ................................................ 195
9.3 OVB: ES habits .................................................... 195
9.4 binary interaction .................................................. 199
9.5 bad panel approach #1, for Mariel boatlift ................. 205
9.6 bad panel approach #2, for Mariel boatlift ................. 206
9.7 bad panel approach #1, for fracking .......................... 206
9.8 bad panel approach #2, for fracking .......................... 206
9.9 parallel trends skepticism ....................................... 209
10.1 OVB with multiple regressors .................................. 219
10.2 sleep and interactions .......................................... 222
10.3 wage and interactions ......................................... 222
10.4 linear-in-variables? ............................................. 223
12.1 external validity: minimum wage .............................. 238
12.2 exercise error .................................................... 241
12.3 measurement error: scrap rate ............................... 243
12.4 program attrition ................................................ 249
12.5 missing salary data ............................................. 249
12.6 health and medical expenditure .............................. 252
13.1 autocorrelation ................................................... 263
13.2 nonstationarity .................................................... 270
14.1 forecast and reality ............................................. 282
14.2 recession-affected coefficient ................................ 285
14.3 long-horizon AR forecast ...................................... 286
14.4 forecast sanity check ............................................ 288
15.1 lag choice for forecasting ...................................... 298
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," meaning estimation of both structural and causal models. Description and prediction (forecasting) with time series are also covered. Students will 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 with any text, this text may be used to teach a few types of classes. In full, the text provides a 15-week semester class that assumes a previous class in probability and statistics. If desired, the prerequisite could be skipped if more time is spent on the "review" material (and later chapters skipped). Calculus is mostly avoided, except for derivatives of polynomials to interpret nonlinear-in-variables regression models; calculus could be skipped entirely, or more could be added (e.g., for expectations of continuous random variables). For the quarter system, a quarter of econometrics could be taught using this text without the time series material, which could be incorporated into a subsequent quarter-long econometric class. Of course, any material may be expanded, condensed, or skipped, as the instructor desires.


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, with usage 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 all the R code and output is in the same .Rtex files alongside the \LaTeX code. Many graphs are generated this way, too, so you can easily change them (formatting, data, function, methods, etc.); most of the rest are generated from a single .R file also provided in the source material. You may see, copy, and download the entire project at: https://www.overleaf.com/read/fszrgmwzftrk

Third, I provide learning objectives for the overall book and for each chapter. This follows current best practices for course design. In future work, I hope to provide a library of (good!) multiple choice questions and empirical exercises, labeled with the corresponding learning objective.

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
Columbia, Missouri, USA
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 $\mathbf{X}$ defined above is

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

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.

- $y$ scalar fixed (non-random) value
- $Y$ scalar random variable
- $\theta$ scalar non-random value
- $\hat{\theta}$ scalar random variable
- $\mathbf{x}$ non-random column vector
- $\mathbf{x}'$ transpose of $\mathbf{w}$
- $\mathbf{X}$ random column vector
- $\beta$ non-random column vector
- $\hat{\beta}$ random column vector
- $\mathbf{w}$ non-random matrix
- $\mathbf{w}'$ transpose of $\mathbf{w}$
- $\mathbf{W}$ random matrix
- $\Omega$ non-random matrix
- $\hat{\Omega}$ random matrix

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
\(\equiv\) 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 3.4.4
\(\Rightarrow\) converges to (like in pre-calculus)
\(\hat{p}_n\) converges in probability to; see Section 3.4.4
\(\equiv\) is defined as
\(\approx\) approximately equals
\(\sim\) is distributed as
\(\tilde{\sim}\) is distributed approximately (or asymptotically) as; see (3.25)
\(X \perp Y\) \(X\) 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 2.1.3
\(f_Y(\cdot)\) PMF of \(Y\) (if \(Y\) is discrete); see Section 2.1.3
\(f_Y(\cdot)\) PDF of \(Y\) (if \(Y\) is continuous); see Figure 2.1
\(\mathds{1}\{\cdot\}\) indicator function; see (2.2)
\(P(A)\) probability of event \(A\)
\(P(A \mid B)\) conditional probability of \(A\) given \(B\); see Section 6.2.3
\(\mathbb{E}(Y)\) expectation (mean) of \(Y\); see Section 2.3.3
\(\mathbb{E}(Y \mid X = x)\) CEF (a function of \(x\)); see Section 6.3
\(\mathbb{E}(Y \mid X)\) conditional expectation of \(Y\) given \(X\); this is a random variable
\[\sum_{i=1}^{n}\] summation from \(i = 1\) to \(i = n\)
\(\text{Var}(Y)\) variance of \(Y\); see Section 2.3.4
\(\text{Var}(Y \mid X = x)\) conditional variance (a non-random value); see Section 6.8.2
\(\text{Var}(Y \mid X)\) conditional variance (a random variable)
\(\text{Cov}(Y, X)\) covariance
\(\text{Corr}(Y, X)\) correlation
\(\{a, b, \ldots\}\) a set (containing elements \(a, b,\) etc.)
\(s \in \mathcal{S}\) element \(s\) is in set \(\mathcal{S}\)
\(\hat{\mathbb{E}}(Y)\) expectation for sample distribution; see Section 3.3.1
\(\hat{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i\); same as \(\hat{\mathbb{E}}(Y)\); see Section 3.3.1
\(\hat{\theta}\) estimator of population parameter \(\theta\); see Section 3.3
\(\text{SE}(\hat{\theta})\) standard error of estimator \(\hat{\theta}\); see Section 3.5.3
\(\arg \min_{g} f(g)\) the value of \(g\) that minimizes \(f(g)\)
\(\arg \max_{g} f(g)\) the value of \(g\) that maximizes \(f(g)\)
\(\mathbf{y}', \mathbf{x}'\) transposes of matrix \(\mathbf{y}\) and vector \(\mathbf{x}\), respectively
\( y^{-1} \) inverse of matrix \( y \)
Chapter 1

Getting Started with R (or Stata)

Depends on: no other chapters

<table>
<thead>
<tr>
<th>Unit learning objectives for this chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. Run statistical software (R/RStudio or Stata) [TLO 7]</td>
</tr>
<tr>
<td>1.2. Write code to do basic data manipulation, description, and display [TLO 7]</td>
</tr>
</tbody>
</table>

You will use R (or Stata) for the empirical exercises in this textbook. 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 usually lots of explicit hints about the code you need to write. If you actually do have previous experience (or have above-average interest in coding), 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 try one of the free DataCamp online courses to further develop your R skills, which are very valuable for many good jobs.

Due to the many excellent resources like DataCamp, 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.
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 (with limited capability).
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.

The examples in the text are all in R, but you can use Stata for the empirical exercises if you have a strong preference (e.g., if your current job requires you to use Stata). The empirical exercise instructions should be fairly explicit (helpful), but you can peek ahead at future Stata exercises to see if you want to try Stata. You may ask questions about Stata on the course website, too.

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, or on another computer like in a campus computer lab.

Note that 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

At the website https://mybinder.org/v2/gh/binder-examples/r/master?filepath=index.ipynb you can select File–New Notebook–R to start your own. You can install any “package” with the `install.packages()` function. This is sufficient to complete the empirical exercises in this textbook. Basically, you can type your code, and then hit Control-Enter to run it and see the output. (If you have any issues, you can look at the help, or Google how to use Jupyter notebooks with R.) One time, there was a problem with the default server from which R was trying to download packages, but it was fixed by running the following code first:

```r
(r <- getOption('repos'))
r['CRAN'] <- 'https://rweb.crmda.ku.edu/cran/
options(repos=r)
```

After running that code, I was able to run `install.packages()` like usual. However, if this doesn’t work either, you could try replacing the `https://rweb.crmda.ku.edu/cran` / with one of the many other options from https://cran.r-project.org/mirrors.html

At my website https://faculty.missouri.edu/~kaplandm/intro_text/DataCamp.html I hope in the future to have something simple yet sufficient for empirical exercises. Currently, it does not have all the necessary packages available, nor can you install packages. Also, I have only tested that it works in Firefox on Windows (and it does not work in Chrome on Windows).

In a Mizzou Computer Lab

The page https://doit.missouri.edu/services/computing-sites/sites-software shows you which on-campus (Mizzou) computer labs (“computing sites”) have certain software. If you scroll down to RStudio, you can see where you can use R with RStudio. Currently (Spring 2019), this is just Middlebush 7 and 8. Also note that 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. The overall list of computer labs is at https://doit.missouri.edu/services/computing-sites and clicking on the name shows more detail, including a daily calendar; e.g., the Middlebush 7 page looks like https://doit.missouri.edu/services/computer-lab/computer-lab-locations/middlebush-7/.

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.

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 R: either Google “R Windows” or try 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
- Double-click the downloaded .exe file and follow instructions to install.

On Mac:

- Download the .pkg file for R: Google “R Mac” or try https://cran.r-project.org/bin/macosx
- Follow the usual Mac installation procedure (double-click, and then drag something into Applications when it tells you, etc.).
- Download the .dmg file for the free version of RStudio Desktop: Google “RStudio download” or try https://www.rstudio.com/products/rstudio/download/#download
- Follow the usual Mac installation procedure (double-click and drag the application into your Applications folder).

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.
1.3. STATA

1.2.2 Installing Packages

Depending how you access R, you may need to install certain packages to complete the empirical exercises. This is very easy: it 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:

\[
\text{install.packages(c('wooldridge','lmtest','sandwich', 'forecast','survey'))}
\]

If you’re running 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.

A bit about the packages:

- **wooldridge** [Shea, 2018] has the datasets we’ll use, 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.


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

The page [https://doit.missouri.edu/services/computing-sites/sites-software](https://doit.missouri.edu/services/computing-sites/sites-software) shows you which on-campus (Mizzou) computer labs (“computing sites”) have certain software. If you scroll down to Stata, you can see which labs have Stata installed; there are many, but not all. Note that 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. The overall list of computer labs is at [https://doit.missouri.edu/services/computing-sites](https://doit.missouri.edu/services/computing-sites) and clicking on the name shows more detail, including a daily calendar; e.g., the Middlebush 7 page looks like [https://doit.missouri.edu/services/computer-lab/computer-lab-locations/middlebush-7/](https://doit.missouri.edu/services/computer-lab/computer-lab-locations/middlebush-7/).

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 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

For pricing, see [https://www.stata.com/order/new/edu/gradplans/student-pricing/](https://www.stata.com/order/new/edu/gradplans/student-pricing/). The cheapest option is (currently) the 6-month Stata/IC license; the student price is $45, although the price may differ in other countries. 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)

At [https://doit.missouri.edu/services/software/software-anywhere](https://doit.missouri.edu/services/software/software-anywhere) click the “Getting Started” tab. Follow the instructions. (Currently: 1) download the client software installer, 2) install it, 3) run it, 4) double-click the icon, 5) enter your log-in credentials.) Note: make sure to use the correct DOMAIN when logging into the server.

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
1.4. OPTIONAL RESOURCES

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), which are described at http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html.

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.

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 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 available free on their website]
5. Courses at datacamp.com like Introduction to R, the first chapter of each course is free
6. Chapter 4 in Kaplan (2018)
1.4.2 R Quick References

These are some references, as an alternative to using Google.


2. [https://cran.r-project.org/doc/contrib/Short-refcard.pdf](https://cran.r-project.org/doc/contrib/Short-refcard.pdf)

1.4.3 Running Code in This Textbook

If you’re curious, you should be able to copy code directly from this textbook (the .pdf version) and paste it into R. In some cases, 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 the package `mgcv`, simply run the command `install.packages('mgcv')` within R.

This text’s R code for graphs refers to certain variables that have been set ahead of time. To run the graphing code yourself, you must first run code to set these parameters, such as:

```
# Graphical parameters for consistent look
CEXAXIS <- 1.0; CEXMAIN <- 1.0; CEXLAB <- 1.0; CEXLEG <- 1.0
WIDTHLARGE <- 7; HEIGHTLARGE <- 5
WIDTHMED <- 5; HEIGHTMED <- 5
WIDTHSMALL <- 4; HEIGHTSMALL <- 3
LWD <- 1
PARFAM <- 'serif'
PARMGP <- c(2.8, 1, 0)
PARMAR <- c(4, 4, 0.1, 0.1)
options(width=62)
```

1.4.4 Stata Resources

For Stata, helpful cheat sheets (quick references) are available for free in a .pdf file at [https://www.stata.com/bookstore/statacheatsheets.pdf](https://www.stata.com/bookstore/statacheatsheets.pdf) and various tutorials are linked from [https://www.stata.com/links/resources-for-learning-stata/](https://www.stata.com/links/resources-for-learning-stata/)
Empirical Exercises

Empirical Exercise EE1.1. If using R, first install the \texttt{wooldridge} package if it’s not already installed. You can print a list of installed packages with the R command \texttt{sort(row.names(installed.packages()))} and can install the package with the command \texttt{install.packages("wooldridge")}. Then, in either R or Stata, create a \texttt{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 package \texttt{wooldridge} with command \texttt{library(wooldridge)} and view the package help with command \texttt{help(package='wooldridge')}. Stata: run \texttt{ssc install bcuse} to ensure command \texttt{bcuse} is installed, and then load the dataset with \texttt{bcuse wine, clear}

b. View basic info for the \texttt{wine} dataset with the R command \texttt{?wine} or Stata command \texttt{describe}

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

d. Rename the \texttt{alcohol} column, which measures liters of alcohol from wine (consumed per capita per year), with R command \texttt{names(wine)[2] <- 'wine'} or alternatively \texttt{names(wine)[names(wine)=='alcohol'] <- 'wine'} or Stata command \texttt{rename alcohol wine}

e. Add a column named \texttt{id} whose value is just 1, 2, 3, 4, 5, etc., using R command \texttt{wine$id <- 1:nrow(wine)} or Stata command \texttt{generate id = _n}

f. Display the countries with fewer than 100 heart disease deaths per 100,000 people with R command \texttt{wine$countrypo[heart<100]} or the Stata command \texttt{list country if heart<100}

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

h. Add a column with the sum of heart and liver disease deaths per 100,000 with R command \texttt{wine$heart.plus.liver <- wine$heart + wine$liver} or Stata command \texttt{generate heartplusliver = heart + liver}

i. Add squared death rate in R with \texttt{wine$deaths.sq <- wine$deaths^2} or Stata command \texttt{generate deaths_sq = deaths^2}

j. Display the sorted death rates with R command \texttt{print(sort(wine$deaths))} or Stata command \texttt{sort deaths} followed by \texttt{list deaths}

k. R: create a vector with the proportion of total deaths (per 100,000) caused by heart disease with command \texttt{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 with R command `hist(wine$liver)` or Stata command `histogram liver`

m. Create a scatterplot of liver death rates (vertical axis) against wine consumption (horizontal axis) with R command `plot(x=wine$wine, y=wine$liver)` or Stata command `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
Part I

Analysis of One Variable
Introduction

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

1. Description (how things are/were)
2. Prediction (guessing an unknown value)
3. Causality (how one thing changes as a result of another changing)

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 individuals in the U.S. have an income below $20,000/yr? What is the median income? What is the difference between the 90th percentile and 10th percentile? How does income differ between two socioeconomic or demographic groups, like male and female, black and white, or college degree and no college degree?

2. Prediction: what’s the best prediction of the income of an unknown person, say for advertising purposes? How does the prediction change if you know more about the person, like their sex, race, and education level? What if you need to predict a range of income that should include the unknown person’s true income half the time?

3. Causality: for a given individual, how much higher would her income be if she had a college degree than if she didn’t, holding everything else about her (even her DNA) constant? How much higher 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 now-ubiquitous task of advertising, where correctly guessing a person’s income and other characteristics is crucial to the success of the advertisement. In other private
sector jobs, you may need to predict future demand to know how many Tesla sedans to start producing, or predict future oil prices for an airline. In government or non-profit work, optimal policy may depend on predicting how much an individual will benefit from a welfare or training program, or predicting how the economy will fare next year.

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 consumer to buy more? Among the three types, causal questions are the most difficult to answer statistically; this has been a primary focus of the field of econometrics.
Chapter 2

One Variable: Population

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 3]

2.5. Compute expected 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: https://www.khanacademy.org/math/ap-statistics

- Expected value (Lambert video): https://www.youtube.com/watch?v=6XqICKT1Kug

- Probability distribution basics (more than you need to know for this class): https://en.wikipedia.org/wiki/Probability_distribution
In Chapter 2, we study a single variable, like income. The setting’s simplicity helps us focus on the complexity of fundamental concepts like sampling, description, and prediction. This fundamental understanding will help us tackle more complex models in the rest of this text, as well as any future statistics or econometrics material you encounter.

Depending on your previous probability and statistics class(es), you may wish to spend more (or less) time in sections with material that you do not (or do) remember well.

2.1 The World is Random

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, say with probability $1/2$ each. (Writing $1/2$, 0.5, or 50% all indicate the same probability.)

The “after” view sees $W$ as a realized value. It is either heads or tails. The actual “value” (heads or tails) may be unknown to us, but 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. In the remainder of the text, values (as opposed to random variables) are usually written as lowercase letters, like $w$ instead of $W$.

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 with different probabilities. In the coin flip example, the possible outcomes are $(h, t)$, and the corresponding probabilities are $(0.5, 0.5)$.

2.1.2 Outcomes and Mechanisms

A realized value like $W = h$ says nothing about how the value was generated, but the random variable $W$ does. For example, suppose two different coins were flipped, and both came up heads. This single result (or realization) being the same does not imply the
coins are identical. Similarly, identical coins could produce different outcomes, just by chance; e.g., you could flip the same coin twice but get two different outcomes. However, identical coins would be described by random variables with the same properties. That is, if coin A is described by random variable $W^A$ with $P(W^A = h) = 0.5$, then an identical coin B would also have $P(W^B = h) = 0.5$. The “before” view with random variables provides a way to describe properties of the coins that we may want to learn about.

The coin flip is a metaphor for more complex mechanisms. In economics, we are interested in many different mechanisms that determine how the world is: how wages are determined, how inflation is determined, etc. Naturally, the more complex mechanisms have many more unknown parameters, too.

### 2.1.3 Probability Distributions

#### Probability Mass Functions

The probability distribution of a random variable can be written as a **probability mass function** (PMF). The PMF states the probability of the random variable equalling each possible value. For example, with the coin flip $W$, the PMF is

$$P(W = h) = 1/2, \quad P(W = t) = 1/2.$$  \hfill (2.1)

Equivalently, the PMF $f_W(\cdot)$ is

$$f_W(h) = 1/2, \quad f_W(t) = 1/2,$$

where $f_W(j) \equiv P(W = j)$ for any possible value $j$ (here, $j = h$ or $j = t$).

#### Binary Random Variables

Representing heads and tails by 1 and 0 facilitates analysis. Define the **indicator function** $1\{\cdot\}$ that equals 1 if the argument is true and 0 if false:

$$1\{A\} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false}. \end{cases} \hfill (2.2)$$

Define a new random variable $Y$, where $Y = 1$ means heads ($W = h$) and $Y = 0$ means tails ($W = t$):

$$Y = 1\{\text{heads}\} = 1\{W = h\} = \begin{cases} 1 & \text{if } W = h \text{ (heads)} \\ 0 & \text{if } W = t \text{ (tails)}. \end{cases}$$  \hfill (2.3)

One interpretation is that $Y$ counts the number of flips that are heads; with a single flip, it can only be 0 or 1.
The following properties apply to any random variable like \( Y \) in (2.3) that takes only 0 and 1 values. Such variables have many names, including binary variable, Bernoulli random variable, indicator variable, and dummy variable. Each follows a Bernoulli distribution with parameter \( p \), indicating the probability of equaling one:

\[
P(Y = 1) = p, \quad P(Y = 0) = 1 - p.
\]

Using (2.4) and the formula for a discrete distribution’s mean (summing products of values and probabilities),

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

The mean \( E(Y) \) is also called the expected value or expectation, hence \( E \). However, these are technical terms with different meaning than in colloquial English (e.g., “I expect the value will be three”), which often causes confusion.

Cumulative Distribution Functions

Instead of using the PMF to describe a distribution, the cumulative distribution function (CDF) (sometimes just “distribution function”) can be used. The letter \( F \) is commonly used for CDFs. Sometimes a subscript naming the random variable is added for clarity. For example, \( F_Y(\cdot) \) is the CDF of random variable \( Y \), whereas \( F_X(\cdot) \) is the CDF of \( X \). The CDF takes any real number \( r \) as its input and reports the probability that the random variable is less than or equal to \( r \). For example, generally, the CDF of \( Y \) is

\[
F_Y(r) \equiv P(Y \leq r), \quad \text{for any } -\infty < r < \infty.
\]

Consider the CDF of the coin flip \( Y \). If \( r < 0 \), then \( P(Y \leq r) = 0 \). If \( r \geq 1 \), then \( P(Y \leq r) = 1 \). If \( r = 0 \), then \( P(Y \leq r) = P(Y = 0) \). If \( 0 < r < 1 \), then \( P(Y \leq r) = P(Y = 0) \), too. Altogether, letting \( p \equiv P(Y = 1) \) and \( 1 - p = P(Y = 0) \), the CDF of \( Y \) is

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

2.1.4 Sums and Averages are Random

Consider multiple coin flips. Imagine flipping two coins, represented by \( Y_1 \) and \( Y_2 \). The sum \( S = Y_1 + Y_2 \) counts how many coins came up heads (0, 1, or 2). Like \( Y \) before,
2.1. THE WORLD IS RANDOM

this $S$ can be seen as a fixed (already observed) number in the “after” perspective, or as a random variable in the “before” perspective. That is, before we flip the coins, $S = 0$, $S = 1$, and $S = 2$ are all possible, with different probabilities. For example, flipping two fair coins independently, $P(Y_1 = 1) = P(Y_2 = 1) = 1/2$ (fair) and $Y_1 \perp Y_2$ (independent), so

$$P(S = 0) = P(Y_1 + Y_2 = 0) = P(Y_1 = 0 \text{ and } Y_2 = 0)$$
$$= P(Y_1 = 0)P(Y_2 = 0) = (1/2)(1/2) = 1/4,$$
$$P(S = 2) = P(Y_1 + Y_2 = 2) = P(Y_1 = 1 \text{ and } Y_2 = 1)$$
$$= P(Y_1 = 1)P(Y_2 = 1) = (1/2)(1/2) = 1/4,$$
$$P(S = 1) = 1 - P(S = 0) - P(S = 2) = 1 - (1/4) - (1/4) = 1/2. \quad (2.8)$$

The PMF in (2.8) can then be used to determine the distribution of the sample proportion of flips coming up heads. Specifically, $\bar{Y} = (Y_1 + Y_2)/2 = S/2$ is the sample average and represents the sample proportion of heads. It can be seen as a random variable in the “before” view. The possible values of $\bar{Y}$ are $0/2 = 0$, $1/2$, or $2/2 = 1$. Given the PMF of $S$, the PMF of $\bar{Y}$ is

$$P(\bar{Y} = 0) = P(S = 0) = 1/4,$$
$$P(\bar{Y} = 1/2) = P(S = 1) = 1/2,$$
$$P(\bar{Y} = 1) = P(S = 2) = 1/4. \quad (2.9)$$

Equivalently, the CDF could be computed from (2.6) and (2.9). Let $F(\cdot)$ denote the CDF of $\bar{Y}$. For any $r < 0$, $F(r) = 0$ since $P(\bar{Y} < 0) = 0$. But since $P(\bar{Y} = 0) = 1/4$, $F(0) = P(\bar{Y} \leq 0) = 1/4$. That is, the CDF jumps from 0 to 1/4 at $r = 0$. It is then flat again for a while; e.g., $F(0.4) = 1/4$ still. Then, it jumps again at $r = 1/2$, and again at $r = 1$. At $r = 1/2$,

$$F(r) = P(\bar{Y} \leq r) = P(\bar{Y} = 0) + P(\bar{Y} = 1/2) = (1/4) + (1/2) = 3/4.$$  

At $r = 1$, $F(r) = 1$; that is, there is 100% probability that the sample proportion of heads is less than or equal to 100%, which is true by definition of a proportion. Altogether, writing the CDF two equivalent ways,

$$F(r) = (1/4) \mathbb{1}_{\{0 \leq r\}} + (1/2) \mathbb{1}_{\{1/2 \leq r\}} + (1/4) \mathbb{1}_{\{1 \leq r\}} = \begin{cases} 
0 & \text{if } r < 0 \\
1/4 & \text{if } 0 \leq r < 1/2 \\
3/4 & \text{if } 1/2 \leq r < 1 \\
1 & \text{if } 1 \leq r \end{cases}.$$  

(2.10)
2.1.5 Random Terminology

There are different words used to describe generating values from a probability distribution. Sometimes \( Y \) is called a random draw (or just draw) from its distribution. Equivalently, \( Y \) is randomly drawn from its distribution. This terminology applies to other random variables above, like \( Y_1, Y_2, S, \) and \( \bar{Y} \). Alternatively, we could say \( Y_1 \) and \( Y_2 \) are (randomly) sampled from their population distribution, or that \( (Y_1, Y_2) \) together form a random sample.

Mathematically, we could simply view \( \bar{Y} \) as a random draw from the distribution in (2.9). The role of \( Y_1 \) and \( Y_2 \) does not matter after (2.9) is determined.

2.1.6 Employment is Random

Instead of a coin flip with heads or tails, imagine \( Y \) represents an individual’s employment status, employed \( (Y = 1) \) or unemployed \( (Y = 0) \): \( Y = 1 \{ \text{employed} \} \). Again, the “before” view sees \( Y \) as a random variable. For example, I don’t know whether I’ll be employed or not next year; both outcomes are possible, with different probabilities.

A slight variation on the “before” view is “before sampled” instead of “before determined.” For example, it has already been determined for all individuals in Missouri whether or not they were employed last year. After we sample some such individuals, we can see all the values: \( Y_i = 1 \) if individual \( i \) was employed, or \( Y_i = 0 \) if not. However, before sampling, row \( i \) in our dataset was empty; we could have sampled somebody with \( Y_i = 1 \), or we could have sampled somebody with \( Y_i = 0 \). In that sense, \( Y_i \) can be seen as a random variable; it depends on who gets randomly chosen to be in the sample. Even after sampling, the “before sampling” view helps us assess certain statistical properties. It is essentially the same to consider properties across a very large number of random samples from the same population; this is detailed later in the text.

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?

2.2 The Population

2.2.1 Population Types

The word population has multiple meanings in econometrics. When you’re trying to answer an economic question with data, you should first consider which type of population is most appropriate for your question. This may in turn determine which econometric methods are most appropriate.
2.2. THE POPULATION

Finite Population

In English (or biology), “population” means all the people (or another species) living in some area, like everybody living in Missouri. In econometrics, this is usually called a finite population; it may also apply to firms, schools, etc. This is appropriate if we want to know how many individuals in Missouri are currently unemployed, for example, especially if we have a very large sample.

Infinite Population

Alternatively, we could imagine the sample came from an infinite population. The advantage of this perspective is that an infinite population can be succinctly described by a probability distribution, instead of needing to keep track of everyone individually. Sampling from a probability distribution is mathematically simpler than sampling from a finite population. Although finite populations are never literally infinite, it may be a good approximation when the population size is much larger than the sample. For example, if we observe only 600 individuals out of 6+ million in Missouri, econometric results based on finite and infinite population are practically the same.

In other cases, 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 are supposed to sound an alarm above 50ppm. Most work properly, but there’s some probability that a monitor is faulty and never alarms. If you are interested in this probability (that’s determined by the manufacturing process), then there is no finite number of monitors that perfectly answers your question. This is an infinite population question.

Superpopulation

As a variation of the infinite population, we could imagine sampling from a probability distribution that represents a superpopulation (coined by Deming and Stephan [1941]). This is like imagining many possible universes, actually an infinite number of possible universes, and seeing our actual universe as just one out of infinity. That is, 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).

Which Population is Most Appropriate?

Imagine the president flips a coin 20 times and then randomly selects 10 observations to report to you; which type of population produced your sample?
It could be the finite population of 20 flips. After all, the sample size is a fairly large portion, \( \frac{10}{20} = 0.5 \), so approximating 20 as infinity seems inappropriate. But a finite population framework would mean you should be fully satisfied by knowing the outcomes of the other 10 flips. 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 \( \frac{10}{20} \) heads”).

However, the finite population framework is not 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, i.e., the data-generating process (DGP). Here, the infinite population framework is more appropriate. In other cases where the process itself is of interest, we need to imagine other universes to make the population infinite, i.e., we need a superpopulation.

As seen above, the appropriate population perspective depends on your research question. For example, it depends whether you want to document the actual percentage of Missouri individuals unemployed each week (finite population), or instead learn something about the processes and mechanisms that determine an individual’s employment status (superpopulation). As another example, it could depend whether you want to learn about the current productivity of employees at your specific company (finite population), or instead learn about mechanisms that determine productivity (superpopulation).

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.”

Beyond this section, this text does not cover finite-population methods. Instead, the population is always modeled as a probability distribution, from which we observe randomly drawn values.

**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.

### 2.2.2 Types of Population Distributions

Even with a single variable, there are qualitatively different types of distributions.

**Binary**

To increase your comfort with the random variable perspective of data, consider another binary example: an indicator variable for an individual’s sex being female, \( Y = \)
1\{female\}. (This means $Y = 0$ for both male and intersex.) As with unemployment, an individual’s sex is determined before the individual is sampled, but it can be modeled as a random variable. Before sampling, there is some probability that the observation will be female, just like there is some probability of heads before a coin is flipped. The random variable view does not mean some sort of quantum state superposition where an individual’s sex is only determined when observed.

### Categorical and Ordinal

Consider another random variable that may have more than two values: education. Perhaps only three categories matter: 1) no high-school degree, 2) high-school degree (but no higher), 3) a degree past high-school (BA, PhD, etc.). Since the values of this education random variable are categories (not numbers), it is called a **categorical** variable. These categories have a clear order: category 1 is less education than category 2, which in turn is less than category 3. Thus, this education variable is **ordinal**. (In contrast, other variables may not have clearly ordered categories, like beef, chicken, pork, and tofu.)

This education variable could also be represented by three dummy variables, say $W = 1\{no \ HS\}, \ X = 1\{HS \ only\}, \ Y = 1\{past \ HS\}$. Then $P(W = 1)$ is the probability of category 1, etc. In a way, $Y$ is redundant since $Y = 1\{W = X = 0\}$, completely determined by $W$ and $X$.

Ordinal random variables have PMFs and CDFs, too, just with categories as “values.” Since the categories are mutually exclusive and exhaustive, their probabilities must sum to one. The population distribution of education specifies these probabilities. (In the finite-population framework, instead of probabilities, the population would consist of $N_1$ individuals in category 1, $N_2$ in category 2, and $N_3$ in category 3, where the total population size is $N = N_1 + N_2 + N_3$, so the population proportions are $N_1/N$, $N_2/N$, and $N_3/N$.)

For example, let $Z$ denote education. The PMF specifies the values of $P(Z = no \ HS)$, $P(Z = HS \ only)$, and $P(Z = past \ HS)$. These are equivalent to $P(W = 1)$, $P(X = 1)$, and $P(Y = 1)$, respectively. Since $W$, $X$, and $Y$ are binary, these could also be written in terms of expected values: $P(W = 1) = E(W)$, $P(X = 1) = E(X)$, and $P(Y = 1) = E(Y)$. The CDF would show, for example, $P(Z \leq HS \ only) = P(Z = no \ HS) + P(Z = HS \ only)$, which could also be written as $P(W + X = 1)$.

### Discrete

Alternatively, years of education could be measured. That is, $Y = 12$ means 12 years of education (through high school), $Y = 16$ means 16 years, etc. This $Y$ is **discrete**: we can still count the number of different possible values, and now each value has a numeric meaning, not just representing a category.
A PMF can be specified. For example, with $Y$ measured in units of years, let $P(Y = 11) = 0.2$, $P(Y = 12) = 0.3$, $P(Y = 16) = 0.4$, and $P(Y = 18) = 0.1$.

From the PMF, the mean of $Y$ can be computed. For a discrete random variable, the mean (or “expectation” or “expected value”) is the probability-weighted average of possible values. Letting $v_1, v_2, \ldots, v_J$ denote the $J$ possible values, the mean is

$$E(Y) \equiv \sum_{j=1}^{J} v_j P(Y = v_j). \quad (2.11)$$

In the education example, in units of years,

$$E(Y) = (11)(0.2) + (12)(0.3) + (16)(0.4) + (18)(0.1) = 14. \quad (2.12)$$

The CDF can also be computed:

$$F_Y(r) = (0.2) \mathbb{1}\{r \geq 11\} + (0.3) \mathbb{1}\{r \geq 12\} + (0.4) \mathbb{1}\{r \geq 16\} + (0.1) \mathbb{1}\{r \geq 18\} = \begin{cases} 0 & \text{if } r < 11 \\ 0.2 & \text{if } 11 \leq r < 12 \\ 0.5 & \text{if } 12 \leq r < 16 \\ 0.9 & \text{if } 16 \leq r < 18 \\ 1 & \text{if } 18 \leq r \end{cases} \quad (2.13)$$

Continuous

Some variables may be approximated as continuous random variables. Consider a person’s height, $H$. It may only be measured in units of 0.010 cm, making observed height a discrete random variable (and perhaps true height also comes in discrete increments of the Planck length?). However, there are still so many thousands of possible values, so close together, it may be reasonable to treat height as taking any positive real number value (below some upper bound of human height). In that case, the CDF of $H$ is a continuous function, instead of a step function like with discrete distributions.

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 $f_Y(·)$ 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 2.1 shows an example of $P(0 \leq Y \leq 1)$: it is the area under the PDF between $y = 0$ and $y = 1$.

One strange property of continuous random variables is that the probability of any specific value is zero, even though that value is possible. For example, if a person’s height $H$ is truly continuous, then $P(H = 150.39284) = 0$, even though there may be an
individual with that exact height. Or with a normal distribution, if $Y \sim N(0, 1)$, then $P(Y = 0) = 0$ even though we may observe $Y = 0$ being sampled. This may be seen in Figure 2.1: $P(Y = 0)$ is the “area” under the PDF between $y = 0$ and $y = 0$; this is just a line, which has zero area, hence the probability is zero.

Since discrete variables are more intuitive and often easier to compute examples with, they will be preferred for illustrations. Many results in this text do not depend on whether a variable is discrete or continuous. Exceptions are noted.

2.3 Description of a Population Distribution

The first of the three major goals in this text is description. Section 2.3 covers many features of a population probability distribution that may be economically interesting to describe.

2.3.1 Units of Measure

Always specify units of measure. 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). Most of the descriptive features below (mean, median, etc.) share the same units of measure of $Y$, but it is still helpful to always specify the units explicitly; and some descriptions
(like variance) have different units. Units always matter greatly, whether for description, prediction, or causality.

2.3.2 Complete Description: Distributions

In some cases, we wish to completely describe a population. That is, we want to fully describe the probability distribution of \( Y \), without losing any information. This discussion will help later when we consider summary features of distributions, too.

Consider again the (un)employment example from Section 2.1: how could the population distribution be quantitatively described? Since the employment indicator variable \( Y \) is binary, the distribution is fully described by a single parameter: \( P(Y = 1) \). This fully describes the PMF (and thus CDF) since \( P(Y = 0) = 1 - P(Y = 1) \). There is nothing else that can be learned about the population distribution of \( Y \).

Imagine another variable with two possible values, but where the values are numbers. For example, maybe a certain (imaginary) population only has one or two children per family, and \( Y \) is how many siblings somebody had as a child (zero or one). Maybe \( p = P(Y = 1) = 1/2 \), but this alone doesn’t fully describe the distribution. That is, another population may have larger families with either five or six children, so then \( Y = 4 \) or \( Y = 5 \). Even if \( p = P(Y = 5) = 1/2 \), this is a very different population distribution than the first population. One description includes the two possible \( Y \) values, \( v_1 \) and \( v_2 \), along with \( p \equiv P(Y = v_2) \): the vector \((p, v_1, v_2) = (1/2, 0, 1)\) fully describes the first population, while \((p, v_1, v_2) = (1/2, 4, 5)\) fully describes the second population.

Imagine (slightly) more complex populations with three possible numbers of siblings. For simplicity, imagine \( P(Y = j) = 1/3 \) for \( j = 1, 2, 3 \): the values 1, 2, and 3 are equally likely. This is the probability mass function (PMF; see Section 2.1.3) that specifies the probability of every possible value of \( Y \).

Figure 2.2 shows two common ways of graphing this PMF. It fully describes the distribution of \( Y \), but it requires many numbers when there are many possible values. In this case, there are three possible values, so the PMF uses six numbers: the three values, and the three corresponding probabilities.

Similarly, the cumulative distribution function (CDF; see Section 2.1.3) includes all information about the distribution. Naturally, it requires just as many numbers as the PMF. Most fundamentally, the CDF \( F_Y(\cdot) \) is defined as

\[
F_Y(y) \equiv P(Y \leq y) \tag{2.14}
\]

for any real number \( y \). In the sibling example of \( P(Y = j) = 1/3 \) for \( j = 1, 2, 3 \), the CDF...
2.3. DESCRIPTION OF A POPULATION DISTRIBUTION

2.3.3 Population Summary: Mean

With description, there is a tradeoff between precision and conciseness. Describing the entire CDF or PMF is very precise, but it may require many numbers, making it difficult to report and understand. Sometimes such detail is required, but other times we prefer a concise summary of a distribution with just a few numbers. This is usually easier to understand, but the risk is missing some important feature of the distribution that the summary does not capture.

Continuing the siblings example, what if we needed to summarize each distribution with a single number (because your boss cannot comprehend multiple numbers at a time, perhaps), to help us see which population tends to have more children? The mean (or
**expectation** or **expected value**) summarizes \((p, v_1, v_2)\) as

\[
E(Y) = (p)(v_2) + (1 - p)(v_1).
\]  

The first population’s mean number of siblings is thus \((1/2)(1) + (1/2)(0) = 0.5\), while the second population’s mean is \((1/2)(4) + (1/2)(5) = 4.5\), much larger. The units of measure for \(E(Y)\) are the same as the units of \(Y\), in this case, number of siblings.

The following are useful formulas related to expected value. Consider a general discrete random variable \(Y\) with possible values \(y_1, y_2, \ldots, y_J\). As in (2.11), the expected value (mean) of \(Y\) is

\[
E(Y) = \sum_{j=1}^{J} y_j P(Y = y_j) = y_1 P(Y = y_1) + \cdots + y_J P(Y = y_J).
\]  

The expected value of a continuous random variable with pdf \(f_Y(\cdot)\) is an integral, \(\int_{-\infty}^{\infty} y f_Y(y) \, dy\), but we won’t worry about that in this text.) The expectation operator \(E(\cdot)\) has a useful property called **linearity**, which implies various equalities. For example, given two random variables \(Y\) and \(Z\) (of any type), and two non-random constants \(a\) and \(b\), it is always true that

\[
E(aY + bZ) = a E(Y) + b E(Z).
\]  

As a special case when \(b = 0\), this implies \(E(aY) = a E(Y)\). Or if \(a = 1\) and \(Y = cW + dX\),
then
\[ E(aY + bZ) = aE(Y) + bE(Z) = (1) E(cW + dX) + bE(Z) = cE(W) + dE(X) + bE(Z). \]  
(2.20)

Extending this, if we have random variables \( Y_i \) for \( i = 1, \ldots, n \), and constants \( c_i \), then
\[ E \left( \sum_{i=1}^{n} c_i Y_i \right) = \sum_{i=1}^{n} c_i E(Y_i). \]  
(2.21)

However, the mean loses a lot of information from the full distribution. Other features like the mode, median, and other percentiles (quantiles) are useful in practice. The single most important feature the mean misses is how spread out the distribution is, which is discussed next.

2.3.4 Population Summary: Standard Deviation

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: \( \sigma_Y = \text{Var}(Y) \) is the variance of \( Y \), while \( \sigma_Y \equiv \sqrt{\sigma_Y^2} \) is the standard deviation.

In the sibling example with \( P(Y = j) = 1/3 \), where \( E(Y) = 2 \),
\[
\sigma_Y^2 = \sum_{j=1}^{3} [j - E(Y)]^2 P(Y = j) = \sum_{j=1}^{3} (j - 2)^2 (1/3) = (1)(1/3) + (0)(1/3) + (1)(1/3) = 2/3, \\
\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{2/3}. 
\]  
(2.22)

For a more general discrete distribution with possible values \( y_j \) for \( j = 1, \ldots, J \),
\[
\sigma_Y^2 \equiv \sum_{j=1}^{J} [y_j - E(Y)]^2 P(Y = y_j), \quad \sigma_Y \equiv \sqrt{\sigma_Y^2}, 
\]  
(2.24)

where as usual \( E(Y) \) is the mean.

The units are easier to interpret for standard deviation than for variance. The units of measure of \( y_j \) and \( E(Y) \) are the same as those of \( Y \), so the units of \( [y_j - E(Y)]^2 \) are the square of the units of \( Y \). Thus, the units of \( \sigma_Y^2 \) are the square of the units of \( Y \), whereas the units of \( \sigma_Y \) are the same as the units of \( Y \). It is usually easier to understand measures in the units of \( Y \) than their square. For example, in the siblings example, the units of \( Y \) and \( \sigma_Y \) are “number of siblings,” whereas the variance has units “squared siblings,” which is difficult to imagine. Thus, standard deviation is usually a more helpful description than variance.
2.3.5 Continuous Distributions

Unlike with a discrete distribution, it’s impossible to summarize a continuous distribution by a list of values and probabilities because there are an infinite number of values. However, we can still summarize a continuous distribution in a picture: a graph of either its cumulative distribution function (CDF) or its probability density function (PDF; see Section 2.2.2). (Sometimes the PDF does not exist; we ignore such cases.)

Just as with discrete distributions, summary features like the mean and standard deviation help describe a continuous distribution with just a few numbers. The interpretation and intuition of all these is the same as in the discrete case, although the math changes some. Also, the units of measure are the same, i.e., they are all the same as the units of Y except for the variance, which is the Y units squared. You don’t need to know the mathematical details for this class, but in case you’re curious, here are the definitions, at least. The mean is again a population probability-weighted average value, but instead of a finite sum, it is an infinite sum (weighted by the PDF instead of PMF), i.e., an integral:

$$E(Y) = \int_{-\infty}^{\infty} yf_Y(y) \, dy.$$  

(2.25)

So far, only the population distribution has been considered; there is no data yet. This section has focused on what may be interesting to learn about the population. To learn about these different population features from data, see Section 3.3.

2.4 Prediction with a Known Distribution

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 to “predict” what will happen in the future (e.g., “Beware the Ides of March!”). In econometrics, prediction shares the qualities of guessing something unknown using something known, but the details differ.

Here, 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, prediction may include predicting taxi demand tomorrow, but also predicting the income of a customer standing right in front of you (who hasn’t yet told you their income).

This section illustrates and formalizes prediction and related concepts. Close connections with description (Section 2.3) are shown. Predicting a single value or an interval of values are both considered, though the former is the main focus.
2.4. PREDICTION WITH A KNOWN DISTRIBUTION

2.4.1 Example: Rain

Imagine you want to predict whether or not it will rain tomorrow. You live in Seattle, where it is just as likely to rain as not rain every single day, regardless of the previous days. How can we formalize this mathematically? Does the “best” prediction depend of the use of the prediction, or is there a single best prediction for all purposes?

Formally, random variable $Y$ represents rain and has the following properties. Let $Y = 1$ if rain and $Y = 0$ if no rain, i.e., $Y = 1\{\text{rain}\}$. “Just as likely to rain as not rain” means $P(Y = 1) = 50\% = P(Y = 0)$, so $Y$ follows a Bernoulli distribution with parameter $p = 0.5$: $Y \sim \text{Bernoulli}(p)$. “Regardless of the previous days” means each day we get a new (independent) random draw of $Y$ from its population distribution. Your goal is to predict the value of $Y$ (whether it rains) drawn randomly from the known distribution (tomorrow’s possible weather).

Predictions and Consequences

First, imagine you bet a friend that you can predict whether or not it rains. Your friend pays you $1 if you predict correctly, and you pay your friend $1 if not. To formalize this, let $g$ be your guess: $g = 1$ for rain, $g = 0$ for no rain. This $g$ is lower case because it is not random: you choose one or the other.

Define a winnings function as follows. Let $W(\cdot)$ compute your winnings (in units of dollars) as a function of your guess and the true realized value $y$: $W(y, g) = 1$ if $y = g$ and $W(y, g) = -1$ if $y \neq g$, i.e., in units of dollars,

$$W(y, g) = 2 \times 1\{y = g\} - 1.$$  

Replacing the fixed value $y$ with the random variable $Y$, $W(Y, g)$ becomes a random variable itself. That is, sometimes $W(Y, g) = W(1, g)$, but other times $W(Y, g) = W(0, g)$, with different probabilities.

You may predict $g = 0$ or $g = 1$; which is best? With $g = 1$, your winnings are (still in dollars) the random variable

$$W_1 \equiv W(Y, g = 1) = 2 \times 1\{Y = 1\} - 1 = \begin{cases} 
\$1 & \text{if rain} \\
-\$1 & \text{if no rain.}
\end{cases} \quad (2.27)$$

Consequently,

$$P(W_1 = 1) = P(Y = 1) = 0.5, \quad P(W_1 = -1) = P(Y = 0) = 0.5, \quad (2.28)$$
i.e., you have a 50% chance of winning a dollar and 50% chance of losing a dollar. With $g = 0$ instead,

$$W_0 \equiv W(Y, g = 0) = 2 \times 1\{Y = 0\} - 1 = \begin{cases} 
-\$1 & \text{if rain} \\
\$1 & \text{if no rain},
\end{cases} \quad (2.29)$$

$$P(W_0 = 1) = P(Y = 0) = 0.5, \quad P(W_0 = -1) = P(Y = 1) = 0.5, \quad (2.30)$$
i.e., you still have a 50% chance of winning a dollar and 50% chance of losing a dollar. The predictions \( g = 1 \) and \( g = 0 \) are different and produce different outcomes on different days, e.g., if it rains tomorrow, then \( g = 1 \) wins and \( g = 0 \) loses. However, they are probabilistically identical: each has a winnings distribution with 50% chance of winning or losing a dollar. In this sense, the two predictions are equally good.

Why are you only allowed to predict \( g = 0 \) or \( g = 1 \), and not other values like \( g = 0.2 \)? Actually, you may predict any number you want: \( g = 0.2 \), \( g = 549 \), \( g = -99 \), etc. However, given the winnings function in (2.26), all of these predictions yield \(-$1\) with certainty since \( \mathbb{1}\{Y = g\} = 0 \) for values of \( g \) that \( Y \) never takes. That is, you will never win unless you pick \( g = 0 \) or \( g = 1 \). However, in other settings, with other winnings functions, it may actually be best to pick a value that is never perfectly correct, if perhaps it is closer to the correct value more often.

What if the winnings function in (2.26) is the same, but the distribution changes to \( P(Y = 1) = 0.1 \)? The random variables \( W_1 \) and \( W_0 \) from (2.27) and (2.29) have the same definition, but their distributions now differ from (2.28) and (2.30). Instead,

\[
P(W_1 = 1) = P(Y = 1) = 0.1, \quad P(W_1 = -1) = P(Y = 0) = 0.9, \\
P(W_0 = 1) = P(Y = 0) = 0.9, \quad P(W_1 = -1) = P(Y = 1) = 0.1.
\]

(2.31)

This formalizes common sense: if rain is unlikely (10% chance), then you’re more likely to win by predicting no rain.

**Prediction to Maximize Expected Winnings**

As in Section 2.3, it may help to summarize the distribution of \( W_1 \) or \( W_0 \) by a single number, like the mean. Using the discrete random variable mean formula and (2.31),

\[
E(W_1) = (0.1)($1) + (0.9)(-$1) = -$0.80, \\
E(W_0) = (0.9)($1) + (0.1)(-$1) = $0.80.
\]

Thus, “expected” winnings are higher with \( g = 0 \) than \( g = 1 \), suggesting \( g = 0 \) (no rain) is the best prediction. (With other values like \( g = 0.1 \), expected winnings are \(-$1\), even worse than \( g = 1 \).) Here (and in econometrics generally), “expected” has the technical meaning of expected value, which is not what we “expect” in the colloquial English sense. Indeed, it would be silly to expect (in the informal sense) to win $0.80 from your bet because the only possibilities are winning $1 or losing $1; winning $0.80 is impossible! “Expectation” is just a way to summarize the different values and probabilities into a single number to aid comparison. A better assessment of prediction would consider not just the monetary values but “utility” (in the technical economic sense) and expected utility. Here, utility is implicitly assumed to be linear in winnings.

In another variation, imagine the distribution is still \( P(Y = 1) = 0.1 \), but the winnings function changes. Now your friend pays you $10 if you correctly predict rain, with the
other values unchanged (\$1 if no rain correctly predicted, \$1 if incorrect). The only change to the definitions of \(W_0\) and \(W_1\) is that \(P(W_1 = 10) = 0.1\) replaces \(P(W_1 = 1) = 0.1\). Consequently, (2.32) becomes

\[
\begin{align*}
E(W_1) &= (0.1)(\$10) + (0.9)(\$1) = \$0.10, \\
E(W_0) &= (0.9)(\$1) + (0.1)(\$1) = \$0.80.
\end{align*}
\] (2.33)

Despite the allure of winning \$10, it is unlikely enough that \(g = 0\) still yields larger expected winnings, so \(g = 0\) is a better prediction from that measure.

To summarize, the “best” prediction that maximizes expected winnings depends on two separate components: the distribution of \(Y\), and the winnings function. Section 2.4.2 formalizes this further and presents more complex examples.

### 2.4.2 Loss Functions

**Losing is Negative Winning**

Instead of optimistically thinking of winnings as in Section 2.4.1 we could equivalently define outcomes in terms of losses. After all, it is equivalent to win \$1 or lose –\$1, and equivalent to win –\$1 or lose \$1. The winnings function in (2.26) can be written as a loss function: in dollars,

\[
L(y, g) = -W(y, g) = -(2 \times 1\{y = g\} - 1) = 1 - 2 \times 1\{y = g\}.
\] (2.34)

More generally, a loss function captures how bad it is to guess \(g\) when the truth is \(y\).

Maximizing expected winnings is equivalent to minimizing expected loss, also called risk; either approach yields the same best prediction. For example, the expected loss analog of the expected winnings in (2.33) is

\[
\begin{align*}
E(L(Y, g = 1)) &= (0.1)(-\$10) + (0.9)(\$1) = -\$0.10, \\
E(L(Y, g = 0)) &= (0.9)(-\$1) + (0.1)(\$1) = -\$0.80.
\end{align*}
\] (2.35)

By this measure, \(g = 0\) minimizes expected loss for the same reasons it maximized expected winnings, and it is the best prediction. Though perhaps less intuitive than winnings, the loss framework dominates statistical prediction, so we use it going forward.

**Formalizing Loss Functions**

There are many (infinite) possible loss functions, but some are used more commonly than others. The two most common are described below.

Define 0–1 loss as

\[
L_0(y, g) = 1\{y \neq g\},
\] (2.36)
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. Although not obvious, this is true in the rain bet example. Maximizing expected winnings turns out to be equivalent to minimizing expected 0–1 loss where

$$L(y, g) = \mathbb{1}\{y \neq g\},$$

i.e., picking $g$ to minimize $E[L(Y, g)] = E[\mathbb{1}\{Y \neq g\}] = P(Y \neq g)$, which is the same as picking $g$ to maximize $P(Y = g)$.

In the examples so far, there have only been two possible values of $Y$. When there are more values, the 0–1 loss approach may not work well because any single value is so unlikely.

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

$$L_2(y, g) = (y - g)^2. \quad (2.37)$$

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 $y = 100$. Then, the guess $g = 99$ is better (or “less bad”) than the guess $g = 90$ since $L_2(100, 99) = (100 - 99)^2 = 1$ is smaller than $L_2(100, 90) = (100 - 90)^2 = 100$.

Quadratic loss is by far the most commonly used loss function, but it does not make sense for every situation. For example, sometimes it may be much worse to over-predict ($g > y$) than under-predict ($g < y$), or vice-versa, but quadratic loss treats them the same. 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 overpredict 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.

2.4.3 Optimal Prediction: Generic Examples

Two Possible Values

The first step is to write out the loss function values for all combinations of $(Y, g)$, where $g$ is our guess of $Y$. Let random variable $Y$ have only 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 row has a single value of $Y$ and each column has a single value of $g$:

$$
\begin{pmatrix}
L(a, a) & L(a, b) \\
L(b, a) & L(b, b)
\end{pmatrix}
= 
\begin{pmatrix}
L(Y = a, g = a) & L(Y = a, g = b) \\
L(Y = b, g = a) & L(Y = b, g = b)
\end{pmatrix}.
\quad (2.38)
$$
2.4. PREDICTION WITH A KNOWN DISTRIBUTION

The second step is to compute the expected loss for each possible guess $g$. In (2.38), each column corresponds to a different $g$. To compute expected loss if $g = a$, look at the first column. The guess $g = a$ is fixed (non-random). The expected value is with respect to the distribution of $Y$. The probabilities $P(Y = a)$ and $P(Y = b)$ do not depend on our guess $g$. They only depend on the population distribution of $Y$. That is, if we guess $g = a$, then we’ll get loss $L(a, a)$ with probability $P(Y = a)$, and we’ll get loss $L(b, a)$ with probability $P(Y = b)$. Similarly for $g = b$. Thus,

$$E[L(Y, a)] = P(Y = a)L(a, a) + P(Y = b)L(b, a),$$
$$E[L(Y, b)] = P(Y = a)L(a, b) + P(Y = b)L(b, b).$$

(2.39)

The second step could also be interpreted in terms of two new random variables. 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 (2.39),

$$E[L(Y, a)] = E(L_a) = P(Y = a)L(a, a) + P(Y = b)L(b, a),$$
$$E[L(Y, a)] = E(L_b) = P(Y = a)L(a, b) + P(Y = b)L(b, b).$$

(2.40)

The third step is to find the $g$ that minimizes $E[L(Y, g)]$. 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) in terms of expected loss.

For example, let

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

$P(Y = a) = 0.7, \quad P(Y = b) = 0.3 = 1 - P(Y = a).

(2.41)

Using (2.39),

$$E[L(Y, a)] = P(Y = a)L(a, a) + P(Y = b)L(b, a) = (0.7)(0) + (0.3)(7) = 2.1,$$
$$E[L(Y, b)] = P(Y = a)L(a, b) + P(Y = b)L(b, b) = (0.7)(5) + (0.3)(0) = 3.5.

(2.42)\quad (2.43)

Since $E[L(Y, a)] < E[L(Y, b)]$, the predictor $g = a$ is better than $g = b$ according to expected 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 $v_1 = 10, \ v_2 = 20, \ v_3 = 22$; or $v_1$ could be “cat,” $v_2$ “dog,” and $v_3$ “echidna.” Like before, assume $g$ must take one of these same values.
CHAPTER 2. ONE VARIABLE: POPULATION

The three steps are the same as with $J = 2$, just with more values to keep track of and compute.

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

\[
\begin{pmatrix}
L(v_1, v_1) & L(v_1, v_2) & \cdots & L(v_1, v_J) \\
L(v_2, v_1) & L(v_2, v_2) & \cdots & L(v_2, v_J) \\
\vdots & \vdots & \ddots & \vdots \\
L(v_J, v_1) & L(v_J, v_2) & \cdots & L(v_J, v_J)
\end{pmatrix}
\] (2.44)

Second, compute all the expected losses:

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

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

\[
\vdots
\]

\[
E[L(Y, v_J)] = P(Y = v_1)L(v_1, v_J) + P(Y = v_2)L(v_2, v_J) + \cdots = \sum_{j=1}^{J} P(Y = v_j)L(v_j, v_J).
\] (2.45)

Third, find the $g$ that minimizes $E[L(Y, g)]$. That is, find the smallest of the $J$ values computed in (2.45); the corresponding $g$ is the optimal predictor.

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

\[
\begin{pmatrix}
L(v_1, v_1) & L(v_1, v_2) & L(v_1, v_3) \\
L(v_2, v_1) & L(v_2, v_2) & L(v_2, v_3) \\
L(v_3, v_1) & L(v_3, v_2) & L(v_3, v_3)
\end{pmatrix}
= \begin{pmatrix}
L(-1, -1) & L(-1, 0) & L(-1, 1) \\
L(0, -1) & L(0, 0) & L(0, 1) \\
L(1, -1) & L(1, 0) & L(1, 1)
\end{pmatrix}
= \begin{pmatrix}
0 & 1 & 4 \\
2 & 0 & 1 \\
8 & 2 & 0
\end{pmatrix}.
\] (2.46)

Let

\[
P(Y = v_1) = P(Y = -1) = 0.2, \\
P(Y = v_2) = P(Y = 0) = 0.3, \\
P(Y = v_3) = P(Y = 1) = 0.5 = 1 - P(Y = -1) - P(Y = 0).
\] (2.47)
2.4. PREDICTION WITH A KNOWN DISTRIBUTION

Then, the expected 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 \(g = v_3 = 1\).

2.4.4 Optimal Prediction: Specific Examples

Example: Carnival Age Game

Instead of rain, imagine predicting (guessing) a person’s age, \(Y\). 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. Since there are only two possible outcomes, the best prediction does not depend on the numerical values we assign the outcomes, so the loss function can be 0–1 loss: \(L_0(y, g) = 1\{y \neq g\}\). (For added challenge, come back to this example and see what changes if they only win when you are more than three years off.)

Starting with a simple distribution for \(Y\), let \(P(Y = 20) = 0.6\) and \(P(Y = 25) = 0.4\). In the carnival, with 0–1 loss, expected losses for \(g = 20\) and \(g = 25\) are, respectively,

\[
E(L_0(Y, 20)) = E(1\{Y \neq 20\}) = (0.6)(0) + (0.4)(1) = 0.4,
\]

\[
E(L_0(Y, 25)) = E(1\{Y \neq 25\}) = (0.6)(1) + (0.4)(0) = 0.6,
\]

so \(g^*_0 = 20\) is the optimal prediction. Equivalently,

\[
E[L_0(Y, 20)] = P(Y \neq 20) = 0.4,
\]

\[
E[L_0(Y, 25)] = P(Y \neq 25) = 0.6.
\]

More generally, since \(E[1\{Y \neq g\}] = P(Y \neq g)\),

\[
E[L_0(Y, g)] = P(Y \neq g) = 1 - P(Y = g),
\]

so the optimal prediction \(g\) minimizes \(P(Y \neq g)\), which is equivalent to maximizing \(P(Y = g)\). Intuitively, it’s best to predict the most likely value when only being right or wrong matters.
If instead we used quadratic loss, we wouldn’t do very well. Comparing \( g = 20 \) 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 can show \( g = 22 \) 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 shows 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, guessing very close but not exactly correct is still very good here. 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. Consequently, we want the loss function to still be very close to zero when the guess is very close to the true age, which was not true with the 0–1 losses.

Another difference with the carnival game is that some incorrect guesses are much worse than others. In the carnival, all incorrect guesses were equally bad. Here, guessing that the 41-year-old is 20 years old is clearly worse than guessing 40. Similarly, guessing that the 41-year-old is 60 is bad, but still better than guessing 80.

In this case, quadratic loss seems much 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
2.4. PREDICTION WITH A KNOWN DISTRIBUTION

quadratic and 0–1 loss lead to different optimal predictions; that result depends only on the mathematical distribution of $Y$, not on the interpretation of a carnival game or advertising. So, the predictions are again different, but we may prefer $g = 22$ over $g = 20$ this time. 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 a more complex example where any value $Y \in \{20, 21, 22, 23, 24, 25\}$ is possible. At the carnival, young people are more likely, so the distribution of $Y$ is

$$P(Y = j) = \frac{(26 - j)}{21}, \quad j = 20, 21, \ldots, 25.$$  \hspace{1cm} (2.50)

For simplicity, the same distribution now applies to the advertising example.

With 0–1 loss, expected loss of a general guess $g \in \{20, \ldots, 25\}$ is

$$E(L_0(Y, g)) = \sum_{y=20}^{25} \mathbb{1}\{y \neq g\} P(Y = y) = \sum_{y \neq g} P(Y = y) = 1 - P(Y = g).$$ \hspace{1cm} (2.51)

This is minimized by choosing $g$ to maximize $P(Y = g)$:

$$\arg \min_g E(L_0(Y, g)) = \arg \min_g [1 - P(Y = g)] = \arg \max_g P(Y = g) = 20,$$ \hspace{1cm} (2.52)

so the best prediction is $g_0^* = 20$.

This result is intuitive given that 0–1 loss only cares whether we’re right or wrong. Guessing $g_0^* = 20$ gives you a $6/21$ probability of being correct, which is the largest possible. Equivalently, it is the smallest possible probability of being wrong. Indeed, $g_0^*$ is always the single most likely value (the mode).

Like before, quadratic loss leads to a different optimal prediction. If we guessed $g = 20$, then

$$E[L_2(Y, g)] = \sum_{y=20}^{25} P(Y = y)(y - 20)^2 = \sum_{y=20}^{25} [(26 - y)/21](y - 20)^2 = 5.$$ \hspace{1cm} (2.53)

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 $Y = 20$. If we try $g = 21,$

$$E[L_2(Y, g)] = \sum_{y=20}^{25} P(Y = y)(y - 21)^2 = \sum_{y=20}^{25} [(26 - y)/21](y - 21)^2 = 8/3 \approx 2.67.$$ \hspace{1cm} (2.54)

Since $2.67 < 5$, guessing $g = 21$ is better than $g = 20$ according to expected 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 ($g < y$).

2.4.5 Mean and Mode as Optimal Predictions

Under quadratic loss, the mean is the optimal predictor that minimizes expected loss. Generally, the optimal predictor $g^*_2$ under quadratic loss minimizes the expected value of the loss function:

$$g^*_2 \equiv \arg \min_g E[(Y - g)^2]. \quad (2.55)$$

Although the details are beyond our scope, calculus can be used to take the derivative of the right-hand side and set it equal to zero (the “first-order condition”), yielding

$$g^*_2 = E(Y). \quad (2.56)$$

This result shows that the mean of a distribution may be interpreted in two separate ways. For description, the mean helps summarize the center of the distribution. For prediction, the mean is the “best” prediction of an unknown value of $Y$ in terms of minimizing expected quadratic loss. When causality is introduced in Chapter 4, we will have cases where the same statistical object like $E(Y)$ has a third, causal interpretation, too.

Under 0–1 loss, the optimal prediction is

$$g^*_0 \equiv \arg \min_g E(L_0(Y, g))$$

$$= \arg \min_g E[1 \{Y \neq g\}]$$

$$= \arg \min_g P(Y \neq g)$$

$$= \arg \min_g [1 - P(Y = g)]$$

$$= \arg \min_g - P(Y = g)$$

$$= \arg \max_g P(Y = g). \quad (2.57)$$
2.4. PREDICTION WITH A KNOWN DISTRIBUTION

That is, the mode (the single most likely value of \( Y \)) is the optimal predictor. This only makes sense if \( Y \) is discrete; otherwise, for any prediction \( g \), \( P(Y = g) = 0 \). As with the mean and percentiles, 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 (expected) profit. Imagine you know the distribution of \( Y \) (banana quantity demanded in one week). a) Do you think the mean \( E(Y) \) is a good “predicted” number of bananas to buy wholesale? Explain why or why not; if not, also explain why you think \( E(Y) \) 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 \( E(Y) \) be good, or too high, or too low, and why? c) What if the retail price were equal to the wholesale price?

2.4.6 Interval Prediction

Only **point prediction** has been discussed so far, i.e., the single number that provides the best guess of the unknown value. Here, we consider **interval prediction**, i.e., guessing an interval of numbers called a **prediction interval**.

The disadvantage of point predictions is that they are almost always wrong. Despite the optimality properties shown above, trying to guess a single number is nearly impossible. For example, if \( Y \sim N(0, 1) \), then the best point prediction under either quadratic or absolute loss is \( \hat{Y} = 0 \), the mean and median. But, this guess will be wrong 100% of the time since \( P(Y = 0) = 0 \).

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 \( P(Y = j) = 1/4 \) for \( j = 1, 2, 3, 4 \). Then, the interval \([2, 3]\) contains a randomly drawn \( Y \) with probability 1/2. If instead \( P(Y = j) = 1/10 \) for \( j = 1, \ldots, 10 \), then the interval \([2, 3]\) only has 1/5 probability of containing \( Y \). There is more uncertainty, so the interval would have to be lengthened to something like \([2, 6]\) to achieve the same 1/2 probability as before.

Even for the same probability, like 90%, there may be multiple (usually infinitely many) possible prediction intervals. Other properties can be used to distinguish among the intervals, but such is beyond our scope.
Chapter 3

One Variable: Sample

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: https://www.khanacademy.org/math/ap-statistics
• Quantifying uncertainty and statistical significance (Masten video): https://www.youtube.com/watch?v=3I0OzqOhttY
• Estimator properties (Lambert video): [https://www.youtube.com/watch?v=UxbY85Cm9SQ](https://www.youtube.com/watch?v=UxbY85Cm9SQ)
• Unbiasedness and consistency (Lambert videos): [https://www.youtube.com/watch?v=21lXGcO2XwM](https://www.youtube.com/watch?v=21lXGcO2XwM) and [https://www.youtube.com/watch?v=6i7mqDJ1CzQ](https://www.youtube.com/watch?v=6i7mqDJ1CzQ)
• iid sampling (Lambert video): [https://www.youtube.com/watch?v=0OoJHh1jmSO](https://www.youtube.com/watch?v=0OoJHh1jmSO)
• Section 2.8 (“Exploratory Data Analysis with R”) in [Kleiber and Zeileis (2008)](https://www.youtube.com/watch?v=21lXGcO2XwM) [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)](https://www.youtube.com/watch?v=6i7mqDJ1CzQ)
• Sections 1.5.4 (“Fundamental Statistics”) and 1.9.3 (“Simulation of Confidence Intervals and t Tests”) in [Heiss (2016)](https://www.youtube.com/watch?v=0OoJHh1jmSO)
• R package boot [Canty and Ripley, 2017] Davison and Hinkley, 1997

Sections 2.3 and 2.4 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), but more details are needed to consider estimation and uncertainty, as in Sections 3.3 and 3.5.

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

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](https://xkcd.com/1132)), most econometricians appreciate and respect both frameworks (and the people who use them), sometimes working with both in turn.

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. Without getting into more technical issues like admissibility and misspecification, there are a few main differences.

First, there are different ways to evaluate whether or not an estimator is “good.” The frequentist approach considers how a method performs across many different (hypothetical) datasets sampled from the same population. In contrast, the Bayesian approach tries to accurately quantify our beliefs about the unknown population. These beliefs combine our prior beliefs with the observed data. Ideally, we want both frequentist and Bayesian properties; both are helpful. Often, especially with simpler models, frequentist and Bayesian approaches lead to similar results with little practical difference. However, other times the differences are important.

Second, related, 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. In contrast, the Bayesian framework treats population features as random to reflect your beliefs about different possible values, whereas it treats the data as non-random values (the “after” view). For example, the Bayesian framework is designed to answer questions like, “Given the observed data values, what do I believe is the probability that the population mean is above 1/2?” The mean \( \mu = E(Y) \) is a random variable, to reflect our uncertainty (even if we think there’s only one true value). In contrast, the frequentist framework treats \( \mu \) as non-random, but answers questions like, “Given the value of \( \mu \), what’s the probability that the sample mean is between \( \mu - 0.1 \) and \( \mu + 0.1 \)?” The sample mean is a function of data, so it is treated as a random variable with its own probability distribution.

Third, frequentist methods use only the data, whereas Bayesian methods can formally incorporate additional knowledge. In practice, though, frequentist results should be interpreted in light of other knowledge; the difference is that this process is less formal and outside 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! See the chapter appendix for more on the Bayesian approach.

This text uses the frequentist framework. Why? Mostly, that’s just how I wrote it; I’ll spare you post hoc rationalization. In the future, I hope to include more Bayesian material.

### 3.2 Types of Sampling

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 \( i = n \). 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.
CHAPTER 3. ONE VARIABLE: SAMPLE

One of the two most important sampling properties is whether observations are independent or not. In Parts I and II, each \( Y_i \) is assumed independent of the others: \( Y_i \perp \perp Y_k \) for all \( i \neq k \).

The second important sampling property is whether the observations are drawn from identical distributions or not. Sampling is called iid when the \( Y_i \) are independent and identically distributed random variables. Sometimes the much more vague phrase random sample refers to iid sampling. This type of sampling is mathematically simplest, but not always realistic. Although iid sampling is the focus here (like other introductory texts), weights are briefly mentioned, and Part III considers dependent (i.e., not independent) data.

With iid sampling, each \( Y_i \) is separately (independently) drawn from the same population distribution \( F_Y \), where \( F_Y(\cdot) \) is the population CDF. Notationally,

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

For \( i \neq k \), the independence property \( Y_i \perp \perp Y_k \) implies \( \text{Cov}(Y_i, Y_k) = 0 \), \( \text{Var}(Y_i + Y_k) = \text{Var}(Y_i) + \text{Var}(Y_k) \), and \( \mathbb{E}(Y_i \mid Y_k) = \mathbb{E}(Y_i) \), among other derived properties. The identically distributed property means \( Y_i \) and \( Y_k \) have the same distribution, which implies \( \mathbb{E}(Y_i) = \mathbb{E}(Y_k) \) and \( \text{Var}(Y_i) = \text{Var}(Y_k) \), and similarly for percentiles and other distributional features.

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

1. Randomly picking a Mizzou student ID number, then randomly picking a 2nd, 3rd, and 4th: this would be both independent and identically distributed (iid). Each number is drawn without consideration of the previous numbers (independent). Each number is drawn from the same population of current Mizzou students (identically distributed). That is, each bucket is equally likely to contain different types of students; e.g., we could switch buckets 2 and 4 (before filling them) without changing anything.

2. Randomly picking 2 econ majors and 2 philosophy majors: for a variable like ACT score (an exam taken before starting college), this is again independent, but not identically distributed (inid). Econ and philosophy majors are systematically different, so their distributions (of whichever variable is observed) are probably also different (not identical). But, couldn’t there be econ and philosophy students in the first example, too, which was claimed to be identically distributed? Yes: the difference is that in the first example, the students’ majors were only known after sampling (after the bucket is filled), whereas here, the buckets themselves
3.2. TYPES OF SAMPLING

specify the major. Put differently: the two econ buckets get filled from a different subpopulation (econ majors) than the two philosophy buckets. Assigning buckets to different groups (“strata”) before sampling is called **stratified sampling**.

For other variables whose value is determined in part by a shared experience of econ majors (or philosophy majors), even independence may not hold. The next example explains why.

3. Randomly picking 2 classes at Mizzou, and then picking 2 students from each class: this is identically distributed, but probably no longer independent (again depending on the variable of interest). It is an example of clustered sampling, where each class is a cluster. Students in the same class are affected by their shared experience in class in similar ways. Here, buckets 1 and 2 are correlated, and 3 and 4 are correlated, but not 1 and 3, or 2 and 4, etc. For example, if we measure 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 somehow it was harmful), than that one is high and one is low, whereas there is no such correlation between \( Y_1 \) and \( Y_3 \). That is, if the first class is Econometrics, then probably both \( Y_1 \) and \( Y_2 \) are high. Of course, it is still possible that \( Y_2 \) is low even in an econometrics class, but the probability is low. However, since the 2 classes are picked randomly, it’s equally likely to get econometrics buckets first as second, so the overall distribution of \( Y_1 \) and \( Y_4 \) (etc.) is the same.

4. Randomly picking 2 students, then observing them this semester and next semester: this is another type of clustering that usually violates independence. For example, imagine 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 are probably both high or both low, rather than one high and one low; and similarly for buckets 3 and 4. That is, buckets (observations) 1 and 2 are correlated, and 3 and 4 are correlated. In this case, we probably do not have identical distributions either, if the overall GPA distribution is different in fall and spring semesters. However, if we instead consider the first two buckets together as one observations, and the other two buckets as a second observation, then we have iid sampling, since the students themselves are sampled iid.

5. One student over her first 4 semesters: this is time series data, where there is often dependence; see Part [III](#).

**Discussion Question 3.1** (rural household sampling). You want to learn about household consumption in rural Indonesia. In an area with 100 villages, you either a) pick 5 villages at random, then survey every household in each of the 5 villages; or b) make a list of all households in all 100 villages, then randomly pick 5% of them. a) Explain why
each approach is or isn’t iid. b) Guess which would be more expensive (financially) to do, and explain why.

There can also be sampling bias. This is not about correlation among observations, but 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 a bit more in Chapter 12 in terms of “missing data” and “sample selection.”

### 3.3 Estimation

Sections 2.3 and 2.4 helped us think about which features of the population are most useful for description and prediction, but in practice we don’t know the true population distribution. Such population features must be estimated using a sample of data. The relevant population feature is called the **estimand** or **object of interest**.

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 true value. In contrast, an interval estimate is a range of numbers, i.e., an interval. The most common interval estimate is a confidence interval; see Section 3.5.

Notationally, estimators are often written as the estimand with a hat added on top. For example, an estimator of the population mean \( \mu_Y \) is often written as \( \hat{\mu}_Y \). More generally, \( \hat{\theta} \) is an estimator of population parameter \( \theta \) (the estimand). Sometimes, a subscript \( n \) is added (like \( \hat{\theta}_n \)) as a reminder that the properties of an estimator depend on the sample size. Since an estimator is calculated using the sample \( Y_i \) values, it is an example of a **statistic**.

#### 3.3.1 The Sample Mean and Least Squares Estimation

**Mean of Sample Distribution**

Just as the population mean is the mean of the population distribution, the **sample mean** can be interpreted as the mean of the **sample distribution**. For simplicity, imagine all the observed \( Y_i \) values are unique. The sample distribution is a discrete probability distribution, where the \( Y_i \) are the possible values, each with probability \( 1/n \). That is, random variable \( S \) follows the sample distribution if

\[
P(S = Y_1) = 1/n, \quad P(S = Y_2) = 1/n, \quad \ldots, \quad P(S = Y_n) = 1/n. \tag{3.2}
\]

This describes the sample distribution in terms of its PMF.
Although the sample PMF is usually more helpful, the sample distribution is usually referred to by its CDF. The sample CDF \( \hat{F}_Y(\cdot) \) is also called the empirical distribution function. Notationally, the “hat” on \( \hat{F}_Y(\cdot) \) distinguishes it as being computed from sample values, rather than the population \( F_Y(\cdot) \) that has no hat.

The mean of the sample distribution can be computed using (3.2). Recall that the mean of a discrete distribution sums the products of the possible values and their corresponding probabilities. Here, the \( n \) possible values of \( S \) are \( Y_1, \ldots, Y_n \), each with \( 1/n \) probability. Thus,

\[
\hat{E}(Y) = E(S) = \sum_{i=1}^{n} (Y_i \cdot \text{probability}) = \sum_{i=1}^{n} (Y_i) \cdot (1/n) = (1/n) \sum_{i=1}^{n} Y_i \equiv \bar{Y}_n. \tag{3.3}
\]

This \( \bar{Y}_n \) is also called the sample average since it is the average of the \( Y_i \) values. If the values of \( Y_i \) are not all unique, then the sample distribution changes, but actually the sample mean in (3.3) remains unchanged.

Notationally, for continuity with future chapters, sometimes the sample mean will be written as \( \hat{\mu}_Y \) or \( \hat{\mu}_{Y,n} \). This reminds us that it is an estimator of the population parameter \( \mu_Y \equiv E(Y) \). However, there are other possible estimators of the population mean, so writing \( \bar{Y}_n \) is more precise. You should become familiar with both notations.

### Least Squares

Use of the sample mean is also motivated by the interpretation of the population mean as the best predictor. In Section 2.4.5, it was seen that \( E(Y) \) is the best predictor of \( Y \) when the loss function is quadratic. That is, (2.55) and (2.56) showed that \( g = g_2^* = E(Y) \) minimizes \( E[(Y-g)^2] \). Replacing the population mean \( (E) \) with the sample mean \( (\frac{1}{n} \sum_{i=1}^{n}) \), the analogous minimization problem in the sample is

\[
\hat{g}_2^* \equiv \arg \min_g \frac{1}{n} \sum_{i=1}^{n} (Y_i - g)^2. \tag{3.4}
\]

Although the technical derivation is optional (see below), the important result is that

\[
\hat{g}_2^* = \bar{Y}_n. \tag{3.5}
\]

Rewriting (3.4) allows the introduction of some terms and concepts used in later chapters. In (3.4), the \( 1/n \) has no effect on the minimization problem because it is unaffected by \( g \). Consequently, it is equivalent to write

\[
\hat{g}_2^* = \arg \min_g \sum_{i=1}^{n} (Y_i - g)^2. \tag{3.6}
\]
To dissect the right-hand side of (3.6), 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

\[
\sum_{i=1}^{n} \hat{U}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{g})^2. \tag{3.7}
\]

Consequently, (3.5) and (3.6) together say that \( \bar{Y}_n \) minimizes the SSR. For this reason, \( \bar{Y}_n \) is a least squares estimator: “least” referring to minimization, and “squares” referring to the second S in SSR.

**Sampling Distribution**

Since the sample mean \( \bar{Y}_n \) is a function of all the \( Y_i \) random variables, it too is seen as a random variable in the frequentist framework. The distribution of an estimator is called its sampling distribution (not to be confused with “sample distribution”). In the special case \( n = 1 \), this is particularly clear since \( \bar{Y}_n = Y_1 \). With \( n = 2 \), consider binary (Bernoulli) \( Y_i \), with \( P(Y_1 = 1) = P(Y_2 = 1) = p \), and \( P(Y_1 = 0) = P(Y_2 = 0) = 1 - p \). Then, \( \bar{Y}_n = (Y_1 + Y_2)/2 \) has three possible values: 0, 1/2, or 1. Further, the distribution of \( \bar{Y}_n \) is fully described by

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

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

\[
P(\bar{Y}_n = 1/2) = 1 - P(\bar{Y}_n = 0) - P(\bar{Y}_n = 1) = 1 - (1 - p)^2 - p^2 = 2p(1 - p). \tag{3.8}
\]

With more complicated distributions of \( Y_i \), it is more difficult to analytically express the distribution of \( \bar{Y}_n \), but hopefully it is clear that \( \bar{Y}_n \) will still be a random variable.

**Discussion Question 3.2** (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.
3.3. ESTIMATION

3.3.2 Non-iid Sampling: Survey Weights

Two of the most common violations of iid sampling are with time series data (Part III) and survey weights.

There are actually many types of weights, although the distinction is usually unimportant for estimation. If you work with weighted data in practice, you should try to understand the particular type of weight in your data. In economics, the most common type is a survey weight, so I briefly discuss those below. Generally, weighted estimation does not depend on the type of weight, but weighted inference (confidence intervals, etc.) does.

If your dataset has survey weights, also called sampling weights, you should use them. Treating weighted data as iid often produces misleading results. Survey weights indicate that the sample is not representative of the population, but they also provide a specific adjustment to allow estimation of features of the population. That is, ignoring weights is like studying a different population, for which external validity may not hold; see Chapter 12.

Skipping the theory, an example is shown in the following code. There are two sub-populations in the population; the first has mean zero, the second has mean 1. In the population, the subpopulations are the same size, so the overall population mean is \((1/2)(0) + (1/2)(1) = 1/2 = 0.5\). However, the second subpopulation is oversampled: each observation has a \(2/3\) probability (instead of \(1/2\)) of coming from the second subpopulation. Thus, there are probably more individuals from the second subpopulation in the sample, which makes the unweighted sample mean larger than \(0.5\). Sampling weights must be used to get the population mean. 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 `svymean()` in the R package `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) #should be ~2/3=0.67

## [1] 0.708

weighted.mean(x=Y, w=1/((itype+1)/3)) #should be ~0.50

## [1] 0.551
```
3.3.3 Brief Bayesian Comparison

The sample mean $\bar{Y}_n$ only uses information from the data; is there other information about $\mu_Y$? If so, a Bayesian approach may be useful. This is especially true if the “other information” is not simply another dataset, but knowledge you (or another expect) have gained from a variety of sources over time. The Bayesian approach is also especially helpful if you need to make a decision (in business or policy) and don’t have lots of data; it combines your expert knowledge with the data’s knowledge. On the other hand, if you do have a lot of data, or your prior knowledge is very limited, or you are trying to summarize the knowledge of one dataset without injecting your own beliefs, then the sample mean may be the better choice.

3.4 Estimator Properties

3.4.1 Bias

Generally, the bias of an estimator $\hat{\theta}_n$ of population parameter $\theta$ is

$$\text{Bias}(\hat{\theta}_n) \equiv E(\hat{\theta}_n) - \theta. \quad (3.9)$$

The subscript $n$ reminds us that the bias may be different for different $n$, even with the same population distribution. This dependence on $n$ makes bias a finite-sample property, which below is contrasted with “large-sample” or “asymptotic” properties. Like $\bar{Y}_n$, any estimator $\hat{\theta}_n$ can be seen as a random variable, and $E(\hat{\theta}_n)$ is the mean of that random variable. So, bias tries to capture whether the estimator is systematically “too high” or “too low.”

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 (3.9),

\[
\text{Bias}(\hat{\theta}) = 0 \iff E(\hat{\theta}) = \theta, \tag{3.10}
\]

where symbol \( \iff \) can be read as “is equivalent to” (or “is true if and only if”).

With iid sampling, the sample mean is unbiased. The bias of \( \bar{Y}_n \) is \( E(\bar{Y}_n) - \mu_Y \), the difference between the mean of the estimator’s sampling distribution and the true value \( \mu_Y \). Unbiasedness means \( E(\bar{Y}_n) = \mu_Y \).

The sample mean’s unbiasedness can be derived as follows. With iid sampling, the mean of each random variable \( Y_i \) equals the mean of the population \( F_Y \), denoted \( \mu_Y \). Since \( Y_i \overset{\text{iid}}{\sim} F_Y \), \( E(Y_i) = \mu_Y \) for all \( i = 1, \ldots, n \). Using the linearity property of the expectation operator repeatedly, i.e., using \( E(aX) = aE(X) \) and \( E(W + X) = E(W) + E(X) \),

\[
E[\bar{Y}_n] = E \left[ \frac{1}{n} \sum_{i=1}^{n} Y_i \right] = (1/n) E \left[ \sum_{i=1}^{n} Y_i \right] = (1/n) E[Y_1 + \cdots + Y_n] = (1/n) \left[ E(Y_1) + \cdots + E(Y_n) \right] = (1/n) [\mu_Y + \cdots + \mu_Y] = (1/n) n \mu_Y = \mu_Y, \tag{3.11}
\]

so \( \bar{Y}_n \) is unbiased for the population mean when sampling is iid.

Although easy to show here, unbiasedness is not very useful in more complex econometric models. First, usually the assumptions required for unbiasedness are much less realistic than those required for another property called consistency (defined below). Second, unbiasedness is often not optimal anyway. For example, imagine the true \( \theta = 0 \), and there are two estimators: \( P(\hat{\theta}_1 = -100) = P(\hat{\theta}_1 = 100) = 1/2 \) and \( P(\hat{\theta}_2 = 0.01) = 1 \). The first estimator is unbiased since

\[
E(\hat{\theta}_1) = (1/2)(-100) + (1/2)(100) = -50 + 50 = 0 = \theta,
\]

whereas the second estimator is biased since \( E(\hat{\theta}_2) = E(0.01) = 0.01 \neq 0 \). However, the first estimator is always wrong by 100 units, whereas the second estimator is always only 0.01 units from the true value. Any reasonable loss function (e.g., quadratic or absolute loss) says \( \hat{\theta}_2 \) is better than \( \hat{\theta}_1 \), as does common sense. So, the biased estimator is clearly better than the unbiased estimator.


3.4.2 Mean Squared Error

Bias is not the only important property. This is shown in the next example, followed by an alternative measure of how bad an estimator is.

Consider two estimators of unknown parameter $\theta$. The first estimator’s distribution is

$$P(\hat{\theta}_1 = \theta - 100) = P(\hat{\theta}_1 = \theta + 100) = \frac{1}{2}. \quad (3.12)$$

The second estimator’s distribution is

$$P(\hat{\theta}_2 = \theta) = P(\hat{\theta}_2 = \theta + 2) = \frac{1}{2}. \quad (3.13)$$

If we only cared about bias, we should prefer the first estimator. The mean of each estimator is

$$E(\hat{\theta}_1) = \frac{1}{2}(\theta - 100) + \frac{1}{2}(\theta + 100) = \frac{\theta}{2} - 50 + \frac{\theta}{2} + 50 = \theta, \quad (3.14)$$

$$E(\hat{\theta}_2) = \frac{1}{2}(\theta) + \frac{1}{2}(\theta + 2) = \frac{\theta}{2} + \frac{\theta}{2} + 1 = \theta + 1. \quad (3.15)$$

Thus, the bias of each estimator is

$$\text{Bias}(\hat{\theta}_1) = E(\hat{\theta}_1) - \theta = \theta - \theta = 0, \quad (3.16)$$

$$\text{Bias}(\hat{\theta}_2) = E(\hat{\theta}_2) - \theta = \theta + 1 - \theta = 1. \quad (3.17)$$

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

Intuitively, it seems like $\hat{\theta}_2$ is clearly better, despite its larger bias. With $\hat{\theta}_1$, our estimate is wrong by 100 every single time: $P(|\hat{\theta}_1 - \theta| = 100) = 100\%$. In contrast, $\hat{\theta}_2$ is exactly right half the time, and only wrong by 2 the other half. Put differently, $\hat{\theta}_2$ is always much closer to $\theta$ than is $\hat{\theta}_1$: $P(|\hat{\theta}_2 - \theta| < |\hat{\theta}_1 - \theta|) = 1$.

Mean squared error (MSE) is a better measure of how bad an estimator is. The idea is analogous to using quadratic loss for prediction. Mean squared error is expected quadratic loss:

$$\text{MSE}(\hat{\theta}) \equiv E[L_2(\hat{\theta}, \theta)] = E[(\hat{\theta} - \theta)^2]. \quad (3.18)$$

There are other loss functions we could use, but this is the most common, and it is generally reasonable.

Continuing our example, we can see $\hat{\theta}_2$ is much better than $\hat{\theta}_1$ according to MSE. Specifically,

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

$$= (1/2)(-100)^2 + (1/2)(100)^2 = 10,000, \quad (3.19)$$

$$\text{MSE}(\hat{\theta}_2) = E[(\hat{\theta}_2 - \theta)^2] = (1/2)(\theta - \theta)^2 + (1/2)(\theta + 2 - \theta)^2$$

$$= (1/2)(0)^2 + (1/2)(2)^2 = 2. \quad (3.20)$$
This matches our intuition: \( \hat{\theta}_2 \) is much better than \( \hat{\theta}_1 \).

MSE can be decomposed into squared bias plus variance. Skipping the algebra,

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

(3.21)

Discussion Question 3.3 (estimator MSE). Consider three estimators of the population mean \( \mu = \text{E}(Y) \), 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. a) Compute the MSE of each estimator. b) Rank the three estimators from best to worst, in terms of MSE. c) Explain why you do or don’t find this ranking intuitive.

3.4.3 Asymptotic Theory

The concept and purpose of asymptotic theory are now mentioned. Conceptually, as in (3.8), estimators like \( \bar{Y}_n \) can be seen as random variables. It is helpful (as in Section 3.5) to know not only the mean of such random variables, but their full distribution. Unfortunately, the distribution of \( \bar{Y}_n \) depends on the distribution of \( Y \), which is unknown. For example, instead of being binary, if \( Y \) has 3 possible values like \( \text{P}(Y = j) = 1/3 \) for \( j = 1, 2, 3 \), then the distribution of \( \bar{Y}_n \) in (3.8) must change; e.g., previously \( \text{P}(\bar{Y}_n = 3) = 0 \), but now \( \text{P}(\bar{Y}_n = 3) > 0 \). So, while it is pleasing to work out the exact distribution of \( \bar{Y}_n \) in simple examples, it is impossible in practice (without resorting to unrealistically strong assumptions, like pretending we known \( Y \) is normal or something).

The purpose of asymptotic theory is to approximate an estimator’s distribution in a way that can be used in practice. The approximate distribution serves two main purposes. The first purpose is to see if an estimator is “good” in the sense of probably being close to the true value (similar to the “probably approximately correct” idea in computer science). The second purpose is to quantify uncertainty about the true value, as in Section 3.5.

To derive such approximations, asymptotic theory relies on taking limits as a mathematical tool. Specifically, the limit as the sample size \( n \) goes to infinity (\( n \to \infty \)) is taken. This should not be interpreted literally; it does not mean that we plan to collect more and more data until we have an infinite amount. It is simply a mathematical tool for approximation.

Approximations can be good or bad. For example, if \( n = 1 \), then an asymptotic approximation based on \( n \to \infty \) is probably be very bad. Even with \( n = 2 \) or \( n = 3 \), it will still be bad. If \( n = 1,000,000 \), then the asymptotic approximation is probably very good. Unfortunately, there is no way to know how good the approximation is for a given \( n \) without knowing the population distribution (which in practice we don’t). For certain distributions of \( Y \), the asymptotic approximation can be very good even for small values like \( n = 10 \). However, for other distributions, and in more complex models like
in Part III the approximation may be bad even with \( n = 500 \). In practice, more data is always better, and then you use the best available method and hope for the best. Although there are many pleasing, precise mathematical results in econometrics, it is never perfect in practice, but for important questions an approximately correct answer is usually better than no answer.

### 3.4.4 Consistency

One important asymptotic property of an estimator is **consistency**. In technical terms, consistency of \( \hat{Y}_n \) for \( \mu_Y \) means that \( \hat{Y}_n \xrightarrow{p} \mu_Y \) as \( n \to \infty \), i.e., the sample mean “converges in probability” to the population mean as \( n \) goes to infinity. More generally, for any estimator \( \hat{\theta}_n \) for population parameter \( \theta \), consistency means

\[
\hat{\theta}_n \xrightarrow{p} \theta \quad \text{as} \quad n \to \infty,
\]

also written \( \text{plim}_{n \to \infty} \hat{\theta}_n = \theta \). The technical details of such convergence are beyond our scope, but the intuitive meaning is within grasp. In contrast to the finite-sample property of unbiasedness, consistency is a **large-sample** property, or asymptotic property, since it involves a (probability) limit as \( n \to \infty \).

Consistency can also be defined as zero asymptotic bias. The primary (though not only) definition of asymptotic bias of an estimator is

\[
\text{AsyBias}(\hat{\theta}) \equiv \text{plim}_{n \to \infty} (\hat{\theta}_n - \theta).
\]

Thus, zero asymptotic bias is equivalent to consistency:

\[
0 = \text{AsyBias}(\hat{\theta}) = \text{plim}_{n \to \infty} \hat{\theta}_n - \theta \iff \text{plim}_{n \to \infty} \hat{\theta}_n = \theta \iff \hat{\theta}_n \xrightarrow{p} \theta \quad \text{as} \quad n \to \infty.
\]

Also, there are the same four types of asymptotic bias as bias (upward, downward, attenuation, away from zero).

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"](https://en.wikipedia.org/wiki/Probably_Approximately_Correct) in computer science.) Unfortunately, again, there are usually no precise, quantitative definitions of “large,” “high,” and “close.” Still, we can see the concept of consistency in some graphs, using the sample mean as the estimator. First, note the subscript \( n \) on \( \hat{Y}_n \). This reminds us that the distribution of the sample mean when \( n = 1 \) is different than the distribution when \( n = 2 \), which differs from the distribution when \( n = 3 \), etc. Seen differently, \( \hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \ldots \) forms a sequence of random variables. Taking the limit of a non-random sequence may sound familiar to you from a previous (precalculus) class, like finding \( \lim_{n \to \infty} 1/n = 0 \) or \( \lim_{n \to \infty} n/(n+1) = 1 \). The probability limit is similar in spirit, but requires more precise, probabilistic characterization. Here, graphs will suffice.
3.4. ESTIMATOR PROPERTIES

Figure 3.1: Simulated sampling distributions of $\bar{Y}_n$. 
Figure 3.1 shows the distribution of $\bar{Y}_n$ for different sample sizes ($n$). The population distribution is $P(Y = 0) = P(Y = 1) = 0.5$, so the true mean is $E(Y) = 0.5$. When $n = 3$, there are only four possible values, but they include zero and one: $\bar{Y}_n \in \{0, 1/3, 2/3, 1\}$. With $n = 10$, there is very little probability that $\bar{Y}_n = 0$ or $\bar{Y}_n = 1$, but it still takes a wide range of values in different samples. With much larger $n = 100$, there is very little probability of randomly sampling a dataset in which $\bar{Y}_n$ is outside the range $[0.4, 0.6]$. That is, although there is still a chance that we sample a dataset with $\bar{Y}_n$ far from the true $E(Y) = 0.5$ (it is random, after all), it is very, very likely that our estimate $\bar{Y}_n$ is at least within 0.1 of the true value, i.e., $P(0.4 \leq \bar{Y}_n \leq 0.6)$ is very close to 100%. With even larger $n$, it will be even more probable that the estimator is even closer to the true value. This is what “consistency” means.

3.5 Quantifying Uncertainty

The point estimates in Section 3.3 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. Section 3.5 discusses the underlying frequentist view of uncertainty, followed by different methods to quantify uncertainty. Although the term is ambiguous, inference often refers to the types of methods in this section (i.e., statistical methods other than point estimation).

3.5.1 Frequentist and Bayesian Views of Uncertainty

As noted in Section 3.1, the frequentist and Bayesian perspectives on uncertainty differ. The Bayesian approach computes a posterior distribution that describes an uncertain belief. Specifically, the posterior assigns different probabilities to different possible values of a parameter (like the mean or median), representing our belief after having seen the data (and given our prior beliefs). For example, maybe 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. After looking at the artifacts more closely, they look most similar to Missouria artifacts, but you are unsure. Quantitatively, you believe there’s a 50% chance they were Missouria, 40% chance Osage, and 10% chance Illini, which is your posterior distribution of $\theta$ after having seen the artifacts: $P(\theta = M) = 0.5$, $P(\theta = O) = 0.4$, $P(\theta = I) = 0.1$. This is the Bayesian way of quantifying uncertainty, and probably the more intuitive one.

Instead of probabilities from a posterior belief, frequentist probabilities consider different random samples (datasets) that could be drawn from the same population. This setup is sometimes called repeated sampling. For intuition, imagine you could ran-
3.5. QUANTIFYING UNCERTAINTY

domly 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. Of particular interest are events involving an estimator like the sample mean, \( \bar{Y}_n \). For example, if \( \bar{Y}_n \leq 0 \) in 50 of the 100 hypothetical samples, then the probability is roughly \( P(\bar{Y}_n \leq 0) = 50/100 = 50\% \). Or, if \( \bar{Y}_n \) falls in the interval \([-0.4, 0.4]\) in 70 of 100 samples, then \( P(-0.4 \leq \bar{Y}_n \leq 0.4) = P(\bar{Y}_n \in [-0.4, 0.4]) \approx 70\% \). More technically, we’ll imagine the long-run probability of events over an ever larger number of repeated samples from the same population, essentially an infinite number of such samples; but to understand the meaning behind these probabilities, it may still help to just imagine 100 samples. A simulated example is shown below in Table 3.1.

Although beyond our scope, the frequentist/Bayesian difference in measuring uncertainty is insightfully and entertainingly illustrated (using cookies and only discrete probability distributions) at [https://stats.stackexchange.com/a/2287](https://stats.stackexchange.com/a/2287)

3.5.2 Sampling Distribution

The sampling distribution of an estimator describes its probability distribution from the repeated sampling perspective of Section 3.5.1. As in Section 3.3.1 an estimator like \( \bar{Y}_n \) can be seen as a random variable. Estimators are random variables because they are computed using sample values, which can be seen as random variables. The probability distribution of the estimator is called the sampling distribution, where the estimator’s probabilities of having different values can be thought of as in Section 3.5.1 for intuition. Instead of individual events like in Section 3.5.1 you could construct a histogram of the 100 values of \( \bar{Y}_n \) themselves to show the (approximate) sampling distribution of \( \bar{Y}_n \).

Such a histogram and probabilities can be simulated in R. In the code below, \( P(Y = j) = 1/5 \) for \( j = -2, -1, 0, 1, 2 \). Sample sizes \( n = 1, 10, 100 \) are considered. In each case, 100 random samples (of \( n \) observations each) are drawn from the population.

```r
n <- c(1, 10, 100) # sample sizes
NREP <- 100 # number of simulated samples
set.seed(112358) # for replicability
Ybars <- matrix(NA, nrow=NREP, ncol=length(n))
for (irep in 1:NREP) {
  # Y[i] ~iid uniform{-2,-1,0,1,2}
  Y <- sample(x=-2:2, size=max(n), replace=TRUE)
  for (jn in 1:length(n)) {
    Ybars[irep, jn] <- mean(Y[1:n[jn]])
  }
}
# plot simulated sampling distributions
```
CHAPTER 3. ONE VARIABLE: SAMPLE

Figure 3.2: Simulated sampling distributions of the sample mean.

```r
par(family=PARFAM, mar=c(4.0,5.0,3.0,0.0),
    mgp=PARMGP, mfrow=c(1,3))
hist1 <- hist(x=Ybars[,1], plot=FALSE)
histmax <- hist(x=Ybars[,length(n)], plot=FALSE)
max.dens <- max(histmax$density)
for (jn in 1:length(n)) {
  tmp <- hist(x=Ybars[,jn], plot=FALSE)
  tmpby <- 1
  if (jn>1) tmpby <- tmp$breaks[2]-tmp$breaks[1]
  b <- seq(from=-2.5, to=2.5, by=tmpby)
  hist(x=Ybars[,jn], breaks=b, freq=FALSE, ylim=c(0,max.dens),
      xlab="Value",
      main=sprintf("Sampling distribution \n n=%d",n[jn]),
      cex.main=1.5, cex.lab=1.5, cex.axis=1.5)
}

# compute simulated probabilities
le0 <- 0+(Ybars <= 0)
Pr.le0 <- colMeans(le0)
a <- 0.4 # check if Ybar in [-a,a]
in.int <- 0+(-a <= Ybars & Ybars <= a)
Pr.int <- colMeans(in.int)
```
Table 3.1: Simulated event probabilities.

<table>
<thead>
<tr>
<th>Sample</th>
<th>$\bar{Y}_n$</th>
<th>$\bar{Y}_n \leq 0$?</th>
<th>$\bar{Y}_n \in [-0.4, 0.4]$?</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</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>-0.20</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Average 0.01 52/100 67/100

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

Table 3.1 shows simulated probabilities to aid intuition about frequentist probabilities generally and the upcoming concept of a confidence interval (Section 3.5.4) specifically. The sample size is $n = 10$. Each row of values in the table corresponds to one of the 100 datasets randomly sampled from the (same) population. From each sample, the value of $\bar{Y}_n$ is computed and shown in the second column of the table. Given the value of $\bar{Y}_n$, each event either occurs (represented by a value of 1) or does not (0). For example, in the first row (Sample #1), $\bar{Y}_n = 0.5$, so the event $\bar{Y}_n \leq 0$ does not occur since $\bar{Y}_n = 0.5 > 0$ (hence 0 entry in that column), and the event $\bar{Y}_n \in [-0.4, 0.4]$ does not occur since $\bar{Y}_n = 0.5 > 0.4$ (hence 0 in that column). In Sample #2, though, both events occur, since $\bar{Y}_n = -0.1 \leq 0$ and $-0.4 \leq \bar{Y}_n = -0.1 \leq 0.4$, so both columns have an entry of 1. For each event column, the final, bottom row of the table shows the total number of 1s divided by the total number of samples (100), i.e., in what proportion of the samples/rows the event occurred. These totals suggest that $P(\bar{Y}_n \leq 0) \approx 52\%$ and $P(\bar{Y}_n \in [-0.4, 0.4]) \approx 67\%$.

Effect of Larger Sample Size

Figure 3.2 shows two main differences in the distribution of $\bar{Y}_n$ for smaller versus larger $n$. First, the distribution’s spread differs. As in Figure 3.1, the distribution is more spread out when $n$ is small, and more concentrated (around the population $E(Y)$) when $n$ is large. This again reflects the property of consistency.

Second, Figure 3.2 shows that the distribution’s shape differs by sample size. When $n = 1$, then $\bar{Y}_n = Y_1/1 = Y_1$, which has the same distribution as the population random variable $Y$, i.e., uniform over the integers between $-2$ and $2$ (inclusive). Thus, the $n = 1$ histogram is very flat. It is not perfectly flat because we only used 100 simulated samples, so there is some simulation error; if we simulated 1,000,000 samples, it would
look nearly perfectly flat, accurately reflecting the true sampling distribution of \(\bar{Y}_n\) with \(n = 1\). Even with 100 samples, the difference in shape among the histograms is clear. With larger \(n\), instead of looking flat, there is a high concentration of probability around the middle values, decreasing for values farther from the middle. Even with \(n = 10\), the shape looks reminiscent of a bell curve, an idiom for the graph of the probability density function (PDF) of a normal distribution. This reflects a central limit theorem: even when the distribution of \(Y\) is not normal, the distribution of \(\bar{Y}_n\) is very close to normal when \(n\) is large.

**Approximation: Central Limit Theorem**

The central limit theorem is seen better in a slightly modified simulation. Before, with the largest \(n\), it is difficult to tell the shape of the distribution because it is so concentrated. The shape is more easily seen after rescaling the graph to spread out the histogram. Specifically, \(\sqrt{n}\) is magically just the right amount to rescale by, after we “center” the estimator at its true value (i.e., subtract the true value). So, the centered mean estimator is \(\bar{Y}_n - \mu_Y\), where as usual \(\mu_Y \equiv E(Y)\), and the scaled and centered estimator is

\[
\sqrt{n}(\bar{Y}_n - \mu_Y). \tag{3.24}
\]

More generally, the scaled and centered estimator \(\hat{\theta}\) of \(\theta\) is \(\sqrt{n}(\hat{\theta} - \theta)\).

The random variable in (3.24) has a distribution that depends on \(n\), but the distribution is basically the same for any large \(n\). For simplicity, let \(R_n \equiv \sqrt{n}(\bar{Y}_n - \mu_Y)\). When \(n\) is small, the distribution of \(R_n\) for different \(n\) can look very different; e.g., \(R_1\) and \(R_2\) may have very different distributions. But with large \(n\), the distributions all look roughly the same; \(R_{100,000}\) and \(R_{1,000,000}\) have nearly identical distributions (you could not tell the difference visually). That is, there is a single random variable \(R\) whose distribution is an excellent approximation for the distribution of all \(R_n\) with “large enough” \(n\). (Unfortunately, again, “large enough” is not usually defined very precisely in practice.)

This idea that a single \(R\) distribution can approximate many different \(R_n\) distributions is the essence of the central limit theorem (CLT), another asymptotic concept. A CLT (there are many variations) provides technical conditions under which \(R\) exists and can be described. More precisely, \(R\) follows a normal (Gaussian) distribution with mean zero and a variance that can be estimated from the data. Although this approximation uses the mathematical trick of taking a limit as \(n \to \infty\), this again does not literally mean “we are collecting more data” or anything; it is just a way to approximate the distribution of \(R_n\) for finite \(n\).

For the sample mean with iid sampling, the approximating distribution is

\[
\sqrt{n}(\bar{Y}_n - \mu_Y) \sim N(0, \sigma_a^2), \tag{3.25}
\]

where “\(a\)” stands for “asymptotic” or “approximate,” and the variance \(\sigma_a^2\) is discussed
3.5. QUANTIFYING UNCERTAINTY

below. Rearranging, the approximate distribution of \( \bar{Y}_n \) is

\[
\bar{Y}_n \overset{a}{\sim} N(\mu_Y, \sigma_a^2/n).
\] (3.26)

That is, \( \bar{Y}_n \) is approximately normally distributed with mean \( \mu_Y \) and variance proportional to \( 1/n \) (i.e., smaller variance for larger \( n \)), where this approximation is good for large \( n \) but possibly (very) bad for small \( n \). The fact that the variance is very small when \( n \) is very large is related to the property of consistency.

There are three important features of the approximate distribution of \( \sqrt{n}(\bar{Y}_n - \mu_Y) \) in (3.25). First, its mean is zero. This should not be surprising given the unbiasedness of \( \bar{Y}_n \). From (3.11),

\[
E(\bar{Y}_n) = \mu_Y \quad \text{so} \quad E(\bar{Y}_n - \mu_Y) = 0
\]

the centering of \( \bar{Y}_n \) at \( \mu_Y \) makes the mean exactly zero, even in finite samples (finite \( n < \infty \)).

Second, the distribution’s shape is Gaussian, regardless of the distribution of \( Y \). (Technically, there are some exceptions, but they are extremely rare in economic data.)

Third, unlike the mean or shape, the variance \( \sigma_a^2 \) does depend on the distribution of \( Y \), but it can be estimated. However, like the finite-sample mean of \( \bar{Y}_n \), the finite-sample variance of \( \bar{Y}_n \) can be computed. Recall the variance properties \( \text{Var}(aX) = a^2 \text{Var}(X) \) and \( \text{Var}(X - a) = \text{Var}(X) \) for constant (non-random) \( a \) and random variable \( X \). Using these properties,

\[
\text{Var}(\sqrt{n}(\bar{Y}_n - \mu_Y)) = \text{Var}(\sqrt{n}\bar{Y}_n - \sqrt{n}\mu_Y) = \text{Var}(\sqrt{n}\bar{Y}_n) = n \text{Var}(\bar{Y}_n). \tag{3.27}
\]

Consequently, \( \sigma_a^2 = n \text{Var}(\bar{Y}_n) \). To compute \( \text{Var}(\bar{Y}_n) \), additional properties of variance combined with the iid sampling properties are used. In particular, the “independence” part of iid implies \( \text{Cov}(Y_i, Y_k) = 0 \) for any \( i \neq k \), and the “identically distributed” part of iid implies \( \text{Var}(Y_i) = \sigma_Y^2 \) for all \( i = 1, \ldots, n \). Using this,

\[
\text{Var}(\bar{Y}_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^{n} Y_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} \text{Cov}(Y_i, Y_k) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(Y_i) = \frac{1}{n^2} n \sigma_Y^2 = \sigma_Y^2/n. \tag{3.28}
\]

Plugging this into (3.27),

\[
\sigma_a^2 = n \text{Var}(\bar{Y}_n) = \sigma_Y^2. \tag{3.29}
\]

Although unknown, \( \sigma_Y^2 \) can be estimated as in Section 3.3.
Figure 3.3: Simulated and approximate distributions of $\sqrt{n}(\bar{Y}_n - \mu_Y)$.

Figure 3.3 is similar to Figure 3.2 but with centered and scaled sample means. There are also more simulated samples to reduce simulation error. Also, on top of each histogram is the PDF of the approximating normal distribution from (3.25). With $n = 1$, this is clearly a poor approximation. With $n = 10$, it is not perfect, but it is very good. With $n = 100$, the approximation is very good. This helps visualize the central limit theorem. However, the approximation’s accuracy for a given $n$ depends on the population distribution of $Y$, so unfortunately we may not extrapolate too much.

Non-iid Sampling

The sampling distribution will differ if sampling is not iid. In addition to weights and time series data (discussed in Section 3.3.2 and Part III), economic data is often not iid due to cluster sampling and/or stratified sampling.

In cluster sampling, individual units in the population are grouped into clusters. Instead of randomly sampling individual units from the population, clusters are randomly chosen, and only individual units within selected clusters may appear in the sample. Sometimes, all individual units within each selected cluster are included; alternatively, units may be randomly sampled within each selected cluster, known as two-stage cluster sampling. Cluster sampling is often used for its logistical and financial benefits, but it may increase the spread of the sampling distribution of estimators.

In stratified sampling, again individual units are considered as parts of larger groups, but now all groups are represented. That is, a certain number of individual units from each group is randomly sampled, to ensure the desired number of units from each group. The goal of stratified sampling is often to reduce uncertainty by ensuring a more represen-
3.5. QUANTIFYING UNCERTAINTY

tative sample, or else to ensure enough observations for important but small groups (e.g.,
homeless), in which case there will also be sampling weights. Some diagrams, discussion,
and examples are at \url{https://en.wikipedia.org/wiki/Stratified_sampling}.

Altogether, combinations of weights, clustering, and/or stratification are often termed
complex sampling design. The first key step is for you to recognize when a dataset
has weights, clustering, and/or stratification. For estimation, there is often a weights
argument in R functions, and the weight type usually does not matter. For inference,
the weight type matters, as do clustering and stratification; how to proceed in such cases
is beyond our scope.

3.5.3 Standard Errors

Definition and Terminology

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

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

(3.30)

Recall from Section 2.3 that the standard deviation has the same units as the variable
itself, so the standard error has the same units as \( \hat{\theta} \). Although personally I prefer to say
“estimated standard error,” sometimes people say “standard error” to refer to an estimate
of the standard deviation in (3.30), so its meaning is ambiguous. In this text at least,
“estimated standard error” will be used in the latter case, with notation \( \hat{\text{SE}}(\hat{\theta}) \), where the
“hat” over SE indicates that it is an estimate computed from the sample. Causing yet
more confusion, \( \hat{\text{SE}}(\hat{Y}_n) \) is often called the standard error of the mean, even though
I would call it the estimated standard error of the sample mean.

Interpretation

In economics, empirical results almost always report standard errors alongside point
estimates. Understanding standard errors (and their use in Section 3.5.4) is critical not
only for producing empirical analysis, but also understanding and critically discussing
empirical analysis.

The standard error helps quantify uncertainty. Qualitatively, the spread of an estimator’s
sampling distribution represents uncertainty. For example, in the (impossible)

extreme where \( \hat{\mu}_Y = \mu_Y \) in every possible dataset, there is no uncertainty: we can learn
\( \mu_Y \) exactly, and we are certain of this because it is true regardless of our random sample.

As an example where the sampling distribution is very spread out, maybe in one random
sample \( \hat{\mu}_Y = \mu_Y + 59 \), but in another \( \hat{\mu}_Y = \mu_Y - 589 \), and in another \( \hat{\mu}_Y = \mu_Y + 3 \), all
occurring with similar probability. If in our sample we compute \( \hat{\mu}_Y = 19 \), then our best
single guess is \( \mu_Y = 19 \), but there is much uncertainty: we don’t know if we have the
dataset where \( \hat{\mu}_Y = \mu_Y + 59 \), in which case \( \mu_Y = -40 \), or if we have the dataset where \( \hat{\mu}_Y = \mu_Y - 589 \), in which case \( \mu_Y = 608 \).

One way to quantify the spread of a sampling distribution is with its standard deviation, which is the standard error. Since “uncertainty” and “spread” are not precise terms, there could be other ways to quantify them, as discussed in Section 2.3. However, the spread of a normal distribution is fully described by its standard deviation, and the sampling distribution of \( \bar{Y}_n \) (and many other estimators) is approximately normal. Further, the standard error is tightly linked to the width of confidence intervals; see Section 3.5.4.

**Estimation of Standard Errors**

To estimate the standard error, first consider the sample mean with iid sampling. From (3.28), the true SE of \( \bar{Y}_n \) is

\[
SE(\bar{Y}_n) \equiv \sqrt{\text{Var}(\bar{Y}_n)} = \sqrt{\frac{\sigma_Y^2}{n}} = \frac{\sigma_Y}{\sqrt{n}}.
\]

(3.31)

This is true for all \( n \) (i.e., not an approximation), but it is less useful when \( n \) is small enough that \( \bar{Y}_n \) may not have an approximately normal distribution. Estimating \( SE(\bar{Y}_n) \) requires estimating \( \sigma_Y \).

```r
set.seed(112358)
n <- 20
Y <- rnorm(n) # iid N(0,1)
Ybar <- mean(Y)
SEhat1 <- sd(Y) / sqrt(n)
retmean <- boot(data=Y, statistic=function(x,i) mean(x[i]),
R=100)
SEhat2 <- sd(retmean$t)
c(SEhat1, SEhat2) #est'd asy. and bootstrap SE for sample mean
## [1] 0.229 0.228
```

**3.5.4 Confidence Intervals**

Like the standard error, a **confidence interval** (CI) helps quantify uncertainty. Essentially, a CI is a range of values that includes the true population value (like \( \theta \) or \( \mu_Y \)) with high probability. Recall that frequentist “probability” is over repeated samples from the same population, like in Table 3.1. So, if we imagine 100 random samples, a frequentist CI would be different in each sample (since it depends on the data) but include the true value in most of the samples, say 90/100. The focus on such frequencies of events underlies the term “frequentist.”
In contrast, Bayesian (posterior) probabilities describe our beliefs after seeing the (actual) data. A Bayesian credible interval is similar to a CI in containing the true value with “high probability,” but reflecting our belief about the true value. The Bayesian idea aligns better with how people discuss uncertainty informally, e.g., “I think there’s a 50% chance that person is older than 60” describes my belief about the person’s age, not a statement about randomly sampling many datasets. Nonetheless, we focus on frequentist CIs, which (for better or for worse) remain more common in practice. Thankfully, with large enough \( n \), the frequentist and Bayesian intervals are very similar in many cases.

The probability that a CI contains the true value is called its **coverage probability** (or **confidence level**). In practice, you specify the nominal coverage probability of your CI. A good CI will have actual coverage probability close to the nominal level, but the nominal and actual probabilities may differ greatly in some cases. 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. It is “two-sided” since (let’s assume) \( \hat{L} \neq -\infty \) and \( \hat{U} \neq \infty \). The coverage probability of this CI for the parameter \( \theta \) is

\[
P(\theta \in [\hat{L}, \hat{U}]) = P(\hat{L} \leq \theta \leq \hat{U}).
\] (3.32)

In the frequentist view, \( \theta \) is fixed (non-random), whereas \( \hat{L} \) and \( \hat{U} \) are random variables since they are computed from the \( Y_i \) sample values, which are different across different datasets.

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 interesting, details are beyond our scope.

```r
library(boot)
set.seed(112358) #for replicability
n <- 50
Y <- rnorm(n=50, mean=0, sd=1) # iid N(0,1)
CIttest <- t.test(x=Y, conf.level=0.95,
    alternative='two.sided')$conf.int
ret <- boot(data=Y, statistic=function(x,i) mean(x[i]), R=100)
tmp <- boot.ci(boot.out=ret, conf=0.95, type=c('basic','bca'))
out.table <- rbind(CIttest,tmp$basic[4:5],tmp$bca[4:5])
rownames(out.table) <- c('Normality','Boot.basic','Boot.BCa')
colnames(out.table) <- c('Lower','Upper')
```
Perils of Ignoring Non-iid Sampling

A CI justified by iid sampling may perform poorly when sampling is not iid. For example, 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 \( Y_i = Y_k \). 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. Seen this way, Samples A and B contain the same amount of information, so they should lead to the same amount of uncertainty. However, if (incorrectly) both are assumed iid, the larger sample size \( n_B > n_A \) is incorrectly interpreted as greater certainty and leads to smaller CIs.

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

\[
\]

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

```r
print(round(out.table,digits=3))
##        Lower Upper
## Normality -0.213  0.370
## Boot.basic -0.234  0.387
## Boot.BCa  -0.233  0.388
```

```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) {
```
3.5. QUANTIFYING UNCERTAINTY

```r
sampleA <- sample(x=vY, size=nA, replace=TRUE, prob=pY)
sampleB <- rep(x=sample(x=vY, size=nA, replace=TRUE, prob=pY), times=2)
CIs[['A']][irep,] <-
t.test(x=sampleA, conf.level=CL,
       alternative='two.sided')$conf.int
CIs[['B']][irep,] <-
t.test(x=sampleB, conf.level=CL,
       alternative='two.sided')$conf.int
}
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        /zero.alt3.9 /zero.alt3.895 /zero.alt3.749
```

The Sample A CI is much better than the Sample B CI; i.e., its coverage probability is closer to the desired confidence level of 90%. As the results show, the simulated coverage probability (CP) of the CI using Sample A is 89.5%, whereas the CP of the CI using Sample B is 74.9%.

3.5.5  \textit{p}-values

Frequentist hypothesis testing and \textit{p}-values are precise, strange, common, and commonly misunderstood. They are usually not the best way to quantify uncertainty and often confuse people. Still, they are ubiquitous, so you should have a basic understanding of the concepts as part of your econometric fluency.

**Interpretation and Use**

A \textit{p}-value measures how likely the observed data would be if a certain hypothesis were true. For example, with notation explain more later, consider the hypothesis $H_0: \mu_Y = 0$. (It will be more interesting in later chapters when the mean is replaced by a causal effect.) The \textit{p}-value is always between zero and one. Small values near zero indicate that the observed dataset would be unlikely if actually $\mu_Y = 0$. Specifically, if the observed dataset has $n$ observations and sample mean $\bar{Y}$,\footnote{Equation (3.33).}

\[ p = P(|\bar{Y}_n| \geq |\bar{Y}_o| \mid \mu_Y = 0). \]  (3.33)
That is, the \textit{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,

\[ p = P(\text{estimate magnitude at least as big as observed} | H_0 \text{ is true}). \] (3.34)

Results with low \( p \)-values are often called “statistically significant,” or having “statistical significance.” This mostly applies to cases where \( H_0 \) is that some effect is zero. Then, a \( p \)-value near zero is interpreted as evidence against a zero effect, i.e., evidence for a non-zero effect. That is, the estimated effect appears more “significant” than would be expected “statistically” (due to random sampling variation) if there were actually zero effect. Historically, \( p < 0.05 \) (5\%) has been a rule-of-thumb for when something is “statistically significant,” but it is more complicated than that. The idea of statistical significance is detailed in Section 3.5.6.

\textbf{Misinterpretation and Misuse}

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.

It is easy to misuse (or abuse) \( p \)-values for many reasons. An entire Wikipedia page (https://en.wikipedia.org/wiki/Misunderstandings_of_p-values) is devoted to this topic. First, as pointed out in an insightful webcomic (xkcd.com/1132), common sense and outside knowledge should be used when interpreting \( p \)-values. A small \( p \)-value alone does not mean \( H_0 \) must be false. Remember, a \( p \)-value below 0.05 is observed 5\% of the time even if \( H_0 \) is always false; this 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 may suggest that it is even less likely that \( H_0 \) is false, as in the comic. A longer, numerical example is given in Section 3.5.7.

A second form of \( p \)-value misuse and abuse relates to \textbf{multiple testing}. Another insightful comic (xkcd.com/882) illustrates this point. In the comic, first somebody runs in, claiming, “Jelly beans cause acne!” (This claim has no reason behind it; it is clearly just one person’s wild idea.) Scientists then (implicitly) run an experiment to measure the effect of eating jelly beans (a candy) on acne (a skin condition). The null hypothesis \( H_0 \) is that jelly beans have zero effect on acne. The \( p \)-value computed from the experimental data is above 0.05, so they decide there is not enough evidence to say jelly beans cause acne.

However, the comic continues, then the first person suggests that it’s not \textit{all} jelly beans, but only a certain color jelly bean that causes acne. The scientists proceed to collect jelly beans of 20 different colors, running 20 experiments and computing 20 \( p \)-values. For 19 of the colors, like purple, red, cyan, and beige, \( p > 0.05 \), which they interpret again as “no link” between jelly beans and acne. But for the green jelly beans,
3.5. QUANTIFYING UNCERTAINTY

$p < 0.05$, which is interpreted as finding a non-zero effect of green jelly beans on acne. The final panel shows the big news story the next day: “Green jelly beans linked to acne! Only 5% chance of coincidence!”

Was it really only 5% chance of coincidence? There is a 5% chance of $p < 0.05$ for a single $p$-value. But what if we compute 20 $p$-values and only consider the smallest one? Note that $5\% = 1/20$. Even without doing formal math, if something happens 1/20 times (on average) and you do it 20 times, there must be a pretty good chance of it happening.

This is the “multiple testing” problem. A $p$-value is intended to measure evidence against a single hypothesis, not testing groups of 20 hypotheses. Usually a null hypothesis is true if there is zero effect; here, null hypothesis $H_0$ is that jelly beans of color $c$ cause acne, where $c$ could be green, blue, etc. If you keep testing enough true null hypotheses (i.e., true effect is zero), you will eventually get false positives (i.e., incorrectly reject a true null hypothesis and conclude there is probably a non-zero effect), approximately 1/20 times if you use $p < 0.05$. There are ways to adjust for this if you are the one doing the research, e.g., use 0.05/20 instead of 0.05 if you run 20 $p$-values. But the newspaper article about jelly beans didn’t mention all the other $p$-values, so the reader would have no way to know to make such an adjustment. Even unintentionally, imagine 20 different labs in 20 different cities, unrelated to each other, all doing an experiment on jelly beans and acne; probably one of them would get $p < 0.05$, just by chance. If only the “exciting” $p < 0.05$ result is published, it misleads us to believe in this result that is false. (This is publication bias.)

The jelly bean experiments also illustrate the importance of remembering what “science” means. The result of a single study is not by itself science. The scientific method rests on replication and repeated testing of hypotheses. A single study, no matter how well done, cannot alone produce a scientific conclusion. If you ever hear, “There was this one new study that found [crazy result]!” you can ignore it; wait till it gets replicated at least a few times. The green jelly bean result provides an interesting hypothesis, but you should wait till other scientists study green jelly beans and replicate the results before avoiding them.

A third misuse of $p$-values is over-emphasizing statistical significance and ignoring the estimated effect magnitude. Statistical significance is essentially about our confidence that an effect is non-zero, but non-zero could still be very close to zero. For example, we could be very, very certain that some program increases annual income by precisely $0.02/yr. The fact that $p = 0.0001$ does not mean this program is good, especially if it costs more than $0.02/yr per participant to operate.

This distinction is detailed further in Section 3.5.6. Unlike the first two issues, this one is mostly solved by simply looking at confidence intervals instead of just $p$-values.

\[\text{This is related to the “replication crisis,” which has its own Wikipedia page: } \text{https://en.wikipedia.org/wiki/Replication_crisis} \]
Discussion Question 3.4 (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?

3.5.6 Statistical and Economic Significance

For description and causality, empirical results have two separate but equally important features: a best guess (point estimate) of a population value, and a quantification of uncertainty. For example, to measure how much a student learns from a class, let \( Y \) be the difference between a student’s final exam score and her baseline score (from before the class). The sample mean \( \bar{Y}_n \) provides a best guess (under quadratic loss) of the average effect of taking the class on score. One way to quantify uncertainty is a \( p \)-value for \( H_0: \mathbb{E}(Y) = 0 \). Roughly speaking, a result is economically significant if \( |\bar{Y}_n| \) is “big” and statistically significant if \( p \) is “small.”

Statistical Significance

The term statistical significance is imprecise; a particular level must be specified. Continuing the exam example, if the \( p \)-value is below 0.05, then the result is “statistically significant at a 5% level.” More precisely still, \( \bar{Y}_n \) is statistically significantly different from zero at a 5% level. This is currently the most common level used in practice, but statistical significance can be defined at any level. For example, if \( p < 0.10 \), then there is statistical significance at a 10% level, and \( p < 0.01 \) means statistical significance at a 1% level, etc. As detailed in Section 3.5.5, all else equal, smaller \( p \)-values correspond to more evidence that the class’s effect is not zero, i.e., that \( \mathbb{E}(Y) \neq 0 \).

Statistical significance can be confusing. As noted in Section 3.5.5, it does not relate to our belief about the probability that \( \mathbb{E}(Y) = 0 \) (like a Bayesian posterior would). Also, “high” statistical significance means a low \( p \)-value (near 0), while “low” significance means a high \( p \)-value.

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,
3.5. QUANTIFYING UNCERTAINTY

and economics) wrote a piece simply titled, “Redefine statistical significance” (Benjamin, Berger, Johannesson, Nosek, Wagenmakers, Berk, Bollen, Brembs, Brown, Camerer, Cesareni, Chambers, Clyde, Cook, De Boeck, Dienes, Dreber, Easwaran, Efferson, Fehr, Fidler, Field, Forster, George, Gonzalez, Goodman, Green, Greenwald, Hadfield, Hedges, Held, Ho, Hoiitink, 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 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 instead of \( p \)-values.

**Economic Significance**

In addition to statistical significance, we care about economic significance: is the estimated effect or difference meaningfully different than zero? Whereas statistical significance only differentiates between zero and non-zero effects, economic significance considers the size of the effect. One way to think about this is: would you personally care about the difference? In the exam example, you could ask yourself: would I care about having an exam score that’s \( \bar{Y}_n \) points higher? If the exam is out of 100 points, and \( \bar{Y}_n = 0.01 \), then you probably don’t care; if \( \bar{Y}_n = 50 \), then you probably do. 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?

For economic significance, we must consider units of measure, and in later chapters, we must consider realistic policy changes. For example, if \( \bar{Y}_n = 10 \) exam points, that may sound like a lot, but what if the exam is scored out of 1000 points, or 10,000 points? In another example, maybe our estimate for an effect on income is 10; is that big? 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. Numbers are meaningless without units. Even with units, there can be some disagreement about what is economically significant ($0.10/hr? $0.01/hr?), but at least it can be a meaningful discussion.

**Examples**

Since statistical and economic significance are two separate concepts, there are four general possibilities: both statistical and economic significance, just statistical, just economic, or neither. Despite conventions with statistical significance, there is no magical threshold for either statistical or economic significance, but rather a continuum between
none and lots. There could be basically no statistical significance, or “marginal” significance, or a very high degree of significance; the \( p \)-value could be any number between 0 and 1, and there are no hard thresholds. The following four examples show more clear versions of the four possibilities.

Continuing the class example, there could be both statistical and economic significance. If \( \bar{Y}_{n} = 30 \) points and the standard errors are small enough that \( p = 0.001 \), then there is statistical significance at a 0.1% level (or a 1% level, or any higher level). If the exam is out of 100 points total, and the baseline scores ranged from 40 to 60 (roughly uniformly, say), then a 30 point increase seems economically significant.

Changing the numbers, there could be neither statistical nor economic significance. Most obviously, if \( \bar{Y}_{n} = 0 \) exactly, then our best guess is that the class does not change the average exam score at all. This is as little economic significance as possible. In that case, \( p = 1 \), so there is not statistical significance at any level.

Another possibility is suggestive evidence where there is economic but not statistical significance. For example, maybe \( \bar{Y}_{n} = 30 \) points like in the first case, but the standard errors are much larger, so \( p = 0.11 \). There is economic significance for the same reasons as before. However, there is no statistical significance at a 5% level since \( p > 0.05 \), or even a 10% level since \( p > 0.10 \), let alone a 0.5% level. That is, in our sample, students scored much better after taking the class, but maybe the sample size was small enough that such a result would occur 10% of the time just by random chance even when \( E(Y) = 0 \). So: does the class help? Because of the large \( p \)-value, there can be no conclusive statement about the class’s effect, but the large point estimate suggests we might want to do a follow-up study with a larger dataset (to improve statistical significance). This is more clearly seen with a 95% CI, which in this case may be something like \([-5, 65]\) points. The CI includes zero and even some negative values, but also some very large values.

The fourth possibility is statistical but not economic significance. For example, maybe \( \bar{Y}_{n} = 0.1 \) points, which is barely even noticeable on the scoring scale used (remember the 40–60 baseline range). However, the standard errors could be so small that \( p = 0.001 \), which earlier we saw was statistically significant at a variety of common levels, even 0.5%. So: does the class help? Because of the small \( p \)-value, we are fairly certain that it does, but we are also fairly certain that it helps only a very small amount. This is more clearly seen in a confidence interval; e.g., a 95% CI may span the very narrow range \([0.09, 0.11]\) points. This is the opposite of the case with economic but not statistical significance: instead of being uncertain about a large effect, here we are certain about a small effect.

**Discussion Question 3.5** (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
3.5. QUANTIFYING UNCERTAINTY

$p = 0.03$. a) Is this estimate economically significant? Hint: consider the units when thinking about if $-0.03$ seems “big.” b) Is this statistically significant? Be precise.

3.5.7 Hypothesis Testing

In the scientific method, theories imply certain hypotheses that can be tested with data. Scientific theories are falsifiable but not verifiable: they can be disproved, but not proved. A theory is maintained until it is disproved, at which point a new theory replaces it, and the new theory is tested empirically.

Sometimes this is the purpose of hypothesis testing in economics, but not often. When it is, clearly we do not wish to falsely “disprove” a true theory, but we do wish to disprove false theories. More often, hypothesis testing is used like the $p$-value, to provide evidence against a statement like $E(Y) = 0$.

Notation and Terminology

Notationally, $H_0$ denotes the null hypothesis 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 to decide whether $H_0$ or $H_1$ is true, but many caveats apply. A somewhat better (but still not nuanced enough) description is that frequentist testing tries to quantify how unlikely a dataset’s evidence against $H_0$ would be if $H_0$ were indeed true.

Much jargon accompanies hypothesis testing. A hypothesis test takes a dataset and computes a decision: either reject $H_0$, or do not reject $H_0$. Sometimes “do not reject” is replaced with “accept,” which seems to be a more obvious antonym for “reject,” but “do not reject” helps emphasize the asymmetry between $H_0$ and $H_1$ in frequentist testing. That is, if there is sufficient evidence, then $H_0$ is rejected by the test, but a lack of rejection only indicates a lack of evidence against $H_0$, not necessarily strong evidence in support of it.

There are four combinations of truth and decision, with more jargon. Either $H_0$ is true (and thus $H_1$ false), or else $H_0$ is false (and $H_1$ true). Either $H_0$ is rejected, or not. If $H_0$ is true and not rejected, then the decision is correct. If $H_0$ is false and rejected, the decision is again correct. If $H_0$ is true and rejected, then the decision is incorrect; this is a type I error, often called a false positive. If $H_0$ is false and not rejected, then the decision is incorrect in a different way; this is a type II error (or “false negative”).

In the frequentist view, the truth about $H_0$ is fixed, and probabilities are again defined in repeated sampling. There is yet more jargon for various such probabilities. Since the

\[^2\text{Pedagogical criticism duly noted: xkcd.com/892}\]
test’s decision depends on the sample, it may reject in some randomly drawn samples but not others, even if sampled from the same population. For example, in 100 imaginary random samples, a test may reject $H_0$ 20/100 times. This corresponds to the **rejection probability** of the test. If $H_0$ is true, then each rejection is a type I error, so the rejection probability is also the **type I error rate**, i.e., the probability of the test making a type I error. If instead $H_0$ is false, then each rejection is correct, and the rejection probability is called the **power** of the test. Alternatively, when $H_0$ is false, each non-rejection is a type II error, and there is a **type II error rate**, i.e., the probability of the test making a type II error. Thus, power is just 100% minus the type II error rate (and vice-versa), but power is good, whereas errors are bad. 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 type I error rate when $\theta = 0$ is identical to the size of the test.

**Frequentist Hypothesis Testing**

The statistical goal of a frequentist test is to maximize power while controlling the type I error rate at the desired level. For example, a **level 5%** test promises to keep the type I error rate below 5%. In practice, such claims are usually based on asymptotic approximations, so they are usually true for large $n$ but may not be true for small $n$; again, unfortunately “large” and “small” have no precise, universal meaning. There may be many possible level 5% tests; some econometricians work to find such a test that has better power than other such tests, in different models.

Why focus on the type I (instead of type II) error rate? If the null hypothesis is an implication of an actual economic theory, like rational expectations or unpredictability of stock returns, then the answer is that our primary concern is not to reject true economic theories very often. That is, if we have much uncertainty, then we’d rather err on the side of maintaining our theory than falsely rejecting it. (You may have also heard the analogy of the criminal court judge wanting to minimize the rate of convicting innocent defendants, but while compelling, it seems of limited use in economics.)

However, the most common hypothesis test in economics is whether the coefficient on some variable is zero. To make it more interesting, let’s call the coefficient an “effect” (like the effect of watching TV on college graduation); the null hypothesis is zero effect. The type I error rate is how often the test incorrectly decides that there is a non-zero effect ($H_1$) when in fact there is no effect ($H_0$). Thus, controlling the type I error rate at 5% means that if we examine lots of datasets where the true effect is zero, then we will incorrectly infer a non-zero effect about one in 20 times. However, that doesn’t really answer the question. Type II error is also bad: if there is a real effect, but we concluded there isn’t, then people may miss out on the benefits of a truly helpful policy. Rather than discussing power analysis and optimal selecting of a hypothesis test’s level and Bayesian alternative, I instead suggest that confidence intervals (or Bayesian credible
3.5. QUANTIFYING UNCERTAINTY

intervals) are more insightful.

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.

**Discussion Question 3.6** (jellybean solution?). Recall from Section 3.5.5 the jelly bean comic from [xkcd.com/882](https://xkcd.com/882). In the comic, they essentially do hypothesis tests with a 5% level, rejecting \( H_0 \) (no effect) if \( p < 0.05 \). Since they do 20 hypothesis tests, even if each individual test is unlikely to make an error, it becomes very likely that one of the tests will make 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? Explain why or why not.

**Discussion Question 3.7** (nova). Consider again the comic from [xkcd.com/1132](https://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?

**Misuse**

Hypothesis testing is subject to the same misuse and abuse as \( p \)-values, as mentioned in Section 3.5.5. One problem discussed there was neglect of how likely \( H_0 \) seems *before* seeing data. This problem is now further illustrated by the classic example of a medical test for an uncommon disease.

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

Table 3.2 shows the disease status and test results for 1,000,000 random people.
The table shows that the disease is uncommon since only 1000 people have it, i.e., \( 1/1000 \) or 0.1%.

Table 3.2 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 \( H_0 \). A type I error is made when the test is positive for somebody without the disease, and a type II error is when the test is negative for somebody with the disease. Table 3.2 shows the type II error rate is actually zero (or, power is 100%): of the 949,050 people who test negative, none of them have the disease. The type I error rate is how often the test is positive (rejects) when testing only individuals without the disease (\( H_0 \) true). Table 3.2 shows that out of 999,000 people who did not have the disease (\( H_0 \) true), the test incorrectly said 49,950 of them had the disease (reject \( H_0 \)). Thus, the type I error rate is \( 49,950/999,000 = 5\% \).

Since a 5% rate of type I errors (false positives) sounds low, should you be worried if your test result comes back positive? Given that you tested positive (but don’t know your true disease status), you could be in one of two boxes in the table: 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. But 49,950 is much larger than 1000! That is, conditional on having tested positive (like 50,950 people in the population), 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. However, the conclusion is not that test results are meaningless; for example, according to the table, if the test comes back negative, then there is 100% guarantee of having no disease. The reason for being skeptical of the positive test result is that \( H_0 \) being false is so unlikely. This is not true for all \( H_0 \) generally; it requires thought in each specific case. In fact, what if you only took the disease test because you displayed all 17 characteristic symptoms of it, whereas Table 3.2 is for a random sample from the full population (most of whom have no symptoms)? If you have all the symptoms, you may think \( H_0 \) being true is very unlikely, and you would probably be reasonable to conclude you have the disease if the test is also positive. This also touches on the idea of external validity explored in Chapter 12: your “population of interest” is people like you who have all the symptoms, whereas the “population studied” in Table 3.2 is mostly people who do not have symptoms.

3.5.8 Mental Math for Statistical Uncertainty

Imagine you see estimate \( \hat{\theta} \) with standard error \( \widehat{SE}(\hat{\theta}) \). At approximately a 5% level (actually 4.55%), the estimate is statistically significant if \( |\hat{\theta}| \geq 2 \widehat{SE}(\hat{\theta}) \), i.e., null hypothesis \( H_0: \theta = 0 \) can be rejected at approximately a 5% level. Further, an approximately 95% (actually 95.45%) confidence interval is from \( \hat{\theta} - 2 \widehat{SE}(\hat{\theta}) \) to \( \hat{\theta} + 2 \widehat{SE}(\hat{\theta}) \).
These calculations are useful for a few reasons. First, they're easy enough to do in your head in many cases. Second, 5% is arbitrary anyway, so 4.55% is equally good (or equally bad!). Third, confidence intervals, $p$-values, and such are based on asymptotic approximations that aren’t exact in finite samples anyway. For 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). The difference between the actual (finite-sample) significance level and the desired significance level (like 5%) is almost always much bigger than the difference between 5% and 4.55%.

Consider an example. Imagine you estimate $\hat{\theta} = -3.2$ and $\hat{SE}(\hat{\theta}) = 1.5$. Then, $|\hat{\theta}| = 3.2$ and $2\hat{SE}(\hat{\theta}) = 3.0$. Since $3.2 > 3.0$, the result is statistically significant at a 5% level, i.e., the $p$-values is below 0.05 and $H_0: \theta = 0$ is rejected at a 5% level. Further, a 95% confidence interval is $[-3.2 - 3.0, -3.2 + 3.0] = [-6.2, -0.2]$.

### 3.5.9 Statistical Decision Theory

If you want to incorporate data more formally into your decisions, then you should learn more about statistical decision theory, though it is beyond our scope. Hypothesis testing is basically never the best way to make a decision using data. For example, see [Berger (1985)](#).

**Discussion Question 3.8** (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 the variable `wage` with R command `mean(card$wage)` or Stata command `mean wage` (which also computes a 95% confidence interval)

d. Estimate the population mean accounting for the sampling weights with R command `weighted.mean(x=card$wage, w=card$weight)` or Stata command `mean wage [pweight=weight]` (which also computes a 95% confidence interval)

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. Beyond the simple average of `wage` from before, compute a version using the natural log function and another with its inverse, the exponential function. Specifically, compute the mean of the log wage observations, and then apply the inverse log (exponential) function. Use R command `exp(mean(log(card$wage)))` or Stata commands:

   ```r
   generate lnwage = log(wage)
   summarize lnwage
   display exp(r(mean))
   ```
Note that \texttt{log}() is natural log, \texttt{ln(·)}, in both Stata and R. (And yes, the values should be different; this illustrates something called Jensen’s inequality.)

h. Compute a weighted, 90\% confidence interval for wage. In R, replace \texttt{level=0.95} with \texttt{level=0.90}; in Stata, add “option” \texttt{level(90)} to get \texttt{mean wage [pweight =weight]}, \texttt{level(90)}

i. Repeat computation of a point estimate and 95\% confidence interval (without and with weights) for the mean of a different variable in the dataset. In R: part \texttt{(c)} computes the unweighted point estimate, part \texttt{(d)} computes the weighted point estimate, part \texttt{(e)} computes the unweighted CI, and part \texttt{(f)} computes the weighted CI. In Stata: part \texttt{(c)} computes both the unweighted point estimate and unweighted CI, and part \texttt{(d)} computes both the weighted point estimate and weighted CI.
Chapter 4

One Variable, Two Populations

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)
- Causal inference intro (Masten video): https://www.youtube.com/watch?v=FNpcwi0me1g
With two populations, we can discuss not only description and prediction, but also causality. Important ideas are introduced here that are extended to more complex regression models 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 could the means $E(Y^A)$ and $E(Y^B)$ be 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) Can we interpret $E(Y^B) - E(Y^A)$ as the causal effect of a college degree on wage? 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. You could similarly imagine any other two socioeconomic, geographic, or demographic groups.

The difference of means is $E(Y^B) - E(Y^A)$. 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,

$$P(Y^A = 0) = 0.8, \quad P(Y^A = 1) = 0.2, \quad P(Y^A = 2) = 0,$$
$$P(Y^B = 0) = P(Y^B = 1) = P(Y^B = 2) = 1/3,$$

(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 $B$. Then,

$$E(Y^B) - E(Y^A) = \left[ \sum_{y=0}^{2} y P(Y^B = y) \right] - \left[ \sum_{y=0}^{2} y P(Y^A = y) \right]$$
$$= [(0)(1/3) + (1)(1/3) + (2)(1/3)] - [(0)(0.8) + (1)(0.2) + (2)(0)]$$
$$= [(1/3) + (2/3)] - 0.2$$
$$= 0.8.$$

Always clarify whether you are subtracting the mean of population $A$ from that of population $B$, or $B$ from $A$. Otherwise, confusion abounds. If I just said, “The difference in mean number of children between the populations is 0.8,” it is unclear whether population $B$’s mean is larger or smaller.

The difference of means is also the mean of the differences. For this reason, “mean difference” could mean either; they’re equal anyway. Because of the linearity of the expectation operator,

$$E(Y^B - Y^A) = E(Y^B) - E(Y^A).$$

(4.2)

But, 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, this difference is itself a random variable. Thus, $E(Y^B - Y^A)$ is the population mean of the child number difference, 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 $A$. 
4.1.2 Estimation

To estimate such differences, we can use the same estimators as in Section 3.3. First, each population mean is estimated separately, just as in Section 3.3. Then, the estimates are simply subtracted from each other. For example, with iid sampling, $E(Y_B)$ and $E(Y_A)$ are consistently estimated by the sample means, $\bar{Y}_B$ and $\bar{Y}_A$, so $E(Y_B) - E(Y_A)$ is consistently estimated by $\bar{Y}_A - \bar{Y}_B$. As long as the individual estimators are consistent, their difference is consistent for the population difference. Further, since $E(Y_B - Y_A) = E(Y_B) - E(Y_A)$, the same estimator can be used for both.

The following code estimates the difference in means $E(Y_B) - E(Y_A)$ from simulated iid samples, suggesting population B has a lower mean.

```r
set.seed(112358)
YA <- 0 + sample(x=1:2, size=4, replace=TRUE, prob=rep(1/11,11))
YB <- 2 + sample(x=1:25, size=3, replace=TRUE, prob=26:1/sum(1:26))
mean(YB) - mean(YA)
```

## [1] -2.99

4.1.3 Quantifying Uncertainty

Consider a simple example to develop intuition for quantifying uncertainty about differences. The population object of interest is the mean difference, $\mu_B - \mu_A \equiv E(Y_B) - E(Y_A)$. Assume independent, iid samples with equal sample size $n$. That is, not only are $Y_i^A$ and $Y_k^A$ independent if $i \neq k$ (iid), but $Y_i^A$ and $Y_k^B$ are independent, even if $i = k$ (independent samples). Let $\sigma_A$ and $\sigma_B$ denote the population standard deviations of $Y^A$ and $Y^B$. From (3.26) and (3.29),

\[ \bar{Y}^A \sim N(\mu_A, \sigma_A^2/n), \quad \bar{Y}^B \sim N(\mu_B, \sigma_B^2/n). \]

Since the observations in $\bar{Y}^A$ are assumed statistically independent of the observations in $\bar{Y}^B$, $\bar{Y}^A \perp \bar{Y}^B$. The difference of two independent normal random variables is also normal, with mean equal to the mean difference and variance equal to the sum of variances. (Technically it’s a little more complicated since these are asymptotic approximations, but this result is correct.) In this case,

\[ \bar{Y}^B - \bar{Y}^A \sim N(\mu_B - \mu_A, \sigma_A^2/n + \sigma_B^2/n). \] (4.3)

More generally, a similar result applies to the difference of any two approximately normal estimators based on independent samples.
As in Section 3.5, once the approximate sampling distribution of an estimator is known, we can construct approximate confidence intervals, \( p \)-values, and hypothesis tests. In practice, statistical software like R takes care of all the computation, including the estimation of the unknown \( \sigma_A \) and \( \sigma_B \) in (4.3). Your job is not to compute, but to tell R what you need (e.g., are the data iid, or are there survey weights, etc.), and to interpret the results properly (e.g., don’t conclude \( E(Y^B) = E(Y^A) \) if \( p = 0.12 \)).

The following R code shows 95% confidence intervals for the difference in mean. Datasets are simulated iid. Here, the median CI contains zero, whereas the mean CI does not; and the IQR CI contains zero, whereas the 90th–10th percentile spread CI does not. This reminds us that not all location measures are equivalent, nor are all spread measures, although appropriate choices should come from economic reasons, not statistical significance.

```r
set.seed(112358)
YA <- 0+sample(x=10:20, size=40, replace=TRUE,
               prob=rep(1/11,11))
YB <- 2+sample(x=0:25, 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,
digits=3)
```

### 4.2 Prediction

For prediction, not much changes with a second population. Usually, it is not interesting to predict the difference between two random people from two populations; describing the difference is sufficient. Prediction for each population by itself is identical to Chapter 2.

However, better prediction than in Chapter 2 is now possible. Imagine you work at a carnival, guessing people’s height. In Chapter 2, there is a single population: everyone who might visit the carnival. Chapter 2 only considered prediction for a random person’s height, without consideration of any other information. Now, we can consider two populations, like male and female, or child and adult (assuming we could perfectly observe these classifications, which is not realistic). Then, we could form a different prediction for each population. For example, we could then predict a greater height for males than females, or adults than children, which should increase accuracy.

Part II extends this idea, exploring how regression models can incorporate additional information to improve prediction accuracy.
4.3 Causality

The framework and concepts of Section 4.3 are fundamental to many econometric methods for learning about cause-and-effect relationships. These concepts appear repeatedly in subsequent chapters.

4.3.1 Some Dichotomies

Structural and Reduced Form Approaches

Generally, there are two approaches to learning about causality: the reduced form approach, and the structural approach. The reduced form approach is also called causal inference or program evaluation (like assessing the effects of a job training program or welfare program), among other names. To isolate causality, the reduced form approach tries to find comparisons that are either randomized or “as good as randomized.” The actual mechanisms driving the effect are left unspecified, a black box. Instead, the structural approach tries to model the behavior of economic agents (like individuals and firms), often using economic theory. Related to this difference, the structural approach tries to estimate structural parameters (or “deep parameters”) that describe such behavior, like elasticities, discount factors, risk aversion, and demand curves. In contrast, the reduced form approach focuses on the effect of a particular variable or policy change. Related to this difference, the reduced form approach often has a single binary treatment variable, with $X = 1$ if an individual (or firm, or city, or other unit) was treated and $X = 0$ if not; other variables are only to help estimate the treatment effect. In contrast, structural models usually have multiple variables and parameters of interest, since they hope to learn about the actual mechanisms underlying causal relationships.

The structural and reduced form approaches have complementary advantages, and often both are helpful (Lewbel 2019). Both have contributed to our understanding of economics. Generally, the structural approach requires stronger (often less realistic) assumptions, but it delivers more informative results in return.

There are three main advantages to the structural approach. First, it enables consideration of more possible counterfactual scenarios and policies. (In the language of Chapter 12, it claims more external validity.) A counterfactual is something like a world where sales tax is 2 percentage points higher (like 10% instead of 8%), or minimum wage is $2/hr higher: not the current, actual world (hence “counterfactual”), but a world we could potentially live in. In contrast, the reduced form approach only estimates the effect of a particular policy change in a particular setting; additional arguments must be used to extrapolate such results to other settings or policies. Second, the structural approach attempts to estimate unobservable variables that are important for policy analysis, like utility or social welfare. In contrast, the reduced form approach can only estimate effects on observable variables. Third, the structural approach is often the only option in
complex settings with features like social interactions.

The reduced form approach is less powerful, but often the assumptions are more realistic. (In the words of Chapter 12, reduced form results often have greater internal validity.) As the alternative name “program evaluation” implies, it is very useful for evaluating a particular program, to estimate the program’s effect. It is often helpful to combine a reduced form estimate that we’re confident in with a structural estimate that can help translate that specific result to other possible settings or policies.

Within this chapter’s setup, we can only talk formally about the reduced form approach. Chapter 6 has more on the structural approach in a simple setting.

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 allows a broader set of counterfactual policies, but it requires stronger assumptions. For example, imagine you were analyzing the impact of public (free) 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 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, general equilibrium 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 general equilibrium effects, which requires deeper, structural understanding and modeling of economic behavior.

4.3.2 Differences with Description and Prediction

How does learning about causality differ from description and prediction? Let \( Y^A \) denote the rainfall when nobody is carrying an umbrella, and \( Y^B \) the rainfall when everybody is
carrying an umbrella. A helpful description of the two distributions is that \( P(Y^A = 0) \) is very close to 100\%, and even the 99th percentile of \( Y^A \) is close to zero, whereas the mean, median, and standard deviation of \( Y^B \) are all significantly larger. A good prediction under absolute loss is the median, which is (let’s assume) zero for \( Y^A \) but a positive value for \( Y^B \), something like 0.250 cm (depending on the location). But for causality, does this mean that if we’re in a drought and want it to rain, we should all walk around with umbrellas to cause it to rain? This is preposterous: rain causes umbrella-carrying, not vice-versa. This is called reverse causality, when our outcome variable (rain) actually has a causal effect on the other variable (umbrellas). This shows one added difficulty for causal inference that does not matter for description or prediction.

Consider another example where a descriptive or predictive difference lacks causal interpretation. Let \( Y^A \) be the time it takes me to get to work when nobody is carrying umbrellas, and let \( Y^B \) be the time it takes when everyone is carrying umbrellas. Naturally, \( \text{median}(Y^B) > \text{median}(Y^A) \): more umbrellas correspond to longer commutes. But this doesn’t mean that you can make me late for class by opening lots of umbrellas. The larger median of \( Y^B \) is from rain causing both umbrella-carrying and longer commute time. That is, rain is a confounder (or confounding variable or confounding factor). Figure 4.1 depicts the true causal relationships.

Figure 4.1: Causal relationships among rain, umbrella-carrying, and commute time.

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. Details about such laws may be found at https://en.wikipedia.org/wiki/Right-to-work_law, for example, but the following discussion about causality does not depend on such details. Before the election, one 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. This $8000/yr was computed as the difference in average annual income between states that had a right-to-work law and those that did not, similar to the median\((Y^B) > \text{median}(Y^A)\) in the commute example. This average income difference is a good estimator of the population difference, as in Section 4.1.
but the population difference does not necessarily have a causal interpretation. (Ads endorsing right-to-work also made errors, but not in econometrics.) For example, maybe having lower income causes states to pass such laws, i.e., causality is in the opposite 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 umbrellas and 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. More assumptions are required to interpret a population difference as a causal effect. Such assumptions are discussed more in Section 4.3.5.

To review: when is causality important, separate from description and prediction? Causality is about what will happen if we make a certain decision or change a policy. This may be confusing since “what will happen” sounds like prediction. But “prediction” in econometrics is essentially guessing the realization of a random variable, as a passive observer. With a causal effect, the decision or policy is actively changing one variable, like passing a right-to-work law or increasing the sales tax.

**Discussion Question 4.2** (description, prediction, causality). Which type of question (description, prediction, causality) is each of the following?

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.?

### 4.3.3 Potential Outcomes Framework

The reduced form approach is built on the potential outcomes framework. This is 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 word treatment is often used in this framework, but it can be interpreted very broadly. Because of the framework’s origins, usually the phrase treatment effect is used instead of “causal effect.” In English, usually “treatment” makes us think narrowly about medicine (or facials and lumber?), so initially it may sound odd for economic applications. For example, in economics, the treatment could be a job training program, and the treatment effect is the resulting change in 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. None of these would be called a “treatment” in common English, but they are within the potential outcomes model.
Potential Outcomes

To understand potential outcomes, imagine two parallel universes. For example, in one universe, a student takes introductory econometrics, and in the other universe, she instead takes Intro Stat II (STAT 3500). Everything else in each universe is identical: her parents, her other classes, her height—everything. (For now, we gloss over the difficulties with “everything”: e.g., what if econometrics is required for your degree?) The outcome variable is her annual income five years after graduation, in thousands of U.S. dollars per year (so $Y = 70$ is $70,000/yr$, etc.). Let $Y^A$ denote income in the universe where a student takes STAT 3500, and $Y^B$ income in the econometrics universe. Consider econometrics the treatment, $Y^B$ is called the treated potential outcome (or just treated outcome), while $Y^A$ is the untreated potential outcome (or untreated outcome). (Notationally, instead of $Y^A$ and $Y^B$, often $Y_0$ and $Y_1$ represent the untreated and treated potential outcomes, or $Y(0)$ and $Y(1)$.) Unlike in Section 4.1, $Y^A$ and $Y^B$ are not always observable. That is, if a student takes STAT 3500 instead of econometrics, then we can observe her $Y^A$ but not $Y^B$; if she takes econometrics instead of STAT 3500, then her $Y^B$ is observable but not $Y^A$. The outcomes for student $i$ are denoted $Y^A_i$ for income in the first universe and $Y^B_i$ in the second universe; either $Y^A_i$ or $Y^B_i$ is observable, but not both. This partial observability makes causal inference more difficult than description or prediction.

To deepen understanding, consider other examples of the potential outcomes framework. In the right-to-work example, for each individual, we could imagine her income if her state had a right-to-work law or if it didn’t. In our universe, her state either does or does not have such a law; in the parallel universe, it is the opposite, but we cannot observe the other universe. More formally, the untreated potential outcome $Y^A_i$ is the income of worker $i$ in the universe where her state does not have a right-to-work law, and the treated potential outcome $Y^B_i$ is her income in the universe that does have such a law but is otherwise completely identical. Again, either $Y^A_i$ or $Y^B_i$ is observable, but not both. As another example, universe $B$ could be where a student won the lottery to enter a popular charter school, and universe $A$ where the student went to the conventional public school. The potential outcomes $Y^A_i$ and $Y^B_i$ measure whether or not the student eventually graduated from college in the two universes. In our universe, we can observe $Y^B_i$ if the student won the school lottery or $Y^A_i$ if he didn’t, but never both.

Treatment Effects

The difference between the two potential outcomes for an individual can be considered a treatment effect. Formally, the treatment effect for individual $i$ is $Y^B_i - Y^A_i$. As usual, “individual” could also be firm, country, school, etc. I try to write “treatment effect” instead of “causal effect” when discussing the reduced form approach, as in this chapter.

Consider this definition of treatment effect in the previous examples. For student $i$
with income $Y_i^B$ after taking econometrics and $Y_i^A$ after instead taking STAT 3500, the
treatment effect of her taking econometrics instead of STAT 3500 is defined as $Y_i^B - Y_i^A$.
That is, the treatment effect on her is how much higher (or lower!) her income is in the
parallel universe that is identical other than her taking econometrics instead of STAT 3500. In the right-to-work example, the treatment effect of the law on an individual
is again $Y_i^B - Y_i^A$. The interpretation now is the difference between her income in the
universe where there is the law and the universe without the law, with everything else held constant. In the charter school example, the individual treatment effect is again
$Y_i^B - Y_i^A$ but with a different interpretation. There, 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, 0 if don’t), the only possible treatment effect
values are $Y_i^B - Y_i^A = 1$ if the student graduates in the charter school universe ($Y_i^B = 1$)
but not the public school universe ($Y_i^A = 0$); $Y_i^B - Y_i^A = -1$ if they only graduate in the
public school universe ($Y_i^A = 1$) but not in the charter school universe ($Y_i^B = 0$);
and $Y_i^B - Y_i^A = 0$ if they graduate in both universes ($Y_i^B = Y_i^A = 1$) or if they don’t
($Y_i^B = Y_i^A = 0$). In all examples, the potential outcomes and treatment effects may be
different for different individuals, one example of **heterogeneity** (people being different).

Just like outcomes $Y_i$ in Chapters 2 and 3, individual treatment effects can be seen
as random variables. Since potential outcomes $Y_i^B$ and $Y_i^A$ are seen as random variables
with some joint probability distribution in the population, the individual treatment effect
$Y_i^B - Y_i^A$ is also a random variable with some population distribution. Consequently, if we
could observe both $Y_i^B$ and $Y_i^A$ for all individuals, then we could observe $Y_i^B - Y_i^A$ for all
individuals. Then, like in Section 3.3, we could simply take the sample mean of individual
treatment effects to estimate the population average treatment effect. But, again, usually
only $Y_i^B$ or $Y_i^A$ is observable, not both, so this simple estimator is impossible in practice.
For now, we focus on understanding treatment effects themselves, without worrying about
estimation yet.

**Table 4.1: Potential outcomes: charter school example.**

<table>
<thead>
<tr>
<th>Student type</th>
<th>$Y^A$</th>
<th>$Y^B$</th>
<th>$Y^B - Y^A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.1 shows a numerical illustration of the charter school example. Recall that $Y^A$
and $Y^B$ are binary, equal to 1 if the student graduates from college in the corresponding
universe ($A$ or $B$) and 0 if not. Thus, there are four types of student since there are four
possible values of $(Y^A, Y^B)$: $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. The are types 1, 2, 3, and 4,
respectively, in Table 4.1. The corresponding treatment effects $Y^B - Y^A$ for each student type are respectively $0 - 0 = 0$, $1 - 0 = 1$, $0 - 1 = -1$, and $1 - 1 = 0$.

Table 4.2: Potential outcomes: right-to-work example.

<table>
<thead>
<tr>
<th>Worker type</th>
<th>$Y^A$ ($/yr$)</th>
<th>$Y^B$ ($/yr$)</th>
<th>$Y^B - Y^A$ ($/yr$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40,000</td>
<td>41,000</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>40,000</td>
<td>38,000</td>
<td>-2000</td>
</tr>
<tr>
<td>3</td>
<td>50,000</td>
<td>51,000</td>
<td>1000</td>
</tr>
<tr>
<td>4</td>
<td>50,000</td>
<td>47,000</td>
<td>-3000</td>
</tr>
</tbody>
</table>

Table 4.2 shows a simplified numerical illustration of the right-to-work example. The outcome of interest is annual income; units are dollars per year. Instead of different types of students like in the charter school example, now there are different types of workers. As in the school example, each “type” of individual corresponds to their pair of potential outcome values, $(Y^A, Y^B)$, which in turn determines their treatment effect $Y^B - Y^A$. To keep it simple, Table 4.2 has only four worker types; more realistically, there are thousands or millions (or more) of worker types. In Table 4.2, the first two worker types both have $40,000/yr annual income without right-to-work, but they have different incomes with right-to-work: the law helps type 1 workers but hurts type 2 workers. The next two worker types both have $50,000/yr income without right-to-work, but again one is helped by the law and the other is hurt.

Assumption: SUTVA

The potential outcomes definition of causality relies on a critical assumption called the stable unit treatment value assumption (SUTVA). SUTVA is formally stated in A6.1 below. Part of SUTVA is that every treated individual receives the same treatment. This seems true in the charter school and right-to-work examples, but there are cases where it fails or at least requires some thought. For example, maybe a program to help teen parents includes one-on-one mentoring; if different individuals have different mentors, are they receiving the same treatment? The second, more critical part of SUTVA is called the “no interference” assumption. This assumes that one person’s treatment (or non-treatment) doesn’t affect the potential outcomes of another 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.

SUTVA can be violated in many ways. One violation is from spillover effects. For example, if the treatment gives treated individuals information (about financial planning, or social services, or risk probabilities, etc.), then untreated individuals may get the same
4.3. CAUSALITY

information if they are friends with a treated individual, and this may benefit them. That is, the benefit of the treatment “spills over” into untreated individuals.

SUTVA may also be violated by general equilibrium effects. For example, maybe the treatment is a new agricultural technology to increase production of cacao. If only one farmer gets this treatment (technology), then she will benefit from the increased production, selling more cacao at the global price. But if all farmers in the world get the technology, then the global cacao supply curve shifts to the right, lowering prices. The number of farmers using the technology affects the equilibrium of the global cacao market, which affects the price and thus affects each individual farmer. Thus, both potential outcomes for all farmers are affected by the treatment assignment to all farmers. Other general equilibrium effects could come through other markets: 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, too.

Note that SUTVA may be violated even with a perfectly randomized and controlled experiment. SUTVA is not about how treatments are actually assigned, but rather different possible assignments and outcomes. Conversely, even if SUTVA is satisfied, there could be problems if treatment is not randomized, as seen later in Section 4.3.5.

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.

4.3.4 Average Treatment Effect

Although the full joint distribution of (YA, YB) contains the most information about causality, usually only certain features of this distribution are studied. This is similar to Sections 2.3 and 4.1 where different features of the full distributions were considered for descriptive purposes. Indeed, as in Section 4.1, here we focus on differences in mean or percentiles between YB and YA. However, the objects of interest from Section 4.1 now have a causal interpretation given our potential outcomes interpretation of YA and YB. This causal interpretation does not come for free, though, as seen in subsequent sections.

Average Treatment Effects

Table 4.3 extends the charter school example of Table 4.1. The four student types now have different population probabilities (or, population proportions) shown in the table. In particular, it is less likely that a student graduates college in universe A after attending public school, but not in universe B after attending the charter school. Such students are only 1/10 of the population, i.e., P(YA = 1, YB = 0) = 0.1. Given the probabilities,
Table 4.3: Potential outcomes: charter school example with probabilities.

<table>
<thead>
<tr>
<th>Student type</th>
<th>Probability</th>
<th>( Y^A )</th>
<th>( Y^B )</th>
<th>( Y^B - Y^A )</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>Mean</td>
<td>0.4</td>
<td>0.6</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

The mean untreated outcome \( E(Y^A) \), mean treated outcome \( E(Y^B) \), and mean treatment effect \( E(Y^B - Y^A) \) are computed:

\[
E(Y^A) = (0.3)(0) + (0.3)(0) + (0.1)(1) + (0.3)(1) = 0.4, \quad (4.4) \\
E(Y^B) = (0.3)(0) + (0.3)(1) + (0.1)(0) + (0.3)(1) = 0.6, \quad (4.5) \\
E(Y^B - Y^A) = (0.3)(0) + (0.3)(1) + (0.1)(-1) + (0.3)(0) = 0.2. \quad (4.6)
\]

The linearity of the expectation operator can be verified using the above numbers:

\[
E(Y^B - Y^A) = 0.2 = 0.6 - 0.4 = E(Y^B) - E(Y^A). \quad (4.7)
\]

Table 4.4: Potential outcomes: right-to-work example with probabilities.

<table>
<thead>
<tr>
<th>Worker type</th>
<th>Probability</th>
<th>( Y^A ) ($/yr)</th>
<th>( Y^B ) ($/yr)</th>
<th>( Y^B - Y^A ) ($/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>Mean</td>
<td></td>
<td>43,000</td>
<td>43,000</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.4 extends the right-to-work example of Table 4.2. The four worker types now have different population probabilities shown. In particular, the lower untreated income value is more likely than the higher one, and in either case, the law is more likely to increase than decrease income. Given the probabilities, the mean untreated outcome \( E(Y^A) \), mean treated outcome \( E(Y^B) \), and average treatment effect \( E(Y^B - Y^A) \) are, in
dollars per year,

\[
E(Y^A) = (0.5)(40,000) + (0.2)(40,000) + (0.2)(50,000) + (0.1)(50,000) = 43,000, \quad (4.8)
\]

\[
E(Y^B) = (0.5)(41,000) + (0.2)(38,000) + (0.2)(51,000) + (0.1)(47,000) = 43,000, \quad (4.9)
\]

\[
E(Y^B - Y^A) = (0.5)(1000) + (0.2)(-2000) + (0.2)(1000) + (0.1)(-3000) = 0. \quad (4.10)
\]

Again, the average of the treatment effects equals the treatment effect on the average:

\[
E(Y^B - Y^A) = $0/yr = $43,000/yr - $43,000/yr = E(Y^B) - E(Y^A). \quad (4.11)
\]

As computed in the above examples, the **average treatment effect** (ATE) is \(E(Y^B - Y^A)\). “Average” refers to the population average, i.e., the mean (expectation), while “treatment effect” refers to \(Y^B - Y^A\). 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 try to use ATE to emphasize this concept is from the potential outcomes framework (not a structural model).

The ATE has another interpretation. Equation (4.7) shows \(E(Y^B - Y^A) = E(Y^B) - E(Y^A)\), where \(E(Y^B)\) is the mean treated potential outcome and \(E(Y^A)\) is the mean untreated potential outcome. Thus, \(E(Y^B) - E(Y^A)\) is the mean difference between the treated and untreated potential outcome distributions. This is similar to the mean difference from Section 4.1, but with potential outcomes instead of observed outcomes. Thus, this second ATE interpretation could be rephrased as the treatment effect on the mean outcome: treatment causes the mean outcome to change from \(E(Y^A)\) to \(E(Y^B)\).

**Limitation of ATE**

Figure 4.2 suggests why the ATE alone is not sufficient to make policy decisions. The figure shows the PDFs of three hourly wage distributions whose means are all identical. Picking any two distributions to be the potential outcomes \(Y^A\) and \(Y^B\), \(E(Y^A) = E(Y^B)\), so the ATE is \(E(Y^B) - E(Y^A) = $0/hr\). However, the distributions are clearly different. We may disagree about which is “better” or “worse,” but we can agree the differences are important. Sometimes this idea is easiest to remember in joke form; as retold by Hansen (2018, p. 25):

> 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.”

It is possible to examine effects on percentiles (“quantile treatment effects”), but they are beyond our scope.
Reminder

As a reminder, there is a difference between causal and statistical objects of interest. The ATE and related objects have causal meaning. The descriptive mean difference and related objects from Section 4.1 have statistical meaning. Since \( Y^A \) and \( Y^B \) are not both observable for all individuals, it is impossible to learn about the ATE directly, but the mean difference can be learned from data. Section 4.3.5 considers conditions under which the two objects are in fact equal, which allows us to indirectly learn about the ATE through the mean difference.

Discussion Question 4.4 (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 \( Y^A \) and \( Y^B \), and causal effect \( Y^B - Y^A \), which must be non-zero. Then compute the ATE to verify it’s zero.

4.3.5 Identification: ATE

The ATE is defined in terms of partially unobservable variables, \( Y^A \) and \( Y^B \). For each individual, only \( Y^A \) or \( Y^B \) is observed, never both. This means we cannot learn about the ATE directly. The ATE must somehow be written in terms of observable variables for us to learn about it.

Identification is a general concept that is central to econometrics and appears throughout this text. In general, the object of interest (like ATE) is something we care about
economically, but it often involves variables that are not always observable (like $Y^A$ and $Y^B$). The object of interest is identified if it can be shown equal to a statistical object (like a mean difference) that can be written in terms of observable variables (like $Y_T$ and $Y_C$). Usually the object of interest is not identified unless certain conditions are met. Such conditions are called identifying assumptions.

**Setup and Identification Question**

For each individual, a single $Y$ is observed. If the individual was actually treated (in our universe), then $Y = Y^B$ is observed. If the individual was actually untreated (in our universe), then $Y = Y^A$ is observed.

The treatment status is also observed for each individual. For now, the treatment and control groups are considered different populations, represented by random variables $Y_T$ and $Y_C$. In this view, $Y_T$ and $Y_C$ are always observable for individuals in the respective populations. For individuals in the untreated “control” population ($C$), $Y_C$ is always observable. For individuals in the treated population ($T$), $Y_T$ is always observable. For example, in the charter school example, $Y_C$ is the college graduation outcome for students who in reality did not attend the charter school. Similarly, $Y_T$ is the graduation outcome for students who did attend the charter school.

For the ATE, the question of identification is whether the ATE equals the mean difference between the actually treated and actually untreated populations. Recall from (4.7) that the ATE can be written equivalently as $E(Y^B - Y^A)$ or $E(Y^B) - E(Y^A)$. In math, the identification question is whether or not

$$E(Y^B) - E(Y^A) = E(Y_T) - E(Y_C).$$

We know how to learn about $E(Y_T)$ and $E(Y_C)$ from data, as in Chapter 3. If (4.12) holds, then this is equivalent to learning about the ATE.

How would (4.12) help us learn the ATE in the charter school and right-to-work examples? For the charter school example, the ATE could be estimated by simply comparing the graduation outcomes for students who attended the charter school and students who did not. For the right-to-work example, the ATE could be estimated by comparing the average income of workers in right-to-work states with average incomes in other states. Of course, if (4.12) is false, then this comparison does not help us learn about the ATE.

**Randomization**

Randomized experiments are often used to estimate treatment effects. In a randomized experiment, also called a randomized controlled trial (RCT), the experimenter can control who is treated and who is not. More formally, for each individual (or firm, school, etc.), the experimenter gets to decide whether to observe $Y^A$ or $Y^B$. Specifically, this is decided randomly. Although there are actually many different ways to “randomly” assign
CHAPTER 4. ONE VARIABLE, TWO POPULATIONS

treatment, we’ll imagine a 50% probability of observing $Y^A$ or $Y^B$ for each individual. At the end of the experiment, the treatment group outcomes and control group outcomes are compared, and differences are interpreted as treatment effects.

Why does randomization allow the mean difference to be interpreted as the average treatment effect? In Sections 4.3.3 and 4.3.4, the value of $(Y^A, Y^B)$ was called the individual’s “type.” For example, there was the type of student who always graduated from college regardless of attending public or charter school, represented by $(Y^A, Y^B) = (1, 1)$, or the type who never graduated either way, $(Y^A, Y^B) = (0, 0)$. Equation (6.29) can be interpreted as the treatment being assigned without regard to the individual’s type. The word “assigned” should be interpreted broadly; e.g., it can include weather conditions or genetic characteristics that are not “assigned” by the researcher (or any human).

Another way to see randomization is as follows. First consider the untreated potential outcome $Y^A$. If we have a random sample of individuals, then the individuals have a random sample of $Y^A$. If nobody were treated, so everybody’s $Y^A$ is observed, then we are back in the simpler world of Chapter 3. Similarly, if everybody were treated, then everybody’s $Y^B$ is observed, and we have a random sample of $Y^B$ from the population. So, if we take one random sample of individuals and treat none of them, and we take another random sample and treat everybody, then we can estimate both $E(Y^A)$ and $E(Y^B)$ and thus the difference, the ATE $E(Y^B) - E(Y^A)$. It is basically the same to take one random sample and then randomly split it into treated and untreated groups: there are still random samples of $Y^A$ and $Y^B$. More formal discussion is in Section 6.4.

Reasons for Identification Failure

When does (4.12) fail? It (generally) fails when SUTVA fails or treatment is not random. SUTVA violations were discussed in Section 4.3.3. In the right-to-work example, treatment is probably not random: the factors leading to a state legislature’s decision to pass a right-to-work law probably affect income, too, and perhaps the current income distribution affects politicians’ decisions to support right-to-work or not. Whatever the specific mechanism, it seems unlikely that right-to-work laws are “as good as random.” For example, just looking at a map, it is notable that 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).

Even with randomized treatment assignment, randomization can fail (what?!). For example, you may tell somebody to show up at a job training program, but they never do; this is called non-compliance. This is a type of self-selection, where individuals decide which group to be in (non-randomly). Maybe lots of people fail to show up. Such people may also fail to show up to their jobs regularly, and thus have lower incomes. They were assigned to the treatment group, but they end up untreated. This means a
lot of low-income individuals who “should have” been in the treatment group are now in the control group, which makes the control group’s average income a lot lower than it “should be” (under randomization), and raises the treatment group’s average income. Even if the training program has zero treatment effect and \( E(Y_B) - E(Y_A) = 0 \), this non-compliance makes it look like the treatment has a big effect: \( E(Y_T) - E(Y_C) \) is very big because the failed-to-show folks are all in \( Y_C \) instead of \( Y_T \).

One way to avoid this incorrect conclusion is to compare the group that was assigned to the treatment with the unassigned group, regardless of whether or not people were actually treated. This is called the intention-to-treat effect since it measured the change in \( Y \) corresponding to the intention to treat (i.e., assignment to treatment group). Sometimes this is more relevant for policy, if the actual policy would not force people to be treated. It can also be used to estimate something like the ATE using an instrumental variables (IV) approach, but this is beyond our scope.

Attrition can sometimes be a problem, too. The term attrition refers to people (or firms, etc.) dropping out of the study (and out of your data) after it starts. So, perhaps people comply and do the job training, but then they move away, so their income cannot be observed. If people drop out of the data randomly, it’s fine, but non-random attrition is problematic. For example, maybe the training program is so good that people then get higher-paying jobs in other states where you can’t collect data on them. Instead, 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 your data because you don’t see all the highest-earning treated individuals.

**Discussion Question 4.5** (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.

### 4.3.6 Estimation and Inference

There is nothing new for estimation and inference. It is identical to Sections 4.1.2 and 4.1.3. The point of treatment effect identification is not to propose a new statistical object, but rather to imbue an existing descriptive statistical object with causal meaning. For example, (4.12) links the ATE to a descriptive mean difference. Estimation and inference on the mean difference proceed as usual. That is, estimation and inference for a mean difference do not depend on how we interpret the mean difference.

However, quantifying uncertainty only covers sampling uncertainty, not uncertainty about identification. For example, a 95% CI for a mean difference contains the mean
difference with 95% probability, but this does not capture uncertainty about whether the mean difference equals the ATE. So, additional caution is advised.
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 commands `trt <- jtrain2[jtrain2$train==1 , ]` and `ctl <- jtrain2[jtrain2$train==0 , ]`

d. Estimate the mean 1978 earnings (in thousands of dollars) for the treatment group minus that of the control group, using R command `mean(trt$re78) - mean(ctl$re78)`, as well as a 95% CI for the difference with R command `t.test(x=trt$re78, y=ctl$re78)` or Stata command `ttest re78 , by(train) unequal` (which 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$married==1 , ]` and `ctl.mar1 <- ctl[ctl$married==1 , ]`

f. Compute the mean difference estimate and 95% confidence interval for the 1978 earnings outcome variable, comparing treated and untreated married individuals, with R commands `mean(trt.mar1$re78) - mean(ctl.mar1$re78)` and `t.test(x=trt.mar1$re78, y=ctl.mar1$re78)` or Stata command `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) with R command `jtrain2$
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 with R command `mean(audit$w) - mean(audit$b)` or Stata command `ttest w == b` (which also computes a 95% confidence interval)

d. Compute the sample mean of all the pairs’ white-minus-black difference with R command `mean(audit$w - audit$b)` or Stata commands `generate wminusb = w-b` followed by `ttest wminusb==0` (which also computes a 95% confidence interval; see the row labeled `diff` for both). 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.

e. R only (since Stata already reported this in the row labeled `diff`): compute a 95% confidence interval for the population mean difference with command `t.test(x=audit$w, y=audit$b, paired=TRUE)` or `t.test(x=audit$w-audit$b)`
Chapter 5

Midterm Exam #1

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). Even most advanced econometric techniques depend on regression. Whether you need to properly interpret others’ research or do your own in your future work, a deep understanding of regression as description, prediction, and causal inference is valuable.

Part II extends the concepts and methods of Part I to the more complex regression setting. In the population, the concepts of description, prediction, and causality from Part I are extended to more complex regression models. In the data, estimation and inference methods also extend those of Part I.

There are different ways to make a regression model more flexible. This can make a model more realistic. We will be careful to understand how interpretation changes with different types of flexibility, and to understand how there can be too much flexibility. This flexibility tradeoff leads to nonparametric and machine learning methods, which are discussed more practically than theoretically.
Chapter 6

Comparing Two Distributions by Regression

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

- James et al. (2013, §3.1)
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 **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.

### 6.1 Logic

Some basic logic is useful for understanding certain parts of econometrics. First, logic is useful for understanding the relationship among different conditions. Often these conditions are assumptions used in various theorems. Second, logic is useful for understanding what a theorem actually claims. Third, logic is helpful for interpreting 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.
6.1.1 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)
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 (or just sufficient) for $B$
9. $B$ is a necessary condition (or just 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. 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. $A$ is necessary and sufficient for $B$ (or equivalently, $B$ is necessary and sufficient for $A$)
4. \( A \) is equivalent to \( B \)

5. It is impossible for \( A \) to be false when \( B \) is true, and impossible for \( A \) to be true when \( B \) is false.

6. 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 \( A \)”:

- \( \neg A \implies \neg B \) 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 \( A \implies B \).

Interestingly, 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: the inverse is the contrapositive of the converse.)

For example, let \( A \) be “\( X \leq 0 \)” and let \( B \) be “\( X \leq 10 \)”.

Here \( A \implies B \): 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 Assumptions

To compare assumptions, the terms “stronger” and “weaker” are most commonly used. Instead of assumptions \( A \) and conclusion \( B \), let \( A \) and \( B \) denote different assumptions. For example, let \( A \) be \( \text{E}(Y^4) < \infty \), and let \( B \) be \( \text{E}(Y^2) < \infty \), as in [A7.3] and [A7.4]. Any random variable \( Y \) with finite \( \text{E}(Y^4) \) also has finite \( \text{E}(Y^2) \), but some have finite \( \text{E}(Y^2) \) and infinite \( \text{E}(Y^4) \). That is, \( A \implies B \). Thus, people say “\( \text{E}(Y^4) < \infty \) is a stronger assumption than \( \text{E}(Y^2) < \infty \),” or equivalently, “\( \text{E}(Y^2) < \infty \) is weaker than \( \text{E}(Y^4) \).”

As another example (that will make sense later), consider the linear projection (LP) and linear conditional expectation function (CEF) models from Sections 6.3 and 7.3. Consider the linear model \( Y = \beta_0 + \beta_1 X + U \). Let assumption \( A \) be \( \text{E}(U \mid X) = 0 \), and let \( B \) be \( \text{E}(U) = 0 \) and \( \text{Cov}(X,U) = 0 \); i.e., \( A \) says \( U \) is a CEF error, whereas \( B \) says \( U \) is a linear projection error. Here, \( A \implies B \), so \( A \) is a stronger assumption than \( B \), and
B is weaker than A. Seen another way, the linear projection model is more general than the linear CEF model: if the CEF is $\beta_0 + \beta_1 x$, then so is the linear projection, but if the linear projection is $\beta_0 + \beta_1 X$, it is still possible to have a nonlinear CEF.

All else equal, it is nicer to have results based on weaker assumptions. In the previous example, the weaker assumption $\mathbb{E}(Y^2) < \infty$ is true for more distributions of $Y$ than is the stronger assumption $\mathbb{E}(Y^4) < \infty$. If $Y$ has a distribution with $\mathbb{E}(Y^2) < \infty$ but $\mathbb{E}(Y^4) = \infty$, then only results based on the weaker assumption can be applied. For this reason, results with weaker assumptions are often called “more general” than results based on stronger assumptions: they apply to more situations.

### 6.1.3 Theorems

Theorems all have the same logical structure: if assumption $A$ is true, then result (conclusion) $B$ is true. Sometimes $A$ and $B$ have multiple parts, like the three assumptions in Theorem 7.1, but the logical structure of a theorem is always the same. The theorem claims that if we can verify that $A$ is true, then we know that $B$ is also true. But what if we don’t know about $A$, or we think it’s 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. We could be somewhere inside $B$ (where $B$ is true) but outside $A$ (where $A$ is false); or we could be outside both, where both are false. The theorem is not equivalent to, “If $A$ is false, then $B$ is false” (the “inverse”). However, it is equivalent to the contrapositive: “If $B$ is false, then $A$ is false.” Again, this is probably seen most easily in the picture.

**Discussion Question 6.1** (median theorem logic). Consider the statement, “If sampling is iid, then the sample median consistently estimates the population median.” 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? Hint: draw a picture.

**Discussion Question 6.2** (mean theorem logic). Consider the statement, “If sampling is iid and the population mean is well-defined, then the sample mean consistently estimates the population mean.” 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? Hint: draw a picture with $A_1$ (iid), $A_2$ (well-defined), and $B$ (consistency).

**Discussion Question 6.3** (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

Before getting to regression, some simpler material may provide intuition. Hopefully it is just review of your prior statistics class. 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 $\mu_Y \equiv \mathbb{E}(Y)$ 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 2 and Chapter 6. First, the mean can be written directly:

$$\mathbb{E}(Y) = \mu_Y. \quad (6.1)$$

Second, the error form of this model is

$$Y = \mu_Y + U, \quad \mathbb{E}(U) = 0, \quad (6.2)$$

where $U$ is the error term.

These two models are equivalent. Taking the expectation of both sides of (6.2), using the linearity property of expectation,

$$\mathbb{E}(Y) = \mathbb{E}(\mu_Y + U) = \mathbb{E}(\mu_Y) + \mathbb{E}(U) = \mu_Y + 0 = \mu_Y. \quad (6.3)$$

In this sense, it is not an assumption that $\mathbb{E}(U) = 0$, but rather a definition of the mean model error: as long as $\mathbb{E}(Y)$ exists, there exists a $U$ satisfying (6.2). 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. But to understand conditional distributions, it helps to understand joint distributions and marginal distributions first.

Consider the joint distribution of two binary variables. 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 (sum)</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>
<tr>
<td>Marginal (sum)</td>
<td>0.30</td>
<td>0.70</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 6.1 shows both joint and marginal probabilities. Let the joint probability values be $p_{xy} \equiv P(X = x, Y = y)$, 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 probability of an individual being married is 0.8 (80%). This is a complete description of the distribution of $X$. Similarly, by itself, $Y$ is a random variable with $P(Y = 0) = 0.30$ and $P(Y = 1) = 0.70$: the population probability of being employed is 0.7 (70%).

Discussion Question 6.4 (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 $P(X = Y) = 0.8 > 0.2 = P(X \neq Y)$. Does this suggest that $X$ has a causal effect on $Y$? (Hint: think about some concrete examples of $X$ and $Y$, like marital status and employment, or whether it’s raining and whether your commute time is longer than average; and then think about switching which variable you label $X$ and which is $Y$: does the joint distribution and/or causal effect change?)

6.2.3 Conditional Distributions

Table 6.1 can also be used to derive conditional distributions. 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. For example, consider the probability of employment ($Y = 1$) conditional on being married ($X = 1$): $P(Y = 1 \mid X = 1)$. This mathematical expression could be read aloud 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.

For intuition, 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 \(60/100 = 0.60 = 60\%\). Parallel to Table 6.1, the number of people with different values is shown in Table 6.2.

Table 6.2: Counts of individuals by marital status \((X)\) and employment status \((Y)\).

<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.

The conditional probabilities \(P(X = 1 \mid Y = 0)\) and \(P(X = 1 \mid Y = 1)\) could also be computed. 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 \(Y\), to avoid confusion, we would re-label marital status as \(Y\) and employment status as \(X\), and then look at \(Y\) conditional on \(X\).

There is a general formula for conditional probabilities. Above, the conditional proportion was the joint divided by the marginal. The same can be done with probabilities:

\[
P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}. \quad (6.4)
\]
6.2. PRELIMINARIES

This requires \( P(X = x) \neq 0 \) to be well defined, so a different formula is required for continuous variables with \( P(X = x) = 0 \). Rearranging (6.4), the joint probability can be written as the product of the conditional and marginal probabilities:

\[
P(X = x, Y = y) = P(Y = y \mid X = x) P(X = x).
\] (6.5)

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

\[
E(Y \mid X = x).
\] (6.6)

To read (6.6) aloud, you would say, “the conditional mean of \( Y \) given \( X = x \),” or “the mean of \( Y \) conditional on \( X = x \),” or related variants.

We can compute a conditional mean for the example in Table 6.1. 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 \( E(Y \mid X = 1) \).

In our running example, 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,

\[
E(Y) = (0) P(Y = 0) + (1) P(Y = 1) = (0)(0.3) + (1)(0.7) = 0 + 0.7
= 0.7,
\]

\[
E(Y \mid X = 1) = (0) P(Y = 0 \mid X = 1) + (1) P(Y = 1 \mid X = 1)
= (0)(0.25) + (1)(0.75) = 0 + 0.75
= 0.75.
\]

Since \( Y \) is binary (with values 0 and 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 \( X \), too. Regardless of how many possible values of \( X \) there are, the conditional probabilities of the \( Y \) values must still sum to one. For example, imagine \( Y \) is hours worked per week, which is either 0, 20, or 40, and \( X \) is years of education, which is an integer in \( 0, 1, 2, \ldots, 25 \). Conditional on any value \( x \), like \( x = 12 \), the conditional probabilities satisfy \( P(Y = 0 \mid X = x) + P(Y = 20 \mid X = x) + P(Y = 40 \mid X = x) = 1 \). The
CHAPTER 6. COMPARING TWO DISTRIBUTIONS BY REGRESSION

conditional mean is

\[ E(Y \mid X = x) = \sum_{j \in \{0,20,40\}} (j) P(Y = j \mid X = x) \]

\[ = (0) P(Y = 0 \mid X = x) + (20) P(Y = 20 \mid X = x) + (40) P(Y = 40 \mid X = x). \]

6.2.5 Comparison of Joint, Marginal, and Conditional Distributions

Which has the most information about the population: the joint, marginal, or conditional distributions? The marginal probabilities can be computed from the joint probabilities, by summing the values in a row or column. The conditional probabilities can also be computed from the joint probabilities, as seen above. Thus, the joint distribution has at least as much information as the marginal or conditional distributions. But, it would still be possible that they all have the same amount of information.

Can the joint probabilities be calculated from the marginal probabilities, or from the conditional probabilities? No. First, consider the marginals in Table 6.1. It is possible to change the joint probabilities without altering the marginal probabilities, implying that the marginals do not uniquely determine the joint distribution. For example, we could move 0.1 probability “up” in the \( Y = 0 \) column and the same amount “down” in the \( Y = 1 \) column, so \( p_{00} = 0.2, p_{01} = 0, p_{10} = 0.1, \) and \( p_{11} = 0.7. \) (This is equivalent to moving 0.1 probability “left” in the \( X = 0 \) row, and the same amount “right” in the \( X = 1 \) row.) Since we are only moving probability around within each column, the column totals (sums) do not change. Since the same amount is moved up and down, the row sums also do not change. Thus, all four marginals stay the same. Second, consider the conditional probabilities. These could also be kept unchanged while certain joint probabilities are changed.

As seen in (6.5), the joint distributions contain the same amount of information as the marginal and conditional distributions combined. That is, if both the marginal and conditional distributions are known, then each joint probability can be calculated from the formula. Thus, knowing the marginal and conditional distributions is equivalent to knowing the joint distribution (i.e., has the same amount of information), but either one alone is less information.

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. Intuitively, if \( X \) is unrelated to \( Y \), then knowing the value of \( X \) has no information about the value of \( Y \). Mathematically, the probability distribution of \( Y \) without knowledge of \( X \) is the marginal distribution of \( Y \), whereas the conditional distribution of \( Y \) given \( X \) shows how knowledge of \( X \) changes the distribution of \( Y \). If \( Y \) is discrete, then independence implies the marginal PMF equals the conditional PMF: for any possible \( y \) and \( x \) values, \( Y \perp \!
\!\!\perp X \implies P(Y = y) = P(Y = y \mid X = x) \). This is also true for the CDF, for discrete or continuous \( Y \): \( Y \perp \!
\!\!\perp X \implies P(Y \leq y) = P(Y \leq y \mid X = x) \).

Consequently, independence implies equality of marginal and conditional means. That is, \( Y \perp \!
\!\!\perp X \implies E(Y) = E(Y \mid X = x) \), for any possible \( x \) value. The term for this latter property is **mean independence**.

Independence implies lots of other properties, too, like \( \text{Cov}(X, Y) = \text{Corr}(X, Y) = 0 \).

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 describes how income changes with education, the 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.

Since there are different types of relationships, there are different types of models. Section 6.3 provides a model describing a statistical relationship between two variables, whereas Sections 6.4 and 6.5 provide different models of causal relationships. In some cases, the statistical and causal models may coincide, but in general they differ.

This section combines Sections 6.2.1 and 6.2.3 to get a conditional mean regression model.
6.3.1 Conditional Expectation Function

Using (6.6), let \( m(\cdot) \) be the **conditional expectation function** (CEF) of \( Y \) given \( X \):

\[
m(x) \equiv \mathbb{E}(Y \mid X = x).
\]  

(6.7)

If \( X \) is binary, then there are two conditional means of interest:

\[
m(0) = \mathbb{E}(Y \mid X = 0), \quad m(1) = \mathbb{E}(Y \mid X = 1).
\]  

(6.8)

These could be studied directly (similar to prior chapters), but they can also be captured in a CEF model. Note \( m(x) \) is a number (like \( m(x) = 7 \)), whereas \( m(X) \) is a random variable. For example, imagine the CEF has values \( m(0) = 3 \) and \( m(1) = 6 \), i.e., \( m(x) = 3 \) when \( x = 0 \), and \( m(x) = 6 \) when \( x = 1 \). Imagine \( X \) is a random variable with \( P(X = 0) = P(X = 1) = 1/2 \). Then, \( m(X) \) is a random variable with

\[
P(m(X) = 3) = P(X = 0) = 1/2, \quad P(m(X) = 6) = P(X = 1) = 1/2.
\]  

(6.9)

6.3.2 CEF Error Term

Consider the random variable

\[
V \equiv Y - m(X).
\]  

(6.10)

This \( V \) is the **CEF error term**, i.e., the difference between an individual’s actual outcome \( Y \) and the CEF evaluated at her \( X \) value, \( m(X) \). Since \( Y \) and \( X \) are random variables, so is \( V \); e.g., \( P(V = 0) = P(Y - m(X) = 0) \). For example, if \( X = 1 \) means having a college degree (and \( X = 0 \) otherwise), and \( Y \) is income, then \( m(0) \) is the average 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, then your \( Y \) is high above \( m(1) \), so your \( V \) is very large and positive. If you didn’t go to college and make exactly the average income for that group, your \( V \) equals \( m(0) \), so your \( V = 0 \). 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 \( V = Y - m(1) = 0 - m(1) \) is negative (in a big way).

The key property of the CEF error is that it has conditional mean zero. For any \( X = x \),

\[
\mathbb{E}(V \mid X = x) = \mathbb{E}(Y - m(X) \mid X = x) = \mathbb{E}(Y \mid X = x) - \mathbb{E}(m(X) \mid X = x)
\]

\[= m(x) - m(x) = 0.\]

Equivalently,

\[
\mathbb{E}(V \mid X) = 0.
\]  

(6.11)

That is, \( \mathbb{E}(V \mid X) \) is a random variable depending on \( X \), but it equals zero for every possible value of \( X \); or, just imagine “\( \mathbb{E}(V \mid X = x) = 0 \) for all \( x \)” every time you see
“E(V | X) = 0.” This conditional mean zero property is not an assumption: it is true by construction for any CEF error defined as in (6.10). By analogy, if we construct an equilateral triangle, then it has the property of three equal side lengths by construction; such a property does not have to be assumed additionally.

### 6.3.3 CEF Model in Error Form

Given (6.10), we could instead write

\[ Y = m(X) + V, \quad \text{E}(V | X) = 0. \]  

(6.12)

This is a CEF model. The statement that \( \text{E}(V | X) = 0 \) is equivalent to saying that \( m(x) = \text{E}(Y | X = x) \), 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 (6.12) can apply to non-binary \( X \), too, although such models are for later chapters.

### 6.3.4 Linear CEF

The model in (6.12) can be rewritten yet again when \( X \) is binary. Specifically, it is equivalent to write

\[ Y = m(0) \mathbb{1}\{X = 0\} + m(1) \mathbb{1}\{X = 1\} + V 
  = m(0)(1 - X) + m(1)X + V 
  = m(0) + [m(1) - m(0)]X + V. \]  

(6.13)

To double-check: when \( X = 0 \), then \( m(1)X = 0 \), so \( Y = m(0) + V \), as in the original (6.12). 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 (6.12). Thus, (6.12) and (6.13) are indeed equivalent.

The CEF model in (6.13) can be rewritten yet again. Following conventional notation, let \( \beta_0 \equiv m(0) \) and \( \beta_1 \equiv m(1) - m(0) \). That is, \( \beta_0 \) and \( \beta_1 \) are respectively the intercept and slope in the linear model. Plugging these definitions into (6.13),

\[ Y = \beta_0 + \beta_1X + V, \quad \text{E}(V | X) = 0. \]  

(6.14)

Greek letters like \( \beta \) are commonly used to denote unknown parameters in the population model. In the frequentist framework, these are seen as unknown but fixed (non-random) values, whereas \( Y \), \( X \), and \( V \) are random variables.

Since \( X \) is binary, no assumptions were required to write (6.14). However, when \( X \) has more than two possible values, it is more complicated, as discussed in Chapter 7. For now, with binary \( X \), as long as \( Y \) has a well-defined mean (which is almost always true for economic variables), the CEF model can always be written as in (6.14).
6.3.5 Interpretation, Description, and Prediction

Discussion Question 6.5 (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 (6.14), what are the units of measure for $\beta_0$ and $\beta_1$, respectively?

To interpret (6.14), 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 as the left-hand side, i.e., the same units as $Y$. Thus, each of the three right-hand side terms must have the same units as $Y$. First, this means $\beta_0$ has the same units as $Y$. Second, $\beta_1 X$ has the same units as $Y$. Thus, the units of $\beta_1$ are the units of $Y$ divided by the units of $X$. 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). Third, $V$ also has the same units as $Y$, although this is less useful than knowing the units of $\beta_0$ and $\beta_1$.

The CEF model in (6.14) has a descriptive interpretation. Specifically, the model parameters $\beta_0$ and $\beta_1$ have interpretations in terms of conditional means. As seen earlier, $\beta_0 = m(0) = E(Y \mid X = 0)$, the mean outcome among all individuals with $X = 0$. Also, the difference between the mean outcome for the $X = 1$ subpopulation and 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.”

The CEF model in (6.14) is also helpful for prediction. Section 2.4.5 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 $X$),

$$m(\cdot) = \arg\min_{g(\cdot)} E[(Y - g(X))^2].$$

In terms of the model parameters in (6.14), 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? 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 $m(12) = E(Y \mid X = 12)$ and $m(13) = E(Y \mid X = 13)$, but $\beta_0$ 
and $\beta_1$ in (6.14) 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)/(#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 (6.14) can now be written in terms 
of $m(12)$ and $m(13)$. Writing the CEF as $m(X) = \beta_0 + \beta_1X$ as in (6.14),

$$
m(13\text{yr}) = \beta_0 + (13\text{yr})\beta_1, \quad m(12\text{yr}) = \beta_0 + (12\text{yr})\beta_1,
$$

$$
m(13\text{yr}) - m(12\text{yr}) = (\beta_0 - \beta_0) + (13\text{yr} - 12\text{yr})\beta_1 = \beta_1,
$$

$$
\beta_1 = [m(13\text{yr}) - m(12\text{yr})]/(13\text{yr} - 12\text{yr}) = [m(13\text{yr}) - m(12\text{yr})]/\text{yr},
$$

$$
\beta_0 = m(12\text{yr}) - (12\text{yr})\beta_1 = m(12\text{yr}) - (12\text{yr})[m(13\text{yr}) - m(12\text{yr})].
$$

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).

Further, the meaning of $\beta_0$ is very different. 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 be very meaningful, unless $X = 0$ is a possible and interesting value 
in the population.

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. Writing the CEF as $m(X) = \beta_0 + \beta_1X$
as in (6.14),

\[ m(16 \text{ yr}) = \beta_0 + (16 \text{ yr})\beta_1, \quad m(12 \text{ yr}) = \beta_0 + (12 \text{ yr})\beta_1, \]

\[ m(16 \text{ yr}) - m(12 \text{ yr}) = (\beta_0 - \beta_0) + (16 \text{ yr} - 12 \text{ yr})\beta_1 = (4 \text{ yr})\beta_1, \]

\[ \beta_1 = \frac{m(16 \text{ yr}) - m(12 \text{ yr})}{4 \text{ yr}}, \quad \beta_0 = m(12 \text{ yr}) - (12 \text{ yr})\beta_1 = m(12 \text{ yr}) - 3[m(16 \text{ yr}) - m(12 \text{ yr})]. \]

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 also different than before, but still meaningful. It takes the mean income difference \( m(16 \text{ yr}) - m(12 \text{ yr}) \) and then divides by \( 4 \text{ yr} \). 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 \( (400 \text{ km})/(5 \text{ hr}) = 80 \text{ km/hr} \). 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.

Finally: is the CEF model in (6.14) helpful for causal inference? See Section 6.6.

### 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 4.3.3. Again, let \( Y^A \) and \( Y^B \) denote the untreated and treated potential outcomes, respectively.

Here, the two observable variables are the observed outcome \( Y \) and the treatment indicator \( X \). That is, \( X = 1 \) if an individual was “treated” and \( X = 0 \) if not; \( X \) is binary. 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

\[ Y = (Y^A)(1 - X) + (Y^B)(X). \] (6.17)

Plugging in \( X = 0 \) yields \( Y = Y^A \), whereas plugging in \( X = 1 \) yields \( Y = Y^B \). So, we observe \( Y = Y^A \) if the individual is untreated and \( Y = Y^B \) if treated.

Sometimes potential outcomes notation writes \( Y \) as a function of \( X \). Specifically, the treated potential outcome is written \( Y(1) \), and the untreated potential outcome \( Y(0) \). In the previous notation, \( Y(1) = Y^B \) and \( Y(0) = Y^A \). We observe \( X \) and \( Y(X) \).

We could also rearrange (6.17) to look more like a linear model in error form. First,

\[ Y = Y^A + X(Y^B - Y^A). \]
This is a “random coefficient” model with intercept $Y^A$ and slope $Y^B - Y^A$; the slope and intercept can differ by individual. Then, define $\beta_0 \equiv \text{E}(Y^A)$, $\beta_1 \equiv \text{E}(Y^B - Y^A)$ (i.e., the ATE), so

\[
Y = [Y^A + \beta_0 - \beta_0] + X[Y^B - Y^A + \beta_1 - \beta_1] \\
= \beta_0 + X \beta_1 + Y^A - \beta_0 + X[Y^B - Y^A - \beta_1].
\]  

(6.18)

Equation (6.18) has the same structure as (6.14), 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 $\text{E}(U | X) = 0$. 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

As an alternative to the potential outcomes model, 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 guess 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 Model

Consider the model

\[
Y = \beta_0 + \beta_1 X + U.
\]  

(6.19)

This has the same form as (6.14), so it seems natural to think it’s a CEF model. However, (6.19) says nothing about the error term $U$, whereas (6.14) said the error term had zero conditional mean, i.e., was a CEF error. (I used $V$ before 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.) Without knowing anything about the meaning or properties of $U$, and without knowing anything about the meaning of $\beta_0$ and $\beta_1$, (6.19) is ambiguous: it might be a CEF model, but it might not.

Instead of a CEF, another possibility is that (6.19) is a structural model. Then, $\beta_1$ is called a structural parameter (as is $\beta_0$). It usually has some economic meaning, like an elasticity or demand curve slope.
Similarly, instead of being a CEF error term, $U$ is called the structural error term. This $U$ can be interpreted as all other variables that help determine $Y$. We can think about $U$ economically, instead of just statistically. In our example where $Y$ is income and $X$ is an indicator of having a college degree, $U$ contains everything else that helps determine a person’s income: their occupation, their different skill levels (human capital), where they live (which city or state or country), etc. It’s possible that $U$ also has conditional mean zero, but that is not true by definition like for the CEF error. Conversely, the CEF error did not have this meaning of other determinants of $Y$; it was simply defined as $Y - m(X)$.

In the extremely simple model of (6.19), it can be difficult to distinguish a structural model from a non-structural causal model. In fact, the structural model in (6.19) looks identical to (6.18), which was derived from potential outcomes such that $\beta_1$ is the ATE, not a parameter with “economic” meaning. Still, the reduced form approach to identifying the ATE involves assumptions about potential outcomes, whereas the structural approach involves assumptions about $X$ and $U$. The distinction is much clearer in more complex models.

Superficially, it appears (6.19) claims the causal effect of $X$ on $Y$ is $\beta_1$ for everybody. For example, it seems it says that getting a college degree will increase anybody’s income by $\beta_1$ ($/\text{yr}$), regardless of that person’s occupation, skills, location, etc. This homogeneity (everyone being the same) would simplify policy analysis, but it is not realistic. Actually, the structural error $U$ can include heterogeneity in the effect of $X$. In that case, $\beta_1$ is something like an average causal effect of $X$. This is discussed more carefully in Section 6.6.

To reiterate the earlier warning once more: if you see a model that says $Y = \beta_0 + \beta_1 X + U$, 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 else 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 being interpreted in very different ways.

At face value, the model in (6.19) seems to assume everyone has the same structural effect of $X$ on $Y$: $\beta_1$. In reality, people probably respond differently to changes in their $X$. Such differences (or “heterogeneity”) ends up in the structural error term, $U$. It’s possible to more explicitly model such differences in a “random coefficient” structural model, but this is beyond our scope.

### 6.5.2 General Model

The average structural effect (ASE) can be defined in a more general structural model that is easier to generalize to more complex models. As defined below, the ASE is sometimes called the average causal effect ([Hansen](2018)), but I write “structural” instead
of “causal” to remind us that it comes from a structural model, not a causal (potential outcomes) model. Although defined differently, the ASE and ATE may be equivalent in some cases.

Consider the structural model

\[ Y = h(X, U), \]  

(6.20)

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 function.

The linear structural model

\[ Y = \beta_0 + \beta_1 X + U \]

is a special case of (6.20). For example, let \( \tilde{U} \) be the vector of unobservable variables from (6.20), and let \( U = g(\tilde{U}) \) be a scalar summary of the unobservables, where \( g(\cdot) \) is the summarizing function. If also \( h(\cdot) \) has the form

\[ h(a, b) = \beta_0 + \beta_1 a + g(b), \]

then plugging everything into (6.20) yields

\[ Y = h(X, \tilde{U}) = \beta_0 + \beta_1 X + g(\tilde{U}) = \beta_0 + \beta_1 X + U, \]

which looks exactly like the linear structural model \( Y = \beta_0 + \beta_1 X + U \) from before.

### 6.5.3 ASE and CEF

Given (6.20), the ASE can be derived. For a single individual, the structural effect on \( Y \) of changing \( X = 0 \) to \( X = 1 \) is

\[ S(U) \equiv h(1, U) - h(0, U). \]

(6.21)

According to the structural model, this describes how \( Y \) changes when \( X \) increases from 0 to 1, all else equal (ceteris paribus), i.e., holding the value of \( U \) fixed. Note the structural effect of \( X \) on \( Y \) may depend on the value of the unobservable \( U \).

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 average value. As usual, there is a tradeoff: the average is not as informative and helpful for policy, but it is easier to learn about. For example, we could simply assume \( S(U) \) is a constant, and then everybody’s structural effect would equal the average (so we could learn everybody’s structural effect), but this is almost always unrealistic.

The ASE takes a weighted average of the individual structural effects. The weights depend on the population distribution of \( U \). Mathematically, this weighted average is the population expected value:

\[ \text{ASE} = \mathbb{E}[S(U)] = \mathbb{E}[h(1, U) - h(0, U)] = \mathbb{E}[h(1, U)] - \mathbb{E}[h(0, U)]. \]

(6.22)

Given (6.20), the CEF can also be derived:

\[ \mathbb{E}(Y \mid X = 0) = \mathbb{E}(h(X, U) \mid X = 0) = \mathbb{E}(h(0, U) \mid X = 0), \]

\[ \mathbb{E}(Y \mid X = 1) = \mathbb{E}(h(X, U) \mid X = 1) = \mathbb{E}(h(1, U) \mid X = 1). \]
The mean difference between the $X = 1$ and $X = 0$ subpopulations is thus

$$E(Y | X = 1) - E(Y | X = 0) = E[h(1, U) | X = 1] - E[h(0, U) | X = 0]. \quad (6.23)$$

**Discussion Question 6.6** (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 $E(U | X = 0) = E(U | X = 1)$? Why/not? c) Do you think $b = d$, $b < d$, or $b > d$? Why? d) What would you guess are reasonable possible values of the parameters $a$, $b$, $c$, and $d$? Explain why.

### 6.6 Identification

This section focuses on identification of parameters with causal or structural meaning. In particular, the question is: when does the slope parameter $\beta_1$ in the CEF model also have a causal or structural interpretation? Recalling that $\beta_1$ in the CEF model is the difference in means between the $X = 1$ and $X = 0$ subpopulations, the identification question is: when is this mean difference not just descriptive, but also causal? These questions are investigated from multiple perspectives.

#### 6.6.1 Structural Model

Under certain conditions, the slope parameter in the structural model

$$Y = \beta_0 + \beta_1 X + U \quad (6.24)$$

is identified. Specifically, $\beta_1$ can be shown to equal $\gamma_1$, the slope of the CEF

$$E(Y | X = x) = \gamma_0 + \gamma_1 x. \quad (6.25)$$

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 $U$). If this is 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 (6.24). Most directly, if $E(U | X) = 0$, then (6.24) is a CEF model, so $\beta_0 = \gamma_0$ and $\beta_1 = \gamma_1$. A weaker
condition leads to $\beta_1 = \gamma_1$, but not $\beta_0 = \gamma_0$; usually we only care about $\beta_1$ (not $\beta_0$) anyway, so this is still useful. Specifically, if $E(U \mid X = 0) = E(U \mid X = 1)$, then

$$E(Y \mid X = 0) = E(\beta_0 + \beta_1 X + U \mid X = 0) = \beta_0 + (\beta_1)(0) + E(U \mid X = 0),$$

$$E(Y \mid X = 1) = E(\beta_0 + \beta_1 X + U \mid X = 1) = \beta_0 + (\beta_1)(1) + E(U \mid X = 1),$$

$$E(Y \mid X = 1) - E(Y \mid X = 0) = [\beta_0 + \beta_1 + E(U \mid X = 1)] - [\beta_0 + E(U \mid X = 0)]$$

$$= (\beta_0 + \beta_1) - \beta_0 + \left[ E(U \mid X = 1) - E(U \mid X = 0) \right] = \beta_1.$$

When $X = 0, 1$, as we’ve seen, the CEF slope is $\gamma_1 = E(Y \mid X = 1) - E(Y \mid X = 0)$, so this shows that $\gamma_1 = \beta_1$. If $X$ and $U$ are independent, then $E(U \mid X = 0) = E(U \mid X = 1)$; i.e., $X \perp U$ is a sufficient condition for $\beta_1 = \gamma_1$. If $X \perp U \implies E(U \mid X) = E(U)$. Also, $\text{Cov}(X, U) = 0$ (or, “$X$ and $U$ are uncorrelated”) implies $\beta_1 = \gamma_1$. With binary $X$, $\text{Cov}(X, U) = 0 \iff E(U \mid X = 0) = E(U \mid X = 1)$; with more than two possible values of $X$, $\text{Cov}(X, U) = 0$ is a weaker condition than $E(U \mid X) = E(U)$. Although each condition is different, each conveys the qualitative idea that the unobserved determinants of $Y$ (that are in $U$) do not systematically vary with $X$.

Theorem 6.1 summarizes these results, which extend from $X = 0, 1$ to any two possible values of $X$.

**Theorem 6.1 (linear structural identification).** Consider the linear structural and CEF models in (6.24) and (6.25), respectively. Assume $X$ has only two possible values, $x_1$ and $x_2$. The structural slope is identified and equal to the CEF slope, $\beta_1 = \gamma_1$, when either i) $E(U \mid X = x_1) = E(U \mid X = x_2)$, or equivalently $E(U \mid X) = E(U)$; ii) $X \perp U$; or iii) $\text{Cov}(X, U) = 0$. If additionally $E(U) = 0$, then the structural intercept is also identified, with $\beta_0 = \gamma_0$.

The linear structural model and conditional mean zero condition are simple to state, but usually difficult to justify in real-world examples. 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 average 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 provides a more formal economic model of why the college graduates should have higher ability, i.e., why $E(U \mid X = 1) > E(U \mid X = 0)$. For a very brief overview of his signaling model, see
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.

If \( E(U \mid X) = E(U) \) but \( E(U) \neq 0 \), then the structural and CEF intercepts differ by \( E(U) \). Starting from the structural model,

\[
Y = \beta_0 + \beta_1 X + U
\]

\[
= \beta_0 + \beta_1 X + U + E(U) - E(U)
\]

\[
= \beta_0 + E(U) + \beta_1 X + U - E(U).
\]

Now \( V \equiv U - E(U) \) is a CEF error since

\[
E[U - E(U) \mid X] = E[U \mid X] - E[E(U) \mid X] = E(U) - E(U) = 0,
\]

so the CEF intercept and slope are respectively

\[
\gamma_0 = \beta_0 + E(U), \quad \gamma_1 = \beta_1.
\]

**Discussion Question 6.7** (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. IDENTIFICATION

6.6.2 Average Treatment Effect

Identification of the average treatment effect (ATE) of $X$ on $Y$ was initially (less formally) discussed in Section 4.3. Here, we formally state identifying assumptions for the ATE.

What does it mean for the ATE to be “identified”? Earlier, the ATE was said to be identified whenever (4.12) is true. Updating (4.12) to this chapter’s notation, the ATE is identified whenever

$$\text{E}(Y^B) - \text{E}(Y^A) = \text{E}(Y | X = 1) - \text{E}(Y | X = 0),$$

(6.28)

where $Y^B$ and $Y^A$ are (still) the potential treated and untreated outcomes, respectively. The important feature of (6.28) 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 $\text{E}(Y | X = 1)$ and $\text{E}(Y | X = 0)$. If (6.28) is true, then learning about these conditional means (on the right-hand side) helps us learn the ATE (left-hand side). Without (6.28), we are stuck, because $Y^B$ and $Y^A$ are not both observable for all individuals, so we cannot directly learn about them. Identification connects the seen and the unseen.

The first identifying assumption, SUTVA, was discussed in Section 4.3.3. We now discuss another assumption before providing formal statements.

It was noted in Section 4.3 that randomized treatment is one way to help identify the ATE. More formally, the key is that randomization satisfies statistical independence between the treatment assignment and the individual’s pair of potential outcomes:

$$(Y^A, Y^B) \perp \perp X.$$ (6.29)

This is another of the three assumptions we’ll use.

Three formal assumptions for treatment effect identification are now stated. These are sufficient, but not necessary: if they are 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 A6.1 is usually just called SUTVA, but the main part of it is often called no interference (or non-interference). Assumption A6.2 has many names: independence, ignorability, or unconfoundedness. The combination of A6.2 and A6.3 is called strong ignorability. For more detail, history, and discussion, see Imbens and Wooldridge (2007).

Assumption A6.1 (SUTVA). Everyone with $X = 1$ receives the same treatment, and one individual’s treatment does not affect any other individual’s potential outcomes.

Assumption A6.2 (unconfoundedness). Treatment is independent of the potential outcomes: $(Y^A, Y^B) \perp \perp X$.

Assumption A6.3 (overlap). There is strictly positive probability of both treatment and non-treatment: $0 < P(X = 1) < 1$. 

Assumption \[A6.3\] 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 \(Y = Y^A\) (the untreated potential outcome) for everybody. We can learn about \(E(Y^A)\), but it’s impossible to learn about \(E(Y^B)\) since \(Y^B\) is literally never observed. Although seemingly obvious, a more general overlap assumption may not always hold in the more complex models later in this text.

Formally, identification of the ATE is shown as follows. The key is that \[A6.2\] allows us to observe representative samples of both \(Y^A\) and \(Y^B\). Mathematically, this independence assumption implies that the means of the potential outcomes do not statistically depend on the treatment \(X\):

\[
E(Y^B) = E(Y^B | X = 1), \quad E(Y^A) = E(Y^A | X = 0).
\] (6.30)

From (6.17), \(Y = Y^B\) when \(X = 1\) and \(Y = Y^A\) when \(X = 0\), so

\[
E(Y^B | X = 1) = E(Y | X = 1), \quad E(Y^A | X = 1) = E(Y | X = 0).
\] (6.31)

Combining (6.30) and (6.31), this says that the population mean of the treated potential outcome, \(E(Y^B)\), equals the mean of the observed outcome in the treated population, \(E(Y | X = 1)\), which in earlier notation was \(E(Y^T)\). Since \(E(Y | X = 1)\) is a feature of the joint distribution of \((Y, X)\), it is identified. Since \(E(Y^B) = E(Y | X = 1)\), it is also identified. Similarly, \(E(Y^A) = E(Y | X = 0)\) is identified, so \(E(Y^B) - E(Y^A)\) is identified, too.

**Theorem 6.2 (ATE identification).** Under \[A6.1\] \[A6.3\] the ATE is identified:

\[
E(Y^B - Y^A) = E(Y^B) - E(Y^A) = E(Y | X = 1) - E(Y | X = 0),
\]

which is the slope \(\beta_1\) in the linear CEF model in (6.14).

**Proof.** The constructive proof of ATE identification links the unobservable ATE with the observable mean difference. Specifically,

\[
\begin{align*}
\text{use linearity} & \quad \text{use (6.30)} \\
E(Y^B - Y^A) &= E(Y^B) - E(Y^A) \\
&= E(Y^B | X = 1) - E(Y^A | X = 0) \\
&= E(Y | X = 1) - E(Y | X = 0).
\end{align*}
\]

This expression equals the slope coefficient \(\beta_1\) from the linear CEF model in (6.14), as shown in (6.15). \(\square\)
6.6.3 Average Structural Effect

The ASE in (6.22) is identified if it equals the CEF difference in (6.23). A sufficient condition for this is

$$U \perp \perp X.$$  (6.32)

When the unobservable determinants are independent of $X$, then $h(1, U) \perp \perp X$ and $h(0, U) \perp \perp X$, so

$$E[h(0, U) | X = 0] = E[h(0, U)], \quad E[h(1, U) | X = 1] = E[h(1, U)],$$

so

$$E(Y | X = 1) - E(Y | X = 0) = E[h(1, U) | X = 1] - E[h(0, U) | X = 0]$$

$$= E[h(1, U)] - E[h(0, U)]$$

$$= E[h(1, U) - h(0, U)] = ASE.$$  (6.33)

The left-hand side of (6.33) is the parameter $\beta_1$ from the linear CEF model. That is, given independence, the ASE (defined with unobservables) equals the CEF slope $\beta_1$ (defined with only observables).

6.7 Estimation: OLS

In Chapter 6, the only statistical model we’ve considered is the CEF model. It has been written in different ways, but all are equivalent in the case where $X$ has only two possible values. For estimation, we focus on (6.14).

Various causal and structural models have been discussed, but the statistical object of interest is always $\beta_1$ in (6.14). Such models on differ in their interpretation of $\beta_1$, but interpretation does not matter for estimation.

When $X$ has only two possible values, it is simple to estimate the two conditional means directly, using Section 3.3. For example, if $X$ is 0 or 1, then $E(Y | X = 0)$ is the mean $Y$ in the subpopulation with $X = 0$, and $E(Y | X = 1)$ is the mean $Y$ in the subpopulation with $X = 1$. With iid sampling, these can be estimated by the sample average $\bar{Y}_i$ for individuals with $X_i = 0$ and the sample average $\bar{Y}_i$ for individuals with $X_i = 1$, respectively. These then determine $\beta_1$ as in (6.15).

Though simple, such an approach does not generalize easily, so we explore an alternative: ordinary least squares (OLS). The intuition behind the least squares approach comes from the characterization of the conditional mean (CEF) as the best predictor of $Y$ given $X = x$ given quadratic loss. The idea is the same as the idea in (3.6) for estimating the unconditional mean of $Y$. In the population, if the CEF is $E(Y | X = x) = \beta_0 + \beta_1 x$, then

$$(\beta_0, \beta_1) = \arg\min_{b_0, b_1} E[(Y - b_0 - b_1 X)^2],$$  (6.34)
showing that the CEF is the best (given quadratic loss) predictor of $Y$ given $X$. In the sample, replacing the population mean ($\mu$) with the sample mean ($\frac{1}{n} \sum_{i=1}^{n}$), the minimization problem analogous to (6.34) is

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2.$$ (6.35)

Notationally, the “hats” on the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ indicate that they are computed from the sample, whereas the true population values $\beta_0$ and $\beta_1$ lack hats. The form of (6.35) explains the L and S in OLS: “least” (L) referring to minimization, and “squares” (S) referring to squaring $Y_i - b_0 - b_1 X_i$. (Explaining the O is a special treat reserved for Econ PhD students.)

Using calculus, formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be derived. Computers can calculate $\hat{\beta}_1$ faster than humans, but formulas can help our understanding and intuition. Consider when $X = 0$ or $X = 1$. Analogous to the population $\beta_1 = \mathbb{E}(Y \mid X = 1) - \mathbb{E}(Y \mid X = 0)$, $\hat{\beta}_1$ is the difference between the sample average $Y_i$ for observations with $X_i = 1$ and the average $Y_i$ when $X_i = 0$. Formally,

$$\hat{\beta}_1 = \hat{\mathbb{E}}(Y \mid X = 1) - \hat{\mathbb{E}}(Y \mid X = 0)$$

$$= \frac{\sum_{i=1}^{n} Y_i \cdot 1\{X_i = 1\}}{\sum_{i=1}^{n} 1\{X_i = 1\}} - \frac{\sum_{i=1}^{n} Y_i \cdot 1\{X_i = 0\}}{\sum_{i=1}^{n} 1\{X_i = 0\}}.$$ (6.36)

Equation (6.35) can be described with other terms. These other terms were introduced around (3.7) in the simpler setting of estimating the unconditional mean of $Y$. Given any estimates $(\hat{\beta}_0, \hat{\beta}_1)$, the fitted values are

$$\hat{Y}_i \equiv \hat{\beta}_0 + \hat{\beta}_1 X_i.$$ (6.38)

These are the estimated CEF evaluated at $X_i$. That is, our CEF was

$$m(x) \equiv \mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x,$$

so our estimated CEF is

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$ 

Plugging in $x = X_i$,

$$\hat{m}(X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i = \hat{Y}_i,$$

the $i$th fitted value. Given $\hat{Y}_i$, the residual is defined as

$$\hat{U}_i \equiv Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$ (6.39)

Consequently, (6.35) 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.
6.8. QUANTIFYING UNCERTAINTY

The OLS estimator is consistent under fairly general conditions. Given that $X$ is a binary variable, the only (trivial) restriction on $X$ is that it is not always zero (for everyone in the population) or always one. The outcome $Y$ must have a finite variance; this is true for almost all economic variables. Besides that, the data must be sampled from the population in a way that doesn’t cause bias; e.g., iid sampling is sufficient, but other types of sampling are also fine. For details on assumptions in a more general regression model, see Section 7.7.2. When these assumptions are met, then OLS provides a consistent estimator of the CEF parameters and thus the CEF itself:

$$\hat{\beta}_0 \xrightarrow{p} \beta_0, \quad \hat{\beta}_1 \xrightarrow{p} \beta_1, \quad \hat{m}(0) \xrightarrow{p} m(0), \quad \hat{m}(1) \xrightarrow{p} m(1).$$  \hfill (6.40)

Although more OLS formulas and discussion will appear in later chapters, it suffices for now to say that computers can easily compute the OLS estimator. Formulas can provide insight, but you will always use a computer to actually run OLS.

The following code shows how to run 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 shows that in the case 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, and $\hat{\beta}_1$ 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)
```

```r
## (Intercept)  X
## 2 5
```

```r
c( mean(df$Y[df$X==0]), mean(df$Y[df$X==1]) )
```

```r
## [1] 2 7
```

6.8 Quantifying Uncertainty

The ways to quantify uncertainty here are the same as in Section 3.5. For the unconditional mean, Section 3.5 discussed confidence intervals, $p$-values, and hypothesis testing. The same can be done for the CEF parameters $\beta_0$ and $\beta_1$. Although we’ll skip the details, the underlying math is similar, too, since the OLS estimator is basically just a more complex sample average, so its sampling distribution is approximately normal. For now, you can take it on faith that the methods quantifying uncertainty actually do what they claim (approximately), and instead focus on the interpretation, which is the same as in Section 3.5.
6.8.1 Types of Uncertainty

The “uncertainty” quantified by confidence intervals (and such) is only the uncertainty from having a random sample instead of knowing the full population. In particular, it does not reflect any uncertainty you have about the identifying assumptions that let $\beta_1$ be interpreted as an average causal effect. The confidence interval is just a confidence interval for the CEF slope $\beta_1$. If $\beta_1$ is not actually an average causal effect, then a 95% confidence interval may contain the true average causal effect less than 95% of the time, maybe even closer to 0% of the time. It is important to consider the identifying assumptions very carefully, because statistical methods for estimation and inference on the CEF slope $\beta_1$ cannot tell us anything about such assumptions.

6.8.2 Heteroskedasticity

Different methods for quantifying uncertainty make different assumptions about the conditional variance. Instead of the condition mean $E(Y \mid X = x)$, which is the mean of the conditional distribution of $Y$ given $X = x$, the conditional variance

$$\sigma^2_Y(x) \equiv \text{Var}(Y \mid X = x)$$

(6.41)

is the variance of the conditional distribution of $Y$ given $X = x$. The term **heteroskedasticity** means $\sigma^2_Y(x) = \sigma^2_Y$, a constant not depending on $x$, whereas **homoskedasticity** 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. The question of homoskedasticity is like the question of “equal variances” for a two-sample $t$-test.

You should always assume there might be heteroskedasticity, for the following reasons. First, with economic data (broadly speaking), there usually is heteroskedasticity. This can be seen in data, and it often makes sense. For example, the variance in annual income for very low-education workers (who generally have low incomes) is smaller than the variance in annual income for high-education workers (who generally have high incomes). Second, methods allowing for heteroskedasticity are still accurate with homoskedasticity, but methods assuming homoskedasticity may be very inaccurate with heteroskedasticity. Even with very large samples that make other approximation errors small, assuming homoskedasticity can lead to incorrect conclusions. In this sense, it is safer to just always use methods that allow heteroskedasticity. Such methods are often called **robust** to heteroskedasticity. (Sometimes the word “robust” is used by itself, but this is ambiguous; you should always ask, robust to what?)

**Discussion Question 6.8** (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. QUANTIFYING UNCERTAINTY

6.8.3 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.”

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 $\text{lm()}$ 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 $\text{lmtest}$ and $\text{sandwich}$ 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.

```r
library(lmtest); library(sandwich)
set.seed(112358)
n <- 1000
df <- data.frame(Y=c(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)
# HC0: original from Hal White (1980)
retVC0 <- vcovHC(ret, type="HC0")
sl.out['HC0','SE'] <- sqrt(retVC0[2,2])
# HC1: matches Stata default, and two-sample t.test below
retVC1 <- vcovHC(ret, type="HC1")
# HC3: recommended/default (and larger SE than HC0, HC1)
```
retVC3 <- vcovHC(ret, type="HC3")

# Default homoskedastic results
sl.out['Homosk.'],c('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],
  alternative='two.sided', mu=0, conf.level=0.95,
  paired=FALSE, var.equal=FALSE)

sl.out['ttest.uneq',-1] <-
  c(t.sl$conf.int, t.sl$statistic, t.sl$p.value)

# For comparison: var.equal=TRUE

t2 <- t.test(x=df$Y[df$X==1], y=df$Y[df$X==0],
  alternative='two.sided', mu=0, conf.level=0.95,
  paired=FALSE, var.equal=TRUE)

sl.out['ttest.eq',-1] <- c(t2$conf.int, t2$statistic, t2$p.value)

sl.out['ttest.uneq',1] <-
  (t.sl$conf.int %*% c(-1,1)) /
  (2*qt(p=1-0.05/2,df=t.sl$parameter))

sl.out['ttest.eq',1] <-
  (t2$conf.int %*% c(-1,1)) /
  (2*qt(p=1-0.05/2,df=t2$parameter))

print(round(sl.out, digits=3))

# SE CI.lower CI.upper t.stat p.value
# ttest.eq  0.128  -0.026   0.476  1.76  0.079
# Homosk.  0.128  -0.026   0.476  1.76  0.079
# ttest.uneq  0.095   0.038   0.412  2.36  0.018
# HC0  0.095   0.039   0.412  2.37  0.018
# HC1  0.095   0.038   0.412  2.37  0.018
# HC3  0.095   0.038   0.412  2.36  0.018
Discussion Question 6.9 (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 $E(Y \mid X = x) = P(Y = 1 \mid X = x)$, 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. Discuss both economic significance and statistical significance in the following possible results. (Economic and statistical significance were introduced in Section 3.5.6) 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) $\hat{\beta}_1 = -0.01$, $\hat{\text{SE}} = 0.01$. (Hint: as a quick way to roughly glean statistical significance, $|\hat{\beta}_1/\hat{\text{SE}}| \geq 2$ means statistical significance at a 5% level, with higher values being more statistically significant.)
Empirical Exercises

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:

```r
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`) with R command `ret <- lm(re78~train, data=jtrain2)` or Stata command `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

```r
coeftest(ret, vcov.=vcovHC(ret, type='HC1'))
coefci( 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 `jt2.mar1 <- jtrain2[jtrain2$married==1,]`

f. Run your previous analysis for the subset of married individuals. Replace `data=jtrain2` with `data=jt2.mar1` in R; run `regress re78 train if married==1, vce(robust)` in Stata

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

Comparing Three (or More) Distributions

 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): [https://www.youtube.com/watch?v=R0LeLaR-17U](https://www.youtube.com/watch?v=R0LeLaR-17U)
- 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)
Surprisingly, many critical issues arise with regression comparing three conditional distributions (three values of $X$) that did not appear in Chapter 6 for two distributions. With two distributions, the regression modeled the conditional means, useful for description, prediction, and causal inference. However, if we use the same regression model for three distributions, we may fail to even model the conditional means, let alone anything about causality. Although easily fixed in this simple setting, this problem is generally difficult to solve, so the related concepts are detailed for the case of three values of $X$.

7.1 Misspecification

Consider the population model

$$Y = \beta_0 + \beta_1 X + U,$$  \hspace{1cm} (7.1)

where supposedly $E(U \mid X) = 0$, and this time $X$ has three possible values: 0, 1, and 2. For example, as in Section 6.3, imagine $Y$ is income ($$/yr) and $X$ is number of degrees, but now counting a high school degree. That is, $X = 0$ indicates not having graduated from high school; $X = 1$ is a high school degree; and $X = 2$ is high school plus another degree.

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, but let’s look more formally.

The more general CEF model from (6.12) is still correct, which is always true. With $V \equiv Y - m(X)$, the same steps as in Section 6.3 again lead to $E(V \mid X) = 0$ as in (6.11). As in (6.12),

$$Y = m(X) + V, \quad E(V \mid X) = 0,$$

is the true CEF model.

The question is then whether

$$m(X) = \beta_0 + \beta_1 X, \quad X = 0, 1, 2.$$ 

The left-hand side, $m(X)$, can pick any three values for $m(0)$, $m(1)$, and $m(2)$. For example, maybe there is a big income gap between only children ($X = 0$) and individuals
7.1. MISSPECIFICATION

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 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 assumed 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.

**Discussion Question 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 \(X = 0, 1, 2\), where \(m(1) = 60, m(2) = m(3) = 40\). Is it possible to write \(m(x) = \beta_0 + \beta_1 x\)
now? Why or why not?

Is a linear CEF model always wrong? Continuing the siblings and income example, what if $m(2) = 20$ instead? Then indeed $m(X) = 60 - 20X$, i.e., $\beta_0 = 60$ (thousands of $/yr) and $\beta_1 = -20$ (thousands of $/yr per sibling). In that case, the linear CEF is properly specified (or correctly specified). However, this is extremely unlikely, essentially impossible. It means that, whatever values $m(0)$ and $m(1)$ take in the population, $m(2)$ must perfectly, precisely equal $m(1) + [m(1) - m(0)]$. Even if $m(2) = 20.001$, the linear CEF model is misspecified. However, with such a small amount of misspecification, a linear model is a very good approximation; this sort of interpretation is discussed in Section 7.4.

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.” See 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.” 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 (the optimal magnitude of 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 reinterpretation is detailed in Sections 7.3–7.5.

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 (see Section 2.1.3) for each possible value of $X$. (See Section 2.1 for additional terminology.) Recall the indicator function from (2.2). Here,

$$1\{X = j\} = \begin{cases} 1 & \text{if } X = j \\ 0 & \text{otherwise} \end{cases}, \quad j = 0, 1, 2. \quad (7.2)$$
Since only three values of $X$ are possible, $1\{X = 0\} = 1 - 1\{X = 1\} - 1\{X = 2\}$. Thus,

$$m(x) = m(0) 1\{x = 0\} + m(1) 1\{x = 1\} + m(2) 1\{x = 2\}$$

$$= m(0)[1 - 1\{x = 1\} - 1\{x = 2\}] + m(1) 1\{x = 1\} + m(2) 1\{x = 2\}$$

$$= m(0) + [m(1) - m(0)] 1\{x = 1\} + [m(2) - m(0)] 1\{x = 2\}$$

$$= \beta_0 + \beta_1 1\{x = 1\} + \beta_2 1\{x = 2\},$$

$$\beta_0 \equiv m(0), \ \beta_1 \equiv m(1) - m(0), \ \beta_2 \equiv m(2) - m(0). \quad (7.4)$$

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, $m(2) - m(0) = 40 - 60 = -20$. If all subpopulations had identical means, then $\beta_0$ could be anything, but all other $\beta_j = 0$, indicating zero difference.

More generally, even if $X$ has more than three possible values, dummy variables could be used similarly to avoid CEF misspecification. For example, 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.)

Discussion Question 7.2 (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.

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.
7.3 Linear Projection

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: Orthogonal projection](image)

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 more formally. 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 \( y \). Formally,

\[
p = \arg \min_{s \in S} d_E(y, s).
\] (7.5)

Linear projection with random variables is the same idea, but with a different definition of distance and a different “shape” to search over. For linear projection, unlike a CEF, an intercept term must be included explicitly. This is equivalent to having a regressor \( X_0 = 1 \) in addition to \( X \), so 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 \( (1, X) \). (Section 8.2.1 discusses linear combinations and linearity in more detail.) Consequently,
7.3. LINEAR PROJECTION

the closest "point" inside the "shape" must be a linear combination of \((1, X)\); it is usually written \(\beta_0 + \beta_1 X\). Mathematically, parallel to (7.5),

\[
\text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X = \arg \min_{a+bX} d(Y, a+bX) = \arg \min_{a+bX} \sqrt{E[(Y - a - bX)^2]},
\]

(7.6)

where Euclidean distance \(d_E(\cdot, \cdot)\) has been replaced by a probabilistic "distance" measure

\[
d(A, B) \equiv \sqrt{E[(A - B)^2]}.
\]

(7.7)

That is, linear projection gets \(\beta_0 + \beta_1 X\) as "close" to \(Y\) as possible, in a probabilistic sense.

From the definition in (7.6), a formula for the linear projection can be derived. In particular, we are interested in each linear projection coefficient (LPC), \(\beta_0\) and \(\beta_1\). In this special case with a single regressor \(X\) and an intercept,

\[
\beta_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}, \quad \beta_0 = E(Y) - \beta_1 E(X).
\]

(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:

\[
\beta_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \frac{\text{Cov}(Y, X)}{\sigma_X^2} \frac{\sigma_Y}{\sigma_Y} = \frac{\text{Cov}(Y, X)}{\sigma_X \sigma_Y} \frac{\sigma_Y}{\sigma_X} = \text{Corr}(Y, X) \frac{\sigma_Y}{\sigma_X}.
\]

(7.9)

Either version of the formula shows how the linear projection slope \(\beta_1\) is related to the linear dependence between \(Y\) and \(X\). Once the slope is determined, the intercept \(\beta_0\) simply moves the linear projection line up or down so that \(E(Y) = \beta_0 + \beta_1 E(X)\). That is, the linear projection always goes exactly through the point \((X, Y) = (E(X), E(Y))\).

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 \(\beta_0 = E(Y) - \beta_1 E(X)\) is not easy to interpret, except when the regressor has been demeaned so that \(E(X) = 0\), in which case \(\beta_0 = E(Y)\). In this case, \(\beta_0\) is called the "centercept" instead of intercept; but despite the better interpretation, it is rarely seen in economics.

Equation (7.8) shows that the LPCs summarize the joint probability distribution of \((Y, X)\). Although we won’t do it, in principle we could take any joint distribution of \((Y, X)\) and then compute the terms \(E(Y), E(X), \text{Cov}(Y, X), \text{and} \text{Var}(X)\), which then determine \(\beta_0\) and \(\beta_1\). Similar to how we summarized the population distribution of \(Y\) using features like \(E(Y)\) and \(\text{Var}(Y)\), the LPCs are one way to summarize the joint distribution of \((Y, X)\) in just two numbers. Although a two-number summary of a complicated joint distribution is very convenient, clearly much information is lost in such a summary. Just as \(E(Y)\) does not fully describe the distribution of \(Y\), the LP does not fully describe the distribution of \((Y, X)\). Just as other features like percentiles...
(quantiles) complement the mean in describing \( Y \), other features like quantile regression complement linear projection in describing \((Y, X)\).

Although you don’t need to know it for this class, the linear projection coefficient formula can also be written in matrix form. This generalizes more easily to more complex models, so matrix math is indispensable for more advanced econometrics. Define vector \( \mathbf{X} \equiv (1, X)' = \begin{bmatrix} 1 \\ X \end{bmatrix} \),

\[
(7.10)
\]

where \((1, X)'\) indicates the transpose of the row vector \((1, X)\), i.e., turning it from a row vector to a column vector as seen in the equation. Then,

\[
\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = [E(XX')^{-1} E(XY) = \begin{bmatrix} 1 \\ E(X) \\ E(X^2) \end{bmatrix}^{-1} \begin{bmatrix} E(Y) \\ E(XY) \end{bmatrix}. \quad (7.11)
\]

### 7.4 “Best” Linear Approximation

For description, the linear projection can also 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 \( a + bX \). Mathematically,

\[
\text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X = \arg \min \limits_{a+bX} \{ (m(X) - (a+bX))^2 \}, \quad m(X) \equiv E(Y \mid X). \quad (7.12)
\]

That is, among all possible functions of the form \( a + bX \), 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 \( m(X) = \gamma_0 + \gamma_1 X \) for some \( \gamma_0 \) and \( \gamma_1 \), then setting \( \beta_0 = \gamma_0 \) and \( \beta_1 = \gamma_1 \) means \( m(X) - (\beta_0 + \beta_1 X) = 0 \). Since this term is squared in (7.12), and squared values must be non-negative, zero is the smallest possible value. Consequently,

\[
\text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X = \gamma_0 + \gamma_1 X = E(Y \mid X).
\]

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 linear projection, which is the BLA. Sometimes the best (closest) is still not good (not close).

The following example of a “bad” best linear approximation comes from \cite{Hansen2018} §2.27. 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 | 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. This can be seen in (7.12), where the E\{·\} refers to expected value with respect to the distribution of X. 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; in particular, their slopes 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 is also the best linear predictor (BLP) of Y given X. As with the BLA, “best” assumes quadratic loss. “Linear” again refers to a function of the form a + bX. If the true value is Y, and the prediction is a + bX, then the prediction error is Y − (a + bX). Given quadratic loss, the “loss” for such an error is its square, [Y − (a + bX)]\^2. As before, the optimal predictor minimizes the expected value of this loss. Mathematically,

\[
\text{LP}(Y | 1, X) = \beta_0 + \beta_1 X = \arg \min_{a+bX} E\{[Y - (a + bX)]^2\}. \tag{7.13}
\]

That is, among all possible functions of the form 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, (7.13) 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 before, “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 were omniscient and fully knew everything about the population. The BLP is something we wish to learn about the population; it is a feature of the joint distribution of (Y, X). Fortunately, the BLP (and thus BLA and LP) is precisely what OLS estimates.

**Discussion Question 7.3 (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 ≤ Y ≤ 70. When X = 1, the mean Y is 40 and 30 ≤ Y ≤ 50. When X = 2, it’s the same as when X = 1: the mean Y is 40 and 30 ≤ Y ≤ 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. If the structural error \( U \) satisfies the CEF error property \( E(U \mid X) = 0 \), then the structural function is the CEF, so the BLA interpretation of the linear projection applies: it is also the best linear approximation of the structural function. However, often we care only about the slope 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 a primary motivation for “nonparametric” CEF estimation (“nonparametric regression”); see Section 8.3.

The ASE identification argument also requires a properly specified CEF, which also motivates nonparametric CEF estimation. When the identifying assumption \( X \perp U \) is true, the ASE (for some change in \( X \)) equals the change in the CEF (for the same change in \( X \)). If we run OLS with a simple linear model, then we only estimate the linear projection, not the CEF. Unfortunately, unless the CEF equals the linear projection, changes in the linear projection do not equal the ASE.

### 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 (6.35) mirrors the BLP definition in (7.13). That is, replacing the population mean \( (E) \) in (7.13) with the sample mean \( (\frac{1}{n} \sum_{i=1}^{n}) \) yields (6.35). 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
\[
\hat{\beta}_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}),
\]
\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \bar{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i, \quad \bar{X} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i.
\]
(7.14)

This matches the formulas from solving the minimization problem in (6.35) 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 \( \hat{E}(Y) = \frac{1}{n} \sum_{i=1}^{n} Y_i \) is the sample analog of the population mean \( E(Y) \). Estimators that are sample analogs of the population estimands are said to use the analogy principle. They also use the plug-in principle since they essentially “plug in” the sample distribution for the population distribution. (See Section 3.3.1 for a reminder.)

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. If you want help visualizing inflating an \( n \)-dimensional beach ball centered at \((Y_1, \ldots, Y_n)\), see \cite{Koenker2005}.

### 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 average causal effect of \( X \) on \( Y \), then the following properties apply to average causal effect 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**

Before stating properties, certain assumptions are stated. These are sufficient for Theorems 7.1–7.3, but not necessary (see Section 6.1).

**Assumption A7.1.** Sampling of \((Y_i, X_i)\) is iid.

**Assumption A7.2.** The regressor \( X \) is not a constant, i.e., there is no single value \( x \) such that \( P(X = x) = 1 \).
Assumption A7.3. The variances of $Y$ and $X$ are finite: $\operatorname{Var}(Y) < \infty$, $\operatorname{Var}(X) < \infty$. 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. 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 earlier in Section 3.2 for a single variable 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 $i \neq k$ (i.e., any two observations are sampled independently), and $W_i$ and $W_k$ have the same distribution. However, there could be dependence among the contents of the vector $W_i$. 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. It only restricts $(Y_i, X_i) \perp \perp (Y_k, X_k)$ for $i \neq k$, which implies $Y_i \perp Y_k$, $X_i \perp X_k$, $Y_i \perp X_k$, and $X_i \perp Y_k$. It also 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, $E(Y_i) = E(Y_k)$, $\operatorname{Var}(X_i) = \operatorname{Var}(X_k)$, $E(Y_i \mid X_i = x) = E(Y_k \mid X_k = x)$, $P(Y_i \leq 0 \mid X_i = x) = P(Y_k \leq 0 \mid X_k = x)$, etc. All this readily generalizes to multiple regressors, just redefining $W_i \equiv (Y_i, X_{1i}, X_{2i}, \ldots)$.

Assumptions [A7.3] and [A7.4] are usually true with economic data, but there are some exceptions. First consider variables like age or education that have clear lower and upper bounds. The lower bound for age and education is zero years. A very safe upper bound is 120 years. Letting $Y$ denote the variable, the value of $Y^4$ is never above $(200)^4$, so the expected value $E(Y^4)$ also cannot exceed $(200)^4$. Since $(200)^4$ is finite, $E(Y^4) \leq (200)^4 < \infty$, so the [A7.4] holds (i.e., is true). More generally, for any variable $Y$ bounded in absolute value by $|Y| < M$, $E(Y^4) < M^4 < \infty$. Since $E(Y^4) < \infty \implies E(Y^2) < \infty$, both [A7.4] and [A7.3] hold true in this case.

Nonetheless, there are some economic variables that 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. If violation of [A7.3] or [A7.4] is a concern for the $Y$ variable, but not the $X$ variable, then median regression can be used: it does not require finite variance of $Y$ (or even a well-defined mean), although it still requires finite variance of $X$ for consistency.

Finally, [A7.2] sounds obvious. In more complex models, it must be generalized to a less-obvious condition, but in either case, it is a condition that your software will check for you. Specifically, if the sample version of [A7.2] is false, then your statistical software will report an error. Although it is technically possible that [A7.2] is false even if your software does not report an error, this distinction is rarely of practical importance. That is, don’t worry about [A7.2] unless you get a software error. For our current, simple
regression model, it just means that $X_i$ cannot have the exact same value for all $i$. For example, you cannot regress wage on education if your sample has $X_i = 12$ for every individual. This has some intuition: there have to be some differences in $X_i$ in order to learn about how $Y_i$ varies with $X_i$, i.e., learn about $\beta_1$.

### 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, written: $\hat{\beta}_0 \xrightarrow{p} \beta_0$, $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

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 we can interpret the linear projection as a CEF, or interpret $\beta_1$ as an average causal effect, are questions of identification, not estimation. OLS estimates the linear projection, and leaves further interpretation up to us.

Theorem 7.1 does not say that OLS is a bad estimator if sampling is not iid. That would be the logical inverse, which does not follow from the theorem’s statement, 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 linear projection coefficients.

**Theorem 7.2** (OLS approximate normality, 1 regressor). If $[A7.1] - [A7.3]$ are true, then the OLS intercept and slope estimators are asymptotically normal, i.e., with large $n$, approximately $\hat{\beta}_0 \sim N(\beta_0, \widehat{SE}_0^2)$ and $\hat{\beta}_1 \sim N(\beta_1, \widehat{SE}_1^2)$, where the true standard errors $\widehat{SE}_0$ and $\widehat{SE}_1$ are unknown but can be estimated, and are proportional to $1/\sqrt{n}$. Further, with large $n$, the joint distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ is approximately (bivariate) normal with mean $(\beta_0, \beta_1)$ and an unknown (co)variance matrix that can be estimated.

Theorem 7.2 is practically useful for constructing confidence intervals. Theorem 7.3 more explicitly states that such confidence intervals have approximately the desired coverage.

**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)
CHAPTER 7. COMPARING THREE (OR MORE) DISTRIBUTIONS

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))
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)
n <- 500;  m012 <- c(60,40,40)
df <- data.frame(X=sample(x=0:2, size=n, prob=3:1/6, replace=TRUE),
                 U=rnorm(n))
ret <- lm(formula=Y~X, data=df)
retVC1 <- vcovHC(ret, type="HC1")
LP1 <- c(coeftest(ret, vcov. = retVC1)['X',],
         coefci(ret, vcov. = retVC1)['X',])

set.seed(112358)
n <- 500;  m012 <- c(60,40,40)
```
df <- data.frame(X=sample(x=0:2, size=n, prob=1:3/6, replace=TRUE),
                 U=rnorm(n))
df$Y <- rnorm(n=n, mean=m012[1+df$X]) + df$U
ret <- lm(formula=Y~X, data=df)
retVC1 <- vcovHC(ret, type="HC1")
LP2 <- c(coeftest(ret, vcov. = retVC1)["X",],
         coefci(ret, vcov. = retVC1)["X",])
tmp <- rbind(CEF, LP1, LP2)
round(x=tmp, digits=3)
## Estimate Std. Error t value Pr(>|t|) 2.5 % 97.5 %
## CEF -19.8 0.077 -257.1 0 -19.98 -19.67
## LP1 -12.3 0.310 -39.6 0 -12.91 -11.69
## LP2 -7.7 0.310 -24.8 0 -8.31 -7.09
#
# Use dummies to estimate nonlinear CEF
df$D0 <- (df$X==0) # not used
df$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)
coefficients(ret)
## (Intercept) D1 D2
## 59.8 -19.8 -19.8

### 7.8 Simple Linear Regression

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 even if X could be any real number between 0 and 1. Misspecification is likely; with even more values of X and thus \( m(x) \), if anything it seems even less likely that the CEF would be exactly linear. 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.

To be more explicit about interpretation, imagine we regress Y on a constant and X to get intercept and slope estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \). If miraculously the true CEF were
linear, i.e., if \( E(Y \mid X = x) = \beta_0 + \beta_1 x \), then \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) can be interpreted as estimates of \( \beta_0 \) and \( \beta_1 \). In the CEF, \( \beta_0 = E(Y \mid X = 0) \), and \( \beta_1 \) is the change in the conditional mean of \( Y \) associated with a one-unit increase in \( X \). Even then, \( \beta_0 \) may not have direct economic meaning if \( X = 0 \) is impossible or rare; e.g., imagine a wage (\( Y \)) regression where \( X \) is age, or years of education. More realistically, if the CEF does not have the form \( \beta_0 + \beta_1 x \), then \( \hat{\beta}_0 + \hat{\beta}_1 x \) can be interpreted as an estimate of the best linear predictor of \( Y \) given \( X = x \) (and quadratic loss), or the best linear approximation of the CEF (again with quadratic loss), or the linear projection of \( Y \) onto \((1, X)\). In that case, we’d say something vague-sounding like, “We estimate that a one-unit increase in \( X \) is associated with a \( \hat{\beta}_1 \) change in \( Y \).”

Structural identification is also essentially the same for more than three possible \( X \) values. Specifically, if the structural model is

\[
Y = \beta_0 + \beta_1 X + U
\]

and \( \text{Cov}(X, U) = 0 \), then \( \beta_1 \) equals the linear projection slope. However, ASE identification requires that the CEF is properly specified.

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 \( X \).

**Discussion Question 7.4 (income dummies).** You want to describe how health outcomes relate to income and have data on \( n = 100 \) individuals. Everyone’s income is different, so you include \( n - 1 = 99 \) dummy variables with coefficients \( \beta_1, \ldots, \beta_{99} \) along with the intercept \( \beta_0 \). a) Describe how such an estimated CEF might look. b) Explain why it might be less realistic or less helpful than a linear CEF.

**Discussion Question 7.5 (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. After you guess these, you can come up with your own puzzles in R; first make a scatterplot like \( Y <- c(1,2,3,4,13); X <- c(1,2,3,4,5); \text{plot}(X,Y) \) and then plot the OLS fit with \( \text{abline}(\text{lm}(Y-X)) \)

**Discussion Question 7.6 (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 mi = 1.600 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 1.600 km increase in \( X \), in terms of \( \gamma_1 \)? In terms of \( \delta_1 \)?)
Figure 7.3: Scatterplots for Discussion Question 7.5.
**Discussion Question 7.7** (regression significance). 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? (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 of $\hat{\beta}_1$. (Economic and statistical significance were introduced in Section 3.5.6) e) Describe one reason you doubt $\hat{\beta}_1$ has a causal interpretation.
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 with R command `library(wooldridge)` or Stata command `bcuse athlet1 , nodesc clear` (assuming you’ve already installed the R package or the Stata command)

b. Keep only data from 1993 with R command `dat <- athlet1[athlet1$year ==1993 , ]` or Stata command `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) with R command `dat$bowl4 <- dat$bowl + dat$finfour` or Stata command `generate bowl4 = bowl + finfour`

d. Display the number of observations with each possible value of `bowl4` (0, 1, or 2) with R command `table(dat$bowl4)` or Stata command `tabulate bowl4`

e. Regress the number of applications (for admission) on the prior year’s athletic success with R command `ret <- lm(apps~bowl4, data=dat)` or Stata command `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')`

g. Estimate and store the three CEF values. In R, use `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. In Stata, use `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, with R command `plot (0:2, m012)` (to plot estimated CEF points) followed by `abline(ret)` (to plot the OLS fit line), or Stata command `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. In R, inside the `plot()` command, add argument `main='Title'` to set the title, and similarly for `xlab='...'` and `ylab='...'` to set the x-axis and y-axis labels, respectively; inside the `abline()` function,
add arguments \texttt{col=2} to change the line’s color, \texttt{lty=2} to change the line style, and \texttt{lwd=3} to change the line width; you can set whatever values you like. In Stata, use \texttt{twoway scatter CEF bowl4 || lfit apps bowl4 , XXX} but replace the \texttt{XXX} with options to change the graph’s appearance (all separated by spaces, not any more commas), like \texttt{title("...") xtitle("...") ytitle("...")} for the title and axis labels, and \texttt{lcolor(red) lpattern(dash)} for the line color and style; again, use whatever values you’d like.

j. Display the numerical values of the OLS fit and the estimated CEF, with R command \texttt{rbind(m012, fit012)} or Stata command \texttt{collapse (mean) meanapps= apps , by(bowl4)} followed by \texttt{predict OLSfit, xb} and \texttt{list}
Chapter 8

Nonlinear and Nonparametric Regression

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): https://www.youtube.com/watch?v=5a6eW6UzTUK
- Log-log example (Lambert video): https://www.youtube.com/watch?v=040N6ZIegGM
- Overfitting (Lambert video): https://www.youtube.com/watch?v=ulHpi6nYcRc
• 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, §§2.9,5.2,7.2,7.3)

• 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.” Other transformations could be used, but we focus on log models here.

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. Each model is described below.

Here in Section 8.1, the distinction among causal, CEF, and linear projection models is unimportant. The interpretation of $U$ is left ambiguous intentionally. The emphasis here is instead on the interpretation of the slope coefficient $\beta_1$. This is phrased as “the change in $Y$ associated with a one-unit increase in $X$,” but this phrase could be replaced to better suit one of the models. For example, for a causal model, we might instead interpret $\beta_1$ in terms of an average causal effect on $Y$ of a one-unit increase in $X$. Or, for a CEF, “change in $Y$” should really be “change in the conditional mean of $Y$.”
8.1. LOG TRANSFORMATION

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 \( \infty \).

![Figure 8.1: The (natural) log function, \( \ln(\cdot) \).](image)

The log function has many properties. First, it is only defined for strictly positive input values; you cannot take \( \ln(x) \) if \( x \leq 0 \). Second, the log function is always increasing with \( x \); i.e., if \( x_2 > x_1 \), then \( \ln(x_2) > \ln(x_1) \). Third, the function increases more slowly with larger \( x \); it is very (infinitely!) steep when \( x \) gets near zero, but less and less steep (i.e., flatter and flatter) as \( x \) increases. Fourth, one useful identity is \( \ln(x^b) = b \ln(x) \), for any \( b \) and any strictly positive \( x \). Fifth, another useful identity is \( \ln(x_2/x_1) = \ln(x_2) - \ln(x_1) \), for any (strictly positive) numbers \( x_1 \) and \( x_2 \). Similarly, \( \ln(x_2x_1) = \ln(x_2) + \ln(x_1) \). Sixth, \( \lim_{x \to 0} \ln(x) = -\infty \) 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 percent changes in values much larger than 1, too. Consider some \( v_2 > v_1 \); 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

\[ \ln(v_2) = \ln(v_1(1 + p)) = \ln(v_1) + \ln(1 + p) \approx \ln(v_1) + p. \tag{8.1} \]

That is, the log difference \( \ln(v_2) - \ln(v_1) \approx p \), so for \( p \) near zero, we can interpret a log difference of \( p \) as an approximate 100\% 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.

Additionally, logs capture compound growth rates if \( v_1 \) and \( v_2 \) are measurements at different points in time. If \( v_1 \) is the starting value, and \( v_2 \) is \( t \) years later, and there is continuously compounded growth at a 100\% annualized rate, then \( v_2 = v_1e^r \). For example, if \( v_1 = 100 \), \( r = 0.02 \) (2\%), and \( t = 1 \), then \( v_2 = 100e^{0.02(1)} = 102.02 \); e.g., if you invest $100 at a 2\% continuously compounded annual interest rate, then you’ll have $102.02 one year later. If \( v_2 = v_1e^r \), then \( v_2/v_1 = e^r \), so \( \ln(v_2/v_1) = \ln(e^r) \), so

\[ r = \frac{\ln(v_2) - \ln(v_1)}{t}. \tag{8.2} \]

See Section 13.6 for more.

8.1.2 The Log-Linear Model

Interpretation

A log-linear model specifies

\[ \ln(Y) = \beta_0 + \beta_1 X + U. \tag{8.3} \]
8.1. LOG TRANSFORMATION

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 an approximate \( 100\beta_1 \)\% change in \( Y \). For example, if \( \beta_1 = 0.02 \), then a one unit increase in \( X \) is associated with 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 \( 1.01Y \), which is different than a 1 percentage point change in \( Y \). “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 \( 100(\beta_1)(50)\% = 100\% \). Why? Imagine the initial value of \( Y \) is 20, and the true increase is \( 100p\% \), so the final value is \( 20(1 + p) \).

A 100\% increase means \( p = 1 \), so the final value should be \( 20(1 + 1) = 40 \). However, \( \ln(40) - \ln(20) = 0.69 \) “log units,” not 1. An increase in 1 log unit would result in a final \( Y \) value of around 54.4: \( \ln(54.4) - \ln(20) = 1 \). That is, solving for the final value \( Y \) given a 1-log-unit increase from the initial value of 20, using the properties in Section 8.1.1, \( \ln(Y) - \ln(20) = 1 \), so \( \ln(Y/20) = 1 \), so \( Y/20 = e^1 \) and \( Y = 20e \approx 54.4 \).

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, this model may not be the best if we are actually interested in \( Y \) and not \( \ln(Y) \). Even the simpler task of predicting \( Y \) given \( X \) is more complicated than it seems. In a linear CEF model, the motivation for estimating the CEF \( E(Y \mid X = x) \) is that it is the best (under quadratic loss) predictor of \( Y \) given \( X = x \). In a log-linear CEF model, estimating \( E(\ln(Y) \mid X = x) \) provides the best predictor of \( \ln(Y) \) given \( X = x \).
but maybe not the best predictor of \( Y \). Taking the exponential function of both sides of (8.3), and using properties of the log and exponential functions,

\[
\exp\{\ln(Y)\} = \exp\{\beta_0 + \beta_1 X + U\},
\]

\[
Y = e^{\beta_0 + \beta_1 x} e^U,
\]

\[
E(Y \mid X = x) = e^{\beta_0 + \beta_1 x} E(e^U \mid X = x).
\]

It is easy to plug in \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), but unless \( E(e^U \mid X) = 1 \), the CEF of \( Y \) given \( X = x \) is not simply \( e^{\beta_0 + \beta_1 x} \). That is, \( e^{\beta_0 + \beta_1 x} \) is generally not the best predictor of \( Y \) given \( X = x \). One alternative is to ignore the problem and do it in logs anyway. Another alternative is to model an exponential CEF directly, although such a CEF is nonlinear-in-parameters (the \( \beta_1 \) is inside the exponential function, which is not linear), and such models are beyond our scope. There are formal “model selection” procedures that can help us choose the best model for prediction, but these are beyond our scope.

### 8.1.3 The Linear-Log Model

**Interpretation**

A **linear-log model** specifies

\[
Y = \beta_0 + \beta_1 \ln(X) + U. \tag{8.4}
\]

Here, \( \beta_1 \) has a different interpretation. When \( X \) increase 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 \( 1.01X \), which is different than a 1 percentage point change in \( X \).

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 \( X \).

However, as noted in Section 8.1.2, the percentage interpretation is inaccurate when considering large changes. For example, a 50\% increase in \( X \) is not associated with a \((\beta_1/100)(50)\) unit change in \( Y \). The reason is \( \ln(1.5X) = \ln(1.5) + \ln(X) = \ln(X) + 0.41 \), not \( \ln(X) + 0.5 \).

Instead, with nonlinear models, it is most reliable to just plug in two values of \( X \). For example, to see the change associated with increasing \( X = 40 \) to \( X = 60 \) (a 50\% increase), just plug in and subtract:

\[
\beta_0 + \beta_1 \ln(60) - [\beta_0 + \beta_1 \ln(40)] = (\beta_0 - \beta_0) + \beta_1 [\ln(60) - \ln(40)] = \beta_1 \ln(60/40) = \beta_1 \ln(1.5) = 0.41\beta_1.
\]
In fact, the same result holds for any 50% increase in $X$ since $\ln(1.5X) - \ln(X) = \ln(1.5X/X) = \ln(1.5)$. So the model always associates a 0.41$\beta_1$ change in $Y$ with a 50% increase in $X$, regardless of the starting value of $X$.

**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 even 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 (on the margin).

**8.1.4 The Log-Log Model**

**Interpretation**

A **log-log model** specifies

$$\ln(Y) = \beta_0 + \beta_1 \ln(X) + U. \quad (8.5)$$

Here, a 1% increase in $X$ is associated with an approximate $\beta_1\%$ change in $Y$. Again, if the percentages are too large, then the approximation is poor. However, the percentage interpretation is particularly nice here: $\beta_1$ represents an **elasticity** of $Y$ with respect to $X$. (Hopefully you remember the concept of elasticity from microeconomics.)

**When to Use It**

There are a few reasons to try a log-log model. First, as noted, it’s a simple way to get an elasticity interpretation. Second, like before, a scatterplot of $\ln(Y)$ against $\ln(X)$ may look roughly linear. Third, the log-log model implies a power law type of relationship between $Y$ and $X$. Exponentiating both sides of (8.5),

$$\exp\{\ln(Y)\} = \exp\{\beta_0 + \beta_1 \ln(X) + U\},$$

$$Y = e^{\beta_0} \exp\{\ln(X)\} e^{U} = e^{\beta_0} X^{\beta_1} e^{U}. \quad (8.6)$$

**Issue with Prediction**

However, as with the log-linear model, $e^{\beta_0} X^{\beta_1}$ is generally not the CEF, since $E(e^{U} \mid X) = 1$ is not implied by $E(U \mid X) = 0$. Consequently, predicting $Y$ as $e^{\beta_0} X^{\beta_1}$ is not necessarily the best.
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

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. 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 $Y$ (in levels) as $X$ increases. This pattern does not come from the data, but from the model itself, regardless of the data.

This self-fulfilling prophecy issue is true of any model with only a couple parameters. The comic at https://xkcd.com/2048 makes this point, as does Figure 8.2. 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 always imposes a constant effect of $X$ on $Y$, which is $\beta_1$. 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 discussed in Section 8.1.3. Clearly, the selection of a model is very important; there are formal “model selection” procedures, but they are beyond our scope.

8.1.6 Code

The following code compares linear, log-linear, linear-log, and log-log estimation given the same dataset. The four fitted functions are plotted on four copies of the same scatterplot in Figure 8.2 in homage to https://xkcd.com/2048. The results illustrate the concerns of Section 8.1.5.

```
par(family=PARFAM, mar=c(3,3,1,1), mgp=PARMGP, mfrow=c(2,2))
set.seed(112358)
n <- 31
X <- sort(runif(n=n,min=1,max=9))
df <- data.frame(X=X, Y=1+pnorm(q=X,mean=5,sd=1.5) +
                 2*(rbeta(n=n,shape1=10-X,shape2=X) - (10-X)/10 ) )
```
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, cex=1, bty='l',
     xlab=XL, ylab=YL, main='',
     cex.axis=CEXAXIS, cex.lab=CEXLAB)
lines(predict(ret.linlin)~df$X, col=2, lwd=LWD)
title("Linear", line=-1, adj=0.1)

# plot(x=df$X, y=df$Y, type='p', pch=16, cex=1, bty='l',
#     xlab=XL, ylab=YL, main='',
#     cex.axis=CEXAXIS, cex.lab=CEXLAB)
lines(predict(ret.loglin)~df$X, col=2, lwd=LWD)
title("Linear-Log", line=-1, adj=0.1)

# plot(x=df$X, y=df$Y, type='p', pch=16, cex=1, bty='l',
#     xlab=XL, ylab=YL, main='',
#     cex.axis=CEXAXIS, cex.lab=CEXLAB)
lines(exp(predict(ret.loglog))~df$X, col=2, lwd=LWD)
title("Log-Linear", line=-1, adj=0.1)

# plot(x=df$X, y=df$Y, type='p', pch=16, cex=1, bty='l',
#     xlab=XL, ylab=YL, main='',
#     cex.axis=CEXAXIS, cex.lab=CEXLAB)
lines(exp(predict(ret.loglog))~df$X, col=2, lwd=LWD)
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 with model \( Y = \gamma_0 + \gamma_1 X + V \) and estimate the function \( \hat{\gamma}_0 + \hat{\gamma}_1 X \). 

a) Approximately what value would you expect \( \hat{\gamma}_1 \) to be? b) What does \( \hat{\gamma}_0 + \hat{\gamma}_1 X \) suggest about the causal relationship between \( X \) and \( Y \)? What features are similar or different compared to the true \( 1 + X^2 \)? (Hint: draw a picture.)
Figure 8.2: Same data, different models.
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. OLS can still be used for estimation as long as the function is “linear-in-parameters.” Again, the distinctions among causal, CEF, and linear projection models are not emphasized here.

There are two types of (non)linearity. They are different, but 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

$$w_1A + w_2B,$$

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 $w_1A + w_2B + w_3C + w_4D$. 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:

$$w_1Y_1 + w_2Y_2 + w_3Y_3 + w_4Y_4 = \sum_{i=1}^{4} w_i Y_i.$$

For example, the expected value formula for discrete random variables is a special case of a linear combination, where the linear combination weights are the probabilities of the different possible values. Also, the sample average is a linear combination of observed $Y_i$ values, with weights $w_i = 1/n$.

The function $\beta_0 + \beta_1X$ is **linear-in-parameters**. That is, it is a linear combination of the parameters $\beta_0$ and $\beta_1$. The weights are $w_1 = 1$ and $w_2 = X$, so the linear combination is

$$w_1\beta_0 + w_2\beta_1 = (1)(\beta_0) + (X)(\beta_1) = \beta_0 + \beta_1X.$$

The function $\beta_0 + \beta_1X$ is also **linear-in-variables**. This is trickier to see since it is not actually a linear combination of $X$ alone. (The function $\beta_0 + \beta_1X$ 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 has weights $w_1 = \beta_0$ and $w_2 = \beta_1$, so the linear combination is

$$(w_1)(X_0) + (w_2)(X) = (\beta_0)(1) + (\beta_1)(X) = \beta_0 + \beta_1X.$$

For this reason, in economics, people often call $\beta_0 + \beta_1X$ “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 condition expectation functions or linear projection. For example, if the CEF is \( E(Y \mid X = x) = \beta_0 + \beta_1 x \), then the CEF is linear-in-parameters and linear-in-variables. Regardless of the CEF, the linear projection of \( Y \) onto \((1, X)\) is 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 linear CEF model in error form: \( Y = \beta_0 + \beta_1 X + U \) with \( E(U \mid X) = 0 \). Despite the +\( U \) at the end, sometimes people say this is linear-in-variables and linear-in-parameters, presumably because the corresponding CEF \( E(Y \mid X = x) = \beta_0 + \beta_1 x \) indeed satisfies both types of linearity. Similarly, consider the linear projection model in error form: \( Y = \beta_0 + \beta_1 X + U \) with \( E(U) = E(XU) = 0 \). Again, despite the +\( U \) at the end, sometimes people say this is linear-in-variables and linear-in-parameters, presumably because the corresponding linear projection \( \text{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X \) 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; they are just ignoring the +\( U \) part. One possible rationale is that if additionally \( E(U \mid X) = 0 \), then the “linear” structural model implies a linear CEF, so perhaps the terminology is rooted in optimism. Another possible rationale is that we could write \( Y = \beta_0 + \beta_1 X + U \) as \( Y = h(X) + U \) in which \( h(X) \) is an affine function of \( X \), which we clever economists call “linear.” 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,

\[
Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U \quad (8.9)
\]

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 \( \beta_0 \) and \( \beta_1 \). However, \( (8.9) \) is still linear-in-parameters. That is, with linear combination weights, 1, \( X \), and \( X^2 \),

\[
(1)(\beta_0) + (X)(\beta_1) + (X^2)(\beta_2) = \beta_0 + \beta_1 X + \beta_2 X^2.
\]

A nonlinear-in-parameters model cannot be written as a linear combination of the parameters. One example is the power law model alluded to in \( (8.6) \) in Section 8.1:

\[
Y = \beta_0 X^{\beta_1} + U. \quad (8.10)
\]

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 much in this text.
8.2.3 Estimation and Inference

Returning to the quadratic model in (8.9): how can the parameters be estimated? OLS can actually estimate nonlinear-in-variables models, as long as they are linear-in-parameters, like (8.9).

Inference on parameters is also the same. For example, the same R code to compute a confidence interval for $\beta_1$ earlier will still work, 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 basically the same, interpretation of parameters changes greatly with nonlinear-in-variables models.

Insufficiency of Linear Coefficient

In (8.9), $\beta_1$ 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. (If you’re familiar with the calculus, then you can see this from the derivative of $\beta_0 + \beta_1 X + \beta_2 X^2$ with respect to $X$: it equals $\beta_1 + 2X\beta_2$, which is not constant.) Not only does the amount of change in $Y$ associated with a unit increase in $X$ now depend on $X$, 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

Although $\beta_1$ no longer alone describes the changes in $Y$ given a unit increase in $X$, there are other ways to summarize the relationship. With only one $X$, the best way 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...). In that case, one simple approach is to plug in changes of $X$ that are relevant to policy or whatever economic question is being studied. For example, if $Y$ is income, $X$ is education, and we want to understand the value of the 12th year of education, then only comparing $X = 12$ to $X = 11$ is relevant. Then, we can just compute

$$[\beta_0 + \beta_1(12) + \beta_2(12)^2] - [\beta_0 + \beta_1(11) + \beta_2(11)^2] = \beta_1(12 - 11) + \beta_2(12^2 - 11^2) = \beta_1 + 23\beta_2.$$ 

Even in nonlinear-in-variables models more complex than the quadratic model, graphing and plugging in values work. Generally, we could write

$$Y = f(X) + U,$$

where in the quadratic model we had $f(X) = \beta_0 + \beta_1X + \beta_2X^2$ as a special case. Once we estimate all the parameters by OLS, we can plug them in to get $\hat{f}(X)$. We can graph this estimated function by just evaluating it at many $X$ values and drawing a line through them. Similarly, to get the predicted change in $Y$ for a change from $X = x_1$ to $X = x_2$, just plug in: $\hat{f}(x_2) - \hat{f}(x_1)$.

**Discussion Question 8.3** (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 8.4** (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. However, the concepts remain essentially the same.
8.2. NONLINEAR-IN-VARIABLES REGRESSION

Description and Prediction

For example, consider the quadratic model from (8.9) 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:

$$\text{LP}(Y \mid 1, X, X^2) = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$= \arg \min_{a + bX + cX^2} \{E[(Y - a - bX - cX^2)^2]\}. \quad (8.11)$$

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 (7.12) and (7.13),

$$\text{LP}(Y \mid 1, X, X^2) = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$= \frac{\text{Bla}}{\text{arg \ min}_{a + bX + cX^2} E\{[E(Y \mid X) - (a + bX + cX^2)]^2\}}$$

$$= \frac{\text{Blp}}{\text{arg \ min}_{a + bX + cX^2} E\{[Y - (a + bX + cX^2)]^2\}}. \quad (8.12)$$

As before, if the true CEF actually is quadratic, the these all equal the true CEF.

Structural Identification

For a structural interpretation of OLS estimates, the key is a properly specified CEF and some sort of independence condition. (Mathematically, we could use conditions like $\text{Cov}(X^3, U) = 0$, but those are hard to think about in real-world examples.) The identification arguments in Section 6.6.3 for a binary $X$ generalize to any type of $X$. A sufficient independence condition is the same as before: $X$ is independent of the structural error term $U$, which combines the effects of all other causal determinants of $Y$. Like before, for ASE identification this can be weakened to “mean independence,” where the conditional mean $E(U \mid X)$ does not depend on $X$.

Given (mean) independence, the ASE and CEF are connected. Specifically, the ASE on $Y$ of changing $X$ from $x_1$ to $x_2$ equals the difference in the CEF at those two points:

$$\text{ASE}(x_1 \rightarrow x_2) = E(Y \mid X = x_2) - E(Y \mid X = x_1) \equiv m(x_2) - m(x_1). \quad (8.13)$$

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

\[
\text{ASE}(x_1 \to x_2) = \hat{m}(x_2) - \hat{m}(x_1) = \hat{\beta}_0 + \hat{\beta}_1 x_2 + \hat{\beta}_2 x_2^2 - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2) = \beta_1(x_2 - x_1) + \beta_2(x_2^2 - x_1^2).
\]  

(8.14)

8.2.6 Other Nonlinear Models

The quadratic model is a special case of a polynomial model, where the right-hand side is a polynomial in \( X \) (plus the error term \( U \)). For example, a cubic model is

\[
Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + U.
\]  

(8.15)

In principle, more and more polynomial terms could be added:

\[
Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \cdots + U.
\]  

(8.16)

The more terms, the more flexible the model, reducing the amount of misspecification. This is good, especially if we care about causal effects: we need a properly specified CEF. However, models can get “too big.” This question of “how many terms” can be decided by formal model selection procedures; see Section 8.3.

Non-polynomial models are also fine. Section 8.1 described various models with log transforms. Less commonly, something like

\[
Y = \beta_0 + \beta_1 \cos(X) + \beta_2 \sin(X) + \beta_3 \cos(2X) + \beta_4 \sin(2X) + U
\]  

(8.17)

could be used if some periodic behavior is suspected, perhaps after rescaling to have \( 0 \leq X \leq 2\pi \). We could replace the transformations of \( X \) with other trigonometric functions, or non-integer powers of \( X \), or polynomials in \( \ln(X) \), or step functions; the possibilities seem endless. It is helpful when there is some guidance from economic theory, like with the Mincer equation, but often economic theory provides only qualitative guidance. For example, economic theory tells us that “demand curves are downward sloping,” but not whether they are linear, quadratic, log-log, etc.

8.2.7 Examples

The following code 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.

\begin{verbatim}
par(family=PARFAM, mar=c(3,3,1,1), mgp=PARMGP, mfrow=c(2,2))
set.seed(112358)
n <- 31
X <- sort(3*rbeta(n=n,shape1=1,shape2=1))
df <- data.frame(X=X, Y=1+10*(X/2-0.5)^2*(X/2-0.5-1) +
                 rnorm(n=n))
ret.poly1 <- lm(Y~X, data=df)
ret.poly2 <- lm(Y~X+I(X^2), data=df)
ret.poly3 <- lm(Y~X+I(X^2)+I(X^3), data=df)
ret.trig <- lm(Y~I(cos(2*pi*(X-0)/3))+I(sin(2*pi*(X-0)/3)),
                data=df)

# XL <- ''; YL <- ''
plot(x=df$X, y=df$Y, type='p', pch=16, cex=1, bty='l',
     xlab=XL, ylab=YL,
     main='', cex.axis=CEXAXIS, cex.lab=CEXLAB, xlim=c(0,3))
lines(predict(ret.poly1)~df$X, col=2, lwd=LWD)
title("Linear",line=-1,adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, cex=1, bty='l',
     xlab=XL, ylab=YL,
     main='', cex.axis=CEXAXIS, cex.lab=CEXLAB, xlim=c(0,3))
lines(predict(ret.poly2)~df$X, col=2, lwd=LWD)
title("Quadratic",line=-1,adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, cex=1, bty='l',
     xlab=XL, ylab=YL,
     main='', cex.axis=CEXAXIS, cex.lab=CEXLAB, xlim=c(0,3))
lines(predict(ret.poly3)~df$X, col=2, lwd=LWD)
title("Cubic",line=-1,adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, cex=1, bty='l',
     xlab=XL, ylab=YL,
     main='', cex.axis=CEXAXIS, cex.lab=CEXLAB, xlim=c(0,3))
lines(predict(ret.trig)~df$X, col=2, lwd=LWD)
title("Trigonometric",line=-1,adj=0.1)
\end{verbatim}
Figure 8.3: Same data, different models.
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 assumed to have a particular (nonlinear) functional form, like a cubic polynomial, and only the coefficient values are unknown.

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. Generally, as it sounds, model selection just means 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 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 \( V_i \) 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. 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 or structural 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 \texttt{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 \texttt{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=PARFAM, mar=c(3,3,1,1), mgp=PARMGP, 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)
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, cex=1, xlab='',
ylab='', main='', cex.axis=CEXAXIS, cex.lab=CEXLAB,
xlim=0:1, ylim=0:1*3.04)
  lines(x=xx, y=CEF(xx), col=1, lwd=1)
  lines(predict(rets[[ifig]], x=xx), col=2, lwd=LWD)
  title(main=titles[ifig], line=-1, adj=0.1)
}
```
8.3. NONPARAMETRIC REGRESSION

Discussion Question 8.5 (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 Exercises

**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 with R command `library(wooldridge)` or Stata command `bcuse lawsch85, nodesc clear` (assuming you’ve already installed that R package or Stata command)

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 command `df <- data.frame(Y=lawsch85$salary, X=lawsch85$cost)` and then `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 `retlm <- 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(retlm, 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')`
EMPIRICAL EXERCISES

h. R only: plot the three estimated CEFs as lines over the scatterplot with

```r
lines(x=xx, y=fitlm, col=1, lty=1)
lines(x=xx, y=fitnl, col=2, lty=5)
lines(fitss, col=4, lty=3)
```

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); you can specify logs in R with `df <- data.frame(Y=log(lawsch85$salary), X=log(lawsch85$rank))` and generate log rank in Stata with `generate lrank = log(rank)`

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 with R command `library(wooldridge)` or Stata command `bcuse sleep75, nodesc clear` (assuming you’ve already installed that R package or Stata command)

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 commands `df <- data.frame(Y=sleep75$slpnaps/7/6/zero.alt3, X=sleep75$hrwage)` and `df <- df[!(is.na(df$Y) | is.na(df$X)), ]` 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 `generate sleephrsdaily = slpnaps/7/60`

d. Stata only: graph linear, quadratic, and nonparametric (local linear) CEF estimates similar to part (b) with command `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; 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); see https://en.wikipedia.org/wiki/Prediction_market
for more. 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:

```r
library(wooldridge)
library(sandwich)
library(lmtest)
?pntsprd
```

b. Stata only: load the data with `bcuse pntsprd , nodesc clear` (assuming `bcuse`
is already installed)

c. For each observation (each game), compute whether the actual score difference
was over, under, or equal to the spread, with R command `overunder <- sign(pntsprd$scrdiff-pntsprd$spread)` or Stata command `generate overunder = sign(scrdiff - spread)` noting that in math the “sign” function (not to be
confused with “sine”) equals +1 for strictly positive values, −1 for strictly negative
values, and 0 for zero

d. Display the frequency of over, under, and equal with R command `table(overunder , useNA='ifany')` or Stata command `tabulate overunder , missing`
e. Regress the score difference on the spread with R command `ret <- lm(scrdiff ~ spread, data=pntsprd)` or Stata command `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) with R command `plot(smooth.spline(x=pntsprd$spread, y=pntsprd$scrdiff))` followed by `abline(a=0, b=1, col=2)` or Stata command `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

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): https://www.youtube.com/watch?v=04EjBeKDE2o
- Potential outcomes and CATE (Masten video): https://www.youtube.com/watch?v=2CSSwKFE7iQ
Perhaps surprisingly, there is a lot to think about with even just two binary regressors. We’ll talk about (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

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 follows a child ($X$). 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

GHOST.1 the ghost ($Q$) follows the child ($X$) and

GHOST.2 the ghost ($Q$) causes a mess ($Y$).

The child is the included regressor. The ghost is the omitted variable. The parents over-estimating how much mess the child causes is OVB.

9.1.2 Formal Conditions

The ghost of OVB can be formalized as follows. Consider the structural model

$$Y = \beta_0 + \beta_1 X + \beta_2 Q + V,$$

(9.1)

where Cov($X, V$) = 0. If we don’t observe $Q$, then instead we have the structural model

$$Y = \beta_0 + \beta_1 X + U, \quad U \equiv \beta_2 Q + V.$$

(9.2)

Here, $X$ is called the included regressor (included in the model; not omitted). If $X$ is binary, then for OLS to estimate $\beta_1$ requires $E(U \mid X = 0) = E(U \mid X = 1)$: 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 $X = 0$. That is, $Q$ 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, i.e., 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 $\beta_2 \neq 0$. Again, 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 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 $Y$.

The variable $Q$ may be called an omitted variable or a confounder.

**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 $X = 1$ if classroom size is above 24 students and $X = 0$ otherwise. Let $Y$ denote the annual earnings of an individual at age 30. 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 (the year after kindergarten in the U.S.) class size. 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 Condition OVB.2. However, if students are just randomly assigned to classes each year, Corr($X, Q$) $= 0$. In that case, Condition OVB.1 does not hold, so this $Q$ would not cause OVB. However, if students are not randomly assigned to classes, then Condition OVB.1 may actually hold; e.g., if “better” students are assigned to smaller classes, and if kindergarten students in smaller classes appear “better,” then students with $X = 0$ are more likely to again have $Q = 0$, and students in large kindergarten classes ($X = 1$) are more likely to remain in large classes ($Q = 1$).

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 Corr($Q, X$) $> 0$, satisfying Condition 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, Condition 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 Corr(Q, X) < 0, satisfying Condition 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, meaning Q is a causal determinant of Y, satisfying Condition OVB.2. Thus, omitting this Q causes OVB.

**Discussion Question 9.1** (assessing OVB). Among public schools 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 each 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.  
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

**Formulas**

OVB can be quantified more precisely. Let \( Y = \beta_0 + \beta_1 X + U \) be the structural model, where U is the structural error term. Then, the OLS estimator of \( \beta_1 \) has the property

\[
\hat{\beta}_1 \overset{p}{\to} \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)} \sqrt{\frac{\text{Var}(U)}{\text{Var}(X)}},
\]

also written

\[
\text{plim}_{n \to \infty} \hat{\beta}_1 = \beta_1 + \text{Corr}(X, U) \sqrt{\frac{\text{Var}(U)}{\text{Var}(X)}}.
\]

That is, for large samples (large \( n \)), the estimator \( \hat{\beta}_1 \) is very close to the right-hand side expression in almost all randomly sampled datasets (or, “with high probability”). (To review \( \overset{p}{\to} \) and consistency, see Section [3.4.4](#)). OVB is not solved by having lots of data. Unless Corr(X, U) = 0, the OLS estimator is not consistent for the structural \( \beta_1 \).

Technically, we are not talking about “bias,” but the practical implication is the same. (Bias and consistency were introduced in Sections [3.4.1](#) and [3.4.4](#)). The bias of \( \hat{\beta}_1 \) is

\[
\text{Bias}(\hat{\beta}_1) \equiv E(\hat{\beta}_1) - \beta_1.
\]

We are instead looking at the bias of the asymptotic distribution...
of $\hat{\beta}_1$ (sometimes called asymptotic bias),

$$\lim_{n \to \infty} \hat{\beta}_1 - \beta_1 = \text{Corr}(X, U) \sqrt{\frac{\text{Var}(U)}{\text{Var}(X)}}. \quad (9.5)$$

Although these are very different mathematically, we won’t worry about such technicalities in this book. The practical problem with OVB is that we systematically over-estimate or under-estimate the true structural parameter.

Direction of Asymptotic Bias

(Recall terms and definitions from Sections 3.4.1 and 3.4.4.)

The direction (+ or −) of the asymptotic bias in (9.5) depends only on $\text{Corr}(X, U)$ since the $\sqrt{\cdot}$ part is always positive. (For simplicity, below just says “bias” instead of “asymptotic bias”; although their mathematical definitions differ, the qualitative ideas are the same.) If $\text{Corr}(X, U) > 0$, then $\text{Corr}(X, U) \sqrt{\text{Var}(U)/\text{Var}(X)} > 0$, so $\lim_{n \to \infty} \hat{\beta}_1 - \beta_1 > 0$. This is upward (positive) 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 upward bias causes $\lim_{n \to \infty} \hat{\beta}_1 = 0$. This is upward (positive) bias since $0 > -9$, 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 $\text{Corr}(X, U) \sqrt{\text{Var}(U)/\text{Var}(X)} < 0$, so $\lim_{n \to \infty} \hat{\beta}_1 - \beta_1 < 0$, meaning downward (negative) bias. Again confusing, downward bias can actually make effects look bigger, e.g., if $\beta_1 = 0$ and $\lim_{n \to \infty} \hat{\beta}_1 = -9$: the true effect is zero, but the downward bias makes it appear like there is an effect.

Results in Terms of $Q$

The results in terms of $U$ can be translated to $Q$, too. Like before, let $U = \beta_2 Q + V$, with $\text{Cov}(X, V) = 0$. Then,

$$\text{Cov}(X, U) = \text{Cov}(X, \beta_2 Q + V) = \beta_2 \text{Cov}(X, Q) + \text{Cov}(X, V). \quad (9.6)$$

Plugging this into the first expression in (9.3),

$$\hat{\beta}_1 \overset{p}{\to} \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)} = \beta_1 + \frac{\text{Cov}(X, Q)}{\text{Var}(X)} = \beta_1 + \beta_2 \text{Corr}(X, Q) \sqrt{\frac{\text{Var}(Q)}{\text{Var}(X)}}. \quad (9.7)$$

Note that $\text{Cov}(X, Q)/\text{Var}(X)$ equals the slope of the population linear projection of $Q$ onto $X$ (and an intercept). We can also see why both Conditions OVB.1 and OVB.2 are required for OVB. Condition OVB.1 says $\text{Corr}(X, Q) \neq 0$, while Condition OVB.2 says
\[ \beta_2 \neq 0 \text{ if either } \beta_2 = 0 \text{ or Corr}(X, Q) = 0 \text{ in (9.7), then } \beta_2 \text{ Corr}(X, Q) = 0, \text{ and the bias disappears: } \hat{\beta}_1 \overset{p}{\rightarrow} \beta_1. \]

The direction of bias can also be interpreted in terms of \( Q \). Using (9.7), the sign of the 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) bias; if \( \beta_2 \text{ Corr}(X, Q) < 0 \), then there is negative (downward) bias.

**Example**

Consider the 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 less than \( \beta_1 \). 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 average causal effect of changing from a smaller class size (\( X = 0 \)) to larger class size (\( X = 1 \)) is lower future earnings (\( \beta_1 < 0 \)). Negative bias means we estimate something even more negative: \( \text{plim}_{n \rightarrow \infty} \hat{\beta}_1 < \beta_1 < 0 \). This makes the size of the effect appear larger than it really is: we estimate something farther away from zero.

**Discussion Question 9.2** (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: \( \beta_1 \), Corr(\( X, Q \)), \( \beta_2 \), OVB. Also: will our estimated effect \( \hat{\beta}_1 \) tend to be larger or smaller than the true effect \( \beta_1 \), and why?

**Discussion Question 9.3** (OVB: ES habits). Recall 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 bias would your omitted variable cause? Explain.

**9.1.4 OVB in Linear Projection**

Outside of a causal model, we can still define omitted variable bias. For linear projection, the formula is actually the same as (9.7), just with different interpretation of \( \beta_1 \) and \( \beta_2 \). Similar results for larger linear projection models are in Hansen (2018, §2.23), 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 structural and causal estimation, not prediction.
9.2 Linear-in-Variables Model

How should we write a CEF regression model with 2 binary regressors? With a single binary regressor $X_1$, our CEF model was

$$E(Y \mid X_1) = \beta_0 + \beta_1 X_1.$$ 

With a second binary regressor $X_2$, it seems natural to write

$$E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \quad (9.8)$$

This is the linear-in-variables model.

Misspecification

Is it possible for (9.8) to be misspecified? Recall that the CEF $\beta_0 + \beta_1 X$ could be misspecified when $X$ had three possible values: then there were three CEF values, but 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:

$$m(0, 0) = E(Y \mid X_1 = 0, X_2 = 0), \quad m(0, 1) = E(Y \mid X_1 = 0, X_2 = 1),$$

$$m(1, 0) = E(Y \mid X_1 = 1, X_2 = 0), \quad m(1, 1) = E(Y \mid X_1 = 1, X_2 = 1). \quad (9.9)$$

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

$$m(0, 0) = \beta_0 + (\beta_1)(0) + (\beta_2)(0) = \beta_0, \quad (9.10)$$

$$m(0, 1) = \beta_0 + (\beta_1)(0) + (\beta_2)(1) = \beta_0 + \beta_2, \quad (9.11)$$

$$m(1, 0) = \beta_0 + (\beta_1)(1) + (\beta_2)(0) = \beta_0 + \beta_1, \quad (9.12)$$

$$m(1, 1) = \beta_0 + (\beta_1)(1) + (\beta_2)(1) = \beta_0 + \beta_1 + \beta_2. \quad (9.13)$$

Consequently, $\beta_1$ has two interpretations. It equals either (9.13) minus (9.11), or (9.12) minus (9.10):

$$m(1, 1) - m(0, 1) = (\beta_0 + \beta_1 + \beta_2) - (\beta_0 + \beta_2) = \beta_1,$$

$$m(1, 0) - m(0, 0) = (\beta_0 + \beta_1) - \beta_0 = \beta_1.$$ 

Thus, the model assumes $m(1, 1) - m(0, 1) = m(1, 0) - m(0, 0)$, which may not be true of the real CEF. For example, if $m(0, 0) = 0$, $m(1, 0) = 1$, $m(0, 1) = 2$, and $m(1, 1) = 4$, then $m(1, 1) - m(0, 1) = 2$, but $m(1, 0) - m(0, 0) = 1$: they are not equal. In that case, the CEF model in (9.8) is misspecified (wrong).

As discussed in Chapter 7, if the CEF model is wrong, then OLS estimates the linear projection. Here, OLS estimates $LP(Y \mid 1, X_1, X_2)$. However, this is not useful for causality, and the misspecification is easily fixed.
9.3. **FULLY SATURATED MODEL**

More Consideration

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 (instead of high-experience). 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

To avoid misspecification without the strong assumption in (9.8), an interaction term can be added:

$$E(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$  (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 wage difference associated with education to depend on the value of experience. More generally, interaction terms allow regressors to interact, meaning the change in $Y$ associated with a unit increase in one regressor is allowed to depend on the value of the other regressor.

The CEF model in (9.14) is also called **fully saturated** since it is flexible enough to allow a different CEF value for each value of $(X_1, X_2)$. The fact that there are four possible values of $(X_1, X_2)$ and four $\beta_j$ parameters is necessary but not sufficient for the model to be fully saturated.

Given the fully saturated CEF model in (9.14), each $\beta_j$ can be written in terms of different CEF values. First, similar to (9.10)–(9.13), each CEF value can be written in terms of the $\beta_j$:

$$m(0, 0) = \beta_0 + (\beta_1)(0) + (\beta_2)(0) + (\beta_3)(0)(0) = \beta_0, \quad (9.15)$$
$$m(0, 1) = \beta_0 + (\beta_1)(0) + (\beta_2)(1) + (\beta_3)(0)(1) = \beta_0 + \beta_2, \quad (9.16)$$
$$m(1, 0) = \beta_0 + (\beta_1)(1) + (\beta_2)(0) + (\beta_3)(1)(0) = \beta_0 + \beta_1, \quad (9.17)$$
$$m(1, 1) = \beta_0 + (\beta_1)(1) + (\beta_2)(1) + (\beta_3)(1)(1) = \beta_0 + \beta_1 + \beta_2 + \beta_3. \quad (9.18)$$
From (9.15)–(9.17) and their differences,

\[ \beta_0 = m(0, 0), \]  
\[ \beta_1 = (\beta_0 + \beta_1) - \beta_0 = m(1, 0) - m(0, 0), \]  
\[ \beta_2 = (\beta_0 + \beta_2) - \beta_0 = m(0, 1) - m(0, 0), \]  
\[ \beta_3 = [\beta_2 + \beta_3] - [\beta_2] = [(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1)] - [(\beta_0 + \beta_2) - (\beta_0)] \]

\[ = [m(1, 1) - m(1, 0)] - [m(0, 1) - m(0, 0)]. \]  

Before interpreting each coefficient economically, note that it is mathematically (though not “economically”) equivalent to write

\[ \beta_3 = [(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2)] - [(\beta_0 + \beta_1) - (\beta_0)] \]

\[ = [m(1, 1) - m(0, 1)] - [m(1, 0) - m(0, 0)]. \]  

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.” This special case is explored further in Section 9.7.

### 9.4 CEF Coefficient Interpretation

Using the results in (9.19)–(9.23), the four \( \beta_j \) in (9.14) have the following interpretations.

- \( \beta_0 = m(0, 0) \) is the mean wage among low-education, low-experience individuals; more generally, the mean \( Y \) in the subpopulation with \( X_1 = 0 \) and \( X_2 = 0 \). Note: generally \( \beta_0 \neq \text{E}(Y) \), the unconditional mean.

- \( \beta_1 = m(1, 0) - m(0, 0) \) is the mean wage difference between high-education and low-education individuals within the low-experience subpopulation; more generally, the mean \( Y \) difference between \( X_1 = 1 \) and \( X_1 = 0 \) individuals within the \( X_2 = 0 \) subpopulation. Note: generally \( \beta_1 \neq \text{E}(Y \mid X_1 = 1) - \text{E}(Y \mid X_1 = 0) \); 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, the mean \( Y \) difference between \( X_2 = 1 \) and \( X_2 = 0 \) individuals within the \( X_1 = 0 \) subpopulation. Note: generally \( \beta_2 \neq \text{E}(Y \mid X_2 = 1) - \text{E}(Y \mid X_2 = 0) \); it additionally conditions on \( X_1 = 0 \).

\( \beta_3 = [m(1, 1) - m(1, 0)] - [m(0, 1) - m(0, 0)] \) is the difference between the (predicted or causal) “effect” on mean wage of experience (high vs. low) in the high-education subpopulation and the “effect” in the low-education subpopulation. More generally, \( \beta_3 \) is the difference between the “effect” of \( X_2 \) in the \( X_1 = 1 \) subpopulation and its “effect” in the \( X_1 = 0 \) subpopulation.

Equivalently, \( \beta_3 = [m(1, 1) - m(0, 1)] - [m(1, 0) - m(0, 0)] \) is the difference between the (predicted or causal) “effect” on mean wage of education (college vs. not) in the high-experience subpopulation and the “effect” in the low-experience subpopulation. More generally, \( \beta_3 \) is the difference between the “effect” of \( X_1 \) in the \( X_2 = 1 \) subpopulation and its “effect” in the \( X_2 = 0 \) subpopulation.

The \( \beta_j \) interpretations can also be seen by considering the regression of \( Y \) on \( X_1 \) separately when \( X_2 = 0 \) and \( X_2 = 1 \). That is, plugging in (conditioning on) \( X_2 = 0 \), the CEF becomes

\[
m(x_1, 0) = \beta_0 + \beta_1 x_1 + (\beta_2)(0) + (\beta_3)(x_1)(0) = \beta_0 + \beta_1 x_1. \tag{9.24}
\]

Plugging in \( X_2 = 1 \), the CEF is instead

\[
m(x_1, 1) = \beta_0 + \beta_1 x_1 + (\beta_2)(1) + (\beta_3)(x_1)(1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1. \tag{9.25}
\]

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. Thinking of the slope on \( X_1 \) as the (predicted or causal) “effect” of \( X_1 \) on \( Y \), the interaction coefficient \( \beta_3 \) describes how this effect of \( X_1 \) 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 “effect” of \( X_2 \) when \( X_1 = 1 \) versus when \( X_1 = 0 \):

\[
m(0, x_2) = \beta_0 + (\beta_1)(0) + \beta_2 x_2 + (\beta_3)(0)(x_2) = \beta_0 + \beta_2 x_2, \tag{9.26}
\]

\[
m(1, x_2) = \beta_0 + (\beta_1)(1) + \beta_2 x_2 + (\beta_3)(1)(x_2) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x_2. \tag{9.27}
\]

**Discussion Question 9.4** (binary interaction). Let \( Y \) be wage (\$/hr), \( D_1 = 1 \) if an individual has a college degree (\( D_1 = 0 \) if not), and \( D_2 = 1 \) if an individual has more than 15 years of experience (and \( D_2 = 0 \) if not). You have a sample of data and run OLS on the fully saturated model (including the interaction term), yielding \( \hat{Y} = 10 + 5D_1 + D_2 + 2D_1D_2 \). a) For the college-educated subpopulation, what is the estimated change
in average wage associated with changing from low to high experience? b) Within the low-experience subpopulation, what’s the estimated difference in average wage between the college and no-college subpopulations? c) What does the 2 mean? (The coefficient on \(D_1D_2\).)

### 9.5 Structural Identification by Exogeneity

Imagine \(Y\) is determined by the linear-in-parameters structural model

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + U. \tag{9.28}
\]

Here, \(U\) has economic meaning (other factors that determine \(Y\)), but generally we don’t know if it also has the statistical properties of a CEF or LP error.

The qualitative condition for identification is the same as in Section 6.6.1. Specifically, if we have some reason to believe that the other factors in \(U\) are unrelated to the regressors, then the structural parameters are identified. Recall that a regressor unrelated to \(U\) is called exogenous; otherwise, it’s endogenous.

In this case, there are multiple ways to mathematically define “unrelated” (or exogenous) that each lead to identification. For example, if “unrelated” means

\[
\text{Cov}(U, X_1) = \text{Cov}(U, X_2) = \text{Cov}(U, X_1 X_2) = 0, \tag{9.29}
\]

then \(\beta_1, \beta_2, \text{ and } \beta_3\) are the linear projection slope coefficients from the linear projection of \(Y\) onto \((1, X_1, X_2, X_1 X_2)\). Since linear projection coefficients are uniquely determined by the distribution of \((Y, X_1, X_2)\), so are these structural coefficients. (If \(E(U) = 0\), too, then \(\beta_0\) is also the linear projection intercept, although often intercepts don’t have much economic importance.) If “unrelated” means \(U \perp X_1, X_2\), then it logically implies \(\text{Cov}(U, X_1) = \text{Cov}(U, X_2) = \text{Cov}(U, X_1 X_2) = 0\). Another possible meaning of “unrelated” is if \(E(U \mid X_1, X_2)\) is a constant (not dependent on \(X_1\) or \(X_2\)), which also implies \(\text{Cov}(U, X_1) = \text{Cov}(U, X_2) = \text{Cov}(U, X_1 X_2) = 0\).

In this case where the structural \(\beta_1, \beta_2, \text{ and } \beta_3\) are linear projection coefficients, they can be estimated by OLS. Put differently: OLS can (usually) consistently estimate linear projection coefficients, and the identifying assumption like \(E(U \mid X_1, X_2)\) being constant allows us to interpret the linear projection coefficients as structural parameters from \(9.28\).

### 9.6 Identification by Conditional Independence

Here, we focus on average structural effects.
9.6. IDENTIFICATION BY CONDITIONAL INDEPENDENCE

9.6.1 Setup

The average structural effect (ASE) definition of Section 6.6.3 can be extended to our model with two (instead of one) binary regressors. Imagine we only care about learning the effect of \( X_1 \) on \( Y \), i.e., \( X_1 \) is the regressor of interest. In contrast, \( X_2 \) is a control variable (or “conditioning variable”); we don’t care about its effect, but it helps us identify the ASE for \( X_1 \). Extending the structural model (6.20) in Section 6.6.3, let

\[
Y = h(X_1, X_2, U),
\]  

(9.30)

where \( U = (U_1, U_2, \ldots) \) is a vector containing all the unobserved causal determinants of \( Y \) besides \( X_1 \) and \( X_2 \). Extending (6.21), the structural effect of \( X_1 \) on \( Y \) may now depend on both \( U \) and \( X_2 \):

\[
S(X_2, U) \equiv h(1, X_2, U) - h(0, X_2, U).
\]  

(9.31)

Similar to before, the ASE of \( X_1 \) on \( Y \) is the expected value of the structural effect. That is, averaging over the population distribution of \((X_2, U)\) yields

\[
ASE = E[S(X_2, U)].
\]  

(9.32)

Alternatively, we could fix the value of the observed \( X_2 \) at \( X_2 = x_2 \) and only average over the unobserved \( U \). To distinguish this from (9.32), this is called the conditional average structural effect (CASE), although elsewhere (e.g., Hansen, 2018) the “conditional” part of the name is not used. The CASE is a function of \( x_2 \):

\[
\text{CASE}(x_2) = E[S(X_2, U) \mid X_2 = x_2].
\]  

(9.33)

With two binary regressors, \( \text{CASE}(1) \) is the ASE of \( X_1 \) on \( Y \) within the subpopulation with \( X_2 = 1 \). Similarly, \( \text{CASE}(0) \) is the ASE of \( X_1 \) on \( Y \) within the \( X_2 = 0 \) subpopulation. Thus, to identify the CASE, the same ASE identification arguments from Section 6.6.3 can be applied separately to the \( X_2 = 0 \) and \( X_2 = 1 \) subpopulations. Specifically, with CEF \( m(x_1, x_2) = E[Y \mid X_1 = x_1, X_2 = x_2] \), the CASEs are identified as

\[
\text{CASE}(0) = m(1, 0) - m(0, 0), \quad \text{CASE}(1) = m(1, 1) - m(0, 1).
\]  

(9.34)

The overall ASE is a function of the CASEs. Specifically, it is the expected value of the CASE, averaging the CASEs according to their probabilities:

\[
ASE = E[\text{CASE}(X_2)].
\]  

(9.35)

If we know all the CASEs and the distribution of \( X_2 \), then we can learn the ASE. However, the opposite is not true: knowing the ASE is not sufficient to learn all the CASEs.
9.6.2 Formal Results

Although the details are beyond our scope, it turns out that the CASEs and ASE are identified under an extension of independence called conditional independence. Previously, with a single \( X \), we saw \( X \perp \perp U \) is sufficient to identify the ASE. With multiple regressors, this generalizes to the conditional independence assumption (CIA). Assumption A9.1 states the CIA for the current structural model, as well as a version in potential outcomes notation. Because of its importance, conditional independence has other names, like unconfoundedness, selection on observables, and ignorability. For more, see Imbens and Wooldridge (2007, p. 6) and references therein.

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 \( Y^1 \) and \( Y^0 \) are the treated and untreated potential outcomes, treatment is independent of the potential outcomes, conditional on the control variable: \( (Y^0, Y^1) \perp \perp X_1 \mid X_2 \). In either case, \( X_2 \) may be replaced by multiple control variables, \( X_2, X_3, X_4, \ldots \).

It is nice that A9.1 is weaker than A6.2 but still strong enough to identify structural or causal effects. Recall that “weaker than” means that A9.1 is true whenever A6.2 is, and there are additional cases where A9.1 is true even though A6.2 is false. So, A9.1 is true in more real-world examples than A6.2.

Theorem 9.1 (ASE Identification). Given A9.1, if \( X_1 \) and \( X_2 \) are binary, then the CASEs and ASE of \( X_1 \) on \( Y \) are identified as in \((9.34)\) and \((9.35)\).

9.6.3 Potential Outcomes and CIA

Consider an example. Let \( Y \) be a student’s Gaokao score, the Chinese college entrance exam (like the SAT or ACT in the U.S.). The maximum score (in a common scoring system) is 750; even 700 is very good. Let \( X_1 = 1 \) if the student took a special program that taught healthy eating, sleep, and stress-management skills, and \( X_1 = 0 \) otherwise. This \( X_1 \) is the “treatment.” Let \( X_2 = 1 \) if the student’s family is low-income (below the World Bank’s poverty line), and \( X_2 = 0 \) otherwise. This \( X_2 \) is just a control variable.

We can consider potential outcomes. Let \( Y^0 \) denote the untreated potential outcome and \( Y^1 \) the treated potential outcome. That is, for each individual, \( Y^0 \) is her Gaokao score in the parallel universe where she does not participate in the program, and \( Y^1 \) is her score in the parallel universe where she does participate. Like before, we only observe one of these two; we observe \( Y = X_1Y^1 + (1 - X_1)Y^0 \).

As usual, the potential outcomes help define the causal objects of interest. An individual’s treatment effect of \( X_1 \) on \( Y \) is \( Y^1 - Y^0 \). The average treatment effect (ATE)
is
\[ \text{ATE} = E(Y^1 - Y^0) = E(Y^1) - E(Y^0), \] (9.36)
as in (4.7). It’s also possible that the effect differs depending on income, i.e., maybe the ATE for low-income students is much larger than the ATE for high-income students. Then, we may want to learn the \textbf{conditional average treatment effect} (CATE) for \( X_2 = 0 \) and \( X_2 = 1 \):
\[ \text{CATE}(x_2) \equiv E(Y^1 - Y^0 \mid X_2 = x_2) = E(Y^1 \mid X_2 = x_2) - E(Y^0 \mid X_2 = x_2). \] (9.37)

The goal is to learn about the effect of the program. Specifically, we want to learn the ATE of \( X_1 \) on \( Y \), and the CATEs, too.

To help measure the program’s effect, participation was randomized. However, low-income students were more likely to be selected for participation. Specifically, among low-income students, their participation probability was 90%, whereas the probability was only 30% for high-income students. Mathematically,
\[ P(X_1 = 1 \mid X_2 = 1) = 0.90, \quad P(X_1 = 1 \mid X_2 = 0) = 0.30. \] (9.38)

If we look only at low-income students, then there is a randomized experiment. That is, the CATE for \( X_2 = 1 \) is just the mean Gaokao score difference between the treated (participating) students and untreated students, within the \( X_2 = 1 \) subpopulation.

Of course, we should still worry. SUTVA may be violated; e.g., maybe there are “spillover effects” from program participants sharing their new knowledge with non-participants, which violates SUTVA. We should also worry that actual program participation may not be randomized: some students may never actually attend the program even if they were assigned to go, or some students may sneak in.

Putting aside those (reasonable) worries, if SUTVA and randomization hold within the \( X_2 = 1 \) subpopulation, then
\[ \text{CATE}(X_2 = 1) = E[Y \mid X_1 = 1, X_2 = 1] - E[Y \mid X_1 = 0, X_2 = 1]. \] (9.39)

This links the causal object CATE(1) with the statistical object on the right-hand side. The statistical object is a conditional mean difference that can be estimated using sample averages from data. Similarly, if SUTVA and randomization hold within the \( X_2 = 0 \) subpopulation, then
\[ \text{CATE}(X_2 = 0) = E[Y \mid X_1 = 1, X_2 = 0] - E[Y \mid X_1 = 0, X_2 = 0]. \] (9.40)

Although we can’t compare treated and untreated average scores to get the ATE, the ATE can be derived from the two CATEs. The probability of a student being low-income
is $P(X_2 = 1)$, which we can learn from the data. Consequently,

$$\text{ATE} = E[Y^1 - Y^0] = E[Y^1 - Y^0 \mid X_2 = 0] P(X_2 = 0) + E[Y^1 - Y^0 \mid X_2 = 1] P(X_2 = 1) = \text{CATE}(0) P(X_2 = 0) + \text{CATE}(1) P(X_2 = 1) = E[\text{CATE}(X_2)].$$

(9.41)

The equality $\text{ATE} = E[\text{CATE}(X_2)]$ is analogous to (9.35). As long as the conditional ATEs are identified, so is the unconditional ATE, since the distribution of $X_2$ is observable. To estimate the unconditional ATE, we can estimate the two CATEs and $P(X_2 = 1)$ and plug these into (9.41).

Besides the SUTVA assumption of A6.1, there was also the overlap assumption of A6.3. Before, it just said the probability of treatment is neither 0% nor 100%. Here, that needs to be true conditional on any value of $X_2$ in order to learn all the CATEs (and thus the ATE). In the Gaokao example, overlap means the program must be offered to at least some higher-income students ($X_2 = 0$), and not all low-income students. Mathematically, $0 < P(X_1 = 1 \mid X_2 = 0) < 1$ means the probability of program participation is neither 0% nor 100% within the higher-income subpopulation; similarly, $0 < P(X_1 = 1 \mid X_2 = 1) < 1$ for the low-income subpopulation. If $P(X_1 = 1 \mid X_2 = 1) = 1$, then all low-income students do the program, so there are no untreated low-income students left to compare with. (And comparing treated low-income students to untreated high-income students won’t work since they differ in ways other than program participation.) Similarly, if $P(X_1 = 1 \mid X_2 = 0) = 0$, then no higher-income students do the program, so we can’t possibly learn how it affects them. We could still learn CATE(1), but not CATE(0), and not the overall ATE.

In this Gaokao example, it is crucial that the CIA is sufficient for identifying the ATE since full independence is not true. In the example, $X_1 = 1$ is much more likely for low-income students ($X_2 = 1$). Most likely, $X_2$ has a negative causal effect on $Y$: low-income students cannot afford other programs and tutors to help their Gaokao score like the high-income students can. So, even if the program has zero effect, the treated group would score worse because those students have lower income:

$$\text{ATE} > E[Y \mid X_1 = 1] - E[Y \mid X_1 = 0].$$

(9.42)

That is, the program would seem much worse than it is. Mathematically, if we try to use the simpler model with $X = X_1$ and $U$ including $X_2$, then the independence assumption A6.2 is violated since $X_1$ and $X_2$ are dependent. So, it is critical that we account for $X_2$ explicitly.
Conditional Mean Independence

Assumption A9.1 is slightly stronger than necessary for CATE identification. It could be replaced by **conditional mean independence**:

$$
E(Y^0 | X_1, X_2) = E(Y^0 | X_2), \quad E(Y^1 | X_1, X_2) = E(Y^1 | X_2).
$$

(9.43)

That is, the mean of $Y^0$ does not depend on $X_1$ once we know (condition on) $X_2$, and similarly for $Y^1$. Here, $Y^0$ is conditionally mean independent of $X_1$ given $X_2$, and $Y^1$ is conditionally mean independent of $X_1$ given $X_2$. This is similar to the structural conditional mean independence assumption that $E(U | X_1, X_2) = E(U)$. However, in practice, it is usually difficult to argue that conditional mean independence is true without arguing A9.1 is true. Further, A9.1 is strong enough to identify conditional quantile treatment effects, too, although we do not discuss these in detail.

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. Here, 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. For example, maybe $Y$ is annual labor income, and we are interested in the effect of minimum wage. (We could also look at (un)employment, or hourly wage, etc.) Specifically, our city (hypothetically) recently implemented a large minimum wage increase. The goal is to learn the effect of the minimum wage increase on $Y$ (income). So, $X_1 = 1$ 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 $X_2 = 0$ for the year before the increase).

9.7.1 Bad Approaches

**Discussion Question 9.5** (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 9.6** (bad panel approach #2, for Mariel boatlift). Consider the same setup as in DQ 9.5. 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 9.7** (bad panel approach #1, for fracking). Discussion Questions 9.7 and 9.8 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 9.8** (bad panel approach #2, for fracking). Consider the same setup as in DQ 9.7, 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 $E(Y \mid X_2 = 1, X_1 = 1) - E(Y \mid X_2 = 1, X_1 = 1)$. 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.

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 interpreting 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 try to combine the before vs. after comparison with the our city vs. another city comparison. That is, maybe we can control for macroeconomic changes over time by seeing how incomes in another city change. Or, maybe we can control for persistent differences between the treated and untreated cities by looking at how different they were before the minimum wage increase took effect.

Conceptually, the goal is to construct a counterfactual: what would our city’s average income have been if there were not a minimum wage increase? Indeed, any reduced form causal inference can be viewed from this perspective: we observe the treated potential outcome \(Y^1\) for treated individuals, so if we could just figure out the corresponding \(Y^0\) (at least on average), we could estimate a treatment effect. The “corresponding \(Y^0\)” are counterfactuals: what would the outcomes have been if those individuals had not been treated (when in fact they were treated)? (Side note: such counterfactual thinking is central to economic analysis in litigation, asking, “What economic outcome would have occurred if [some illegal action] had not happened?”) Here, we observe the actual average income, but we want to compare it to this counterfactual parallel universe where the minimum wage did not pass, like the parallel universes in the potential outcomes framework. Without additional assumptions, we cannot do this.

The key for causal identification is called the parallel trends assumption. The cool thing is that this assumption can replace independence (A6.2), which we may not often believe outside randomized experiments. That is, parallel trends can let us learn causal effects from observational (non-randomized) data, which is important for the many economic cases where randomization is not possible (e.g., we cannot tell cities to randomize their minimum wage laws). The SUTVA assumption (A6.1) can be relaxed a little bit (e.g., allowing for wage effect spillovers within our city, since we’re looking at the effect of “treating” the entire city), but it is still necessary to preclude spillovers between the treated and untreated units. For example, our city’s minimum wage law would probably affect wages in all bordering cities, so those would not be a good choice for the comparison city.

Parallel trends can be understood in the following way. Conceptually, parallel trends says that without the minimum wage law, our city’s average income would have increased by exactly the same amount as the other city’s average income increased. Mathematically, the other city’s average income increase is

\[
E(Y \mid X_1 = 0, X_2 = 1) - E(Y \mid X_1 = 0, X_2 = 0). \tag{9.44}
\]

Parallel trends assumes that adding this increase to the baseline average income in our city, \(E(Y \mid X_1 = 1, X_2 = 0)\), gives us the counterfactual income for our city in the later
CHAPTER 9. REGRESSION WITH TWO BINARY REGRESSORS

time period. That is, it assumes we can learn a causal effect by comparing
\[
E(Y \mid X_1 = 1, X_2 = 1) \quad \text{vs.} \quad E(Y \mid X_1 = 1, X_2 = 0) + E(Y \mid X_1 = 0, X_2 = 1) - E(Y \mid X_1 = 0, X_2 = 0).
\]

This is formalized further in Section 9.7.3. The basic idea is that the “actual” average income is an average of treated potential outcomes \(Y^1\), while the “counterfactual” equals the same average but for untreated potential outcomes \(Y^0\).

We can draw a picture of (9.45) after rearranging some. Using the usual notation \(m(x_1, x_2) \equiv E(Y \mid X_1 = x_1, X_2 = x_2)\), the causal effect (we’ll be more specific about what type of causal effect later) is
\[
\begin{align*}
\text{actual} & \quad m(1, 1) - \{m(1, 0) + [m(0, 1) - m(0, 0)]\} \\
\text{counterfactual} & \quad \{m(1, 0) + [m(0, 1) - m(0, 0)]\} - [m(0, 1) - m(0, 0)] = \beta_3,
\end{align*}
\]

the interaction term coefficient from (9.14). Figure 9.1 visualizes this treatment 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, \(m(0, 1) - m(0, 0)\).

Figure 9.1: Difference-in-differences.
9.7. CAUSAL IDENTIFICATION: DIFFERENCE-IN-DIFFERENCES

Discussion Question 9.9 (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: is a BAC law the only way states try to reduce fatal accidents?)

9.7.3 Identification

Population Objects of Interest

Most fundamentally, the difference-in-differences approach only learns the average treatment effect for the group that was actually treated (in our universe). This is called the average treatment effect on the treated (ATT) (or sometimes ATTE or ATET). Recall that mathematically, ATE meant \(E(Y^1 - Y^0)\), where \(Y^1\) is the treated potential outcome (somebody’s outcome in the parallel universe where they’re treated) and \(Y^0\) is the untreated potential outcome (somebody’s outcome in the parallel universe where they’re not treated). ATT is the same comparison, but for the subpopulation who was actually treated in our universe. In our diff-in-diff example, \(X_1 = 1\) if somebody is in the treated group (i.e., in our city where there was a minimum wage increase). Thus, the ATT is

\[
ATT \equiv E(Y^1 - Y^0 \mid X_1 = 1).
\]

(9.46)

If we assume that the average effect is the same in the \(X_1 = 1\) and \(X_1 = 0\) groups, then the ATE and ATT are the same, but sometimes we might think the effect is different. 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. This is essentially a question of external validity; see Chapter 12 for more.

Identification of ATT

Why does diff-in-diff identify the ATT? The key identifying assumption is “parallel trends”:

\[
E(Y^0 \mid X_1 = 1, X_2 = 1) - E(Y^0 \mid X_1 = 1, X_2 = 0) = E(Y^0 \mid X_1 = 0, X_2 = 1) - E(Y^0 \mid X_1 = 0, X_2 = 0).
\]

(9.47)

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. Since the observed \(Y\) is \(Y = Y^1T + Y^0(1 - T)\), where \(T = I\{X_1 = 1, X_2 = 1\}\) (or equivalently \(T = X_1X_2\)), three of the four terms in (9.47) are observable, i.e., other than the first term, \(Y^0 = Y\) since \(T = 0\). It is only \(E(Y^0 \mid X_1 = 1, X_2 = 1)\) that we cannot observe data for. Indeed, this is the counterfactual: what would the average wage have been in
our city \((X_1 = 1)\) in the later time period \((X_2 = 1)\) if we had not passed the minimum wage law \((Y^0, \text{not } Y^1)\)? Rearranging (9.47) thus gives us the counterfactual in terms of things we can learn about:

\[
E(Y^0 \mid X_1 = 1, X_2 = 1) = E(Y^0 \mid X_1 = 1, X_2 = 0) + [E(Y^0 \mid X_1 = 0, X_2 = 1) - E(Y^0 \mid X_1 = 0, X_2 = 0)] \\
= m(1,0) + [m(0,1) - m(0,0)].
\] (9.48)

The ATT is thus

\[
\text{ATT} = E(Y^1 - Y^0 \mid X_1 = 1, X_2 = 1) \\
= E(Y^1 \mid X_1 = 1, X_2 = 1) - E(Y^0 \mid X_1 = 1, X_2 = 1) \\
= m(1,1) - \{m(1,0) + [m(0,1) - m(0,0)]\} \\
= \beta_3.
\]

**Skepticism About Parallel Trends**

The parallel trends condition may not hold, for various reasons. For example, maybe our city was experiencing great economic and wage growth, whereas the comparison city was declining (maybe due to reliance on different industries). In fact, 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.

You can try to see if parallel trends seems plausible, although it is not testable. Empirically, you could check the trends for a few periods before the treatment takes place, to see if it seems plausible. 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. And 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 economic 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 $\beta_j$; remember to use sample/survey weights if they are available in the data. (We could separately compute the four subsample averages, but that would be more work, and we could not add additional regressors easily.) The same heteroskedasticity-robust methods from earlier sections like Section 7.7.3 can be used to compute confidence intervals if sampling is iid.

```r
library(sandwich); library(lmtest)
n <- 4*8
set.seed(112358)
m00 <- 10; m10 <- 15; m01 <- 16; m11 <- 25
df <- data.frame(X1=c(rep(0,n/2),rep(1,n/2)),
                  X2=rep(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)
```
### 2.5 % 97.5 %

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>9.30</td>
<td>10.77</td>
</tr>
<tr>
<td>X1</td>
<td>4.31</td>
<td>6.04</td>
</tr>
<tr>
<td>X2</td>
<td>4.97</td>
<td>7.27</td>
</tr>
<tr>
<td>X1:X2</td>
<td>2.20</td>
<td>5.27</td>
</tr>
</tbody>
</table>
Empirical Exercises

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-difference 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: use http://faculty.missouri.edu/~kaplandm/intro_text/driving , clear to load the data


d. Create a dummy variable for the “after” period (year 1990) with R command `df$after <- (df$year==1990) or Stata command generate after = (year ==1990)`

e. Create a variable `bac` equal to 1 (or TRUE) if there is any BAC law for the entire year, with R command `df$bac <- (df$bac08+df$bac10>=1)` or Stata command `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. In R, use `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 , ]` and in Stata use:

```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. Use R command `treatst <- unique(df$state[df$ bac])` followed by `df$ treat <- (df$state %in% treatst)` or Stata command `bysort state : egen treat = max(bac)`

h. Run a difference-in-difference regression with the intercept, “after” dummy, treatment dummy, and interaction term. Use R command `ret <- lm(wkndfatrte ~treat*after, data=df)` or Stata command `regress wkndfatrte treat##after, vce(robust)` where the * in R and the ## in Stata automatically generate the desired interaction term. In R, you can view the results with `coeftest(ret, vcov.=vcovHC(ret, type='HC1'))` or `coefci(ret, vcov.=vcovHC(ret, type='HC1'))`

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, with R command (`agg <- aggregate(wkndfatrte~treat*after, data=df, FUN=mean)`) or Stata command `tabulate treat after, summarize(wkndfatrte) means missing`

j. To display the CEF-based replication of the OLS estimates, use R command `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. In Stata, run `collapse (mean) wkndfatrte, by(treat after) and then`

   `display wkndfatrte[1]`
   `display wkndfatrte[3]-wkndfatrte[1]`
   `display wkndfatrte[2]-wkndfatrte[1]`
   `display (wkndfatrte[4]-wkndfatrte[3])-(wkndfatrte[2]-wkndfatrte[1])`

k. Repeat your analysis but with a different end year than 1990, to see how sensitive the interaction coefficient estimate is to the time period chosen (i.e., does it change a lot when the end year changes by a few years or less?).

l. Repeat your analysis but replacing your bac treatment variable with a treatment dummy equal to 1 if perse (a different driving law) equals 1 (and equal to 0 otherwise).

m. Repeat your analysis 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.
Empirical Exercise EE9.2. You will analyze wage data for different types of individuals from the 1976 Current Population Survey (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 non-whites 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)
?wage1
```

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 with R command (agg <- aggregate(wage ~nonwhite*female, data=wage1, FUN=mean)) or Stata command `tabulate female nonwhite , summarize(wage) means missing`

d. Run a “difference-in-difference” type of regression with the intercept, non-white dummy, female dummy, and interaction term. Use R command `ret <- lm(wage ~nonwhite*female, data=wage1)` or Stata command `regress wage female##nonwhite , vce(robust)` and further view the results in R with

```r
coeftest(ret, vcov.=vcovHC(ret, type='HC1'))
coefficients(ret, vcov.=vcovHC(ret, type='HC1'))
```

e. Compute the OLS coefficient estimates manually from the four conditional means.

In 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. In 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 your analysis but using south instead of female

g. Repeat your analysis 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

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)
- Average structural effects and their identification: Hansen (2018, §2.29)
Allowing multiple regressors opens a multitude of combinations, especially when combined with nonlinear functions like in Chapter 8. As the possibilities increase exponentially, the importance of thinking about your model carefully increases exponentially.

When not otherwise stated, 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. Even with 100 regressors, OVB can still be a big problem.

Consider the linear structural model

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + V. \]  \hspace{1cm} (10.1)

For OLS to consistently estimate \( \beta_j \) for \( j = 1, 2, 3 \) (the slope coefficients) requires
Cov($X_j, V) = 0$ for $j = 1, 2, 3$. Imagine this is true, but $X_3$ is omitted, so
\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U, \quad U \equiv \beta_3 X_3 + V. \quad (10.2) \]

In order for OLS to consistently estimate both $\beta_1$ and $\beta_2$, we need Cov($X_1, U) = Cov(X_2, U) = 0$. This requires either $\beta_3 = 0$ (i.e., $X_3$ is not a causal determinant of $Y$) or Cov($X_1, X_3) = Cov(X_2, X_3) = 0$. If we had more included regressors, then they would also need to have zero covariance with $X_3$ (if $\beta_3 \neq 0$) in order to avoid OVB.

If we only care about $\beta_1$, then the weaker conditional (mean) independence assumption is required, instead of Cov($X_j, U) = 0$ for all $j$. Even then, if we omit an important control variable, then conditional independence won’t hold, and $\hat{\beta}_1$ won’t consistently estimate $\beta_1$.

**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 the conditions for OVB.

### 10.2 Linear-in-Variables Model

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 $X_1, X_2, \ldots, X_J$, the model is
\[ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J + U = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + U \equiv g(X_1, \ldots, X_J) + U. \quad (10.3) \]

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 or not.

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 $X_1 = x_1, X_2 = x_2, \ldots, X_J$ changes the function by
\[
g(x_1 + 1, x_2, \ldots, x_J) - g(x_1, x_2, \ldots, x_J)
= [\beta_0 + \beta_1 (x_1 + 1) + \sum_{j=2}^{J} \beta_j x_j] - [\beta_0 + \beta_1 x_1 + \sum_{j=2}^{J} \beta_j x_j] = \beta_1 (x_1 + 1 - x_1) = \beta_1. \quad (10.4)\]
For example, if \( Y \) is wage in \$/hr, and \( X_1 \) is years of education, and \( \beta_1 = (\$5/\text{hr})/\text{yr} \), then each additional year of education is associated with a \((\$5/\text{hr})/\text{yr}\) change.

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 (10.4),

\[
g(x_1 + \Delta_1, x_2, \ldots, x_J) - g(x_1, x_2, \ldots, x_J) = \beta_0 + \beta_1(x_1 + \Delta_1 - x_1) = \beta_1 \Delta_1.
\]

(10.5)

While pleasingly simple, these formulas may not be realistic. That is, the predicted/average change in \( Y \) may depend on not only \( \Delta_1 \), but the starting value \( x_1 \). For example, let \( Y \) be wage and \( X \) years of experience. 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 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 another variable (like education) implies some sort of interaction term(s). Nonlinearity in a single variable is discussed in Section 8.2; interactions are discussed in Sections 9.3 and 10.3. Nonparametric models with multiple regressors are beyond our scope.

The following code shows a simple linear-in-variables regression with simulated data.

```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))
df$Y <- CEF(df$X1, df$X2, df$X3) + rnorm(n)
ret <- lm(Y~X1+X2+X3, data=df)
retVC1 <- vcovHC(ret, type="HC1")
round(cbind(coeftest(ret, vcov. = retVC1)[,1:2],
            coefci(ret, vcov. = retVC1)), digits=2)
```

## Estimate Std. Error 2.5 % 97.5 %
## (Intercept)  -0.44  0.47  -1.39  0.52
## X1           0.75  0.40  -0.06  1.57
## X2           1.77  0.49   0.78  2.76
## X3           4.04  0.47   3.10  4.99
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 and $X$ the other regressor. Assume $X$ is the regressor of interest. Again, don’t worry about causality for now.

10.3.1 Limitation of Linear-in-Variables Model

With a linear-in-variables model,

$$Y = g(X, D) + U = \beta_0 + \beta_1 X + \beta_2 D + U.$$  \hspace{1cm} (10.6)

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 $D$ only has 2 possible values, we can take (10.6) and plug in $D = 0$ and $D = 1$.

Without worrying about its interpretation (CEF or LP), we’ll focus on $g(X, D)$. With $D = 0$ and $D = 1$,

$$g(X, D = 0) = \beta_0 + \beta_1 X,$$  \hspace{1cm} (10.7)

$$g(X, D = 1) = \beta_0 + \beta_1 X + (\beta_2)(1) = (\beta_0 + \beta_2) + \beta_1 X.$$  \hspace{1cm} (10.8)

Comparing the two functions, both have the same slope $\beta_1$, but the intercepts are $\beta_0$ and $\beta_0 + \beta_2$, respectively. That is, $\beta_2$ describes how the $D = 1$ intercept differs from the $D = 0$ intercept.

10.3.2 Interpretation of Interaction Term

To allow both the intercept and slope to differ when $D = 0$ and $D = 1$, an interaction term can be used. Specifically, a term with the product $DX$ can be added. This allows the regressors to “interact” in the sense that the slope with respect to $X$ can depend on $D$. Mathematically, adding this term to (10.6),

$$Y = g(X, D) + U = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 DX + U.$$  \hspace{1cm} (10.9)

The model in (10.9) is strictly more general than before since setting $\beta_3 = 0$ yields the linear-in-variables model (10.6). Given the model in (10.9), instead of (10.7) and (10.8),

$$g(X, D = 0) = \beta_0 + \beta_1 X + (\beta_2)(0) + (\beta_3)(0)(X) = \beta_0 + \beta_1 X,$$  \hspace{1cm} (10.10)

$$g(X, D = 1) = \beta_0 + \beta_1 X + (\beta_2)(1) + (\beta_3)(1)(X) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X.$$  \hspace{1cm} (10.11)
Now, the slope differs (by $\beta_3$), too. Note that $\beta_3 > 0$ or $\beta_3 < 0$ is allowed, or even $\beta_3 = 0$, just as $\beta_2 > 0$, $\beta_2 < 0$, and $\beta_2 = 0$ are all possible.

Figure 10.1 illustrates the interpretation of the function from (10.9). In the figure’s example, $\beta_2 > 0$ and $\beta_3 > 0$, but in general either or both could be negative (or zero). 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 $\beta_2$. Think of the model as

$$g(X,D) = (\beta_0 + \beta_1) + D(\beta_2 + \beta_3 X), \quad (10.12)$$

so the coefficient on $D$ is $\beta_2 + \beta_3 X$. For example, even if $\beta_2 = -2$, it could be that $\beta_2 + \beta_3 X > 0$ if $\beta_3 = 2$ and $X > 1$. That is, the coefficient on $D$ could be positive for all values of $X$ bigger than 1, even if $\beta_2$ is negative. The opposite is also possible, e.g., if $\beta_2 = 5$, $\beta_3 = -1$, and $X > 5$.

**Discussion Question 10.2** (sleep and interactions). Let $Y$ be a person’s hours of sleep per night, $D = 1$ if living in the same house as children under 8 years old (and $D = 0$ if not), and $X$ is age. Consider the model from (10.9). a) What would you guess for the signs (+, −, or zero) of $\beta_2$ and $\beta_3$? b) Describe another nonlinear 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. INTERACTION TERMS

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?

Even if neither regressor were binary, an interaction term can allow the slopes to depend on the other regressor’s value. Replacing $X = X_1$ and $D = X_2$ in (10.9),

$$Y = g(X_1, X_2) + U = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + U. \quad (10.13)$$

We could again focus on the slope of the function $g(X_1, X_2)$ with respect to $X_1$, at different values of $X_2$. Considering two possible values $X_2 = a$ and $X_2 = b$, similar to (10.10) and (10.11),

$$g(X_1, X_2 = a) = \beta_0 + \beta_1 X_1 + (\beta_2)(a) + (\beta_3)(a)(X_1) = (\beta_0 + a\beta_2) + (\beta_1 + a\beta_3) X_1, \quad (10.14)$$

$$g(X_1, X_2 = b) = \beta_0 + \beta_1 X_1 + (\beta_2)(b) + (\beta_3)(b)(X_1) = (\beta_0 + b\beta_2) + (\beta_1 + b\beta_3) X_1. \quad (10.15)$$

Changing $X_2$ from $a$ to $b$ changes the intercept from $\beta_0 + a\beta_2$ to $\beta_0 + b\beta_2$, and 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 $g(X_1 = a, X_2)$ and $g(X_1 = b, X_2)$ as functions of $X_2$, where again both the intercept and slope may change.

Don’t get fooled by looking at $\hat{\beta}_1$ alone (a common mistake). For example, imagine $X_1$ is experience and $X_2$ is years of education, and $Y$ is wage ($$/hr$). Imagine

$$\hat{Y} = 5 - 15 X_1 + 2 X_2 + 2 X_1 X_2. \quad (10.16)$$

If we ignore the interaction term, then having $\hat{\beta}_1 = -15$ makes it look like there is 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 $X_1$. Specifically, that slope is $2X_2 - 15$. Further, imagine everyone in the data has at least 10 years of education. That means $X_2 \geq 10$, so $2X_2 - 15 \geq (2)(10) - 15 = 5$: the slope is at least 5, and it is even more positive for higher education levels. Even though $\hat{\beta}_1 < 0$, the predicted values of $Y$ are always increasing with $X_1$. When there is an interaction term (or other nonlinear terms), the linear term’s coefficient does not have any meaning by itself. You always have to look at the whole function.
The interaction model is again more general than the linear-in-variables model, but it is 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 an $X_1X_2$ interaction term. Additional interaction terms could be added, or a nonparametric approach could be taken, although that is beyond our scope.

10.3.4 Code

The following code illustrates estimation, 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 (10.9). 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 (10.9).

```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

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.

With such complicated models, it is better to look at predicted changes using the full model instead of looking at individual coefficients. This is easy in R with the `predict()` function.

10.5 Assumptions for Linear Projection

The following 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, let alone a structural model.

The assumptions are basically the same as in Section 7.7.2, with one exception. 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.

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.

A related concept is imperfect multicollinearity. This refers to regressors being strongly correlated, but not perfectly correlated. 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.)

The assumptions and results refer to the linear projection model

$$\text{LP}(Y \mid X_1, \ldots, X_J) = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J.$$  

(10.17)

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): there are no constants $c_j$ such that $X_J = c_0 + c_1 X_1 + \cdots + c_{J-1} X_{J-1}$, and similarly for other $X_j$.

**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: $E(Y^2) < \infty$, $E(X_j^2) < \infty$.

**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 (10.17), if A10.1–A10.3 are true, then $\hat{\beta}_j \xrightarrow{p} \beta_j$ for $j = 0, 1, \ldots, J$.

**Theorem 10.2 (OLS approximate normality).** Given the linear projection model in (10.17), if A10.1–A10.3 are true, then the OLS coefficient estimators are asymptotically normal, i.e., with large $n$, approximately $\hat{\beta}_j \sim \text{N}(\beta_j, \hat{\text{SE}}_j^2)$, where the true standard error $\hat{\text{SE}}_j$ can be estimated and is proportional to $1/\sqrt{n}$. Further, with large $n$, the joint distribution of $(\hat{\beta}_0, \ldots, \hat{\beta}_J)$ is approximately jointly normal with mean $(\beta_0, \ldots, \beta_J)$ and an unknown (co)variance matrix that can be estimated.

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

This section extends Section 9.6. I cover only a couple out of many interesting structural identification results for regression models.

First, an obvious result. Imagine the structural model is

$$ Y = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + U. \quad (10.18) $$

It is possible that some $X_j$ are nonlinear functions of regressors, including interaction terms. If $E(U) = 0$ and $\text{Cov}(U, X_j) = 0$ for all $j = 1, \ldots, J$, then $U$ satisfies the properties of a linear projection error, so the $\beta_j$ are linear projection coefficients. This links the structural model to a statistical model that we can estimate with OLS.

Second, an extension of Section 9.6. Recall the conditional independence assumption (CIA) of A9.1: conditional on controls variables $X_2, \ldots, X_J$, the regressor of interest $X_1$ is independent of the vector of unobserved determinants $U$: $U \perp X_1 \mid X_2, \ldots, X_J$. The CIA is violated (false) if within a subpopulation of individuals with the same values of $(X_2, \ldots, X_J)$, $X_1$ and $U$ are correlated.

Extending (9.30), consider the structural model

$$ Y = h(X_1, X_2, \ldots, X_J, U), \quad (10.19) $$

where $U = (U_1, U_2, \ldots)$ is a vector containing all the unobserved determinants of $Y$ besides the $X_j$. If $X_1$ is binary, the structural effect of $X_1$ on $Y$ given values of the other $X_j$ and $U$ is

$$ S(X_2, \ldots, X_J, U) \equiv h(1, X_2, \ldots, X_J, U) - h(0, X_2, \ldots, X_J, U), \quad (10.20) $$

and the conditional average structural effect (CASE) is

$$ \text{CASE}(x_2, \ldots, x_J) = E[S(X_2, \ldots, X_J, U) \mid X_2 = x_2, \ldots, X_J = x_J]. \quad (10.21) $$

Given A9.1, the CASE values are identified by differences in CEF values, again linking the causal world (CASE) with the statistical world (CEF):

$$ \text{CASE}(X_2 = x_2, \ldots, X_J = x_J) = E[Y \mid X_1 = 1, X_2 = x_2, \ldots, X_J = x_J] - E[Y \mid X_1 = 0, X_2 = x_2, \ldots, X_J = x_J]. \quad (10.22) $$
This result generalizes to non-binary $X_1$, too. Instead of using $X_1 = 1$ and $X_1 = 0$, we can compare $X_1 = b$ and $X_1 = a$, for any values $a$ and $b$.

If we can guess the correct CEF functional form, then we can easily estimate CASEs. 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.
Empirical Exercises

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 children on 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 (heduc) and the woman’s age (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:

```r
library(wooldridge)
library(sandwich)
library(lmtest)
?fertil2
```

b. Stata only: load the data with `bcuse fertil2, nodesc clear` (assuming `bcuse` is already installed)

c. Run a simple regression of children on educ with R command `ret1 <- lm(children~educ, data=fertil2)` or Stata command `regress children educ, vce(robust)`

d. Repeat but adding heduc as a control variable regressor with R command `ret2 <- lm(children~educ+heduc, data=fertil2)` or Stata command `regress children educ heduc, vce(robust)`

e. Repeat but adding yet another regressor (woman’s age) with R command `ret3 <- lm(children~educ+heduc+age, data=fertil2)` or Stata command `regress children educ heduc age, vce(robust)`

f. Repeat but add even more regressors (in addition to educ, heduc, and age): agesq, knowmeth, usemeth, electric, urban, and catholic, as well as interactions between age and knowmeth and between age and usemeth. In R, store the result as `ret4`, and you can simply write `knowmeth:age` and `usemeth:age` in the regression formula to generate the interactions. In Stata, first create the two interaction variables like with `generate know_age = knowmeth*age` and then add those new variables in your list of regressors.
g. R only (since already displayed by Stata): output the four sets of estimated regression coefficients with

```r
c coef(ret1)
c coef(ret2)
c coef(ret3)
c 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)](http://example.com), 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:

```r
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` with R command `ret1 <- lm (log(price)~log(nox), data=hprice2)` or Stata command `regress lprice lnox , vce(robust)`

d. Repeat but adding `rooms` as a control variable regressor with R command `ret2 <- lm(log(price)~log(nox)+rooms, data=hprice2)` or Stata command `regress lprice lnox rooms , vce(robust)`
e. Repeat but adding yet another regressor (crime rate per capita) with R command `ret3 <- lm(log(price)~log(nox)+rooms+crime, data=hprice2)` or Stata command `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
   
   ```r
   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

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

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): [https://www.youtube.com/watch?v=qEyrtBnJKo8](https://www.youtube.com/watch?v=qEyrtBnJKo8)
- External validity (Masten video): [https://www.youtube.com/watch?v=XIVk8uwptCw](https://www.youtube.com/watch?v=XIVk8uwptCw)
- Missing data approaches, including worst-case bounds (Masten video): [https://www.youtube.com/watch?v=a13TBySF024](https://www.youtube.com/watch?v=a13TBySF024)
- Reverse causality and simultaneity (Masten video): [https://www.youtube.com/watch?v=ROLiLaR-17U](https://www.youtube.com/watch?v=ROLiLaR-17U)
In your future work, when you see econometric results, you’ll ask two questions. First, are the econometric methods appropriate for the setting and sample? That is, do you believe the assumptions needed for consistency of the standard error and coefficient estimators? Second, if so, can you use the results for your work, which might be in a different setting? These questions concern internal validity and external validity, respectively.

12.1 Terminology

There are many valid ways to define these ideas more specifically; Definition 12.1 provides one. The population studied refers to the population from which the data was sampled. The population of interest is the one that you care about. For example, you may see an econometric study of the causal effect of a minimum wage change in California, but your job is to advise Missouri about a possible minimum wage increase. The study is internally valid if it properly estimates causal effects for California (or possibly some subpopulation in California), i.e., for the population studied; it is externally valid if the effects in California are the same as those in Missouri, your population of interest.

Definition 12.1 (internal and external validity). A study has internal validity if its estimators are consistent for the values of interest in the population studied, and the corresponding confidence intervals have (approximately) the desired coverage probability. A study has external validity in a different setting if the population values are the same in the population studied and the population of interest.

Unfortunately, there are many threats to validity, both internal and external. Previous chapters already discussed some threats to internal validity that prevent consistent estimation of average causal effects. Internal validity is usually less of an issue for description and prediction, although some threats to internal validity in time series are discussed.
12.2. THREATS TO EXTERNAL VALIDITY

in Part [III] Still, threats to external validity can debilitate even the most sophisticated prediction techniques, and they can render useless even the most careful causal inference. The rest of Chapter 12 describes specific threats to internal and external validity that you should look for.

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. To highlight this, 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, which gives excellent internal validity by determining the California median exactly, we learn basically nothing about Missouri. This problem is a lack of external validity.

The Lucas critique (Lucas, 1976) can also be interpreted in terms of external validity. When macroeconomic policy changes, that fundamentally changes the setting. Even if our estimates 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.

12.2.1 Different Place

Different places have different legal, political, cultural, and economic settings. The house price example highlights just one of many important 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 (hypothetical) 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.

But 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.

Instead of deciding “close enough,” we could also model how values changed over time historically, in order to predict how they have changed since the data sample was collected; see Chapter 14.

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, 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, then your number will be much too big because the studied population (all owners) differs greatly from the population of interest (first-time owners).

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 average 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, including individuals who already got college degrees even without the additional subsidy. You are instead interested in individuals who currently do not (or cannot) choose to graduate from college, but who could 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, from $10/hr to $15/hr. 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

Although there are many more threats to internal validity for causality, there can be threats to internal validity for description and prediction, too. All of these are also threats to internal validity for causal inference (but not vice-versa), so details are found in Section 12.4. In particular, see Items 1–6 in the list in Section 12.4.

12.4 Threats to Internal Validity for Causality

(For the below discussions of asymptotic bias, it may help to review Sections 3.4.1 and 3.4.4.)

The following threats to internal validity for causal and structural analysis are detailed below.

1. Functional form misspecification
2. Measurement error
3. Survey weights
4. Non-iid sampling
5. Sample selection
6. Missing data
7. Omitted variables
8. Simultaneity and reverse causality

Additionally, violation of SUTVA (as discussed earlier) is another threat to internal validity for causal analysis.
12.4.1 Functional Form Misspecification

Misspecifying the functional form leads to inconsistent estimates of the CEF. This is bad news for description, prediction, and causality alike.

Previous chapters show different ways a model’s function form can be misspecified (wrong). Chapters 7 and 8 discuss first how a linear function of a single variable can be wrong even if the regressor has only three possible values, and then how even complex nonlinear functions can still be misspecified if the regressor has many possible values. Chapters 9 and 10 show how there are even more ways to get the CEF’s form wrong when there are multiple regressors; e.g., even with only two binary regressors, forgetting the interaction term can cause misspecification.

These and other chapters also discuss how to mitigate the problem of functional form misspecification. Chapters 7 and 8 discuss how to add nonlinear terms with a single regressor. Chapters 9 and 10 discuss adding interaction terms involving multiple regressors. To more carefully avoid functional form misspecification, nonparametric methods are required.

12.4.2 Measurement Error in the Outcome

Without good data, it’s hard to get valid econometric results. As they say, “Garbage in, garbage out” (https://en.wikipedia.org/wiki/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. Let $Y^*$ be how much exercise somebody does, in minutes per week. This $Y^*$ is a the latent (unobserved) true value. However, the observable 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. Uppercase $M$ is used to remind us it is a random variable (although in other books lowercase letters like $e$ are often used for error terms). This means 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.” So, $Y^*$, $Y$, and $M$ are all
random variables, with different values for different individuals, having some population probability distribution that describes the probabilities of different possible values.

**Discussion Question 12.2** (exercise error). Consider the example where \(Y^*\) is true exercise minutes per week and \(Y\) is the value somebody reports. Would you guess \(E(M) = 0, < 0\) or \(> 0\)? Why? Note: even to guess, you’ll probably have to make additional assumptions and definitions; e.g., what does “exercise” mean?

**Regression**

Imagine the true linear projection is

\[
Y^* = \beta_0 + \beta_1 X + U, \quad E(U) = \text{Cov}(X,U) = 0. \tag{12.1}
\]

We want to learn \(\beta_1\). Substituting in \(Y^* = Y - M\),

\[
Y - M = \beta_0 + \beta_1 X + U, \\
Y = \beta_0 + \beta_1 X + (U + M). \tag{12.2}
\]

Can we just regress \(Y\) on \(X\) to estimate \(\beta_1\)? It depends. For \(\beta_1\) to be the slope of LP(\(Y\mid 1, X\)) requires

\[
0 = \underbrace{\text{Cov}(X,U + M)}_{\text{by (12.1)}} = \text{Cov}(X,U) + \text{Cov}(X,M) = \text{Cov}(X,M). \tag{12.3}
\]

If \(\text{Cov}(X, M) \neq 0\), then \(\beta_1\) from (12.1) is identified and equal to the slope of the population LP of the observed \(Y\) onto \((1, X)\). If \(\text{Cov}(X, M) \neq 0\), then the slope changes when we replace \(Y^*\) with \(Y\).

So, what do we actually estimate when regressing \(Y\) on \(X\)? Define the LP of \(Y\) onto \((1, X)\) as

\[
\text{LP}(Y \mid 1, X) = \gamma_0 + \gamma_1 X. \tag{12.4}
\]

OLS consistently estimates \(\gamma_1\):

\[
\hat{\gamma}_1 \xrightarrow{p} \gamma_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}. \tag{12.5}
\]

The problem is, if \(\text{Cov}(X, M) \neq 0\), then \(\gamma_1 \neq \beta_1\).

We can compute the asymptotic bias. Recall the asymptotic bias is the difference between what we estimate and the true value, when we have lots of data:

\[
\text{plim}_{n \to \infty} \hat{\gamma}_1 - \beta_1 = \gamma_1 - \beta_1. \tag{12.6}
\]
From (12.1),

$$\beta_1 = \frac{\text{Cov}(Y^*, X)}{\text{Var}(X)}.$$  \hfill (12.7)

Thus, using (12.5),

$$\gamma_1 \xrightarrow{p} \frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \frac{\text{Cov}(Y^* + M, X)}{\text{Var}(X)} = \frac{\text{Cov}(Y^*, X) + \text{Cov}(M, X)}{\text{Var}(X)}.$$

Thus, the asymptotic bias is

$$\text{plim}_{n \to \infty} \gamma_1 - \beta_1 = \beta_1 + \frac{\text{Cov}(M, X)}{\text{Var}(X)} - \beta_1 = \frac{\text{Cov}(M, X)}{\text{Var}(X)} = \frac{\text{Corr}(M, X)}{\sqrt{\text{Var}(X) \text{Var}(M)}} \sqrt{\frac{\text{Var}(M)}{\text{Var}(X)}}.$$  \hfill (12.8)

From (12.8), we can describe the direction of bias. Since \(\text{Var}(X) > 0\), the sign (+ or −) of the asymptotic bias is the sign of \(\text{Cov}(M, X)\). If \(\text{Cov}(M, X) > 0\), then there is positive (upward) bias. If \(\text{Cov}(M, X) < 0\), then there is negative (downward) bias. Recall that if \(\beta_1 < 0\), then positive bias can make the magnitude look smaller by estimating a value closer to zero \((\hat{\beta}_1 < \beta_1 < 0)\). Similarly, negative bias can make the magnitude look larger if \(\hat{\beta}_1 > \beta_1 < 0\).

From (12.8), we can also assess the magnitude of bias. The bias equals the slope of \(\text{LP}(M \mid X)\), which can also be written as \(\text{Corr}(M, X) \sqrt{\text{Var}(M) / \text{Var}(X)}\). Given a fixed correlation, this is larger in magnitude if the variance of \(M\) is larger relative to the variance of \(X\). Given fixed variances, the bias is larger in magnitude when the correlation is larger. So, it matters both how “big” the measurement error is (in terms of \(\text{Var}(M)\), relative to \(\text{Var}(X)\)) as well as how closely related (correlated) it is with \(X\).

**Example**

When might this type of measurement error cause problems? That is, when might \(\text{Cov}(X, M) \neq 0\), or equivalently, when might \(X\) and \(M\) be correlated?

 Continuing with \(Y^*\) as weekly exercise, let \(X = 1\) if somebody has a gym membership and \(X = 0\) otherwise. It’s fine (no asymptotic bias) if everyone overreports \((\text{E}(M) > 0)\) or underreports \((\text{E}(M) < 0)\) as long as it’s the same for both gym members and non-members. It would also be fine if \(\text{E}(M \mid X = 0) = \text{E}(M \mid X = 1) = 0\) but \(\text{Var}(M \mid X = 1) < \text{Var}(M \mid X = 0)\), 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, we could imagine reasons for systematic overreporting by 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., $\text{Cov}(X, M) > 0$. Such positive correlation 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 $\beta_1$.

**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 scrape rates on December 31; this is $Y$. a) Describe a reason why treated firms might systematically overreport
(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?

**Fixing 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. For now, I only hope you can identify when measurement error in Y may be a serious problem.

### 12.4.3 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 is

\[
Y = \beta_0 + \beta_1 X^* + R, \quad E(R) = \text{Cov}(X^*, R) = 0. \tag{12.9}
\]

Substituting in \(X^* = X - M\),

\[
Y = \beta_0 + \beta_1 (X - M) + R = \beta_0 + \beta_1 X + (R - \beta_1 M). \tag{12.10}
\]

The critical question is whether X is correlated with the unobserved “error term” \(R - \beta_1 M\). Using (12.9) and linearity,

\[
\text{Cov}(X, R - \beta_1 M) = \text{Cov}(X, R) - \text{Cov}(X, \beta_1 M)
= \text{Cov}(X^* + M, R) - \beta_1 \text{Cov}(X, M)
= \text{Cov}(X^*, R) + \text{Cov}(M, R) - \beta_1 \text{Cov}(X, M).
\]

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 \(\text{Cov}(X, M) \neq 0\).

Unfortunately, \(\text{Cov}(X, M) = 0\) is very unlikely. Consider the best-case scenario: M is just random noise unrelated to the true value \(X^*\), so \(\text{Cov}(X^*, M) = 0\). For intuition, consider 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 \((2, 2)\),
12.4. THREATS TO INTERNAL VALIDITY FOR CAUSALITY

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 function $f(X) = X$, which was the linear projection of $Y$ on $(1, X)$, is always off by 1: $|Y - X| = 1$ at each point. Thus, $E[(Y - X)^2] = 1$. But we can do better by connecting the opposite corners of the parallelogram, with the function $g(X) = 1 + X/3$. At the first and fourth points, $Y - g(X) = 0$; at the middle two points, $|Y - g(X)| = 2/3$. So, $E[(Y - g(X))^2] = (1/4)(0) + (1/4)(0) + (1/4)(2/3)^2 + (1/4)(2/3)^2 = 2/9$, a much closer (population) “fit” than $f(X) = X$. That is, when we add horizontal noise (errors in $X$), the slope of the linear projection $LP(Y | 1, X)$ is flatter (closer to zero) than the slope of $LP(Y | 1, X^*)$. Figure 12.2 illustrates this.

This intuition holds more generally. Using $\text{Cov}(X^*, M) = 0$,

$$\text{Cov}(X, M) = \text{Cov}(X^* + M, M) = \text{Cov}(X^*, M) + \text{Cov}(M, M) = \text{Var}(M). \quad (12.11)$$

If there’s actually no measurement error and $M = 0$ for everybody, then $\text{Var}(M) = 0$ and thus $\text{Cov}(X, M) = 0$. Otherwise, $\text{Var}(M) > 0$, so $\text{Cov}(X, M) > 0$, implying (12.10) is not a linear projection model. Consequently, OLS does not consistently estimate $\beta_1$. (More generally, technically, if $\text{Cov}(X^*, M) = -\text{Var}(M)$, then $\text{Cov}(X, M) = 0$, but such an assumption is difficult to support in practice.)
In this case with $\text{Cov}(X, M) > 0$, 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. (Confidence intervals would be similarly biased.)

Unfortunately, outside this very special case, the type of bias may differ. It is not necessarily attenuation bias. In the case of a simple linear projection, using (12.10), the slope of the population linear projection $LP(Y \mid 1, X)$ is

$$
\frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \frac{\text{Cov}(\beta_0 + \beta_1 X + R - \beta_1 M, X)}{\text{Var}(X)}
= \frac{0 + \text{Cov}(\beta_1 X, X) + \text{Cov}(R, X) - \text{Cov}(\beta_1 M, X)}{\text{Var}(X)}
= \frac{\beta_1 \text{Cov}(X, X) + \text{Cov}(R, X^* + M) - \beta_1 \text{Cov}(M, X^* + M)}{\text{Var}(X)}
= \beta_1 + \frac{0 + \text{Cov}(R, M) - \beta_1 \text{Cov}(M, X^*) - \beta_1 \text{Var}(M)}{\text{Var}(X)}
= \frac{\text{asymptotic bias}}{\text{Var}(X)}.
$$

When $\text{Cov}(R, M) = \text{Cov}(M, X^*) = 0$, as we had earlier, the second term (the asymptotic bias) simplifies so that there is always attenuation bias. Without assuming $M$ is uncorrelated with $R$ and $X^*$, other types of asymptotic bias could result.

Unfortunately again, the bias due to measurement error does not go away even with lots of data. That is, $\hat{\beta}_1$ is not consistent for $\beta_1$, i.e., $\hat{\beta}_1 \not\xrightarrow{p} \beta_1$ as $n \to \infty$ (or, $\text{plim}_{n \to \infty} \hat{\beta}_1 \neq \beta_1$).

There are methods that address measurement error in $X$, but these are beyond our scope.
12.4.4 Survey Weights

As advised in Section 3.3.2 if your dataset has survey weights (a.k.a. sampling weights), then you should use them. Most built-in statistical estimation functions in R easily allow for such weights.

It’s true that in some cases you don’t actually need to use weights, but it’s safer to just always use them. For example, imagine a simple case where the sample contains a higher proportion female than the population, and the weights help “undo” this. If interest is in \( E(Y \mid X = x) \), where \( X \) is the dummy for female, then we don’t really need the weights since \( P(X = 1) \) doesn’t matter. That is, even though we have “too many” females, the sample average \( Y \) among females is still a consistent estimator of the population \( E(Y \mid X = 1) \), and similarly for the \( X = 0 \) subpopulation. But, if we had an \( X \) with more values and misspecified the CEF, so OLS estimates the linear projection coefficients, then the weights do matter. Regardless: you should just use weights whenever available.

12.4.5 Non-iid Sampling

Sampling may be non-iid for reasons other than weights. Clustered and/or stratified sampling can cause non-iid sampling, as discussed in Section 3.5.2. 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 <- 100*runif(n)+10
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)
out <- data.frame(est.method=c('lm/unwgtd','lm/wgtd',
    'svyglm/wgtd','lm/unwgtd','lm/wgtd'),
    SE.method=c('lm/un','lm/w','svyglm/w','coeftest/un','coeftest/w'),
    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)

## est.method  SE.method est  SE
## 1 lm/unwgtd  lm/un  2.07  0.776
## 2 lm/wgtd   lm/w  1.55  0.824
## 3 svyglm/wgtd  svyglm/w  1.55 1.426
## 4 lm/unwgtd coeftest/un  2.07 1.259
## 5 lm/wgtd coeftest/w  1.55 1.465

12.4.6 Missing Data

Like with measurement error in Y (Section 12.4.2), 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 regression 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 with regression.
12.4. THREATS TO INTERNAL VALIDITY FOR CAUSALITY

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.

Many of the big concepts about missing data originally come from Rubin (1976).

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? b) How/does this bias your regression of salary on a dummy for being a professor in a STEM field? c) Discuss your previous answers in terms of external validity.

12.4.7 Sample Selection

If we had missing data, but then our dataset only included complete cases, then we would have a sample selection problem. “Sample selection” refers to whether or not we even see individual $i$ in our data. This can be trickier since usually it is not obvious just from the dataset that sample selection is a problem. It is also trickier since we don’t know how many individuals we “should” be observing but aren’t, which precludes approaches like the worst-case bounds. (In some cases, like people not answering an invitation to complete a survey, we may know the exact number of non-responders and apply worst-case bounds.)

As with missing data, the reason behind the sample selection is crucial for whether it results in sample selection bias. For example, analogous to MCAR, 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 (iid) sample of a random sample, so we can just proceed as normal. However, analogous to non-ignorable missing data, if individuals are selected into (or out of) the sample based on their $Y_i$, then OLS (and other estimators) can be very biased. For example, imagine $Y$ is wage, and individuals with high wage are less likely to take a survey at all (since they have higher opportunity cost). 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 different in important ways compared to people who do not answer the survey.

An important economic example of sample selection is with wages. We may want to learn what determines the wage an individual is offered by a firm. However, some individuals may have a reservation wage above the wage offered by a firm, in which case they choose not to work. But if they don’t work, then we can’t observe the wage they were offered (but declined). We can only observe the wage offer if the individual accepted it and started working at that 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 should at least try to think about whether sample selection might be an issue in practice.
12.4. Threats to Internal Validity for Causality

12.4.8 Omitted Variables

Omitted variable bias is discussed in Section 9.1. 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; e.g., see Chapter 10. 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.

12.4.9 Simultaneity and Reverse Causality

When we regress \( Y \) on \( X \), we often (perhaps subconsciously) assume that \( X \) has a causal effect on \( Y \) (maybe big, maybe small, maybe even zero), but that \( Y \) does not have an effect on \( X \). However, the labels \( Y \) and \( X \) are just labels. We could have labeled the outcome variable \( Y \) and labeled the regressor \( X \). That is, it is quite possible that \( Y \) affects \( X \), in addition to \( X \) affecting \( Y \). If so, 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 \( X \). 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 better to say that price and quantity are determined simultaneously by the same economic system (simultaneity).

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. (This could be a specific type of crime, or any type of crime; the below argument still applies.) 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). In this case, OLS estimates a positive slope, but that does not mean that more police cause more crime; indeed, we assumed the effect was zero.

There are methods like instrumental variables that can (in some cases) solve the problem of simultaneity or reverse causality, but they are beyond our scope. For now, you should just try to think about whether simultaneity or reverse causality is a problem or not.

**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 status ($Y$, 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 Exercises

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 with R commands library(wooldridge) and ?voucher or Stata commands use http://faculty.missouri.edu/~kaplandm/intro_text/voucher, clear and describe

b. R only: create a data frame named df with a copy of the dataset with command df <- voucher

c. Display the total number of observations (rows) in the dataset with R command nrow(df) or Stata command count

d. Display summary statistics of mnce90 and mnce, including the number of missing observations, with R command summary(df[,c('mnce','mnce90')]) or Stata commands 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) with R command (ret1 <- lm(mnce~select, data=df)) or Stata command regress mnce select, vce(robust)

f. Repeat but adding the 1990 math test score mnce90 as a regressor, with R command (ret2 <- lm(mnce~select+mnce90, data=df)) or Stata command regress mnce select mnce90, vce(robust) noting that observations with missing mnce90 are automatically (and silently) omitted from the regression. Stata shows the number of observations actually used, which you can compare to the number in the full dataset; in R you can run length(ret2$residuals) or summary (ret2) to see this information.

g. To try to see how much of the estimate’s change is due to controlling for 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 mnce90, with R command (ret2b <- lm(mnce~select, data=df[!is.na(df$ mnce90),])) or Stata command regress mnce select if !missing(mnce90), vce(robust)
CHAPTER 12. INTERNAL AND EXTERNAL VALIDITY

h. Repeat the above three regressions but with selectyrs (number of years eligible for voucher program) instead of the binary select

i. Repeat the first three regressions but with additional regressors like female to see if they further change the coefficient on select

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 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 with R commands `library(wooldridge)` and `?card` or Stata command `bcuse card, clear`

b. R only: create a data frame named `df` with a copy of the dataset with command `df <- card`

c. Display the total number of observations (rows) in the dataset with R command `nrow(df)` or Stata command `count`

d. Show how many observations are missing IQ with R command `table(is.na(df$IQ))` or Stata command `count if missing(IQ)`

e. Run a simple regression of log wage on years of education with R command `(ret1u <- lm(log(wage)~educ, data=df))` or Stata command `regress lwage educ`, vce(robust)

f. Run the same regression but with the provided weights, with R command `(ret1w <- lm(log(wage)~educ, data=df, weights=weight))` or Stata command `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; simply replace `df` with `df[!is.na(df$IQ),]` in R, and add `if !missing(IQ)` after `educ` in the Stata command

h. Regress log wage on education and IQ (which automatically uses only observations where IQ is non-missing) with R command `(ret2w <- lm(log(wage)~educ+IQ, data=df, weights=weight))` or Stata command `regress lwage educ IQ [pweight=weight]`, vce(robust)

i. Repeat the previous three weighted regressions, but with additional regressors of your choice. “Three” being 1) without IQ, full sample; 2) without IQ, restricted to observations with non-missing IQ; 3) with IQ.
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 may be found in Diebold (2018b), Hanck et al. (2018, Ch. 14), and https://campus.datacamp.com/courses/introduction-to-time-series-analysis
Chapter 13

Time Series: One Variable

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): [https://www.youtube.com/watch?v=yCM6N8sRtPY](https://www.youtube.com/watch?v=yCM6N8sRtPY)
- Chapter 14 (“Univariate Time Series”) in Hansen (2018)
- 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)
A time series of a single variable is written as $Y_t$ for $t = 1, \ldots, T$. In this text, we assume this means a variable $Y$ is observed in $T$ different periods of equal length. (This is a “discrete time” view, as opposed to “continuous time” where $t$ has infinite decimal precision; continuous time is omitted from this text.) 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, so $T = 40$. Or, $Y$ could be the weekly return on a certain stock observed over a single calendar year, $t = 1, \ldots, 52$.

Although we’ll mostly ignore the issue, it is important in practice to know precisely when and how the “time $t$” observation is measured. For example, imagine annual GDP, 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 $Y_{t+1}$; it may make more sense to treat $X_t$ like $X_{t+1}$.

Some terminology describes 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; the first difference is $\Delta Y_t = Y_t - Y_{t-1}$. Similarly, the $j$th lag is $Y_{t-j}$. Looking to
13.1 Populations, Randomness, and Sampling

We continue the perspective of $Y_t$ as a random variable, just as $Y_i$ was earlier. It may help to review Sections 2.1 and 2.2. 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 other universes, an infinite population of 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; you could imagine running a regression using the value pairs $(Y_t, Y_{t+1})$ observed across many universes.

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, we only have a few chapters left, and we want to learn how to make some practical descriptions and predictions, so we won’t worry too much about these “deeper” issues in this book.

13.2 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 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 do any of the percentiles;
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 strict stationarity; a weaker concept called covariance stationarity requires only the means and autocovariances (of any lag) to be the same at all $t$, not the full joint distributions.

As noted, stationarity implies $\text{E}(Y_t)$ is fixed over time, as are the percentiles. Thus, an estimate of $\text{E}(Y_t)$ from historical data can be interpreted as an estimate of the future $\text{E}(Y_{t+1})$, 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 practice, you should not simply 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 13.5 contains more on nonstationarity.

### 13.3 Autocovariance and 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. Sometimes this autocorrelation is near zero, like with stock returns; today’s price change is not systematically related to yesterday’s price change. Sometimes this autocorrelation is positive: usually high quarterly GDP growth follows high growth, and low follows low, rather than jumping around randomly each quarter. A 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.

The sampling frequency, like yearly (one observation each year) or quarterly, affects the first autocorrelation. Generally, first autocorrelations are closer to positive one with a high frequency and closer to zero with a 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 frequency data. However, with yearly data (lower frequency), the first autocorrelation is lower; even more extreme, if we had decade-frequency data, the first autocorrelation may be near zero.

More generally, the $j$th autocorrelation (or autocorrelation coefficient) $\rho_j$ describes the relationship between $Y_t$ and $Y_{t-j}$, as does the related $j$th autocovariance $\gamma_j$. Stationarity is assumed, so that these values do not vary with $t$, only $j$ (the lag).
13.4. ESTIMATION

Consequently, it is the same (statistically) if we look $j$ periods in the past or $j$ periods in the future:

\[
\begin{align*}
\gamma_j &\equiv \text{Cov}(Y_t, Y_{t-j}) = \text{Cov}(Y_{t+j}, Y_t) = \gamma_{-j}, \\
\rho_j &\equiv \text{Corr}(Y_t, Y_{t-j}) = \text{Corr}(Y_{t+j}, Y_t) = \rho_{-j}, \\
\gamma_0 &\equiv \text{Cov}(Y_t, Y_t) = \text{Var}(Y_t) = \sigma_Y^2, \quad \rho_0 = \text{Corr}(Y_t, Y_t) = 1,
\end{align*}
\]

(13.1) (13.2)

\[
\text{Corr}(Y_t, Y_{t-j}) = \frac{\text{Cov}(Y_t, Y_{t-j})}{\sqrt{\text{Var}(Y_t) \text{Var}(Y_{t-j})}} = \frac{\gamma_j}{\sqrt{\sigma_Y^2 \sigma_Y^2}} = \frac{\gamma_j}{\sigma_Y^2}.
\]

(13.4)

Under stationarity, the denominator simplifies since $\text{Var}(Y_{t-j}) = \text{Var}(Y_t) = \sigma_Y^2$, a constant not dependent on $t$. Further, $\sigma_Y^2 = \gamma_0$. Thus,

\[
\sqrt{\text{Var}(Y_t) \text{Var}(Y_{t-j})} = \sqrt{\sigma_Y^2 \sigma_Y^2} = \sigma_Y^2 = \gamma_0,
\]

so

\[
\rho_j \equiv \text{Corr}(Y_t, Y_{t-j}) = \frac{\gamma_j}{\sigma_Y^2} = \gamma_j/\gamma_0.
\]

(13.5)

**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 (return), annual.

13.4 Estimation

13.4.1 Mean

With stationarity, the mean is the same $\mu = \text{E}(Y_t)$ 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 for the population mean,

\[
\frac{1}{T} \sum_{t=1}^{T} Y_t \overset{p}{\rightarrow} \mu \equiv \text{E}(Y_t).
\]

(13.6)

See [DasGupta (2008, p. 40)] for more, and Proposition 7.5 in [Hamilton (1994)] for slightly different sufficient conditions.
13.4.2 Autocovariances and Autocorrelations

With stationarity, similarly, autocovariances and autocorrelations do not depend on \( t \), so the sample autocovariances and sample autocorrelations seem promising as estimators. Since \( \rho_j = \gamma_j / \gamma_0 \), we simply estimate \( \hat{\rho}_j = \hat{\gamma}_j / \hat{\gamma}_0 \), so the following focuses on the autocovariance estimators, \( \hat{\gamma}_j \).

To estimate the \( j \)th autocovariance \( \text{Cov}(Y_t, Y_{t-j}) \), the sample \( j \)th autocovariance is

\[
\hat{\gamma}_j = \frac{1}{T} \sum_{t=1}^{T} (Y_t - \bar{Y})(Y_{t-j} - \bar{Y}).
\]  

(13.7)

This estimator is often consistent, meaning \( \hat{\gamma}_j \xrightarrow{p} \gamma_j \) 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
\[ Y_1 = -1, \quad Y_2 = -1, \quad Y_3 = 1, \quad Y_4 = 1, \] so \( T = 4 \). The sample average is zero:
\[ \bar{Y} = \frac{((-1) + (-1) + 1 + 1)}{4} / 0. \] Thus, (13.7) simplifies to
\[ \hat{\gamma}_j = \frac{1}{4} \sum_{t=1}^{4} Y_t Y_{t-j}. \]

For \( j = 0 \), \( \hat{\gamma}_0 = \frac{1}{4}(Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2) = \frac{1}{4}(4) = 1 \). For \( j = 1 \), \( \hat{\gamma}_1 = \frac{1}{4}(Y_2 Y_1 + Y_3 Y_2 + Y_4 Y_3) = \frac{1}{4}(1 + 1) = 1/4 = 0.25 \). For \( j = 2 \), \( \hat{\gamma}_2 = \frac{1}{4}(Y_3 Y_1 + Y_4 Y_2) = \frac{1}{4}((-1) + (-1)) = -2/4 = -0.5 \). For \( j = 3 \), \( \hat{\gamma}_3 = \frac{1}{4}(Y_4 Y_1) = -1/4 = -0.25 \). These match the following output.

```r
c(acf(x=c(-1,-1,1,1), type='covariance', plot=FALSE)$acf)
```

As noted, given the estimated autocovariances \( \hat{\gamma}_j \), the estimated autocorrelations are

\[ \hat{\rho}_j = \hat{\gamma}_j / \hat{\gamma}_0. \]  

(13.8)

Difficulties

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. Intuitively, an average of a single number is a bad estimator; arguments about consistency of averages assume a large number of values being averaged. This implies limits to what we can learn from data. 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, `acf()` estimates autocovariances and autocorrelations, as seen in the following code. The result \( \hat{\rho}_{12} > \hat{\rho}_6 \) seems surprising at first: \( Y_t \) is more strongly correlated with
13.5 Nonstationarity

Like endogeneity in cross-sectional data, there are many possible sources of nonstationarity in time series data. The following are some of the most common in economic data.

13.5.1 Trends

Stochastic Trends

A random walk generates nonstationary \( Y_t \). This is a special case of a more general unit root process, which all share qualitatively similar properties (including nonstationarity).
266  

CHAPTER 13. TIME SERIES: ONE VARIABLE

It is also sometimes called a **stochastic trend**. Let \( Y_0 \) be the initial value. Let

\[
Y_t = Y_{t-1} + \epsilon_t,
\]

where the increments \( \epsilon_t \) are white noise. For now, imagine the \( \epsilon_t \) have mean zero, are iid, and are independent of all past values \( Y_s \) for \( s \leq t-1 \); see Section 14.1 for more.

One way to see the nonstationarity is that

\[
\text{Var}(Y_t) = \text{Var}(Y_{t-1} + \epsilon_t) = \text{Var}(Y_{t-1}) + \text{Var}(\epsilon_t),
\]

(13.10)

since \( \text{Cov}(Y_{t-1}, \epsilon_t) = 0 \) is implied by \( \epsilon_t \perp Y_{t-1} \). Since variance is by definition non-negative, meaning \( \text{Var}(\epsilon_t) \geq 0 \), then \( \text{Var}(Y_t) \geq \text{Var}(Y_{t-1}) \), with equality only possible in the silly case when all \( \epsilon_t = 0 \) and thus \( Y_0 = Y_1 = Y_2 = \cdots \). Excluding such silliness, \( \text{Var}(Y_t) > \text{Var}(Y_{t-1}) \), which violates the property of stationarity that the variance is the same at all points in time \( t \).

The values \( Y_t \) can all be written in terms of an initial value \( Y_0 \) and the history of \( \epsilon_t \) values. Specifically,

\[
\begin{align*}
Y_1 &= Y_0 + \epsilon_1, \\
Y_2 &= Y_1 + \epsilon_2 = Y_0 + \epsilon_1 + \epsilon_2, \\
Y_t &= Y_{t-1} + \epsilon_t = \cdots = Y_0 + \sum_{s=1}^{t} \epsilon_s.
\end{align*}
\]

The first-period “shock” \( \epsilon_1 \) affects \( Y_t \) in the far future as much as \( Y_1 \), although by period \( t \) the effects of other shocks have also accumulated.

For prediction, given (13.9), it seems that the best guess (under quadratic loss) of next period’s \( Y_{t+1} \) is the current period’s \( Y_t \). That is, given \( Y_t = y_t \), \( \epsilon_{t+1} \) is mean zero, so

\[
E(Y_{t+1} \mid Y_t = y_t) = E(Y_t + \epsilon_{t+1} \mid Y_t = y_t) = E(Y_t \mid Y_t = y_t) + E(\epsilon_{t+1} \mid Y_t = y_t) = y_t + 0 = y_t.
\]

(13.11)

From before, we know 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 \( Y_{t+1} \).

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 zero new information.

Although nonstationary, the random walk can be transformed into a stationary process by taking a **first difference**, i.e., looking at the difference between two consecutive periods, \( \Delta Y_t \equiv Y_t - Y_{t-1} \). Subtracting \( Y_{t-1} \) from both sides of (13.9),

\[
\Delta Y_t \equiv Y_t - Y_{t-1} = \epsilon_t,
\]

(13.12)

and \( \epsilon_t \) is iid, which is a special case of stationarity. When a first difference of a time series produces a stationary series, the original time series is called **difference stationary**.
Deterministic Trends

Time trends also violate stationarity. The random walk is sometimes referred to as a stochastic trend; alternatively, there could be a deterministic trend, like $Y_t = t$, where 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. However, analogous to the difference stationary property of a random walk, this time series is trend stationary. Specifically, removing the deterministic trend leaves a stationary process: $Y_t - t = \epsilon_t$, which is (here assumed) stationary.

Distinguishing Trend Types

Sometimes it is difficult to distinguish a stochastic trend from a deterministic trend. For example, in climate econometrics\footnote{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/} 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, Kauppi, and Stock (2010), Chang, Kaufmann, Kim, Miller, Park, and Park (2020), and references therein.

However difficult, it is important to distinguish stochastic and deterministic trends because they affect forecasts. Roughly, a time series with deterministic trend is expected to jump back to the trend line 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. Diebold (2018c §8.1) shows a similar example with U.S. gross national product.

13.5.2 Seasonality

Another common source of nonstationarity in economics is seasonality. Seasonal time series always have higher values some seasons of the year than others. For example, over one calendar year, retail sales are always highest near the Christmas holiday season, and some agricultural crops are only harvested in one season. Residential energy use is also seasonal, with the highest heating in the winter and highest 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 often refers to seasons within one calendar year, but it can be any period of time. 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, you can add dummy variables for such events. For Easter specifically, the function `easter()` in the `forecast` package is helpful.

Seasonality is often visually apparent. Plotting \( Y_t \) over \( t \) shows a cyclical up-and-down pattern that repeats every year. For example, `plot(AirPassengers)` generates the left graph of Figure [13.1]. This graph 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., 2018; Hyndman and Khandakar, 2008).

With the air travel data plotted in Figure [13.1] (left graph), seasonality is still apparent after taking the log of the values, \( \ln(Y_t) \). The plot of \( \ln(Y_t) \) in the right graph of Figure [13.1] is generated by `plot(log(AirPassengers))`. Although both show seasonality, there is an important difference. With the original \( Y_t \) (left graph), the height of the seasonal peak each year seems to increase as \( t \) and \( Y_t \) increase. That is, the magnitude of seasonality seems proportional to the level of \( Y_t \). In contrast, with \( \ln(Y_t) \) (right graph), all the peaks seem approximately the same height. See Section [13.5.4] for more.

### 13.5.3 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
13.5. NONSTATIONARITY

Figure 13.1: Seasonality in international air travel.

component that includes cycles, too. However, it can be helpful to explicitly split out cycles; for more on cycles, see for example Diebold (2018b, §§6–7).

13.5.4 Decomposition

The observed time series $Y_t$ can be written in terms of unobserved components of trend, seasonality, and a remainder (Diebold 2018b, §2.10); more generally, the cyclical component could be separated from the trend. 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. Using the notation of Hyndman and Athanasopoulos (2019, §6), let $T_t$ denote trend, $S_t$ seasonality, and $R_t$ remainder. Then,

$$ R_t \equiv Y_t - T_t - S_t, \quad (13.13) $$

so

$$ Y_t = T_t + S_t + R_t. \quad (13.14) $$

For some series, it may make more sense to think of a multiplicative model, where the trend and seasonality are percent changes. For example, the trend could be 5% annual growth, rather than a $500 annual increase. The percentage change often makes more sense with economic data. Then,

$$ Y_t = T_t \times S_t \times R_t. \quad (13.15) $$
Actually, (13.15) is equivalent to an additive model after taking logs:

\[ \ln(Y_t) = \ln(T_t) + \ln(S_t) + \ln(R_t). \]  
(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?

For intuition, the following roughly describes “classical” additive decompositions (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.

For additive decomposition into trend (trend–cycle), seasonality, and remainder, try \texttt{decompose()} in R (in the built-in \texttt{stats} package). The following code produces Figure 13.2.

```
par(family=PARFAM, mgp=PARMGP)
ret <- decompose(co2, type='additive')
plot(ret, lwd=1, cex.axis=CEXAXIS, cex.lab=CEXLAB)
```

If instead seasonality is multiplicative, then simply replace \texttt{type='additive'} with \texttt{type='multiplicative'}.

```
par(family=PARFAM, mgp=PARMGP)
ret <- decompose(AirPassengers, type='multiplicative')
plot(ret, lwd=1, cex.axis=CEXAXIS, cex.lab=CEXLAB)
```

Other R decomposition functions to try (or Google) include \texttt{stl()}, \texttt{HoltWinters()}, and the \texttt{forecast} package’s \texttt{mstl()} (multiple seasonal).

**Discussion Question 13.2** (nonstationarity). Why specifically do you doubt the strict stationarity of each of the following time series? a) GDP, annual. b) Stock market index, annual. c) World population, annual. d) Residential water usage, monthly.
Figure 13.2: Decomposition of monthly atmospheric CO$_2$ (ppm).
Figure 13.3: Decomposition of monthly international airline passengers (thousands).
13.6 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.

One common transformation is the **first difference** (FD). That is, instead of looking at \( Y_t \) directly, we analyze the first differences,

\[
\Delta Y_t \equiv Y_t - Y_{t-1}.
\]  

(13.17)

One motivation is Section 13.5: for some nonstationary \( Y_t \), the first difference series \( \Delta Y_t \) 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 \( Y_t \).

Log transformations are also common. Sometimes it is easier to interpret time series after taking a (natural) log. That is, instead of \( Y_t \), we analyze \( Z_t = \ln(Y_t) \). Then, changes may be interpreted approximately as percent changes, if they are not too big; see Section 8.1.

Taking a log difference yields the compound growth rate. For example, for compound annual growth rates, the formula for the final level \( A \) after continuously compounded growth at 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 identity for the last equality. Thus, with annual data, the log difference \( \ln(Y_t) - \ln(Y_{t-1}) \) represents the compound annual growth rate. This is the first difference of the log-transformed series: letting \( Z_t = \ln(Y_t) \), then \( \Delta Z_t = Z_t - Z_{t-1} = \ln(Y_t) - \ln(Y_{t-1}) = \ln(Y_t/Y_{t-1}) \).
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 urate is in percent units, so 5.2 means 5.2%.

a. Load and see a description of the data with R commands `library(wooldridge)` and `?beveridge` or Stata command `bcuse beveridge , clear`

b. Tell your software that you have monthly time series data. Use R command `tsdat <- ts(data=beveridge$urate, frequency=12, start=c(2,12))` to create a time series variable named `tsdat` that’s a time series (ts) with the unemployment rate data (urate) starting in year 2000 month 12 (the first value of beveridge$month). Argument `frequency=12` says there are 12 “seasons” before getting back to the first one; in this case, 12 different months per year, but sometimes daily data would use `frequency=7` to allow day-of-week “seasonality.” In Stata, `tsset ym , monthly`

c. R only: decompose (additively) the unemployment rate time series into trend(-cycle), seasonal, and remainder components with `tsdec <- decompose(tsdat)` to compute and `plot(tsdec)` to plot. You can also see that the magnitude of the seasonal component is relatively small with `max(abs(tsdec$seasonal))`

d. Stata only: to additively decompose the time series, first estimate the trend component with a nonparametric “moving average smoother” with command `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 `tsline urate furate , name(furate) ylabel(#3)`

e. 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 generate `month = month(dofm(ym))` and compute the within-month averages with `bysort month : egen seasadd = mean(urate-furate)`

f. 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

```
sort ym
replace seasadd = seasadd - normadd
```
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 at [https://fred.stlouisfed.org/series/IPG3273N](https://fred.stlouisfed.org/series/IPG3273N) 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 with R commands `library(wooldridge)` and `?cement` or Stata command `bcuse cement, clear`

b. Tell your software that you have monthly time series data. Use R command

```r
tsdat <- 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`). In Stata,

```stata
generate yrmo = ym(year, month)
format yrmo %tm
tsset yrmo
```
c. R only: plot a multiplicative decomposition and display the seasonal effects, to see how important seasonality is for industrial cement production:

```r
tsdec <- decompose(tsdat, type='mult')
plot(tsdec)
window(tsdec$seasonal, start=c(1964,1), end=c(1964,12))
```

d. R only: compute and store a multiplicative decomposition with `tsdec <- decompose(tsdat, type='mult')` and plot it with `plot(tsdec)`

e. Stata only: estimate the trend with `tssmooth ma fipcem1=ipcem, weights(1 2 2 2 2 2 <2> 2 2 2 2 2 1)` and plot it against the raw data with `tsline ipcem fipcem1, name(fipcem1) ylabel(#3)`

f. Stata only: compute multiplicative seasonal effects with `bysort month : egen seasmult = mean(ipcem/fipcem1)` (but don’t worry about normalizing these to average to 1 like is sometimes done)

g. Stata only: compute the multiplicative remainder as the observed value divided by the trend value, divided yet again by the seasonal effect, with command `generate rem1mult = ipcem/fipcem1/seasmult`

h. 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)
```

i. Plot the autocorrelation function (ACF) of the raw data up to 48 months lag with R command `acf(tsdat, lag.max=48, ci=0, na.action=na.omit)` or Stata command `ac ipcem , level(95) lags(48)`

j. Plot the ACF of the seasonally-adjusted data with R command `acf(tsdat /tsdec$seasonal, lag.max=48, ci=0, na.action=na.omit)` or Stata commands `generate saipcem = ipcem / seasmult` and then `ac saipcem , level(95) lags(48)`

k. Repeat the decomposition plot and ACF plots for a different variable in the dataset.
Chapter 14

First-Order Autoregression

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): https://www.youtube.com/watch?v=AN0a58F6cxA
- AR(1) series with different autocorrelations (Lambert video): https://www.youtube.com/watch?v=v70-1kB3BLM
- Chapter 12 (“Serial Correlation”) in Diebold (2018a)
- Parameter stability: Hanck et al. (2018 §14.8), Diebold (2018a §12.4–5)
- AR(1) model and properties: Hamilton (1994 §3.4)
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:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t,$$  \hspace{1cm} (14.1)

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 of white noise is if the $\epsilon_t$ are iid, with mean zero and finite variance,

$$\mathbb{E}(\epsilon_t) = 0, \quad \sigma^2_\epsilon \equiv \text{Var}(\epsilon_t) < \infty,$$  \hspace{1cm} (14.2)

with all $\epsilon_t$ independent of all $Y_t$. More generally, the “independent” part of iid may be replaced by “serially uncorrelated,” meaning $\text{Corr}(\epsilon_t, \epsilon_{t-1}) = 0$, and “identically distributed” is unnecessary as long as (14.2) still holds. Diebold (2018a, §13.6) and Diebold (2018b, §6.2) have many more details on white noise (that are beyond our scope).

Assuming stationarity, the mean of $Y_t$ can be solved for in terms of $\phi_0$ and $\phi_1$. For the AR(1), $Y_t$ is stationary if and only if $|\phi_1| < 1$. Let $\mu \equiv \mathbb{E}(Y_t)$, which is the same for all $t$ if $Y_t$ is stationary. Using (14.1),

$$\mu = \mathbb{E}(Y_t) = \mathbb{E}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t) = \phi_0 + \phi_1 \mathbb{E}(Y_{t-1}) + \mathbb{E}(\epsilon_t) = \phi_0 + \phi_1 \mu.$$  \hspace{1cm} (14.3)
That is, \( \mu = \phi_0 + \phi_1 \mu \). Solving this for \( \mu \),

\[
\phi_0 = \mu (1 - \phi_1), \quad \mu = \frac{\phi_0}{1 - \phi_1}.
\] (14.4)

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, so that the result is mean zero. Here, \( Y_t - \mu \) is demeaned since \( E(Y_t) = \mu \) and \( E(Y_t - \mu) = E(Y_t) - \mu = 0 \). Similarly, since \( \mu = E(Y_{t-1}) \), \( E(Y_{t-1} - \mu) = E(Y_{t-1}) - E(Y_{t-1}) = 0 \), and similarly \( E[Y_{t-j} - \mu] = 0 \) for any \( j \) since \( E(Y_{t-j}) = E(Y_t) = \mu \) due to stationarity.

The demeaned AR(1) model is

\[
Y_t - \mu = \phi_1 (Y_{t-1} - \mu) + \epsilon_t.
\] (14.5)

Note that (14.5) is equivalent to

\[
Y_t = \mu + \phi_1 Y_{t-1} - \phi_1 \mu + \epsilon_t = \mu (1 - \phi_1) + \phi_1 Y_{t-1} + \epsilon_t,
\] (14.6)

which is the same as (14.1) since \( \phi_0 = \mu (1 - \phi_1) \) by (14.4).

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.

For the mean, assuming stationarity, then \( E(Y_t) = E(Y_{t-1}) \). Again, let \( \mu = E(Y_t) = E(Y_{t-1}) \). Taking the expectation of both sides of (14.1), and using \( E(\epsilon_t) = 0 \) from (14.2),

\[
E(Y_t) = E(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t)
= E(\phi_0) + \phi_1 E(Y_{t-1}) + E(\epsilon_t)
= \phi_0 + \phi_1 \mu.
\]

Rearranging to solve for \( \mu \),

\[
\mu = \phi_0 + \phi_1 \mu \implies \mu (1 - \phi_1) = \phi_0 \implies \mu = \frac{\phi_0}{1 - \phi_1}.
\] (14.7)

This expression is well-defined if \( |\phi_1| < 1 \), but if \( \phi_1 = 1 \), then the denominator is zero. (Explosive processes with \( \phi_1 > 1 \) are possible but not considered in this text.) Indeed, the \( Y_t \) are stationary if and only if \( |\phi_1| < 1 \). The case of \( \phi_1 = 1 \) is called a unit root,
a more general phrase for the type of nonstationarity seen in the random walk model in Section 13.5.

Similarly, the variance is determined by taking the variance of each side of (14.1). Assuming stationarity, letting $\sigma_Y^2 \equiv \text{Var}(Y_t) = \text{Var}(Y_{t-1})$, using variance identities and $\text{Cov}(Y_{t-1}, \epsilon_t) = 0$ since they’re independent,

\[
\begin{align*}
\sigma_Y^2 &= \text{Var}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t) \\
&= \phi_1^2 \text{Var}(Y_{t-1}) + \text{Var}(\epsilon_t) + 2 \text{Cov}(\phi_1 Y_{t-1}, \epsilon_t) \\
&= \phi_1^2 \sigma_Y^2 + \sigma_\epsilon^2 + 0 \\
&= \phi_1^2 \sigma_Y^2 + \sigma_\epsilon^2.
\end{align*}
\]

Rearranging to solve for $\sigma_Y^2$,

\[
\sigma_Y^2 = \phi_1^2 \sigma_Y^2 + \sigma_\epsilon^2 \implies \sigma_Y^2 (1 - \phi_1^2) = \sigma_\epsilon^2 \implies \sigma_Y^2 = \frac{\sigma_\epsilon^2}{1 - \phi_1^2}.
\] (14.8)

Again, $\phi_1 = 1$ leads to division by zero, as does $\phi_1 = -1$ now. With $\phi_1 = \pm 1$, $\text{Var}(Y_t)$ grows toward infinity as $t$ increases, violating stationarity.

Autocovariances can also be calculated. Substituting for $Y_t$ using (14.1), and using the same properties from above,

\[
\begin{align*}
\gamma_1 &\equiv \text{Cov}(Y_t, Y_{t-1}) \\
&= \text{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-1}) \\
&= \text{Cov}(\phi_1 Y_{t-1} + \epsilon_t, Y_{t-1}) \\
&= \phi_1 \text{Cov}(Y_{t-1}, Y_{t-1}) \\
&= \phi_1 \sigma_Y^2.
\end{align*}
\] (14.9)
Using (14.9) recursively,
\[
\gamma_2 \equiv \text{Cov}(Y_t, Y_{t-2}) = \text{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-2}) = 0 \text{ since } \phi_0 = \text{const} \\
= \text{Cov}(\phi_0, Y_{t-2}) + \phi_1 \text{Cov}(Y_{t-1}, Y_{t-2}) + \text{Cov}(\epsilon_t, Y_{t-2}) = 0 \text{ by } \epsilon_t \perp Y_{t-2} \\
= \phi_1 \gamma_1 = \phi_1 \phi_1 \sigma_Y^2 = \phi_1 \sigma_Y^2.
\]

(14.10)

More generally, by induction, if \(\gamma_{j-1} = \phi^{j-1} \sigma_Y^2\), then
\[
\gamma_j \equiv \text{Cov}(Y_t, Y_{t-j}) = \text{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-j}) = 0 \text{ by } \epsilon_t \perp Y_{t-j} \\
= \text{Cov}(\phi_0, Y_{t-j}) + \phi_1 \text{Cov}(Y_{t-1}, Y_{t-j}) + \text{Cov}(\epsilon_t, Y_{t-j}) = 0 \\
= \phi_1 \phi_1^{j-1} \sigma_Y^2 = \phi_1 \sigma_Y^2, \\
\]

(14.11)

which holds for all \(j \geq 0\).

The autocorrelations combine (14.11) with (13.5):
\[
\rho_j \equiv \text{Corr}(Y_t, Y_{t-j}) = \frac{\gamma_j}{\sigma_Y^2} = \frac{(\phi_1 \sigma_Y^2)}{\sigma_Y^2} = \phi_1^j.
\]

(14.12)

Plugging in \(j = 1\), the first autocorrelation \(\rho_1\) is \(\phi_1\), the autoregressive coefficient from (14.1).

### 14.3 Prediction (Forecasting)

For time series, “prediction” usually means forecasting future values of \(Y_t\) given the current and past values, which can be done in the AR(1) model as follows. Replacing \(t\) with \(t + 1\) in (14.1) yields
\[
Y_{t+1} = \phi_0 + \phi_1 Y_t + \epsilon_{t+1}.
\]

(14.13)

A forecast of \(Y_{t+1}\) considers the four right-hand side variables: \(\phi_0, \phi_1, Y_t, \text{ and } \epsilon_{t+1}\). The easiest is \(Y_t\), which we observe directly. The hardest is \(\epsilon_{t+1}\), a value that is not determined until the next period. Last, parameters \(\phi_0\) and \(\phi_1\) are unknown but assumed fixed (although this assumption may be wrong; see Section 14.5). They can be consistently estimated by OLS, as in Section 14.4.
Under quadratic loss, results from Part II show the CEF to be the best predictor. The white noise property implies (14.1) is a CEF model:

\[
E(Y_t \mid Y_{t-1} = y_{t-1}) = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t = \phi_0 + \phi_1 (\epsilon_t + \epsilon_{t-1}) = \phi_0 + \phi_1 \epsilon_{t-1}.
\]

The third equality above uses the fact that if two random variables are independent, \( V \perp W \), then \( E(V \mid W) = E(V) \); that is, independence implies \( W \) contains no information about the mean of \( V \). Replacing \( t \) with \( t+1 \) yields the CEF

\[
E(Y_{t+1} \mid Y_t = y_t) = \phi_0 + \phi_1 y_t. \tag{14.15}
\]

Consequently, the OLS-estimated \( \hat{\phi}_0 + \hat{\phi}_1 Y_t \) is a reasonable forecast of \( Y_{t+1} \) given \( Y_t \). However, accuracy may improve by using additional lags and/or other variables; Chapter 15 describe such extensions.

Even if the AR(1) model is wrong, if the time series is strictly stationary, then OLS estimates the best linear predictor (BLP) of \( Y_t \) given \( Y_{t-1} \). But “best” does not mean “good,” and allowing nonlinearity may improve the forecast. See Section 7.5 for more on the BLP.

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.

**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?

**14.4 Estimation**

Section 13.4 showed consistency of the OLS estimator in an intercept-only regression; what about the AR(1) autoregression? Specifically, Section 13.4 said that under stationarity (and ergodicity), a good estimator of the population mean is the sample mean, which is the OLS estimator for the intercept-only model

\[ Y_t = \mu + \epsilon_t. \]
14.4. ESTIMATION

Skipping the technical details, indeed, the OLS estimator $\hat{\phi}_1$ is consistent in many cases. If $|\phi_1| < 1$, then OLS is consistent, meaning $\hat{\phi}_1 \xrightarrow{p} \phi_1$; 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) can be computed. We have observed $Y_1, \ldots, Y_T$ in the sample, and now we want to forecast the unknown $Y_{T+1}$. To do this, plug in the estimated coefficients and the observed $Y_T$. For the model in (14.1),

$$\hat{Y}_{T+1} = \hat{\phi}_0 + \hat{\phi}_1 Y_T.$$  \hspace{1cm} (14.16)

For the demeaned model in (14.5),

$$\hat{Y}_{T+1} = \hat{\mu} + \hat{\phi}_1 (Y_T - \hat{\mu}).$$  \hspace{1cm} (14.17)

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 ret$x.mean*(1-ret$ar$). The predicted value in pr$pred[1]$ is shown to be equivalent to (14.16) and (14.17). (Alternatively, with the 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)). So, if we assume Gaussian $\epsilon_t$, then a 95% (for example) forecast interval could be constructed by adding and subtracting $2*pr$se from pr$pred. But, there are easier ways to make forecast intervals with the forecast package (Hyndman et al. 2018; Hyndman and Khandakar 2008). See Section 14.7 for more on forecast intervals.

```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("PhiHat0=%5.3f, PhiHat1=%5.3f\n", ret$x.mean*(1-ret$ar), ret$ar))

## PhiHat0=0.024, PhiHat1=0.143

cat(sprintf("SE(PhiHat1)=%5.3f\n", sqrt(ret$asy.var.coef)))
```
### SE(\(\hat{\phi}_1\))=0.100

```r
pr <- predict(ret, n.ahead=1)
# output point forecast and prediction SE
pr$pred, pr$se)
round(as.numeric(pr$se), digits=3))
```

```r
## 0.196 0.966
```

# check prediction against formulas
```r
c(pr$pred[1], ret$x.mean + ret$ar*(Y[n]-ret$x.mean),
ret$x.mean*(1-ret$ar) + Y[n]*ret$ar)
```

```r
## 0.196 0.196 0.196
```

# Not run: OLS to estimate \(\phi_1\) directly.
# (Note: \(\hat{\phi}_1\) estimate differs slightly.)
```r
retols <- ar(x=Y, aic=FALSE, order.max=1, method='ols',
              demean=FALSE, intercept=TRUE)
c(retols$x.intercept, retols$ar) # OLS est'd \(\phi_0, \phi_1\)
```

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? 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 yields better results.

This issue of parameter stability is related to the Lucas critique (Lucas, 1976). Even if we have a structural model, it’s possible that the true parameter values change over time. For example, a partial equilibrium model’s parameters may change when there is a new macroeconomic policy with general equilibrium effects. The policy’s effects might also change optimal predictions and descriptions of time series. The only way around this is to estimate truly “deep” parameters and economic models that are not affected by policy, which of course is extremely difficult.

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.

### 14.6 Multi-Step Forecast

Instead of forecasting \( Y_{t+1} \) given \( Y_t \), you may need to forecast \( Y_{t+h} \) for \( h > 1 \). 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 \( Y_{t+h} \) for \( h = 1, \ldots, 12 \) (the next 12 months).

One approach is to use the estimated \( \hat{\phi}_0 \) and \( \hat{\phi}_1 \) that we already estimated from the AR(1) model. We can proceed iteratively. To gain intuition, imagine \( \phi_0 = 0 \), so \( Y_t = \phi_1 Y_{t-1} + \epsilon_t \). Then

\[
E(Y_{T+1} \mid Y_T) = E(\phi_1 Y_T + \epsilon_T \mid Y_T)
= \phi_1 E(Y_T \mid Y_T) + E(\epsilon_T \mid Y_T)
= \phi_1 Y_T.
\]

Iterating once,

\[
E(Y_{T+2} \mid Y_T) = E(\phi_1 Y_{T+1} + \epsilon_{T+1} \mid Y_T)
= E[\phi_1 (\phi_1 Y_T + \epsilon_T) + \epsilon_{T+1} \mid Y_T]
= \phi_1^2 E(Y_T \mid Y_T) + \phi_1 E(\epsilon_T \mid Y_T) + E(\epsilon_{T+1} \mid Y_T)
= \phi_1^2 Y_T.
\]

This pattern continues:

\[
E(Y_{T+h} \mid Y_T) = \phi_1^h Y_T. \tag{14.18}
\]

Thus, given quadratic loss, \( \phi_1^h Y_T \) is the optimal forecast of \( Y_{T+h} \) given \( Y_T \):

\[
\hat{Y}_{T+h} = \phi_1^h Y_T. \tag{14.19}
\]

Recall that \( |\phi_1| < 1 \) if \( Y_t \) is stationary. Multiplying by a number less than one (in absolute value) makes things smaller, so \( \phi_1^h \) gets smaller as \( h \) gets bigger. In fact, it approaches zero as \( h \) increases. For example, \( (0.5)^{10} \approx 0.001 \). Thus, the farther into the future we forecast, the closer our (optimal) forecast gets to zero, which we assumed was the mean \( E(Y_t) \).
If instead \( \mu \neq 0 \), then the \( h \)-step ahead forecast can be computed in terms of \( \hat{\mu} \) and \( \hat{\phi}_1 \). For \( h = 1 \), the forecast \( \hat{Y}_{T+1} \) given \( Y_T \) satisfies
\[
\hat{Y}_{T+1} - \hat{\mu} = \hat{\phi}_1(Y_T - \hat{\mu}),
\]
equivalent to
\[
\hat{Y}_{T+1} = \hat{\mu} + \hat{\phi}_1(Y_T - \hat{\mu}) = \hat{\mu}(1 - \hat{\phi}_1) + \hat{\phi}_1 Y_T,
\]
equivalent to (14.16). For general \( h \), extending (14.19),
\[
\hat{Y}_{T+h} - \hat{\mu} = \hat{\phi}_h(Y_T - \hat{\mu}),
\]
so
\[
\hat{Y}_{T+h} = \hat{\mu}(1 - \hat{\phi}_h^h) + \hat{\phi}_h^h Y_T,
\]
which is equivalent to the forecast using \( \hat{\phi}_0 \) after plugging in \( \hat{\mu} = \hat{\phi}_0/(1 - \hat{\phi}_1) \).

Alternatively, to forecast \( Y_{T+h} \) given \( Y_T \), simply regress \( Y_{t+h} \) on \( Y_t \) and a constant. This makes sense: such a regression estimates the best linear predictor (BLP) of \( Y_{t+h} \) given \( Y_t \). This regression estimates the parameters in
\[
Y_{t+h} = \phi_0 + \phi_1 Y_t + \epsilon_{t+h}.
\]
The forecast of \( Y_{T+h} \) is then
\[
\hat{Y}_{T+h} = \hat{\phi}_0 + \hat{\phi}_1 Y_T.
\]
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 a constant and \( Y_t \) in our quarterly data, and predicting \( \hat{Y}_{T+4} = \hat{\phi}_0 + \hat{\phi}_1 Y_T \). The forecast in (14.16) showed the special case with \( h = 1 \).

There are functions in R that do multi-step forecasts automatically. In particular, the forecast function in the forecast package [Hyndman et al., 2018] [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 (14.23), 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 2.4.6 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 (14.16), an interval of numbers can help show how much uncertainty there is around the point forecast.

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. That is, imagine your job is to create 95% interval forecasts, and you make one every day for 1000 days. Each day, you make an interval forecast for the next day’s value of $Y_t$. 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 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; we do not know the true population parameters. As before, the standard error helps quantify this type of uncertainty.

Knowing the true parameters 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 $Y_{t+1} = \epsilon_{t+1}$, but 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 $\epsilon_{t+1}$.

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 most interval forecasts; e.g., see [Diebold] (2018, §7.3.3).

Continuing to assume normally distributed $\epsilon_{t+1}$, the 95% interval forecast can be stated more generally. This covers the general AR(1) $Y_{t+1} = \phi_0 + \phi_1 Y_t + \epsilon_{t+1}$, as well as any other forecast. The 95% forecast interval becomes $\hat{Y}_{T+1} \pm 1.96\hat{\sigma}_\epsilon$, i.e.,

$$ [\hat{Y}_{T+1} - 1.96\hat{\sigma}_\epsilon, \hat{Y}_{T+1} + 1.96\hat{\sigma}_\epsilon], \quad (14.26) $$

where $\hat{Y}_{T+1} = \hat{\phi}_0 + \hat{\phi}_1 Y_T$ is the point forecast. Note that 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. (Many 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 (14.26) 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.

It’s possible to estimate the distribution of $\epsilon_{t+1}$, and use the estimated percentiles to construct forecast intervals, but that approach is beyond our scope.

The following code shows basic interval forecasts using the \texttt{forecast} package (which also shows the point forecasts). The argument \texttt{h=12} specifies forecasting values for the next 12 time periods, so the results include multi-step interval forecasts. Argument \texttt{level=c(80,95)} specifies both 80% and 95% prediction intervals.

\begin{verbatim}
library(forecast)
ret <- ar(AirPassengers, aic=FALSE, order.max=1)
forecast(ret, h=10)
\end{verbatim}

<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>

\textbf{Discussion Question 14.4} (forecast sanity check). Do these point forecasts pass a sanity check? E.g., 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

The following code simulates data from an AR(1) model, and then computes various estimates and forecasts. Note that \( T = 100 \) (\( n \approx 100 \)), \( \phi_1 = 0.8 \) (\( \text{RHO} \)), \( \mu = \mathbb{E}(Y_t) = 5 \), and \( \sqrt{\text{Var}(\epsilon_t)} = 1 \) (from the \texttt{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).

```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\n", ret$ar, sqrt(ret$asy.var.coef)))
## PhiHat1=0.685, SE(PhiHat1)=0.074

(fc <- forecast(ret, h=20, level=c(80,95)))
plot(fc)
```

Figure 14.1 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 sense: the farther in the future, the less certainty we have.

The next example uses the `stlf()` function from the `forecast` package. It does much better than the earlier forecast that ignored seasonality and trend.

```r
library(forecast)
ret <- stlf(y=log(AirPassengers), h=24)
par(family=PARFAM, mar=PARMAR, mgp=PARMGP)
plot(ret)
```
Figure 14.2: Air travel data and forecast (in logs) using seasonality and trend.
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. Note that 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.”) 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 before that 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. Use R command `tsdat <- ts(data=nyse$price, frequency=52.18)` or Stata command `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, with R command `holdout <- 20` or Stata command `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 as `tsdattrain <- subset(tsdat , start=1, end=midpt)` and `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, with R command `ret <- ar (x=tsdattrain, aic=FALSE, order.max=1, method='yw')` or Stata command `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) with R command (\texttt{fc <- forecast(ret, h=holdout, level=c(80,95))) or Stata command predict fmulti', y dyn(t[_N-\text{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 with R commands \texttt{plot(fc)} and \texttt{lines(window(tsdattest), col=1)} or Stata command \texttt{twoway tsline price || tsline fmulti if _n>_N-\text{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 \texttt{cprice} ("c" for "change"). In R, when you create \texttt{tsdat}, use \texttt{data=nyse$cprice[-1]} to exclude the first value of \texttt{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. In Stata, just use \texttt{cprice} and make sure to name a different new variable in your \texttt{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 \texttt{D.price} means “take the first difference of the variable \texttt{price}” and \texttt{L.D. price} means “lag of first difference of \texttt{price}”
Chapter 15

Higher-Order Autoregression and Autoregressive Distributed Lag Regression

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 $j = 1, \ldots, p$.

Theoretical details and properties are mostly omitted, 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. Specifically: which $p$ yields the best forecasts?

15.2 Model Selection: How Many Lags?

In practice, which $p$ should we use? This is a question of model selection. Choosing $p$ is also equivalent to setting $\phi_j = 0$ for $j > p$, instead of estimating those $\phi_j$.

15.2.1 Problems and Intuition

Recall the intuition from Section 8.3. If $p$ is too small, then our model may be a poor approximation of the true CEF. Even if we estimated the coefficients perfectly, our estimated model may not predict very well. If $p$ is too big, then our model may be too flexible, overfitting the data. This also causes poor predictions. We want the “just right” $p$ that balances these two sources of error.
Only looking at in-sample fit leads to overfitting. 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 measure 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). 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 be used for model selection with other types of models, too.

### 15.2.2 AIC and BIC Formulas

There are many different but equivalent formulas for AIC and BIC. This is because only the relative values matter, not the absolute values. Thus, we could add 5 to all values, or multiply by $T$, or take the log, etc., since 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,

\[
\text{AIC}(p) = T \ln(\text{SSR}) + 2(p + 1).
\]

Intuitively, we’d like our models to fit the data well (small SSR), but we also prefer smaller models if they can get (almost) just as good fit. The penalty prevents overfitting, where a model fits the data sample “too well” because it fits all the noise, which makes its out-of-sample forecasts poor.
The BIC also involves the SSR and a penalty. Specifically,

\[
BIC(p) = T \ln(\text{SSR}) + (p + 1) \ln(T).
\]

When comparing models with different lag lengths, care is required to ensure a fair comparison. For example, if we observed \(Y_t\) for \(t = 1, \ldots, T\) and wanted to compare AR(1) and AR(2) models, each should use the same number of “data points” for estimation. For the AR(1), hypothetically we could get \(T - 1\) data points: \((Y_t, Y_{t-1})\) for \(t = 2, \ldots, T\). But for the AR(2), only \(T - 2\) data points are available: \((Y_t, Y_{t-1}, Y_{t-2})\) for \(t = 3, \ldots, T\). To make it fair, the AR(1) should only use \(T - 2\) data points, too: \((Y_t, Y_{t-1})\) for \(t = 3, \ldots, T\). More generally, when comparing models with maximum lag \(p\), only \(T - p\) data points should be used, \(t = p + 1, \ldots, T\); in the above example, \(p = 2\). This restriction is only for model selection: for estimation and forecasting, all available data should be used.

### 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 good if the true model is small but bad if the true model is large. 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. Of course, 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, AIC(1) = 11 + 4 = 15, AIC(2) = 8 + 6 = 14, BIC(1) = 11 + 7.8 = 18.8, and BIC(2) = 8 + 11.7 = 19.7. Since AIC(1) > AIC(2), \(p = 2\) is better according to AIC. However, BIC(1) < BIC(2), so the AR(1) model 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 \(Y_{T+1}\).
15.2. MODEL SELECTION: HOW MANY LAGS?

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.

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$order[1], 0, 0))
ret2 <- arima(Y, order=c(ret$order[1]-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. 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. These 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.

More generally, you can keep the intercept and include all but the first seasonal dummy. For example, 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.

```r
library(AER)
library(forecast)
data('USMacroSWQ')
GDPgr <- diff(x=log(USMacroSWQ[, 'gdp'])) # GDP growth
Tblags <- cbind(Tblag1=lag(USMacroSWQ[, 'tbill'],-1),
                 Tblag2=lag(USMacroSWQ[, 'tbill'],-2),
                 Tblag3=lag(USMacroSWQ[, 'tbill'],-3),
                 Tblag4=lag(USMacroSWQ[, 'tbill'],-4))
Tblags <- subset(Tblags, end=NROW(GDPgr))
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-0, '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))

## Point.Forecast  Lo.80  Hi.80  Lo.95  Hi.95
## 2005 Q1   0.0103 -0.0014  0.0219 -0.0076  0.0281
Empirical Exercises

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 clear` (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. Use R command `tsdat <- ts(phillips[phillips$year<=thisyr, ], frequency =1, start=yr1)` or Stata command `tsset year , yearly`

e. Stata only: define `scalar holdout = 1` and `scalar endyr = year[_N]`

f. Plot the unemployment and inflation time series with R command `plot(tsdat[, c('unem','inf')])` or Stata command `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, with R command `ret <- ar(tsdat[, 'unem '], aic=TRUE, order.max=4)` or Stata commands `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 $p = 0, 1, 2, 3, 4$ by adjusting the AIC values with `ret$aic+(log(ret$n.used)-2)*1:length(ret$aic)`

i. Using the estimates based on data years up to 1995, compute (dynamic) forecasts for the next ten years, 1996–2005, and plot them, with R commands (`fcARp <- forecast(ret, h=10)`) and Stata commands: `tsappend , add(9)`. 
predict fcARp if year>endyr-holdout, y
order year unem fcARp
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 with R commands

```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)
r/zero.alt3/zero.alt3 <- lm(Y~1, data=dat1)
r/zero.alt31 <- lm(Y~L1X, data=dat1)
r/zero.alt32 <- lm(Y~L1X+L2X, data=dat1)
r1/zero.alt3 <- lm(Y~L1Y, data=dat1)
r11 <- lm(Y~L1Y+L1X, data=dat1)
r12 <- lm(Y~L1Y+L1X+L2X, data=dat1)
r2/zero.alt3 <- lm(Y~L1Y+L2Y, data=dat1)
r21 <- lm(Y~L1Y+L2Y+L1X, data=dat1)
r22 <- lm(Y~L1Y+L2Y+L1X+L2X, data=dat1)
AICs <- data.frame(L/zero.alt3.inf=c(AIC(r/zero.alt3/zero.alt3),AIC(r1/zero.alt3),AIC(r2/zero.alt3)),
L1.inf=c(AIC(r/zero.alt31),AIC(r11),AIC(r21)),
L2.inf=c(AIC(r/zero.alt32),AIC(r12),AIC(r22)) )
rownames(AICs) <- c("L0.unem","L1.unem","L2.unem")
print(AICs, digits=4)
```

or Stata commands

```r
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, with R commands
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)

or Stata commands

predict fcADL if year>endyr-holdout, y
order year unem fcARp fcADL
list year unem fcARp fcADL if year>=endyr-holdout
Chapter 16

Final Exam

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, 278
AR(p) model, 296
ACE, see average causal effect
ADL, see autoregressive distributed lag
affine function, 173
AIC, see Akaike information criterion
Akaie information criterion, 297
alternative hypothesis, 75
analogy principle, 153
ASE, see average structural effect
asymptotic bias, 56
asymptotic theory, 55
ATE, see average treatment effect
ATT, see average treatment effect on the treated
attenuation bias, 246
attrition, 101
autocorrelation, 262
partial, 296
autocorrelation coefficient, 262
autocovariance, 262
autoregressive coefficient, 278
autoregressive distributed lag, 300
autoregressive parameter, 278
average causal effect, 97
average structural effect, 128
average treatment effect, 97
average treatment effect on the treated, 209
base category, 147
Bayesian, 44, 52, 58
Bayesian information criterion, 297
bell curve, 62
Bernoulli distribution, 18
Bernoulli random variable, 18
best linear approximation, 150
best linear predictor, 151
bias
attenuation, 52
downward, 52
negative, 52
positive, 52
toward zero, 52
upward, 52
BIC, see Bayesian information criterion
binary variable, 18
BLA, see best linear approximation
black swan, 284
BLP, see best linear predictor
CASE, see conditional average structural effect
CATE, see conditional average treatment effect
categorical, 23
causal inference, 88
CDF, see cumulative distribution function
CEF, see conditional expectation function
CEF error term, 122
centered, 62
central limit theorem, 62
CI, see confidence interval
CIA, see conditional independence assumption
classical, 44
CLT, see central limit theorem
cluster sampling, 64
complex sampling design, 65
conditional average structural effect, 201
conditional average treatment effect, 203
conditional expectation function, 122
conditional independence assumption, 202
conditional mean, 119
conditional mean independence, 205
conditional probability, 117
conditioning variable, 121
confidence interval, 66
two-sided, 67
confidence level, 67
confounder, 90, 192
confounding factor, 90
countfounding variable, 90
consistency, 56
OLS, 155
sample mean, 56
continuous random variables, 24
contrapositive, 114, 115
control variable, 201
converse, 114
correctly specified, 146
counterfactual, 88, 207
covariance stationarity, 262
covariate, 121
coverage probability, 67
cubic model, 178
cumulative distribution function, 18
data, 44
data-generating process, 22
dataset, 44
dependence, 121
dependent variable, 121
deterministic trend, 267
detrended, 270
DGP, see data-generating process
diff-in-diff, see difference-in-differences
difference stationary, 266
difference-in-differences, 198, 205
discrete, 23
dummy, 146
dummy variable, 18
economic significance, 73
elasticity, 169
empirical distribution function, 49
endogenous, 130
error form, 116
error term, 116
CEF, 122
structural, 128
estimand, 48
exogenous, 130
expectation, see expected value, 28
expected loss, 33
expected value, 18, 28
external validity, 51, 236
false positive, 75
FD, see first difference
finite-sample, 52
first difference, 260, 266, 273
first lag, 260
first lead, 261
first-order autoregressive model, 278
fitted value, 50
fitted values, 136
forecast
h-step ahead, 285
multi-step, 285
forecast interval, 287
forecasting, 257, 281
fourth moments, 154, 226
frequentist, 44, 58
fully saturated, 147, 197

GE, see general equilibrium
general equilibrium, 89
general equilibrium effects, 89

hat, 48
heterogeneity, 93
heteroskedasticity, 138
homoskedasticity, 138
hypothesis testing, 75

identification, 99
identified, 99
identifying assumptions, 99
if, 113
if and only if, 113
ignorability, 133, 202
iid, 46
implied by, 113
implies, 113
in-sample, 261
included regressor, 191
independence, 133
independent, 120
independent and identically distributed, 46
independent variable, 121
indicator function, 17
indicator variable, 18
inference, 58
information criterion, 297
intention-to-treat, 101
interaction term, 197, 221
internal validity, 236
interval estimation, 48
interval forecast, 287
interval prediction, 41
invariant, 127
inverse, 114
joint distribution, 116
joint probability, 117
large-sample, 56
latent, 240
least squares, 50
left-hand side variable, 121
level (of a hypothesis test), 76
linear combination, 173
linear projection, 148
linear projection coefficient, 149
linear-in-parameters, 173
linear-in-variables, 173, 196
linearity, 28
log function, 165
log units, 167
loss function, 33, 41
0–1, 33
$L_2$, 34
quadratic, 34
squared, 34
squared error, 34
LP, see linear projection
LPC, see linear projection coefficient

machine learning, 181
marginal distribution, 117
marginal probability, 117
Markov chain, 266
Markov property, 266
mean, 27
mean independence, 121
measurement error, 240
misspecification
functional form, 145
misspecified, 145
model, 121
linear, 164
linear-log, 168
log-linear, 166
log-log, 169
model selection, 181
multicollinearity
  imperfect, 225
  perfect, 225
multiple testing, 70
necessary, 113, 153
Neyman–Rubin causal model, 91
no interference, 133
nominal, 67
non-compliance, 100
non-interference, 133
non-random attrition, 101
non-response bias, 250
nonlinear-in-parameters, 174
nonlinear-in-variables, 174
nonparametric regression, 181
null hypothesis, 75
object of interest, 48
OLS, see ordinary least squares
omitted variable bias, 191
only if, 113
ordinal, 23
ordinary least squares, 135
out-of-sample, 261
outcome variable, 121
OVB, see omitted variable bias
overfitting, 182
oversampled, 51
p-value, 70
parallel trends, 207
partial equilibrium, 89
PDF, see probability density function
PE, see partial equilibrium
plug-in principle, 153
PMF, see probability mass function
point estimation, 48
point forecast, 283
point prediction, 41
polynomial model, 178
population, 20
  finite, 21
  infinite, 21
  super-, 21
population of interest, 236
population studied, 236
posterior distribution, 58
potential outcomes, 91
power, 76
predicted value, 50
prediction interval, 41
predictor, 121
probability density function, 24
probability mass function, 17
program evaluation, 88
projection, 148
properly specified, 146
publication bias, 71
quadratic model, 174
random draw, 20
random sample, 20, 46
random variable, 16
random walk, 265
randomized controlled trial, 99
randomized experiment, 99
RCT, see randomized controlled trial
realization, 16
realized value, 16
reduced form, 88
regressand, 121
regression, 112
regressor, 121
regressor of interest, 201
rejection probability, 76
repeated sampling, 58
residual, 50, 136
response variable, 121
reverse causality, 90, 251
right-hand side variable, 121
risk, 33
robust, 138
see also Neyman–Rubin causal model
sample, 44
sample analog, 153
sample average, 48
sample distribution, 48
sample mean, 48
sample selection, 250
sample values, 44
sampled, 20
sampling bias, 48
sampling distribution, 50, 59
sampling frequency, 262
sampling weights, 51
scaled and centered, 62
script, 9
SE, see standard error
seasonal dummies, 300
seasonality, 267
second moments, 154, 226
selection on observables, 202
self-selection, 100
serial correlation, 262
simulation error, 61
simultaneity, 251
simultaneous causality, 251
size (of a hypothesis test), 76
spillover effects, 94
SSR, see sum of squared residuals
stable unit treatment value assumption, 94
standard deviation, 29
standard error, 65
standard error of the mean, 65
stationary, 261
statistic, 48
statistical significance, 72
stochastic trend, 267
stratified sampling, 47, 64
strict stationarity, 262
strong ignorability, 133
stronger, 113
structural, 88
structural error term, 128
structural model, 127
structural parameter, 127
sufficient, 113, 153
sum of squared residuals, 50
survey weights, 51
SUTVA, see stable unit treatment value assumption
threats to validity, 236
time series, 260
time-varying coefficients, 284
transpose, 150
treated potential outcome, 92
treatment, 91
treatment effect, 91, 92
trend
stochastic, 266
trend stationary, 267
trend–cycle component, 269
two-stage cluster sampling, 64
type I error, 75
type I error rate, 76
type II error, 75
type II error rate, 76
unbiased, 53
unconditional probability, 117
unconfoundedness, 133, 202
unit root, 265, 279
units, 45
units of measure, 25
untreated potential outcome, 92
variance, 29
weaker, 113
white noise, 278