Notes for Honors Econometrics

This Version: October 7, 2012

Notes for Honors Econometrics

J. Anthony Cookson

Lulu Press

United States of America 2012

© 2012 by J. Anthony Cookson

All rights reserved. Aside from those exceptions noted below, no part of this document may be reproduced or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior written permission of J. Anthony Cookson.

Copyright Conditions

Paper copies of chapters, whole sections, or subsections of this textbook are prohibited without prior express written permission of J. Anthony Cookson.

Reproduction of practice problems, homework and review questions is allowed for non-profit educational purposes,* provided that the work is attributed to J. Anthony Cookson.

*Non-profit educational purposes include (a) reproduction of questions in assignments and homework, and (b) reproduction of questions for quizzes and examinations. Any other use of textbook does not fall under these categories and is, therefore, prohibited without the express written consent of J. Anthony Cookson.

Contents

Preface

I developed this set of notes for teaching Honors Econometrics at University of Chicago in Spring Quarter 2011 and Winter Quarter 2012. The present draft of these notes owes much to the comments from students and my TA, Xan Vongsathorn. Any remaining typos or errors are my own. This set of notes is still a work in progress, but I have published these notes in the hope that they will be helpful beyond the courses I teach. Please e-mail questions, suggestions or corrections to cookson@uchicago.edu.

This course is an introductory econometrics course that assumes familiarity with probability and mathematical statistics at the multivariate calculus level. These notes cover topics that are appropriate for a 10-week course for advanced undergraduates at an honors level. The material presented in these notes is a preview/ bridge to the study of econometrics at the graduate level as well as an advanced introduction to econometrics. For this reason, advanced undergraduates and first-year graduate students in economics (master's students and Ph.D. students looking for a succinct econometrics review) are the appropriate target audience for this text.

When I teach this course, Chapters 1 and 2 are covered before the midterm, which goes through Ordinary Least Squares in multiple regression. Chapters 3 and 4 comprise the remainder of the course, which surveys topics in generalized least squares, maximum likelihood and instrumental variables.

Each chapter has exercises throughout the text that involve proving theorems, claims or important results in econometrics. At the end of each chapter, there are exercises that are usually more involved. Some of these require statistical computation. When I teach the course, I have a dual-language requirement in Stata and R. All of the computational exercises can be done in R, but doing them in both languages can be a useful exercise in the practice of econometric methods. Although there are a variety of introductions to R available, I have made a YouTube playlist of R video tutorials to cover the basics.[1] You may access any data referenced in the text from the notes website metrics.tonycookson.com.

There is much more to the study of econometrics than is contained in these notes. I have found that pairing these notes with Angrist and Pischke's *Mostly Harmless Econometrics* is an effective combination for an advanced introduction to econometrics. Some excellent references that go beyond the material in these notes are Wooldridge's *Introductory Econometrics* and Wooldridges's *Econometric Analysis of Cross Section and Panel Data* (Wooldridge, 2002, 2003). A separate comprehensive resource is Greene's Econometric Analysis (Greene, 2003). Finally, a good textbook for background on probability and mathematical statistics is Casella and Berger's *Statistical Inference* (Casella and Berger, 2002).

[1] http://www.youtube.com/playlist?p=PL27C2ADEE810BEC09

CHAPTER 1

A Condensed Review of Probability and Statistics

This chapter provides an overview of probability concepts that serve as the foundation for econometrics. A more expansive treatment of these concepts is given in Casella and Berger (2002). For a video introduction to many of the topics covered here, check out my econometrics math YouTube playlist.[1]

1.1. Probability

DEFINITION 1.1.1. An **experiment** is any process that produces an outcome that is unknown in advance. The realizations of an experiment are called **outcomes**. Denote outcomes with ω. The set of all outcomes is called the **sample space**. Denote the sample space as Ω. Sets of realizations of an experiment are called **events**. Use capital letters near the beginning of the alphabet – $A, B, C...$– to denote events. Events are subsets of the sample space... $A \subset \Omega$. The sample space is an event, so is the null set (denoted, $\varnothing$; the set with no outcomes). The collection of all events is called the **sigma algebra**, denoted $\mathcal{B}$.

Formally, we define probability of an event in two ways.

DEFINITION 1.1.2. The **frequentist interpretation of probability** is to imagine running the same experiment repeatedly (until the end of time). The probability of an event is the long-run fraction of the number of times that the event occurs among all of the times you repeatedly run the experiment.

Although this is the most useful interpretation of probability, there are other interpretations of probability. Most intuitively, a probability is a likelihood between zero and one that we ascribe to the occurrence of an event. An event with probability equal to one are certain to occur, while an event with a probability equal to zero are certain to not occur. Note: This is a **subjective interpretation of probability** that need not match precisely with the long-run interpretation. It is correct to think of probability in subjective terms, but this suggests a more technical definition than our frequentist interpretation.

If you want a technical definition for probability, here's one:

DEFINITION 1.1.3. Given an experiment with a sigma algebra $\mathcal{B}$, a **probability measure** is a function P that takes events ($A \in \mathcal{B}$) and assigns to these events a number between 0 and 1. It must satisfy three properties that are known as the Kolmogorov Axioms:

(1) If $A \in \mathcal{B}$, $P[A] \leq 1$.
(2) $P[\Omega] = 1$
(3) If $A_1, A_2, ...$ is a sequence of mutually disjoint events in $\mathcal{B}$, $P[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i]$

The important thing to note about this definition is that probability is a function that takes sets of outcomes and turns them into numbers that we call **probabilities**. Intuitively, these probabilities allow us to assess the likelihood of the events that they describe.

[1] http://www.youtube.com/playlist?p=PLFB79C7150CFA622E

1.2. Random Variables, Distributions and Moments

DEFINITION 1.2.1. A **random variable (RV)** X is a function that takes outcomes from the sample space Ω and maps them into real numbers.

Random variables bring all of the fun of probability to the real line. We can think of the range of values that the random variable can take on as the outcomes (in the real line, denoted x; lower case of the RV), sets of these outcomes are events (still denoted $A, B, C...$), the set of all outcomes is a sample space (denoted S in the real line).

All of this is to say that once we have a random variable that we are interested in studying, we can think of the RV as our summary of the experiment. Other dimensions of the experiment (some that do not show up in our RV) may exist, but for the purposes of econometrics, we just need to know that these dimensions might exist.

There are several types of RVs worth distinguishing.

DEFINITION 1.2.2. A **discrete random variable** is a random variable that can take on a finite (technically, countable) number of points. That is, it maps into a countable subset of the real line. Each of these countable number of points takes on a strictly positive probability.

Points in the real line for which the random variable takes on positive probability are called **mass points** (or **atoms**).

DEFINITION 1.2.3. A **continuous random variable** is a random variable that can take on values in a continuum (technically, an uncountable number of points), and there are no outcomes for which the random variable takes on positive probability. That is, there are no mass points.

Why the second condition? There's a third type of RV:

DEFINITION 1.2.4. A **mixed (continuous-discrete) random variable** is a random variable that takes on a continuum of values, but there is at least one value for which the RV takes on positive probability.

We'll focus on discrete RVs and continuous RVs as two separate cases because it is easier to characterize these cases, but most of the results of probability extend to mixed RVs as well.

For any type of RV, we will want to have a way to summarize the likelihoods of the various outcomes. Distribution functions, densities and mass functions allow us to characterize the random variable in a succinct manner.

DEFINITION 1.2.5. A **distribution function** (or cumulative distribution function, CDF) is the probability of the event $(-\infty, x]$. It is a function of x, given by $F(x) = P[X \leq x]$. Intuitively, the value of the distribution function tells the probability that the random variable X takes on a value less than or equal to x. CDFs have four properties:

(1) $\lim_{x \to -\infty} F(x) = 0$
(2) $\lim_{x \to \infty} F(x) = 1$
(3) $F(x)$ is non-decreasing.
(4) $F(x)$ is "right continuous" (property stems from the $\leq$ sign in the definition)

Any function $F(\cdot)$ that satisfies these four properties is the CDF for some random variable.

CLAIM 1.2.6. On a useful note, if you want to know the probability that X is in an interval $(a, b]$, $P[a < X \leq b]$, compute $P[a < X \leq b] = F(b) - F(a)$. Convince yourself that this is true.

CDFs are useful, but it will often be convenient to summarize the distribution of the random variable by a density or mass function.

DEFINITION 1.2.7. For a continuous RV, the **probability density function** (or **pdf**) is given by $f(x) = \frac{dF(x)}{dx}$, the derivative of the CDF. By the Fundamental Theorem of Calculus, this implies $F(x) = \int_{-\infty}^{x} f(t)\,dt$.

Using Claim 1.2.6, we can obtain a formula for $P[a < X \leq b]$ in terms of pdfs.

$$\begin{aligned} P[a < X \leq b] &= F(b) - F(a) \\ &= \int_{-\infty}^{b} f(t)\,dt - \int_{-\infty}^{a} f(t)\,dt = \int_{a}^{b} f(t)\,dt \end{aligned}$$

That is, areas under pdfs are probabilities. If we want to compute the $P[X = a]$, compute the probability $X \in (a, a)$. Integrate from a to a. There's no area. Hence, there's no probability for the event $X = a$ (and there shouldn't be; continuous RVs have no mass points).

DEFINITION 1.2.8. For a discrete RV, the **probability mass function** (or **pmf**) is given by $f(x) = P(X = x)$. Note that I am using the same notation for pmfs as I use for pdfs.

To relate the pmf to its CDF, denote $A(x) \equiv \{t : P[X = t] > 0, t \leq x\}$ as the set of points with positive probability that are less than or equal to the real number x. Then, we can express the CDF as the sum of these probabilities.

$$F(x) = \sum_{t \in A(x)} f(t)$$

DEFINITION 1.2.9. Let $S = \{x : f(x) > 0\}$ be the set of outcomes for which $X = x$ has a strictly positive pmf or pdf. This set S is called **the support set**.

Both pdfs and pmfs have two important properties:

(1) $f(x) \geq 0$
(2) For discrete RV, $\sum_{x \in S} f(x) = 1$. For continuous RV, $\int_S f(x)\,dx = 1$.

Often a distribution will have a few key characteristics that are of interest. For example, we may wish to know what is a typical (or average) value for the random variable to take. We might also want to know something about the amount of dispersion or whether it is close to symmetric. These aspects of the random variable are captured by its **moments**. The simplest moment is the expectation (or mean) of the random variable:

The **expectation of a random variable** X is given by $E[X] = \sum_{x \in S} x f(x)$ if X is discrete and $E[X] = \int_{-\infty}^{\infty} x f(x)\,dx$ if X is continuous. To emphasize that $E[X]$ is constant with respect to X, we will often use the notation $\mu_X = E[X]$.

The expectation measures the central tendency. If you imagine that a density is made out of some solid material that is uniformly dense (say plywood), $E[X]$ would be the balancing point for the distribution. For discrete RVs, the expectation is literally a probability weighted average. That is a good intuition for how to think about expectations of RVs (discrete, continuous, or otherwise).

CLAIM 1.2.10. Any function of a random variable $g(X)$ is a random variable itself. Hence, we can compute its expectation.

DEFINITION 1.2.11. The **expectation of a function of a random variable** $g(X)$ is given by $E[g(X)] = \sum_{x \in S} g(x) f(x)$ if X is discrete and $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,dx$ if X is continuous.

There are lots of important examples of expectations of functions of RVs:

The **variance of a random variable** X is just the expectation of a special function of X, $\sigma_X^2 \equiv Var[X] = E\left[(X-\mu_X)^2\right]$.

The variance measures the spread of the RV. The square root of the variance of X is called the standard deviation, σ_X. If X is measured in miles, what are the units of variance? What are the units of standard deviation?

Note that the variance is just the expectation of a function of the random vector where the function is $g(X) = (X-\mu_X)^2$.

CLAIM 1.2.12. Computational Form. Another way to write the variance is $Var[X] = E\left[X^2\right] - \mu_X^2$. Can you verify that this is true?

FACT 1.2.13. *Linear Formulas. Let $Y = a+bX$, where a and b are constants. Then, $E[Y] = a+b\mu_X$ and $Var[Y] = b^2 Var[X]$. Make sure you can verify this using prior results.*

This fact implies that the expectation of a constant is just the constant and that the variance of the constant is zero.

CLAIM 1.2.14. An indicator function $g(x) = 1_A(x)$ is a piecewise function that equals 1 if $X \in A$ and equals 0 otherwise. An important result is $E[1_A(X)] = P[X \in A]$. Can you show this?

DEFINITION 1.2.15. The r^{th} **moment** of a random variable X is given by $E[X^r]$. Similarly, the r^{th} **central moment** is given by $E[(X-\mu_X)^r]$. The third central moment is called **skewness** and the fourth central moment is called **kurtosis**.

Roughly speaking, skewness measures the amount of asymmetry in the distribution of the RV; kurtosis measures the "fatness" of the distribution's tails.

DEFINITION 1.2.16. The **moment generating function (MGF) of a random variable** equals $M_X(t) = E\left[e^{tX}\right]$. The r^{th} derivative with respect to t of the MGF evaluated at $t=0$ equals the r^{th} moment (hence, the name) – $M_X^{(r)}(t)\,|_{t=0} = E[X^r]$.

1.3. Random Vectors, Joint Distributions and Covariance

Econometrics is concerned with understanding relationships between random variables. To do this, we need to introduce the idea of a random vector.

DEFINITION 1.3.1. A k-dimensional **random vector X** is a function that takes outcomes from the sample space of an experiment Ω and maps them into $\mathbb{R}^k$.

To a first approximation, you can think of a random vector as a vector of random variables, but it is more than that. The random vector includes all of the information about the relationships between the random variables in the random vector. This is the primary advantage to working with random vectors.

Analogously to random variables, we need a way to compactly describe the distribution of a random vector.

DEFINITION 1.3.2. The **joint CDF** of a k-dimensional random vector $\mathbf{X} = (X_1, X_2, ..., X_k)$ is given by the probability statement: $F(x_1, x_2, ..., x_k) = P[X_1 \le x_1, X_2 \le x_2, ..., X_k \le x_k]$.

Just as with random variables, it is often easier to work with joint pdfs or pmfs. These are defined analogously to the univariate pdfs and pmfs.

For a while, we will specialize to the case of the two-dimensional random vector $\mathbf{X} = (X, Y)$. The definitions extend naturally to k-dimensions, but it is somewhat tedious to write down.

REMARK 1.3.3. A **joint pdf** $f(x, y)$ of a continuous random vector $\mathbf{X} = (X, Y)$ must satisfy the following useful properties:

(1) $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$
(2) $P[a \leq X \leq b, c \leq Y \leq d] = \int_c^d \int_a^b f(x, y) \, \partial x \partial y$
(3) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, \partial x \partial y = 1$

You may also encounter a **discrete random vector**, which we can describe using a joint pmf

REMARK 1.3.4. The **joint pmf** $f(x, y)$ of a discrete random vector must satisfy similar properties, but with sums over the appropriate (x, y) combinations instead of integrals.

Joint pdfs and pmfs are directly related to the univariate pdfs and pmfs, which are called marginal pdfs or pmfs.

For $\mathbf{X}$ continuous, **the marginal pdf** of X, $f(x) = \int_{-\infty}^{\infty} f(x, y) \, \partial y$. For $\mathbf{X}$ discrete, **the marginal pmf** of X, $f(x) = \sum_{y \in S_Y} f(x, y)$ (where S_Y is the support set for the random variable Y)

CLAIM 1.3.5. An indicator function $g(x) = 1_A(x)$ is defined in Claim 1.2.4. Using indicator function notation, can you write the analogous properties of the joint pmf of $\mathbf{X} = (X, Y)$?

FACT 1.3.6. *A function of a random vector, $g(X, Y)$, is a random variable. That is, it (typically) has a univariate distribution, expectation, variance, and other moments from the previous section.*

DEFINITION 1.3.7. The **expectation of a function of a random vector** $E[g(X, Y)]$ is given by $E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \, \partial x \partial y$.

Relating this to the expectation of a function of a random variable, we make three subtle changes (a) we insert $g(x, y)$ instead of $g(x)$, (b) we use the joint pdf $f(x, y)$ instead of the ordinary (marginal) pdf $f(x)$, and (c) we integrate over the additional dimension.

EXERCISE 1.3.8. Using the above definitions, how would you write the **variance of a function of a random vector**, $Var[g(X, Y)]$? Hint: Because $W = g(X, Y)$ is a random variable, we don't need a new definition.

FACT 1.3.9. *Expectation of a linear combination. Let $Z = aX + bY$ where a and b are constants and $\mathbf{X} = (X, Y)$ is a random vector. Then, $E[Z] = aE[X] + bE[Y]$.*

This formula generalizes naturally to an arbitrary linear combination $Z = \sum_{i=1}^N c_i X_i$ implies $E[Z] = \sum_{i=1}^N c_i E[X_i]$.

THEOREM 1.3.10. *Jensen's Inequality. Let $g(X)$ be a convex function of a random variable. Then, $E[g(X)] \geq g(E[X])$. If $g(X)$ is concave, the inequality is reversed.*

This result formalizes the intuition that averages of a convex function are greater than extremes. Can you provide a graphical proof of this intuition?

FACT 1.3.11. *Variance of a linear combination. Let $Z = aX + bY$ where a and b are constants and $\mathbf{X} = (X, Y)$ is a random vector.*

Then, $Var[Z] = a^2 Var[X] + b^2 Var[Y] + 2abCov[X, Y]$. Can you verify this?

There are plenty of important special cases of the expectation of a function of a random vector. Covariance is one of these:

DEFINITION 1.3.12. Given a random vector $\mathbf{X} = (X, Y)$, the **covariance** between X and Y ($Cov[X, Y]$) is defined to be $\sigma_{XY} = Cov[X, Y] \equiv E[(X - \mu_X)(Y - \mu_Y)]$.

EXERCISE 1.3.13. Covariances have many useful properties. As we will be working with covariances frequently, you should understand these properties and why they are true:

(1) $Cov[X, X] = Var[X]$
(2) $Cov[X, Y] = Cov[Y, X]$
(3) Let $Z = a + bX$. Then, $Cov[Z, Y] = bCov[X, Y]$
(4) Computational Form. $Cov[X, Y] = E[XY] - E[X]E[Y]$.

The magnitude of σ_{XY} depends on the overall variability of X and Y. If we change units, this will change the observed magnitude of covariance. A rescaled/unitless measure of the association between two random variables is given by the correlation.

DEFINITION 1.3.14. The **correlation** between X and Y, $\rho_{XY} = Corr[X, Y] = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$.

THEOREM 1.3.15. *For any random vector $\mathbf{X} = (X, Y)$, the correlation $\rho_{XY} \in [-1, 1]$ where $|\rho_{XY}| = 1$ iff Y is a linear function of X.*

Proof. Hint 1. In terms of σ_X^2, σ_Y^2 and σ_{XY}, what is the minimum that $Var[aX + Y]$ can be? Hint 2. Can $Var[aX + Y] < 0$?

Covariance is one way to describe the relationship between random variables. Conditional expectation is another, but this depends on the conditional pdf or pmf.

DEFINITION 1.3.16. Given a random vector $\mathbf{X} = (X, Y)$, the **conditional pdf (or pmf)** of Y given X is defined to be $f(y|x) = \frac{f(x,y)}{f(x)}$

Using the conditional pdf (or pmf), we can compute the conditional expectation (I'll do the continuous case; you do the discrete case).

DEFINITION 1.3.17. Given a random vector $\mathbf{X} = (X, Y)$, the **conditional expectation** of Y given X is defined to be $E[Y|X] = \int_{-\infty}^{\infty} y f(y|x)\, dy$.

Note that $E[Y|X]$ is still a function of X. Functions of random variables are random variables themselves. Why is that the case here? Without being told which realization x to use for the conditioning, we don't know the value in advance of running the experiment. The notation for referring to a *particular* conditional expectation is $E[Y|X = x]$. That is, $E[Y|X = x]$ is a realization while $E[Y|X]$ is a random variable and we can compute its expectation.

THEOREM 1.3.18. ***Law of Iterated Expectations***. *For any random vector $\mathbf{X} = (X, Y)$, $E_Y[Y] = E_X\left[E_{Y|X}[Y|X]\right]$. The subscripts on the expectation operators tell which pdf (or pmf) to use for the expectation.*

This is an incredibly useful theorem in econometrics. The proof uses Definition 1.3.16. Can you prove this?

DEFINITION 1.3.19. We say that Y is **mean independent** of X if $E[Y|X] = c$ where c is some constant.

EXERCISE 1.3.20. *Consequences of Mean Independence.* If Y is mean independent of X,

(1) $E[XY] = E[X]E[Y]$
(2) $Cov[X, Y] = 0$

DEFINITION 1.3.21. We say that two random variables X and Y are **independent** if the joint pdf (or pmf) factors into the product of its marginals $f(x, y) = f(x) f(y)$.

EXERCISE 1.3.22. *Consequences of Independence.* If X and Y are independent,

(1) $f(y|x) = f(y)$
(2) Y is mean independent of X.
(3) $Cov[X, Y] = 0$

1.4. Sampling and Estimators

Using these previous results, we can define a random sample from a population.

DEFINITION 1.4.1. A **population** random variable embodies an experiment whose properties we would like to study.

In practical applications, the population random variable has a distribution (CDF, pdf, pmf) with tangible features we would like to study.[2] We call features of the population random variable **parameters** of the population.

DEFINITION 1.4.2. A **random sample** of size N $(X_1, X_2, ..., X_N)$ from a population is N independent and identical draws (iid) from the population distribution. A common shorthand notation for a random sample is $X_i \sim^{iid} X$ for $i = 1, ..., N$.
Identical means that the distribution for X_i is the same as the population distribution X. Accordingly, all of the parameters of X_i equal the corresponding parameters of X.

We will conceptually think of X_i as the i^{th} observation from our random sample. From this perspective, it is clear that X_i *is not equal* to X. It as the i^{th} observable run of the experiment represented by X.

To connect with some of the previous concepts, our random sample is a really useful example of a random vector. Assuming a continuous population, the random sample has a joint density $f_{\mathbf{X}}(x_1, x_2, ..., x_N)$.

FACT 1.4.3. *Given the definition of a random sample,* $Cov[X_i, X_j] = 0$ *for* $i \neq j$. *Why?*

EXERCISE 1.4.4. Show why the joint density from a random sample can be expressed as follows:

$$f_{\mathbf{X}}(x_1, x_2, ...x_N) = \prod_{i=1}^{N} f_X(x_i)$$

In this expression, note that X denotes the population random variable while $\mathbf{X}$ is shorthand for the random sample in vector notation $\mathbf{X} = (X_1, X_2, ..., X_n)$.

Sometimes it is useful to think of the joint density of the random sample as a function of the parameters of the population. Denote the vector of these parameters as θ. It is common to use a semicolon to indicate that the density is a function of the parameters, $f_{\mathbf{X}}(x_1, x_2, ..., x_N; \theta)$. The following definition introduces some terminology.

[2]For example, let X be the population random variable. We may wish to know about the mean, median, variance or 95th percentile of X.

DEFINITION 1.4.5. Given a random sample with joint density (or pmf) $f_{\mathbf{X}}(x_1, x_2, ..., x_N; \theta)$, we call the likelihood function $L(\theta|\mathbf{X}) = f_{\mathbf{X}}(x_1, x_2, ..., x_N; \theta)$ is the joint density (or pmf) thought of as a function of the parameters of the population.

We will see later that one way to estimate these parameters is to pick parameter values that maximize the likelihood function (which is just the joint density). This is called maximum likelihood estimation. More on this later.

DEFINITION 1.4.6. Given a random sample $X_i \sim^{iid} X$, a **statistic** $T = g(X_1, ..., X_n)$ is a random variable that is solely function of a random sample.

We will use statistics[3] to learn about parameters of the population distribution. The next definition introduces some terminology we will use extensively.

DEFINITION 1.4.7. Given a population with a parameter θ, an **estimator** $\hat{\theta}$ is a statistic that we use to estimate θ.

1.5. Small Sample Properties of Estimators

As an estimator is just a function of the random sample (which is a random vector), the estimator is a random variable itself. Estimators have distributions, expectations and all of the properties of random variables.

DEFINITION 1.5.1. We call distribution of an estimator $\hat{\theta}$ the **sampling distribution** of the estimator.

The term *sampling* distribution refers to the fact that estimators take on different values for different samples of size N from the population. This is a conceptual reminder that the variation from our estimator arises from our sampling procedure.[a]

[a] An important lesson here is that our estimator's properties will be sensitive to how we take the sample. Although this is not emphasized when we analyze the properties of our estimators, the sampling plan should be well designed. If our sampling plan introduces additional sources of variability to the experiment or if the sample is not truly a set of N independent draws from the population, our estimators will have worse properties. Once we study the basic properties of estimators (and there is a lot to explore!), we will consider some of these topics.

In practice, we won't take multiple samples. For this reason, we will not observe the sampling distribution of the estimators we use. This won't constrain our analysis of sampling distributions much. With some moderate assumptions, we can learn a remarkable amount of information about the properties of estimators from what we know about sampling distributions. One important property of estimators is **bias**.

DEFINITION 1.5.2. An estimator $\hat{\theta}$ is **unbiased** for θ if $E\left[\hat{\theta}\right] = \theta$. It is biased otherwise.

Conceptually, there are two strategies to show that an estimator is biased or unbiased.

(1) Derive the sampling distribution of $\hat{\theta}$ – denoted $f_{\hat{\theta}}\left(\hat{\theta}\right)$ – and use it to compute the expectation $E\left[\hat{\theta}\right] = \int \hat{\theta} f_{\hat{\theta}}\left(\hat{\theta}\right) d\hat{\theta}$.
(2) Use properties of expectations, covariances, conditional expectations and / or the law of iterated expectations to relate what you know about moments of the population distribution to the expectation of $\hat{\theta}$.

[3] In case it isn't clear, the main text uses "statistics" to mean "functions of random samples" rather than the "field of statistics" or "subject of statistics."

Strategy 1 is not practical for most applications. For this reason, we will usually use strategy 2 to show that an estimator is (or is not) unbiased.

EXERCISE 1.5.3. Consider two independent random samples of the same size N ($X_i \sim^{iid} X$, $Y_i \sim^{iid} Y$), where $\mu_X = E[X], \mu_Y = E[Y]$. Define the estimators $\hat{\mu}_X = \bar{X} = \frac{\sum_{i=1}^N X_i}{N}$ and $\hat{\mu}_Y = \bar{Y} = \frac{\sum_{i=1}^N Y_i}{N}$

(1) Show $\bar{X}$ is unbiased for μ_X. Show $\bar{Y} - \bar{X}$ is unbiased for $\mu_Y - \mu_X$.
(2) Compute the bias of $\hat{\theta} = (1-\rho) C + \rho \bar{X}$ relative to $\theta = \mu_X$.
(3) Show $\hat{\theta} = \bar{X}\bar{Y}$ is unbiased for $\mu_X \mu_Y$. Hint: What is $Cov\left[\bar{X}, \bar{Y}\right]$?
(4) Show $\hat{\theta} = \frac{1}{\bar{X}}$ is biased for $\frac{1}{\mu_X}$. Hint: Is $g(z) = \frac{1}{z}$ concave or convex?
(5) Show $\hat{\theta} = \left(\bar{X}\right)^2$ is biased for μ_X^2. Provide two distinct proofs (one similar to 3; another similar to 4).
(6) Derive conditions when $\hat{\theta} = \left(\bar{X}\right)^2$ is unbiased for σ_X^2.
(7) Show $S_X^2 = \frac{1}{N-1} \sum_{i=1}^N \left(X_i - \bar{X}\right)^2$ is unbiased for $\sigma_X^2 = Var[X]$.
Hint 1: $X_i - \bar{X} = (X_i - \mu_X) - \left(\bar{X} - \mu_X\right)$
Hint 2: $Var\left[\bar{X}\right] = \frac{\sigma_X^2}{N}$. By itself, this is an important result. Can you show this?
Hint 3: $Cov\left[X_i, \bar{X}\right] \neq 0$, but $Cov\left[W, \sum_{i=1}^k a_i Y_i\right] = \sum_{i=1}^k a_i Cov[W, Y_i]$. Do you remember showing something like this?

EXERCISE 1.5.4. What changes in the previous exercise when we take a random sample from a joint distribution, $(X_i, Y_i) \sim^{iid} (X, Y)$ where $\sigma_{XY} = Cov[X, Y]$. In addition, try the following:

(1) Show $S_{XY} = \frac{1}{N-1} \sum_{i=1}^N \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$ is unbiased for σ_{XY}. Use similar techniques as you used for S_X^2.
(2) Is $\hat{\theta} = \frac{S_{XY}}{S_X^2}$ unbiased for $\frac{\sigma_{XY}}{\sigma_X^2}$ in general?[4]

Another way to evaluate the quality of an estimator $\hat{\theta}$ is to compute its mean squared error (MSE). All else constant, we will prefer estimators with lower mean squared error.

DEFINITION 1.5.5. The **mean squared error** of an estimator $\hat{\theta}$ for θ is $MSE = E\left[\left(\hat{\theta} - \theta\right)^2\right]$.

We can rewrite mean squared error as $MSE = Var\left[\hat{\theta}\right] + \left(bias\left(\hat{\theta}\right)\right)^2$. Can you show this?

MSE allows us to compare the performance of biased estimators and unbiased estimators. Because MSE captures both the variance of the estimator and its bias in one measure, a comparison of MSE will tell us that a low-variance biased estimator is better than an unbiased estimator with sufficiently high variance.

EXERCISE 1.5.6. Take a random sample of size $N > 5$, $X_i \sim^{iid} X$. Compute the MSE of the following estimators of the population mean μ_X. Assume $Var[X] = \sigma_X^2$.

(1) $\hat{\theta} = \frac{X_1 + X_2}{2}$
(2) $\hat{\theta} = (1-\rho) C + \rho \bar{X}$ where C is a constant.
(3) $\hat{\theta} = \frac{3}{2}(X_1 + X_2) - \frac{2}{3}(X_3 + X_4 + X_5)$
(4) $\hat{\theta} = \bar{X} = \frac{\sum_{i=1}^N X_i}{N}$
(5) $\hat{\theta} = \frac{\sum_{i=1}^N X_i}{N+1}$

[4]We will revisit this question. This is an important estimator.

(6) $\hat{\theta} = 5$

If $\hat{\theta}$ is unbiased for θ, computing $Var\left[\hat{\theta}\right]$ is a good way to evaluate the estimator relative to other unbiased estimators. If we further restrict the class of estimators to linear estimators that are unbiased, we get the following definition.

DEFINITION 1.5.7. The estimator $\hat{\theta}$ is a **Best Linear Unbiased Estimator** (BLUE) for θ is the linear unbiased estimator θ with the lowest variance among all linear unbiased estimators $\tilde{\theta}$. That is, for any $\tilde{\theta}$ that is both linear and unbiased for θ, $Var\left[\hat{\theta}\right] \leq Var\left[\tilde{\theta}\right]$.

Consider the univariate setting for some intuition. Let $X_i \sim^{iid} X$ be a random sample. Then,$\hat{\theta} = \sum_{i=1}^{N} \hat{a}_i X_i$ is BLUE for θ if for any other linear unbiased estimator $\tilde{\theta}$ ($\tilde{\theta} = \sum_{i=1}^{N} a_i X_i$ such that $E\left[\tilde{\theta}\right] = \theta$), $Var\left[\hat{\theta}\right] \leq Var\left[\tilde{\theta}\right]$.

EXERCISE 1.5.8. Show that $\bar{X}$ is BLUE for μ_X.

Hint: First show that $\sum_{i=1}^{N} a_i = 1$ implies that $\tilde{\theta}$ is unbiased for μ. Then minimize the variance of $\tilde{\theta}$ by picking the constants in the linear combination. It may be useful to re-express $Var\left[\sum_{i=1}^{N} a_i X_i\right]$ using the special properties of the random sample (use Fact 3.3... but you know more).

1.6. Large Sample Properties of Estimators

It does not take an especially complicated econometric setting to make it difficult to verify that our estimator has good small sample properties. In particular, determining the form of the sampling distribution of $\hat{\theta}$ is usually impractical.[5] Impractical or not, we need to know *something* about the form of sampling distribution.

It seems like we are backed into a corner, but we can still make progress as long as we are willing to use *approximate* methods. Our solution to this problem is to appeal to some powerful results on convergence of sequences of random variables. Some of those results are summarized in this section.

We will study three types of convergence of random variables: convergence in distribution, convergence in probability and convergence in r^{th} mean.[6] Each type of convergence builds on what it means for a sequence of real numbers to converge to a constant.

DEFINITION 1.6.1. A sequence of real numbers $\{a_i\}_{i=1}^{\infty}$ **converges** to a real number a if for any positive constant $\epsilon > 0$, there is a point in the sequence of real numbers N^* such that after that point in the sequence $N \geq N^*$, all of the elements in the sequence are within that positive constant of the limit $|a_N - a| < \epsilon$.

We denote convergence of real numbers with limit notation:

$$\lim_{N \to \infty} a_i \quad = \quad a$$

[5] As an exercise to convince yourself of the difficulty, try deriving the exact sampling distribution of $\bar{X}$ when $\{X_i\}_{i=1}^{N}$ is drawn iid from a uniform distribution on the interval $[0, 1]$.

[6] There is another stronger type of convergence called convergence in measure or convergence almost surely. For our purposes, the three types of convergence in the main text will be enough.

1.6.1. Convergence in Probability and Consistency of Estimators. Convergence in probability is an important tool in understanding the large sample properties of estimators.

DEFINITION 1.6.2. A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ **converges in probability** to a limiting random variable X if for any positive constant $\delta > 0$, there is a point in the sequence of RVs after which $N \geq N^*$ the probability that X_N is farther than δ away from X is arbitrarily small.[a] More compactly, convergence in probability means, for any $\delta > 0$,

$$\lim_{N\to\infty} P\left[|X_N - X| > \delta\right] = 0$$

We denote convergence in probability of $\{X_n\}_{n=1}^{\infty}$ to X with the notation:

$$X_n \xrightarrow{P} X$$

[a] "Arbitrarily small" means $\forall\delta > 0, \forall\epsilon > 0$, $\forall N \geq N^*$ $P\left[|X_N - X| > \delta\right] < \epsilon$

Most econometric applications of convergence in probability are for **convergence in probability to a constant**: $X_n \xrightarrow{P} c$, which can analogously be verified:

$$\lim_{N\to\infty} P\left[|X_N - c| > \delta\right] = 0$$

DEFINITION 1.6.3. Let $X_1, ...X_n$ be a sample of size n from the population X with a parameter θ. Consider an estimator $\hat{\theta}_n = g\left(X_1, ..., X_n\right)$ computed for successively larger sample sizes. We say that the estimator $\hat{\theta}_n$ is **consistent** for θ if $\hat{\theta}_n \xrightarrow{P} \theta$.

On an intuitive level, consistency embodies two ideas: (1) The estimator is asymptotically unbiased and (2) The variance of the estimator goes to zero as the sample size increases. These ideas are captured precisely as an application of a related (but weaker) form of convergence, convergence in r^{th} mean.

DEFINITION 1.6.4. A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ **converges in r^{th} mean** to X if

$$\lim_{n\to\infty} E\left[(|X_n - X|)^r\right] = 0$$

We denote convergence in r^{th} mean with the notation $X_n \xrightarrow{r} X$. A special case of convergence in r^{th} mean is **convergence in mean square** ($r = 2$):

$$\lim_{n\to\infty} E\left[(X_n - X)^2\right] = 0$$

We denote convergence in mean square with the notation $X_n \xrightarrow{m.s.} X$.

FACT 1.6.5. *Consider a sequence of random variables constructed from taking successively larger samples and computing an estimator $\hat{\theta}_n$. Convergence in mean square to a constant θ is equivalent to asking whether $MSE_n = E\left[\left(\hat{\theta}_n - \theta\right)^2\right]$ converges to zero. Because we have the formula $MSE_n = Var\left[\hat{\theta}_n\right] + bias^2\left(\hat{\theta}_n\right)$. This implies that convergence in mean square requires both $Var\left[\hat{\theta}_n\right] \to 0$ and $bias\left(\hat{\theta}\right) \to 0$.*

THEOREM 1.6.6. *Markov's Inequality. Let X be a random variable and $h(X)$ be a non-negative function. Then, for any $k > 0$, $P[h(X) \geq k] \leq \frac{E[h(X)]}{k}$.*

PROOF. Start by writing out the expectation.

$$\begin{aligned} E[h(X)] &= \int_{-\infty}^{\infty} h(x) f(x) dx \\ &\geq \int_{\{x:h(x)\geq k\}} h(x) f(x) dx \end{aligned}$$

integrating over a subset of the support reduces the value of the integral:

$$\begin{aligned} &\geq \int_{\{x:h(x)\geq k\}} k f(x) dx \\ &= k \int_{-\infty}^{\infty} 1_{\{h(x)\geq k\}} f(x) dx \\ &= kP[h(X) \geq k] \end{aligned}$$

Putting this string of inequalities together, we obtain the desired result:

$$P[h(X) \geq k] \leq \frac{E[h(X)]}{k}$$

□

An important application of the Markov inequality relates convergence in mean square to convergence in probability.

THEOREM 1.6.7. *$X_n \xrightarrow{m.s.} X$ implies $X_n \xrightarrow{P} X$.*

EXERCISE 1.6.8. Prove this theorem. Hint: $k = \delta^2 > 0$. Can you prove a stronger result? Namely, $X_n \xrightarrow{r} X$ implies $X_n \xrightarrow{P} X$ when $r = \frac{1}{s}$ for any s even.

THEOREM 1.6.9. *$X_n \xrightarrow{r_1} X$ implies $X_n \xrightarrow{r_2} X$ for $r_1 > r_2 > 0$.*

EXERCISE 1.6.10. Prove this theorem. Hint: $(X_n - X)^{r_2} = ((X_n - X)^{r_1})^{\frac{r_2}{r_1}}$ and notice that $g(z) = z^{\frac{r_2}{r_1}}$ is concave if $r_2 < r_1$.[7]

These previous two theorems imply that for any $r > 0$, convergence in r^{th} mean implies convergence in probability. This is because for any $r > 0$, there exists an even s such that $\frac{1}{s} < r$.

THEOREM 1.6.11. *Weak Law of Large Numbers. Let $X_i \sim^{iid} X$ where $\sigma^2 = Var[X] < \infty$. Denote $\mu = E[X]$ and consider $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$. Then, $\bar{X}_n \xrightarrow{P} \mu$. That is, the sample mean is consistent for μ.*

PROOF. From above, the proof is to show that $\bar{X}_n \xrightarrow{m.s.} \mu$. Note: $E\left[\left(\bar{X}_n - \mu\right)^2\right] = Var\left[\bar{X}_n\right] = \frac{\sigma^2}{n}$, which clearly converges to zero as $n \to \infty$. By Theorem 1.6.2, $\bar{X}_n \xrightarrow{P} \mu$. □

[7] Concavity plus expected value should get you thinking about Jensen's Inequality

Another useful result for verifying consistency of estimators is the continuous mapping theorem.

THEOREM 1.6.12. *Continuous Mapping Theorem. Let* $X \xrightarrow{P} a$ *and* $Y \xrightarrow{P} b$ *and* $g(x,y)$ *be continuous at* (a,b)*. Then,* $g(X,Y) \xrightarrow{P} g(a,b)$*. That is, continuity preserves convergence in probability.*

EXERCISE 1.6.13. Let $(X_i, Y_i) \sim^{iid} (X,Y)$ with $0 < Var[X] < \infty$, $Var[Y] < \infty$, $Cov[X,Y] \neq 0$ Show the following:

(1) $\bar{X}_n \bar{Y}_n$ is consistent for $\mu_X \mu_Y$.
(2) $\frac{1}{n}\sum_{i=1}^n X_i^2$ is consistent for $E\left[X^2\right]$ as long as $Var\left[X^2\right] < \infty$.
(3) $\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$ is consistent for $Var[X]$ as long as $Var\left[X^2\right] < \infty$. Hint: As a preliminary result, show $\hat{\sigma}_X^2 = \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) - \left(\bar{X}_n\right)^2$.
(4) $S_X^2 = \frac{n}{n-1}\hat{\sigma}_X^2$ is consistent for $Var[X]$ under the same conditions as (3).
(5) $\hat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)\left(Y_i - \bar{Y}_n\right)$ is consistent for $Cov[X,Y]$ as long as $Var\left[X^2\right] < \infty$ and $Var\left[Y^2\right] < \infty$. Hint 1: As a preliminary result, show $\frac{1}{n}\sum_{i=1}^n X_i Y_i \xrightarrow{P} E[XY]$. Hint 2: Show $\hat{\sigma}_{XY} = \left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right) - \left(\bar{X}_n \bar{Y}_n\right)$. Then, apply similar techniques as above.
(6) $\hat{\theta} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$ is consistent for $\frac{Cov[X,Y]}{Var[X]}$ as long as the previous hypotheses are satisfied.

1.6.2. Convergence in Distribution and Asymptotic Normality of Estimators. An important type of convergence of random variables that is useful for hypothesis testing is convergence in distribution.

DEFINITION 1.6.14. Denote the CDF of X_n as $F_{X_n}(x)$ and the CDF of X as $F_X(x)$. A sequence of random variables $\{X_n\}_{n=1}^\infty$ **converges in distribution** to a random variable X if for all points of continuity of $F_X(x)$,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

We denote convergence in distribution of $\{X_n\}_{n=1}^\infty$ to X with the notation:

$$X_n \xrightarrow{d} X$$

DEFINITION 1.6.15. In this statement of convergence in distribution, the distribution of the random variable X is called the **limiting distribution**. Alternatively, we may refer to it as the **asymptotic distribution**. Knowledge of the asymptotic distribution of an estimator can be useful in constructing hypothesis tests.

THEOREM 1.6.16. *Central Limit Theorem (CLT). For each* n*, let* $X_1, ..., X_n$ *be a random sample from a population* X *with expectation* $E[X] = \mu < \infty$ *variance* $\left(\sigma^2 = Var[X]\right)$ *and* $0 < \sigma^2 < \infty$*. Form the sample mean for each random sample* $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$*. Under these conditions, the sequence of standardized sample means converges in distribution to a standard normal distribution:*

$$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0,1)$$

where $N(0,1)$ *is shorthand for the standard normal distribution. The standard normal distribution a continuous random variable with a mean of 0 and variance of 1 with the pdf*

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

The normal distribution is generally denoted as $Z \sim N\left(\mu, \sigma^2\right)$ with pdf:

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

This is an incredibly useful result. Together with the following facts about normal distributions (and some other related results to be developed later), the CLT gives us an approximate distribution for many sampling distributions we will want to study. We don't typically use the CLT as stated. Rather, we use the asymptotic distribution as an approximation to the sampling distribution of $\bar{X}_n$:

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

FACT 1.6.17. *Useful Properties of Normal RVs. Let $X_1, X_2, ..., X_n$ be independent normal RVs with mean μ_i and variance σ_i^2, and c_i $i = 1, ..., n$ are constants.*

(1) $Z = \sum_{i=1}^n c_i X_i$ has is distributed normally with mean $\mu_Z = \sum_{i=1}^n c_i \mu_i$ and variance $\sum_{i=1}^n c_i^2 \sigma_i^2$.
(2) $Z = \frac{X_i - \mu_i}{\sigma_i} \sim N(0,1) \iff X_i \sim N\left(\mu_i, \sigma_i^2\right)$.
(3) Suppose that (X_i, X_j) has bivariate normal (BVN). That is, (X_i, X_j) has joint density:

$$f_{X_i,X_j}(x_i, x_j) = \frac{1}{2\pi\sigma_{x_i}\sigma_{x_j}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_i-\mu_i)^2}{\sigma_{x_i}^2} + \frac{(x_j-\mu_j)}{\sigma_{x_j}^2} - \frac{2\rho(x_i-\mu_i)(x_j-\mu_j)}{\sigma_{x_i}\sigma_{x_j}}\right]\right)$$

If $Cov[X_i, X_j] = 0$, then X_i and X_j are independent. This is one of the special cases when zero covariance implies independence. Can you show that zero covariance implies independence in the case of a bivariate normal?

As long as the hypotheses are satisfied, the Central Limit Theorem gives us that sample means of random variables converge in distribution to a normal distribution. Although we will not literally be computing sample means of random samples in every application, this result is far more general than it first appears.

FACT 1.6.18. *We can apply the CLT to both of the following expressions $\frac{1}{n}\sum_{i=1}^n X_i Y_i$ or $\frac{1}{n}\sum_{i=1}^n X_i \exp[Y_i - Z_i]$ because they are sample means of some random variable. The first one is $\bar{Z}_n$ when $Z_i = X_i Y_i$ and the second one is $\bar{W}_n$ when $W_i = X_i \exp[Y_i - Z_i]$. Can you give the conditions under which the CLT applies to these settings?*

Beyond thinking about sample means more broadly, we will often be interested in combining convergence in distribution with convergence in probability. Slutsky's Theorem gives us some additional techniques to extend the Central Limit Theorem logic (you can think of this as an extension of the continuous mapping theorem).

THEOREM 1.6.19. *Slutsky's Theorem. Let $X_n \xrightarrow{d} X$, and $Y_n \xrightarrow{p} c$. Then, the following conditions are true:*

(1) $X_n + Y_n \xrightarrow{d} X + c$
(2) $Y_n X_n \xrightarrow{d} cX$
(3) $X_n / Y_n \xrightarrow{d} X/c$ as long as $c \neq 0$

Slutsky's Theorem extends convergence in distribution from the CLT to statistics that we can relate to sample means through adding, multiplying and dividing. It turns out that there is a much more general extension of these convergence in distribution results, given by the Delta Method.

THEOREM 1.6.20. *Delta Method. Suppose that* $\sqrt{n}\left(Z_n - z_0\right) \xrightarrow{d} N\left(0,1\right)$ *and that* $g\left(z\right)$ *is a continuously differentiable function at* z_0*. Then,*

$$\sqrt{n}\left(g\left(Z_n\right) - g\left(z_0\right)\right) \quad \xrightarrow{d} \quad N\left(0, \left(g'\left(z_0\right)\right)^2\right)$$

PROOF. Taking the first order Taylor expansion of $g\left(Z_n\right) - g\left(z_0\right)$ about z_0, we know:

$$\sqrt{n}\left[g\left(Z_n\right) - g\left(z_0\right)\right] \quad = \quad \sqrt{n}\left(\left[g\left(z_0\right) - g\left(z_0\right)\right] + g'\left(\hat{Z}\right)\left(Z_n - z_0\right)\right)$$

for some $\hat{Z}$ between $Z_n - z_0$. This simplifies:

$$\sqrt{n}\left[g\left(Z_n\right) - g\left(z_0\right)\right] \quad = \quad g'\left(\hat{Z}\right)\sqrt{n}\left(Z_n - z_0\right)$$

Once we establish that $g'\left(\hat{Z}\right) \xrightarrow{P} g'\left(z_0\right)$,[8] we can apply Slutsky's Theorem to obtain the result directly. □

In practice, $g\left(Z_n\right)$ will be an estimator whose properties we wish to study. In the statement of the Delta Method, we will typically think of $g\left(Z_n\right) \xrightarrow{P} g\left(z_0\right)$. The Delta Method gives us an approximation of the distribution for this estimator:

$$g\left(Z_n\right) \quad \approx \quad N\left(g\left(z_0\right), \frac{\left(g'\left(z_0\right)\right)^2}{n}\right)$$

In the language of econometrics, we say that $g\left(Z_n\right)$ is a consistent estimator of $g\left(z_0\right)$ and we say that $\sqrt{\frac{\left(g'\left(z_0\right)\right)^2}{n}}$ are the asymptotic standard errors of this estimator for $g\left(z_0\right)$.

EXERCISE 1.6.21. Convergence in Distribution Practice.

(1) Let $X_n \xrightarrow{d} X$ and $X_n - Y_n \xrightarrow{P} 0$. Prove that $Y_n \xrightarrow{d} X$.
(2) Let $\{X_i\}_{i=1}^n$ be a random sample from a population with mean $\mu = 3$ and standard deviation $\sigma = 1$. Give the asymptotic properties (limiting distribution) of $W = \sqrt{n}\frac{\bar{X}_n}{S_X^2}\left(\bar{X}_n - 3\right)$. Justify your answer.
(3) Suppose you take a random sample from a population X with mean $\mu \neq 0$ and variance σ^2. What is the limiting distribution of $\frac{1}{\bar{X}_n}$? Can you apply the Delta Method to an expression involving $\frac{1}{\bar{X}_n}$ to obtain some information about the asymptotic properties of $\frac{1}{\bar{X}_n}$?

[8] Here's the proof of this fact. $Z_n - z_0 \xrightarrow{d} 0$ because $\frac{1}{\sqrt{n}}\sqrt{n}\left(Z_n - z_0\right) = \frac{1}{\sqrt{n}}N\left(0,1\right) \xrightarrow{d} 0$. Convergence in distribution to a constant implies convergence in probability. Hence, $Z_n \xrightarrow{P} z_0$. This further implies $\hat{Z} \xrightarrow{P} z_0$ because $\hat{Z}$ is between z_0 and Z_n. Moreover, $g'\left(\cdot\right)$ is continuous at z_0 and we know that $\hat{Z} \xrightarrow{p} z_0$. This implies that $g'\left(\hat{Z}\right) \xrightarrow{P} g'\left(z_0\right)$ by the continuous mapping theorem.

1.7. Statistical Inference

The convergence in probability results (WLLN and CMT) are useful in justifying our choice of an estimator while the convergence in distribution results (CLT and Delta Method) help us to learn about the shape of the sampling distribution of our estimator.

We can use the sampling distribution of the estimator to make **inference**. That is, use the outcome of the sample and some assumptions about the sampling process to make statements about population parameters, either by providing a set of plausible values for the parameter (**confidence intervals**) or by assessing the validity of one particular parameter (**hypothesis tests**). For simplicity, we will consider inference for means, but we will do so in a way that will quickly extend our inference results to other population parameters.

1.7.1. How do all of the results fit together? Take a random sample $X_1, X_2, ..., X_n$ from a population random variable X and suppose we wish to test a claim about the population mean μ. We know from the weak law of large numbers (WLLN) that a natural estimator for μ is the sample mean $\bar{X}$ because it is consistent for μ. We know from the central limit theorem (CLT) that

$$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \quad \xrightarrow{d} \quad N(0,1)$$

If we knew (or could reliably assume) the values of μ and σ^2, we could use the realization of our sample $\bar{x}$ in place of the random variable $\bar{X}_n$, and we could treat $z_n = \frac{\bar{x}_n - \mu}{\sqrt{\sigma^2/n}}$ as a draw from Z_n. This allows us to make probability statements about the values of μ that are most likely given our sample.

When we conduct a hypothesis test about the mean, the null hypothesis implies a value for μ, but in practice, we will not know the value of σ^2. Instead, we plug in a sample analog to σ^2, the sample variance S^2. Given a sample, we can compute its realization and therefore compute

$$t_n \quad = \quad \frac{\bar{x}_n - \mu}{\sqrt{s^2/n}}$$

which is a realization from the random variable:

$$T_n \quad = \quad \frac{\bar{X}_n - \mu}{\sqrt{S^2/n}}$$

Assuming the population random variable X is distributed normally, T_n is distributed as a t distribution with $n-1$ degrees of freedom at every sample size n. In many econometric applications, the assumption of normality in the population is too strong of an assumption. It is rarely the case that we can assume that the population is normally distributed, even approximately so.[9] For this reason, we will not assume a normal population and we will not obtain a t distribution for inference.

[9]For example, the normal distribution has infinite support. Prices, wages, earnings and quantities are strictly positive. For this reason, they cannot be normal, and could only be approximately normal. If we are going to use an appoximate method of inference, we might as well use a method that is grounded in a formal proof of convergence in distribution. That's the motivation to move toward asymptotic inference.

Leaning on asymptotic theory and the CLT, we will typically use the normal distribution in place of the t distribution because it is a valid asymptotic distribution.[10]

CLAIM 1.7.1. As long as $E\left[X^4\right] < \infty$ along with the assumptions made before, we can show (using the asymptotic results in the previous section) that $T \xrightarrow{d} N(0,1)$.

Given this result and plugging in s^2 for S^2, the approximate distribution of $\bar{X}$ is normal:

$$\bar{X} \approx N\left(\mu, \frac{s^2}{n}\right)$$

This leads to the simplest form of inference: confidence intervals.

1.7.2. Confidence Intervals. Suppose that $\hat{\theta}$ is a consistent estimator for θ and that we have used the results on converence in distribution to demonstrate that $T = \frac{\hat{\theta}-\theta}{se\left(\hat{\theta}\right)} \xrightarrow{d} N(0,1)$, where $se\left(\hat{\theta}\right)$ is called the standard error, which is a consistent estimator of the standard deviation of the sampling distribution of $\hat{\theta}$.[11]

Because T is asymptotically standard normal, we can use it to construct a (two-sided) confidence interval for θ. The symmetry of the normal distribution suggests that we will want a confidence interval that is symmetric about our estimate for the parameter θ. Given this motivation, consider the probability statement

$$P\left[-c_\alpha \leq T \leq c_\alpha\right] = 1-\alpha$$

The constant c_α is called a two-sided critical value and its value depends on α, the level of significance. For example, if T is standard normal and $\alpha = 0.05$, $c_\alpha = 1.96$. For a two-sided confidence interval c_α is actually the $1-\frac{\alpha}{2}$ quantile of T's sampling distribution.[12]

Inside the probability operator, we can apply algebraic steps without changing the probability

$$\begin{aligned} P\left[-c_\alpha \leq \frac{\hat{\theta}-\theta}{se\left(\hat{\theta}\right)} \leq c_\alpha\right] &= P\left[-c_\alpha se\left(\hat{\theta}\right) \leq \hat{\theta}-\theta \leq c_\alpha se\left(\hat{\theta}\right)\right] \\ &= P\left[-c_\alpha se\left(\hat{\theta}\right) \leq \theta-\hat{\theta} \leq c_\alpha se\left(\hat{\theta}\right)\right] \\ &= P\left[\hat{\theta}-c_\alpha se\left(\hat{\theta}\right) \leq \theta \leq \hat{\theta}+c_\alpha se\left(\hat{\theta}\right)\right] \end{aligned}$$

In other words, before we compute the realizations of our random sample the interval

$$\left(\hat{\theta}-c_\alpha se\left(\hat{\theta}\right), \hat{\theta}+c_\alpha se\left(\hat{\theta}\right)\right)$$

contains the true population parameter θ with probability $1-\alpha$.

[10] As long as the sample size is "large enough," this approximation is good. Using an asymptotic approximation for a small sample will produce invalid estimates. Still, we do not want to make an assumption about the shape of the population distribution if it is unwarranted. If there is time this quarter, we will discuss some techniques for inference that work well in small samples.

[11] An alternative method to find the formula for the standard error is to make sure it satisfies two properties: (1) is possible to compute given the sample and (2) is the "right" statistic to use in the denominator to achieve $T \xrightarrow{d} N(0,1)$.

[12] Therefore, if you find yourself in a non-standard situation with a sampling distribution that is not normal, you can always look up the critical value using software. In R, $qnorm(0.975)$ equals 1.96. In addition, there is a host of other common distributions available.

DEFINITION 1.7.2. More generally, let $\mathcal{C}_{1-\alpha}$ be a set of plausible values for θ that depends on the realization of a random sample. To be a $(1-\alpha)\times 100\%$ confidence interval, $C_{1-\alpha}$ must satisfy the coverage property $\lim_{n\to\infty} P\left[\theta \in \mathcal{C}_{1-\alpha}\right] = 1-\alpha$.

EXAMPLE 1.7.3. Let $\{X_i\}_{i=1}^n$ be a random sample from X. Suppose we want a 95% confidence interval for μ. Let's translate the general notation to this specific case, $\hat{\theta} = \bar{X}$, $\theta = \mu$ and $se\left(\hat{\theta}\right) = \frac{S}{\sqrt{n}}$. Given this, a 95% confidence interval for μ is given by $\bar{X} \pm 1.96 \times \frac{S}{\sqrt{n}}$.

Given the realizations from a random sample $\bar{X} = 25$, $S = 14$ and $n = 49$, we compute the interval $25 \pm 1.96 \times \frac{14}{\sqrt{49}} = (21.08, 28.92)$.

EXERCISE 1.7.4. In the previous example, does $(21.08, 28.92)$ have a 0.95 probability of containing the true population mean μ? Explain.

1.7.3. T-based Hypothesis Testing in General. In hypothesis testing, we would like to assess the validity of a paricular claim about the population parameter θ. We refer to the claim about the parameter we want to test as the ***null hypothesis*** and use the notation H_0. We refer the the claim that may be true if the null hypothesis is false as the ***alternative hypothesis***, and use the notation H_1 or H_a. The null hypothesis is the hypothesis of the status quo or no change. It is usually the conservative thing to conclude. The alternative hypothesis is often called the research hypothesis. This is the statement we would like to make about the population.

To test a null hypothesis $H_0 :\ \theta = \theta_0$ against the alternative $H_1 :\ \theta \neq \theta_0$, we need some measure of distance of our sample data from the hypothesized value of the parameter. If $\hat{\theta}$ is a consistent estimator for θ, we will use the T-ratio that we used to motivate confidence intervals as our measure of distance between our estimate $\hat{\theta}$ and our hypothesized value θ_0.

$$T \quad = \quad \frac{\hat{\theta} - \theta_0}{se\left(\hat{\theta}\right)}$$

Intuitively, the computed value of T tells us (approximately) how many standard deviations of the sampling distribution the estimate is from the hypothesized value. If $|T|$ is large enough (above some threshold), the data are inconsistent with the null hypothesis being true. Formally, we will say that we **reject** the null hypothesis in these cases. If $|T|$ is small enough that we think θ_0 to be a plausible value for θ we will say that we **fail to reject** the null hypothesis.[13]

As long as T has a known distribution that does not depend on any unknown parameters, we can use it for inference. In general, if a test statistic's distribution is free of unknown parameters the distribution is called **pivotal**.

1.7.3.1. *Type I and Type II error.* A decision rule may incorrectly accept or reject H_0 because there is still uncertainty in the process. We have uninformative names for the two types of errors we can make. **Type I error** occurs when the null hypothesis H_0 is true, but we mistakenly reject it. **Type II error** occurs when H_1 is true, but we fail to reject the null hypothesis H_0. Given a sample size, there is a trade-off between these two types of errors. To decrease Type I error, we need to reduce the threshhold for rejecting (rejecting fewer hypothesis, failing to reject more). This increases Type II error. The only way to reduce both types of errors at the same time is to increase the sample size.

[13]If we reject H_0, we can conclude H_1. If we fail to reject H_0, we cannot conclude H_1 unless we are really confident in our data and estimation plan. If we had little hope of detecting a meaningful deviation from the null hypothesis (due to loads of sampling variability), failing to reject might reflect more information about our sample than about the null hypothesis.

Following standard practice, we will choose decision rules so that the probability of Type I error does not exceed α. We refer to α as the ***level*** of the test when

$$P\left[\text{Type I Error}\right] = \alpha \ .$$

Moreover, we will choose our decision rule make this happen. The level of the test dictates our cutoff rule, but it is also the rate at which our testing procedure makes Type I errors. We do not control Type II error explicitly in our testing procedure. If we cannot control the sample size, we cannot control the rate at which we make Type II errors. In practice, Type II error can be quite high and this is a problem (but it is only a problem if we fail to reject the null hypothesis).

1.7.3.2. *Two-Sided Tests for μ.* In a **two-sided test** regarding μ, we consider null and alternative hypotheses of the form

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0, \end{aligned}$$

and define a test statistic

$$T = \left| \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \right|,$$

Technically, the denominator of T could be any consistent estimator of the variance of $\bar{X}_n$. We call this denominator the ***standard error*** of $\bar{X}$ or $se(\bar{X})$. It is an estimate of the standard deviation of the sampling distribution of $\bar{X}$.

The test statistic T is a random variable for which, if H_0 is true, large realizations are very unlikely. Let t represent a realization of the test statistic (random variable) T. Thus our decision rule will reject H_0 if the realization t of is too large.

Recall that we derive our decision rule by controlling the Type I error probability to be some predetermined level α. If H_0 is true, we want to reject the hypothesis with probability α. If H_0 is true, the test statistic T is approximately normal, a symmetric distribution. For this reason, we can select the critical value c such that the probability of rejecting when $T > c$ is $\alpha/2$. Formally, we pick c so that

$$\begin{aligned} P\left[|T| > c\right] &= \alpha \\ &\Rightarrow \\ P\left[T > c\right] &= \frac{\alpha}{2} \end{aligned}$$

With some rearranging, we obtain $P\left[T < c\right] = 1 - \frac{\alpha}{2}$. In other words, c is the $1 - \frac{\alpha}{2}$ quantile of the sampling distribution of T. We sometimes label the critical value for a two-sided test of level α as $c_{1-\alpha/2}$.

EXAMPLE 1.7.5. Suppose that we wish to test the claim that $\mu = 6$ against the two-sided alternative at the five percent level. The null and alternative hypotheses are $H_0 : \ \mu = 6$ and $H_1 : \ \mu \neq 6$. Given a random sample $\{X_i\}_{i=1}^n$, we can compute the test statistic $T = \frac{\bar{X}-6}{se\left(\bar{X}\right)}$ and we will reject the null hypothesis whenever $|T| > 1.96$ (because 1.96 is the 97.5th quantile of the standard normal distribution).

An important point in this example is that one can set up the hypothesis test without actually having data to carry it out.

- If we were given that the sample mean is $\bar{X} = 4$ with a standard error of 0.5, the computed value of $T = \frac{4-6}{0.5} = -4$. This is greater in magnitude than the critical value. Hence, we would reject.
- If we were given that the sample mean is $\bar{X} = 5.5$ with a standard error of 0.5, the computed value of $T = \frac{5.5-6}{0.5} = -1$, which is smaller in magnitude than the critical value. Hence, we fail to reject.

Finally, there is another important aspect of a hypothesis test: the p-value.

DEFINITION 1.7.6. The **p-value of a test** is the probability of observing a test statistic more extreme (in the direction of the alternative hypothesis) than the computed value of the test statistic. Formally, given a computed value $T_n = t$

$$p - value = P\left[|T| > t\right]$$

An alternative way to conduct a hypothesis test is to examine the p-value. If $p - value < \alpha$, reject the null hypothesis.

1.7.4. One-Sided Tests regarding μ. In a ***one-sided test*** about μ the null and alternative hypotheses are:[14]

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0, \end{aligned}$$

and now define the test statistic

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} .$$

Again, we know the large sample approximation of the distribution of T_n under H_0, so we can find a critical value c where

$$P\left[T > c\right] = \alpha .$$

under the asymptotic distribution.

This time, given our definition of T_n, and its approximate distribution under H_0 we know

$$P\left[\frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} > c\right] \to 1 - \Phi(c)$$

where $\Phi\left(z\right)$ is the standard normal CDF. So we find c that solves

$$(1 - \Phi(c)) = \alpha \implies \Phi(c) = 1 - \alpha.$$

We sometimes label the critical value for a one-sided test of level α as $c_{1-\alpha}$. To compute the p-value for a one-sided test, just compute the probability under one tail of the sampling distribution (in the direction of the alternative hypothesis). In this case, we can compute the p-value as

$$p - value = P\left[T_n > t\right] \approx 1 - \Phi(t)$$

[14]We could have also chosen $H_1 : \mu < \mu_0$ if the context warrants that hypothesis, but we would need to select a negative cutoff value rather than a positive cutoff value because only negative values of the test statistic will convince us that the null hypothesis is wrong in the direction of the alternative.

1.7.5. Two-Sided tests about $\mu_X - \mu_Y$. We can extend the idea of hypothesis testing to the simple difference between two parameters. We continue to use the sample $X_1, \ldots, X_n$ as described. Now we add an additional *independent* sample from Y, and let $Y_1, \ldots, Y_m$ be random sample of size m from Y. We will also assume that $E[Y^4] < \infty$ and $Var[Y] > 0$. We define the null and alternative hypothesis as follows:

$$\begin{aligned} H_0 &: \mu_X - \mu_Y = d_0 \\ H_1 &: \mu_X - \mu_Y \neq d_0 \end{aligned}$$

For a two-sided test, we define the test statistic in the usual way

$$T = \left| \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\widehat{Var}\left[\bar{X} - \bar{Y}\right]}} \right| .$$

This should be a straightforward extension once we notice that we are now interested in a parameter about the distribution $X - Y$, for which we have collected two separate (and independent) samples on X and on Y.

EXERCISE 1.7.7. Explain why a natural choice for the variance estimator in the denominator is

$$\hat{Var}\left[\bar{X} - \bar{Y}\right] = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$$

Namely, prove that $\hat{Var}\left[\bar{X} - \bar{Y}\right] \xrightarrow{P} Var\left[\bar{X} - \bar{Y}\right]$. How does the validity of this estimator depend on the independence of the two samples? If the sample means are positively correlated, how would this estimator be wrong (too large or too small)? Justify.

Once we have a consistent estimator for the estimator's standard deviation, all we need to do is apply the general form of the hypothesis test to this specific setting. When we cover topics in single and multiple regression, we will cover the details of inference in more detail.

1.8. Homework Exercises

(1) **Consistency and Convergence in Probability**. Complete Exercise 1.6.3 and Fact 1.6.1 from the probability and math review notes. Use the notes as your guide to provide a proof that convergence in r^{th} mean ($r > 0$) implies convergence in probability.

(2) Consider a test of the null hypothesis $H_0 : \theta = \theta_0$ against its two-sided alternative. The **power** of a hypothesis test is defined to be the probability of rejecting the null hypothesis (assuming a particular value of the parameter $\theta = \tilde{\theta}$ is true). Use the notation $\kappa\left(\tilde{\theta}\right)$ to denote the power of a hypothesis test as a function of the assumed-true parameter value $\tilde{\theta}$. For the following questions, assume that we construct a test statistic $T_n = \frac{\hat{\theta}_n - \theta_0}{se\left(\hat{\theta}_n\right)}$ and suppose that we have a large enough sample that T_n is well-approximated by a standard normal distribution.

   (a) How is power related to the probability of making a Type I error α? Hint: It is possible for $\tilde{\theta} = \theta_0$. What if $\tilde{\theta} \neq \theta_0$?

   (b) How is power related to the probability of making a Type II error $\beta\left(\tilde{\theta}\right)$?

   (c) Graph of the power function (You may hand draw or use R to plot it). To make your graph precise, assume that $\alpha = 0.05$, $\theta = \mu_X$, $\theta_0 = 6$, $n = 64$, $S^2 = 16$ and $\hat{\theta}_n = \bar{X}_n$. Plot for a range spanning the points $\tilde{\theta} \in \{3, 4, 5, 6, 7, 8, 9\}$. If you are using R, you should plot more points than this (and use the `lines()` command) to produce a smoother plot.

(d) On the same plot as (c), redraw the power function if $n = 256$ instead of $n = 64$.

(3) Prove or give a counterexample.
   (a) Can an estimator be unbiased for all sample sizes, but still be inconsistent?
   (b) Can an estimator be consistent, but be biased for all sample sizes?
   (c) Can a sequence of random variables converge in probability, but not converge in mean square?
   (d) Does Y is mean independent of X imply that X is mean independent of Y?
   (e) Does $Cov[X, Y] = 0$ imply X and Y are indepdendent?

(4) **Chebyshev's Inequality**. Use Markov's Inequality (Theorem 1.6.1) to prove $P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$ for any positive constant k.
   (a) Suppose that $X \sim Expon(1)$. In this example, compute $P[|X - \mu| \geq 2\sigma]$ and compare with the bounds given by the Chebyshev inequality.

(5) Let X be a random variable that denotes the amount of exposure a student has to mathematics in high school (in thousands of hours). Y is a random variable denoting a student's performance on an economics aptitude test.
   (a) Suppose that $X \sim Expon(1)$, where the exponential distribution $Expon(\mu)$ is parametrized with pdf $f_X(x) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)$. Show that the mean $E[X] = \mu$ and that the r^{th} raw moment can be computed as $E[X^r] = \mu^r r!$. *Hint: Integration by parts.*
   (b) Prove the population ANOVA theorem: $Var[Y] = E[Var[Y|X]] + Var[E[Y|X]]$.
   (c) Suppose that the conditional distribution of Y given X is $N\left(60 + 2X^3, 10X\right)$. Compute the expected value and variance of Y. Your answers should be numbers, not random variables.
   (d) Use the Chebyshev inequality and your answer for Y to obtain a bound on 75% of the data. If someone told you that the test score Y can only be a positive number, would you be surprised? Why or why not?

(6) We know that S_X^2 is unbiased for σ_X^2. Show that S_X is a biased estimator of σ_X. Is the expectation of S_X too large or too small? Justify. Is S_X consistent for σ_X? Provide a proof.

(7) **Simulated Asymptotics**. Using R, perform the following Monte Carlo exercise.[15]
   (a) Generate 1000 independently drawn samples of size 100 $(X_1, ..., X_{100})$ from a $Beta(1, 60)$ distribution. For each sample, store the following statistics:
      (i) $\hat{\theta}_1 = min\{X_1, ..., X_{100}\}$
      (ii) $\hat{\theta}_2 = \bar{X}$
      (iii) $\hat{\theta}_3 = max\{X_1, ..., X_{100}\}$
      (iv) $\hat{\theta}_4 = X_1$
      (v) $\hat{\theta}_5 = median\{X_1, ..., X_{100}\}$
      (vi) $\hat{\theta}_6 = \frac{\hat{\theta}_3 - \hat{\theta}_1}{2}$
   (b) Plot the histograms of the statistics you generated in part (a). Comment on the shape of the distributions. Do any of these shapes surprise you?
   (c) Repeat the exercise in (a) and (b), but draw instead from a $Uniform[0, 60]$.
   (d) Repeat the previous two exercises, but use samples of size 15. What changes?

[15] I have posted some R tutorials that should be useful for getting R to do this exercise.

(8) **Computation using matrices**. Let $\mathbb{X}$ denote a 5×2 matrix and $\mathbf{Y}$ denote a 5×1 vector. Let these be given by:

$$\mathbb{X} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 4 \\ 1 & 5 \\ 1 & 2 \end{bmatrix}$$

and

$$\mathbf{Y} = \begin{bmatrix} 14 \\ 17 \\ 8 \\ 16 \\ 2 \end{bmatrix}$$

respectively. Perform the following computations by hand.

(a) $Q = \mathbb{X}'\mathbb{X}$; $\det(Q)$; Q^{-1}.
(b) $P = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$. Is P full rank? Show that P is idempotent.
(c) $M = I - P$. Is M full rank?

(9) **Manipulating regression matrices**. Let $\mathbb{X}$ denote an $(n \times k)$ matrix, whose rank is k and let

$$\begin{aligned} Q &= \mathbb{X}'\mathbb{X} \\ A &= Q^{-1}\mathbb{X}' \\ P &= \mathbb{X}A \\ M &= I - P \end{aligned}$$

Also, let

$$\begin{aligned} \widehat{\beta} &= A\mathbf{Y} \\ \widehat{\mathbf{Y}} &= P\mathbf{Y} \\ \widehat{\mathbf{U}} &= M\mathbf{Y} \end{aligned}$$

Show as concisely as possible that:

(a) $AP = A$, $AM = 0$, $MP = 0$, $PM = 0$.
(b) $P\mathbb{X} = \mathbb{X}$, $M\mathbb{X} = 0$.
(c) $P\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}$, $P\widehat{\mathbf{U}} = 0$.
(d) $M\widehat{\mathbf{Y}} = 0$, $M\widehat{\mathbf{U}} = \widehat{\mathbf{U}}$.
(e) $\mathbb{X}'\widehat{\mathbf{Y}} = \mathbb{X}'\mathbf{Y}$
(f) $\mathbf{Y}'\widehat{\mathbf{Y}} = \mathbf{Y}'\mathbb{X}\widehat{\beta} = \widehat{\beta}'\mathbb{X}'\mathbf{Y} = \widehat{\beta}'Q\widehat{\beta} = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}$.
(g) $\widehat{\mathbf{U}}'\widehat{\mathbf{U}} = \mathbf{Y}'M\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}$.

(10) **Constructing data matrices**. Suppose that we interview $n = 100$ people and we ask them whether they are males or females. Suppose that n_F is the number of female responses and n_M is the number of male responses. Obviously, $n_F + n_M = n$. We create three variables. Let X_1 denote a variable that is recorded as a 1 for everybody. Let X_2 denote a variable that is recorded as 1 if the respondent is male and 0 if the respondent is female. Let X_3 denote a variable that is recorded as 1 if the respondent is female and 0 if the respondent is male.

(a) Write out a (1×3) row vector that records the typical male response.
(b) Write out a (1×3) row vector that records the typical female response.
(c) Write out the entire matrix in partitioned form, if we partition across variables. Be sure to be very explicit. Label it $\mathbb{X}_1$.

(d) Write out the entire matrix in partitioned form, if we partition across observations (we group males and females). Be sure to be very explicit. Label it $\mathbb{X}_2$.
(e) Compute $\mathbb{X}'\mathbb{X}$ in both cases ($\mathbb{X}$ is equal to $\mathbb{X}_1$ and $\mathbb{X}_2$ respectively).
(f) Is $\mathbb{X}$ a full column rank matrix?

(11) **Raw Statistical Calculation in R**. Use R to verify your work on questions 8 and 10. What happens when you try to invert $\mathbb{X}'\mathbb{X}$ from Question 10? Comment and submit your code with this assignment. For Question 8, you may assume that $n_F = 27$. You may find a use for the rep(), cbind() (and/or rbind()), t(), matrix multiplication (%*%) and matrix inversion (solve()) commands in R.

In addition, use R to perform the following calculations using the setting from Question 9.
(a) $SSE = \hat{\mathbf{U}}'\hat{\mathbf{U}}$, $SSR = \left(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right)' \left(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right)$ and $SST = \left(\mathbf{Y} - \bar{\mathbf{Y}}\right)' \left(\mathbf{Y} - \bar{\mathbf{Y}}\right)$ where $\bar{\mathbf{Y}}$ is a vector with the sample mean of $\mathbf{Y}$ in every element.
(b) $R^2 = {}^{SSR}/_{SST}$
(c) $MSE = {}^{SSE}/_{(n-2)}$ where $n =$# of rows in $\mathbf{Y}$. $MSR = {}^{SSR}/_{k}$ where $k =$# of columns in $\mathbb{X}$ minus 1.
(d) $F = \frac{MSR}{MSE}$
(e) $VC = MSE \times Q^{-1}$. Compute $\sqrt{VC_{ii}}$ for $i = 1, 2$. Store these calculations in a vector $se\left(\hat{\beta}\right)$.
(f) Perform element-by-element division of $\hat{\beta}$ by $se\left(\hat{\beta}\right)$.

(12) **Canned Statistical Calculation in R.** Use the setting from Question 8 to verify that the calculations in Question 11 are correct.
- Bind the matrix $\mathbb{X}$ to the vector $\mathbf{Y}$ using the cbind() command.
- Coerce the resulting matrix object to be of type "data.frame" using the function as.data.frame().
- Use the lm() command to estimate a regression of Y on X. Summarize the regression output using the lm() command the the anova() command on your linear models object.
- Comment on the relationship of the canned output to the raw statistical calculations in Question 11.

(13) **Simulation of random data from a CDF**. This result is incredibly useful to provide simulation from a known distribution.

If F is a strictly increasing distribution function on the interval $[a, b]$ of the real line (potentially infinite endpoints), then the inverse function F^{-1} is a well-defined mapping from $[0, 1]$ to $[a, b]$. Suppose $X \sim Uniform[0, 1]$. Then, $F^{-1}(X) \sim F$.

This result is useful because computing packages always have the capacity to generate n random draws from a uniform distribution, but the functionality may not be implemented for other distribution functions. Take a random variable X with distribution function F. If we know the CDF and can solve for its inverse, this result says that we can simulate random draws from it.[16] Implement this algorithm on the following common distribution functions.
(a) Use R and this algorithm to simulate $n = 1000$ draws from the distributions $Exponential(6)$, $Cauchy(0, 1)$, $Logistic(2, 4)$ and $Pareto(1, 2)$. If you do not know them off hand, you can look up the distribution functions from a reliable statistics textook.

[16]Here's an example of how to implement the algorithm in R (http://novicemetrics.blogspot.com/2010/05/simulating-from-t1ev-gumbel.html).

(b) Plot the histograms of the reandom draws from these distributions. Whenever possible, verify that these draws produce a similar histogram plot to the canned algorithms for random number generation in R.

CHAPTER 2

Linear Regression

Chapter 1 was a quick review of the concepts of mathematical statistics used in econometrics. This chapter begins the study of regression, which allows us to understand relationships between economic variables. We will draw on the results of the introductory chapter as needed. This chapter fills in some details (and skips others) found in Chapter 3 of Angrist and Pischke's *Mostly Harmless Econometrics* (Angrist and Pischke, 2009).

2.1. The Statistical Interpretation of Linear Regression

Start by thinking very carefully about what we are trying to estimate: the population regression.

DEFINITION 2.1.1. **The Population Regression.** Let $(\mathbf{X}, Y, U)$ be a random vector, where $\mathbf{X} = \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$. A linear regression model imposes a linear relationship between Y and $\mathbf{X}$.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + U$$

where we call the random variable Y the **dependent variable** (regressand, response). We call each random variable X_j is an **independent variable** (regressor, explanatory variable). We call the random variable U the **error term** and we call the $\beta_0 + \beta_1 X_1 + ...\beta_k X_k$ part the **regression line** or the systematic component of the population regression model.. The constants $\beta_0, \beta_1, ..., \beta_k$ are called regression coefficients, with β_0 called the **intercept** and β_j $(j \geq 1)$ called the **slope** coefficient of X_j.

In this definition of the population regression, the first element of $\mathbf{X}$ is 1, which technically isn't random, but its purpose is apparent when we write the regression model in matrix notation. This vector form of the population regression is usually more useful in proofs.

DEFINITION 2.1.2. **Vector Form of Population Regression**. Using some vector arithmetic, we can express this population regression much more compactly.

$$Y = \mathbf{X}'\beta + U$$

where β is a $(k+1) \times 1$ vector, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$.

For simple concepts, it is sometimes easier to focus on the case of one regressor.

DEFINITION 2.1.3. **The Population Regression with One Regressor (Simple Linear Regression).** Let (X, Y, U) be a random vector, where X is a random variable. A linear regression model imposes a linear relationship between Y and X.

$$Y = \beta_0 + \beta_1 X + U$$

There are two interpretations of this regression model. Both are used in practice, but they are very different ways to look at a regression model. We will draw a sharp distinction between these two.

(1) **Statistical Interpretation** ("Reduced Form," "Descriptive").
(2) **Causal Interpretation** ("Structural").

For much of this chapter, we will stick with the statistical interpretation of regression because it shines light on how linear regression works mechanically.

2.1.1. Statistical Interpretation. If (X, Y) is a random vector, there is always a statistical relationship between the random variables X and Y.

Let's think about using X to predict Y. A natural way to predict Y given X is to use the conditional expectation, $E[Y|X]$. Not only is it a natural way to predict Y, it is the best prediction of Y using the information in X.

FACT 2.1.4. *The conditional expectation of Y given $\mathbf{X}$ minimizes the mean squared error among all other predictors of Y that are merely a function of $\mathbf{X}$. That is,*

$$E[Y|\mathbf{X}] = \arg\min_{f(\mathbf{X})} E\left[(Y - f(\mathbf{X}))^2\right]$$

Note: $f(\mathbf{X})$ is a function of a vector $\mathbf{X}$. Alternatively, we could have written this $f(X_1, X_2, ..., X_k)$.

An ideal regression model will perfectly match $E[Y|\mathbf{X}]$, but this ideal may be unattainable. The conditional expectation $E[Y|\mathbf{X}]$ can be an arbitrary nonlinear function of $\mathbf{X}$. For this reason, it is often not possible to perfectly rationalize $E[Y|\mathbf{X}]$ using only a linear function of $\mathbf{X}$.

The regression line in the statistical interpretation of regression is the best linear approximation of the conditional expectation $E[Y|\mathbf{X}]$.

DEFINITION 2.1.5. **Best Linear Approximation to $E[Y|\mathbf{X}]$.** Given a random vector $(\mathbf{X}, Y, U)$, the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + U$$

is the best linear approximation to $E[Y|\mathbf{X}]$ if

$$(\beta_0, \beta_1, ..., \beta_k) = \arg\min_{b_0, b_1, ..., b_k} E\left[(E[Y|\mathbf{X}] - (b_0 + b_1 X_1 + ... + b_k X_k))^2\right]$$

or more compactly using vector notation:

$$\beta = \arg\min_{b} E\left[\left(E[Y|X] - \mathbf{X}'b\right)^2\right]$$

Together with a definition of U, this best linear approximation property defines our statistical interpretation of the linear regression model.

DEFINITION 2.1.6. A linear regression model has **the statistical interpretation** if it satisfies two properties.

(1) The **systematic component** $\mathbf{X}'\beta$ is the best linear approximation to $E[Y|\mathbf{X}]$.
(2) The **error term** $U \equiv Y - \mathbf{X}'\beta$ is the difference of the response variable Y from the systematic component of the regression model.

We may also refer to this as a **statistical regression model** or a **reduced form model**.

FACT 2.1.7. *The solution to the best linear approximation (BLA) of* $E[Y|\mathbf{X}]$ *equals the solution to the best linear prediction (BLP) for* Y*. Formally,*

$$\underset{b_0,b_1,...,b_k}{\arg\min}\ E\left[\left(E[Y|\mathbf{X}] - (b_0 + b_1X_1 + ... + b_kX_k)\right)^2\right] \quad = \quad \underset{b_0,b_1,...,b_k}{\arg\min}\ E\left[\left(Y - (b_0 + b_1X_1 + ... + b_kX_k)\right)^2\right]$$

PROPOSITION 2.1.8. *Under the statistical interpretation of regression,* ***the orthogonality conditions*** $E[\mathbf{X}U] = \mathbf{0}$ *are satisfied.*

PROOF. Consider the case of one regressor. By the statistical interpretation, we know that the choice of regression coefficients solves the best linear approximation problem.

$$\beta = (\beta_0, \beta_1) \quad = \quad \underset{b_0,b_1}{\arg\min}\ E\left[\left(E[Y|X] - (b_0 + b_1X_1)\right)^2\right]$$

By the previous fact, we also know the parameter vector solves the best linear predictor problem:

$$\beta = (\beta_0, \beta_1) \quad = \quad \underset{b_0,b_1}{\arg\min}\ E\left[\left(Y - (b_0 + b_1X_1)\right)^2\right]$$

Taking first order conditions of this problem, we obtain.

$$\begin{aligned} [b_0]: \ -2E[(Y - b_0 - b_1X)] &= 0 \\ [b_1]: \ -2E[(Y - b_0 - b_1X)X] &= 0 \end{aligned}$$

Because β_0 and β_1 solve this system of first order conditions, we know

$$\begin{aligned} E[(Y - \beta_0 - \beta_1X)] &= 0 \\ E[(Y - \beta_0 - \beta_1X)X] &= 0 \end{aligned}$$

These are precisely the orthogonality conditions $E[U] = 0$ and $E[XU] = 0$ in the case of one regressor. This argument extends naturally to the case of k regressors. □

PROPOSITION 2.1.9. ***Linear Conditional Expectation.*** *Suppose that* $E[Y|\mathbf{X}]$ *is a linear function of* X.[1] *Under the statistical interpretation of linear regression,* $E[Y|\mathbf{X}] = \mathbf{X}'\beta$*, and as a consequence,* $U = Y - E[Y|\mathbf{X}]$*. In the case of simple linear regression, these expressions are* $E[Y|X] = \beta_0 + \beta_1X$ *and* $U = Y - E[Y|X]$*.*

[1] In general, this property is satisfied when (Y, X) is a bivariate normal random vector or when X takes on only two values. Otherwise, this is an assumption.

PROOF. Sketch. Consider the single regression case. The statistical interpretation implies that the regression coefficients satisfy

$$(\beta_0, \beta_1) = \underset{b_0,b_1}{\arg\min}\, E\left[\left(E[Y|X] - (b_0 + b_1 X)\right)^2\right]$$

Because $E[Y|X]$ is linear, we can write it as an arbitrary linear function $E[Y|X] = \tilde{\beta}_0 + \tilde{\beta}_1 X$. In this special case, the statistical interpretation implies

$$(\beta_0, \beta_1) = \underset{b_0,b_1}{\arg\min}\, E\left[\left(\tilde{\beta}_0 + \tilde{\beta}_1 X - (b_0 + b_1 X)\right)^2\right]$$

Clearly, $b_0 = \tilde{\beta}_0$ and $b_1 = \tilde{\beta}_1$ implies that the $E\left[(\cdot)^2\right]$ term equals zero, which is as small as it can be. Hence, $\beta_0 = \tilde{\beta}_0$ and $\beta_1 = \tilde{\beta}_1$. In other words, the statistical regression line given by $\beta_0 + \beta_1 X$ equals the conditional expectation of Y given X, $E[Y|X] = \tilde{\beta}_0 + \tilde{\beta}_1 X$ □

REMARK 2.1.10. **Properties of the Statistical Interpretation under Linear Conditional Expectation.** A couple of properties fall out immediately from this proposition.

(1) U is mean independent of X. That is, $E[U|X] = 0$.
(2) Without appealing directly to the first order conditions of the best linear approximation problem, the orthogonality conditions $E[U] = 0$ and $E[XU] = 0$ are satisfied.

FACT 2.1.11. *Under the statistical interpretation of linear regression, $Cov[X_j, U] = 0$ where X_j is a regressor and U is the statistical error term. This is true regardless of whether the conditional expectation $E[Y|\mathbf{X}]$ is linear.*

Statistical Interpretation for Single Regression. Consider the case of a single regressor. Under the statistical interpretation, we have the following properties:

(1) The systematic component of the linear regression model $p(X) = \beta_0 + \beta_1 X$ represents both the best linear approximation to $E[Y|X]$ and the best linear predictor of Y. The error term $U = Y - \beta_0 - \beta_1 X$ is uncorrelated with the regressor, $Cov[X, U] = 0$.
(2) For this reason, we interpret the slope coefficient β_1 as the unit change in *our prediction* of Y (or *our approximation* to $E[Y|X]$) from increasing X by one unit. Using some unnecessary math, $\beta_1 = \frac{dp(X)}{dX}$.
(3) We interpret the error term U as the statistical discrepancy of the dependent variable from the best possible linear prediction given X. This is what Angrist and Pischke mean when they say that U does not have a "life of its own" under the statistical interpretation.

Linear Special Case. If the conditional expectation of Y given X is linear, the best approximation to the conditional expectation becomes exact. Therefore, we refine all of the above properties:

(1) The systematic component of the linear regression model $p(X) = \beta_0 + \beta_1 X$ equals the conditional expectation of Y given X. The error term $U = Y - \beta_0 - \beta_1 X$ is mean independent of the regressor $E[U|X] = 0$.
(2) For this reason, we interpret the slope coefficient β_1 as the unit change in *the conditional expectation of Y given X*, $E[Y|X]$, from increasing X by one unit. Using some unnecessary math, $\beta_1 = \frac{dE[Y|X]}{dX}$.
(3) We interpret the error term U as the statistical discrepancy of the dependent variable from the conditional expectation of Y given X.

These properties hold for multiple regression as well.

Statistical Interpretation for Multiple Regression. Consider the case of a single regressor. Under the statistical interpretation, we have the following properties:

(1) The systematic component of the linear regression model $p(X_1, X_2, ..., X_k) = \mathbf{X}'\beta$ represents both the best linear approximation to $E[Y|\mathbf{X}]$ and the best linear predictor of Y. The error term $U = Y - \mathbf{X}'\beta$ is uncorrelated with the regressor, $Cov[X, U] = 0$.
(2) For this reason, we interpret the slope coefficient β_1 as the unit change in *our prediction* of Y (or *our approximation* to $E[Y|X]$) from increasing X_j by one unit, holding constant the values of the other regressors. In mathematics, $\beta_1 = \frac{\partial p(X_1, X_2, ..., X_k)}{\partial X_j}$.
(3) As in single linear regression, we interpret the error term U as the statistical discrepancy of the dependent variable from the best possible linear prediction given X. This is what Angrist and Pischke mean when they say that U does not have a "life of its own" under the statistical interpretation.

The linear special case generalizes analogously.

2.2. Identification of Population Regression Parameters

The algebra of regression is simplest and the intuition is clearest if we start with one regressor.[2] Throughout this section, the goal is to describe the statistical regression model more fully. Specifically, we would like to solve for the regression parameters in terms of estimable features of the joint distribution of *observable* random variables.

2.2.1. Single Linear Regression. In our single linear regression model, we have (X, Y, U) a random vector where

$$Y = \beta_0 + \beta_1 X + U$$

with the orthogonality conditions:

$$\begin{aligned} E[U] &= 0 \\ E[XU] &= 0 \end{aligned}$$

We want to derive an expression for β_0 and β_1 in terms of the features of the joint distribution of X and Y. Because X and Y are random variables whose outcomes we can observe, this will be useful in estimating β_0 and β_1.

Let's try to compute $Cov[X, Y]$:

$$\begin{aligned} Cov[X, Y] &= Cov[X, \beta_0 + \beta_1 X + U] \\ &= \beta_1 Cov[X, X] + \underbrace{Cov[X, U]}_{=0} \end{aligned}$$

where $Cov[X, U] = 0$ by the orthogonality conditions, which implies

$$\beta_1 Var[X] = Cov[X, Y]$$

[2] I will use the terminology simple linear regression, single linear regression and linear regression with one regressor to mean the same thing.

DEFINITION 2.2.1. To solve for β_1, we need to assume $0 < Var[X] < \infty$. Assuming a finite and positive variance in our regressor amounts to asserting that there is some variability in X that we can use to obtain information about how Y changes with X. Such variability is necessary to identify β_1 in the data. For this reason, we call this assumption on variance an **identification assumption**.

As long as the identification assumption is satisfied, $\beta_1 = \frac{Cov[X,Y]}{Var[X]}$.

To get an equation for β_0 in terms of parameters of the joint distribution use $E[U] = 0$ and the definition of U from a statistical model.

$$\begin{aligned} 0 = E[U] &= E[Y - \beta_0 - \beta_1 X] \\ &= E[Y] - \beta_0 - \beta_1 E[X] \end{aligned}$$

Solving for $\beta_0 = E[Y] - \beta_1 E[X]$. Now, we have an expression for β_0 and β_1 in terms of parameters of the joint distribution of (X, Y).

$$\begin{aligned} \beta_1 &= \frac{Cov[X,Y]}{Var[X]} \\ \beta_0 &= E[Y] - \left(\frac{Cov[X,Y]}{Var[X]}\right) E[X] \end{aligned}$$

When we estimate the parameters of statistical regression model, these are the features of the population that we estimate.

2.2.2. Multiple Linear Regression. Similar to the single linear regression model, suppose we study the statistical model with $(\mathbf{X}, Y, U)$ a random vector where

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + U \\ &\text{or} \\ Y &= \mathbf{X}'\beta + U \end{aligned}$$

with the orthogonality conditions:

$$\begin{aligned} E[U] &= 0 \\ E[X_j U] &= 0 \quad \forall j \in \{1, 2, ..., k\} \end{aligned}$$

Or, in matrix form $E[\mathbf{X}U] = \mathbf{0}$.

As with simple regression, we want to derive an expression for $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$ in terms of the features of the joint distribution of $\mathbf{X}$ and Y because this will be useful to motivate an estimator for the parameter vector β.

Using matrix notation, recall that the statistical error term is defined as $U = Y - \mathbf{X}'\beta$. To solve for β, just plug this expression in for U in the orthogonality conditions.

$$\mathbf{0} = E[\mathbf{X}U] \quad = \quad E\left[\mathbf{X}\left(Y - \mathbf{X}'\beta\right)\right] = E[\mathbf{X}Y] - E\left[\mathbf{X}\mathbf{X}'\right]\beta$$

where the simplifying steps above use the linearity of expectation. Reorganizing this expression, our algebra reduces to the **normal equations**.

$$E\left[\mathbf{X}\mathbf{X}'\right]\beta \quad = \quad E[\mathbf{X}Y]$$

As with simple linear regression, the form of the normal equations leads naturally to an identifying assumption.

DEFINITION 2.2.2. **Identification Assumption**. To solve for β, we need to assume that $E[\mathbf{X}\mathbf{X}']$ is invertible. Intuitively, this identification assumption means that the there is no redundant information in the set of variables. Namely, each variable X_j has variability unto itself that can be used to identify an independent relationship between X_j and Y.

Although it looks quite different, this identification assumption reduces to the assumption we made in simple linear regression for $k = 1$. As long as the identification assumption holds, we can solve for the parameter vector in terms of the features of the joint distribution.

$$\beta \quad = \quad E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E[\mathbf{X}Y]$$

It is not intuitive to motivate an assumption about the invertibility of $E[\mathbf{X}\mathbf{X}']$ directly. A supporting result leads to an easier-to-interpret identifying assumption: no multicollinearity.

DEFINITION 2.2.3. A vector of regressors $\mathbf{X}$ is **perfectly collinear** if there exists a non-zero vector of constants $c = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_k \end{pmatrix}$ such that $0 = c'\mathbf{X}$ with probability one.

If the regressors are perfectly collinear, this is known as **perfect multicollinearity**. Perfect multicollinearity manifests itself if the regressors have redundant information. Rearranging the previous definition, we can express X_j as a linear combination of all of the other X_i's. As the following theorem demonstrates, no perfect multicollinearity is a sufficient condition for identification.

THEOREM 2.2.4. *Suppose that $E[\mathbf{X}\mathbf{X}']$ is finite. As long as the vector of regressors $\mathbf{X} = (1, X_1, ..., X_k)$ is not perfectly collinear, the columns of $E[\mathbf{X}\mathbf{X}']$ are linearly independent. Hence, $E[\mathbf{X}\mathbf{X}']$ is invertible and the identification assumption is satisfied.*

In summary, we solved for the parameter vector in terms of potentially observable random variables for both simple and multiple linear regression. This solution of the regression parameters expressed as features of the joint distribution of $(\mathbf{X}, Y)$ is worth repeating:

One Regressor: In the single-regressor statistical model,

$$Y = \beta_0 + \beta_1 X + U$$

our choice of the coefficients as the best linear approximation to the conditional expectation $E[Y|X]$ implies

$$\beta_1 = \frac{Cov[X,Y]}{Var[X]}$$
$$\beta_0 = E[Y] - \left(\frac{Cov[X,Y]}{Var[X]}\right) E[X]$$

Multiple Regressors: In the multiple-regressor statistical model

$$Y = \mathbf{X}'\beta + U$$

where $\mathbf{X} = (1, X_1, X_2, ..., X_k)$, our choice of the coefficients as the best linear approximation to the conditional expectation $E[Y|X]$ implies

$$\beta = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E[\mathbf{X}Y]$$

In the case of one regressor ($k = 1$), the multiple regression formula simplifies to the single regression formula. The matrix algebra is efficient and useful for estimation, but it obscures the intuitive and relatively easy to manipulate expressions from single regression. Fortunately, as the following theorem demonstrates, the single regression intuition extends to multiple regression. After we discuss estimation, this theorem will help us interpret regression estimates in multiple regression.

THEOREM 2.2.5. *Population Frisch-Waugh Theorem. In the multiple regression model,*

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + U$$

we can express the regression coefficient simply $\beta_j = \frac{Cov[Y,\tilde{X}_j]}{Var[\tilde{X}_j]}$*, where* $\tilde{X}_j$ *is the error term from the statistical regression of* X_j *onto the other regressors*

$$X_j = \alpha_0 + \alpha_1 X_1 + ... + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + ... + \alpha_k X_k + \tilde{X}_j$$

Aside from giving a more intuitive interpretation of multiple regression coefficients, the Frisch-Waugh Theorem also makes clear that the residual variation of X_j – that is, the part of X_j that is uncorrelated with other predictors – is what identifies the parameter β_j in the data.

2.3. Regression Estimation

For the calculations in this section, let $(X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n)$ be a random sample of size n from (X, Y) where $Y = \beta_0 + \beta_1 X + U$ is the (statistical) population regression.

Also, let $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), ... (\mathbf{X}_n, Y_n)$ be a random sample of size n from $(\mathbf{X}, Y)$ where $\mathbf{X} = (1, X_1, ..., X_k)$ where $Y = \mathbf{X}'\beta + U$ is the (statistical) population regression. In the multiple-regressor setting, this notation can be a little confusing, so I will stick with the convention that X_{ij} is the i^{th} observation on the j^{th} regressor.[3]

Note that because these are statistical models, the orthogonality conditions are satisfied. Finally, assume that there is a positive, finite variance to each regressor and that there is no perfect multi-collinearity in the regression model.

[3]This convention has some intuition to it. If you were to look at your data set in an Excel file, each variable (potential regressor) would be a column and each row would be an observation. We'll try to stick with analogous notation.

2.3.1. Ordinary Least Squares. In this section, we derive the Ordinary Least Squares (OLS) regression estimator $\hat{\beta}$ for both single and multiple regression. The OLS estimator gets its name because it solves the sample least squares problem, but there are other ways to motivate using the estimator that are easier to apply. Nevertheless, solving the OLS problem is an important exercise. To ground our intuition on how to solve the problem, we will consider the single-regressor case.

2.3.1.1. *OLS with one regressor.* When there is one regressor, the ordinary least squares (OLS) estimator solves the least squares problem:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)^2$$

Let's derive the OLS estimators in this simple case. First, take the first order conditions of the least squares problem:

$$\begin{aligned} [b_0]: \ -2 \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right) &= 0 \\ [b_1]: \ -2 \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right) X_i &= 0 \end{aligned}$$

Solving the $[b_0]$ FOC, $b_0 = \bar{Y} - b_1 \bar{X}$. Plugging this equation into the $[b_1]$ FOC, we obtain:

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y} - b_1 \left(X_i - \bar{X}\right)\right) X_i = 0$$

Solving this equation for b_1, we obtain $\hat{\beta}_1 = \frac{S_{X,Y}}{S_X^2}$ together with $\hat{\beta}_0 = \bar{Y} - b_1 \bar{X}$.

EXERCISE 2.3.1. Fill in the details of this derivation of the OLS estimator. Also, compute $Cov\left[\hat{\beta}_0, \hat{\beta}_1 | \mathbf{Y}, \mathbf{X}\right]$. What is its sign? Can you compute $Cov\left[\hat{\beta}_0, \hat{\beta}_1\right]$?

Next, turn to multiple regression where we use matrix algebra to simplify the solution method.

2.3.1.2. *OLS with multiple regressors.* In multiple regression, the ordinary least squares (OLS) estimator solves the least squares problem:

$$\hat{\beta} = \underset{b}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i' b\right)^2$$

where $b = (b_0, b_1, ..., b_k)$ is an arbitrary candidate coefficient vector.

CLAIM 2.3.2. If we define $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ and $\mathbb{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix}$, the objective function can be reexpressed in matrix notation as

$$\begin{aligned} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i' b\right)^2 &= (\mathbf{Y} - \mathbb{X}b)' (\mathbf{Y} - \mathbb{X}b) \\ &= \mathbf{Y}'\mathbf{Y} - 2b'\mathbb{X}'\mathbf{Y} + b'\mathbb{X}'\mathbb{X}b \end{aligned}$$

Using a result from matrix calculus (see Hansen) the system of first order conditions is given by:

$$-2\mathbb{X}'\mathbf{Y} + 2\mathbb{X}'\mathbb{X}b = 0$$

Solving for b, we obtain the OLS estimator in the multiple-regressor setting

$$\hat{\beta}^{ols} = \left(\mathbb{X}'\mathbb{X}\right)^{-1} \left(\mathbb{X}'\mathbf{Y}\right)$$

This is an important formula when it comes to understanding the OLS estimators. We will spend a considerable amount of time with it.

REMARK 2.3.3. To put this estimation procedure differently, the OLS estimator minimizes the sum of squared residuals. You may think to take the absolute value of residuals and add them up instead. That's a perfectly valid procedure to produce an estimator (called median regression), but it will produce an estimator for β that is different than the OLS estimator.

As this remark suggests, there is a variety of methods that we can use to obtain estimators of regression coefficients. These other ways of deriving estimators can be useful in a more complicated analysis, but they can also teach us more about the OLS estimator for β. For this reason, we will consider several of these methods.

2.3.2. Analogy Principle. The analogy principle allows us to derive estimators for population parameters as long as we can express the parameters as estimable features of the joint distribution of our random sample. It is easiest to see the principle in action.

Single Regression: By analogy to our expressions for β_0 and β_1, we can plug in the sample version of covariance, variance and means into the slots they occupy in the population to obtain estimators for β_0 and β_1.

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{XY}}{S_X^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

Multiple Regression: Using the same analogy principle that we used in single regression, we can construct a sample version of $\beta = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{X}Y\right]$. Denote the i^{th} observation vector as

$\mathbf{X}_i = (1, X_{i1}, X_{i2}, ..., X_{ik})$. To use the analogy principle, replace expectations with sample averages across observations as in

$$\begin{aligned}\hat{\beta}^a &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i Y_i\right) \\ &= \left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i Y_i\right)\end{aligned}$$

Compared with the matrix formula for OLS, this first formula for $\hat{\beta}$ is easier to motivate if we are thinking about the parameter vector β as an estimable feature of the joint distribution of $(\mathbf{X}, Y)$.

Not only is it easier to motivate, but the two expressions are identical. Using the definitions of $\mathbb{X}$ and $\mathbf{Y}$ from the previous section,[4] it is straightforward to show that the analogy estimator $\hat{\beta}^a$ reduces to the OLS estimator $\hat{\beta}^{ols}$

$$\hat{\beta}^a = \hat{\beta}^{ols} = \left(\mathbb{X}'\mathbb{X}\right)^{-1}\left(\mathbb{X}'\mathbf{Y}\right)$$

The matrix formula has two distinct advantages.

(1) On a practical level, $\mathbb{X}$ is just the data matrix with a column of 1's appended. This means that the matrix formulas are attractive for implementation in a computational package like R.
(2) On a conceptual level, the matrix form of the OLS estimator allows us to think about regression as projection onto the column space of the $\mathbb{X}$ data matrix.

2.3.2.1. *A Detour into Projection Methods.* Consider a slight detour to introduce the theory of **regression as projection**. Use a matrix equation to construct the **fitted values** $\hat{\mathbf{Y}} = \mathbb{X}\hat{\beta}^{ols}$ and plug the matrix solution for $\hat{\beta}^{ols}$ into this expression:

$$\hat{\mathbf{Y}} = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\left(\mathbb{X}'\mathbf{Y}\right)$$

Exercise 2.3.4. Write out the matrix multiplication to show that the i^{th} element of $\hat{\mathbf{Y}}$ equals $\hat{Y}_i = \mathbf{X}_i'\hat{\beta}^{ols}$.

We call $P = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'$ the **projection matrix** onto the columns of $\mathbb{X}$. Projection matrices have several useful properties for the theory of regression.

Proposition 2.3.5. ***Some Properties of Projection Matrices****. Let P be a projection matrix defined by $P = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'$, where $\mathbb{X}$ is a $n \times (k+1)$ data matrix.*

(1) *P projects $\mathbb{X}$ to $\mathbb{X}$. That is, $P\mathbb{X} = \mathbb{X}$.*
(2) *P projects any linear combination of the columns of $\mathbb{X}$ to itself. That is, if $W = \mathbb{X}\mathbf{c}$ with $\mathbf{c}$ a conformable vector of constants, then $PW = W$.*
(3) *The fitted values of OLS regression are a projection onto the column space of $\mathbb{X}$. In other words, $\hat{\mathbf{Y}} = \mathbb{X}\hat{\beta} = P\mathbf{Y}$.*
(4) *P is idempotent. That is, $P = PP$.*
(5) *Applying the projection matrix again does not change the fitted values $\hat{Y} = P\hat{Y}$.*

[4]The notation is beginning to get a little cumbersome, but this is unavoidable to some degree. In this set of notes, we have needed to distinguish X (a random variable) from x (a realization of that random variable), from $\mathbf{X}$ (a vector of explanatory random variables) from $\vec{\mathbf{X}}$ (a matrix that contains a random sample from the vector of explanatory random variables).

Another useful matrix in regression computation is often called the **residual maker matrix**, $M = I - P$, where I is the $(k+1) \times (k+1)$ identity and P is the projection matrix onto the column space of $\mathbb{X}$. We call this matrix the residual maker because premuliplying by it produces a vector of OLS **residuals**:

$$\hat{\mathbf{U}} \quad = \quad (I - P)\,\mathbf{Y} = \mathbf{Y} - \mathbb{X}\hat{\beta}^{ols}$$

The residual maker matrix and the projection matrix are intricately related.[5]

PROPOSITION 2.3.6. ***Relationship Between the Projection Matrix and the Residual Maker.*** *Let P be a projection matrix defined by $P = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'$, where $\mathbb{X}$ is a $n \times (k+1)$ data matrix. Define $M = I - P$.*

1. *The projection matrix and the residual maker are orthogonal $PM = \mathbf{0}$.*
2. *The residual maker is idempotent $M = MM$.*
3. *The sum of the projection matrix and the residual maker is the identity matrix $P + M = I$.*

Applying this theorem, we obtain the following important result about the fitted values and residuals from OLS regression.

THEOREM 2.3.7. ***OLS Orthogonal Decomposition.*** *For any $n \times 1$ vector $\mathbf{Y}$, OLS regression (projection) of $\mathbf{Y}$ on the columns of $\mathbb{X}$ decomposes $\mathbf{Y}$ into two orthogonal components.*

$$\mathbf{Y} \quad = \quad P\mathbf{Y} + (I - P)\,\mathbf{Y} = \mathbb{X}\hat{\beta}^{ols} + \hat{\mathbf{U}}$$

We will return to projection methods whenever they allow us to apply a more elegant framework for a proof (or for understanding an important concept).

2.3.3. Method of Moments. From our work on the population regression, the normal equations give us a set of population **moment restrictions**:

$$E\left[\mathbf{X}\left(Y - \mathbf{X}'\beta\right)\right] \quad = \quad 0$$

The derivation of the method of moments estimator $\hat{\beta}^{mm}$ is to replace the population moments with sample moments in this expression, then solve for β. In the case of estimating linear regression coefficients, if we replace the population moments by sample moments, we obtain the first order conditions from the OLS problem, which implies $\hat{\beta}^{mm} = \hat{\beta}^{ols}$.

DEFINITION 2.3.8. **Sample Moment Conditions**. If we define $\hat{U}_i = Y_i - \mathbf{X}_i'\hat{\beta}$ to be the **residual** for the i^{th} observation. The sample moment conditions are given by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\hat{U}_i \quad = \quad \mathbf{0}$$

[5]The residual maker is also a projection matrix, but it projects vectors onto a different subspace of $\mathbb{R}^n$. The projection matrix projects onto the $k+1$-dimensional column space of $\mathbb{X}$ whereas the residual maker projects onto the null space of $\mathbb{X}'$. By the fundamental theorem of linear algebra, these two subspaces are orthogonal complements. That is, they do not overlap, they fill the space $\mathbb{R}^n$ and each vector in the column space of $\mathbb{X}$ is orthogonal to each vector in the null space of $\mathbb{X}'$. As a practical matter, this means that the OLS residuals are orthogonal to the OLS fitted values. If you wish to know more, take a rigorous linear models course.

This is a vector of $k+1$ equations: one for each element of $\mathbf{X} = \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_k \end{pmatrix}$. The first moment restriction is special in that it implies that the sum of the residuals is zero: $\frac{1}{n}\sum_{i=1}^{n}\hat{U}_i = 0$. For an arbitrary regressor j, the moment condition is $\frac{1}{n}\sum_{i=1}^{n}X_{ij}U_i = 0$.

It will often be useful to express the sample moment conditions in matrix form.

EXERCISE 2.3.9. With $\mathbb{X}$ defined as before and $\hat{\mathbf{U}} = \begin{pmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_n \end{pmatrix}$, show that we can write the sample moment conditions as $\mathbb{X}'\hat{\mathbf{U}} = \mathbf{0}$.

The OLS estimator solves these sample moment conditions.

EXERCISE 2.3.10. Use the residual maker matrix to show that the sample moment conditions hold using the OLS estimator.

2.3.4. Maximum Likelihood. There is an alternative motivation for the OLS estimator, but to get to this motivation, we should review some basic theory on maximum likelihood estimation. In Chapter 3, we take up the topic of maximum likelihood estimation in more detail.

2.3.4.1. *Detour into MLE Theory.* For clarity of exposition, assume that we are studying random sample from a continuous population random variable X. Because random samples are defined to be iid, we can think about the **joint density of the random sample** (before we observe it) as follows:

$$\begin{aligned} f_{\mathbf{X}}(x_1, x_2, ..., x_n) &= f_{X_1}(x_1) f_{X_2}(x_2) \times ... \times f_{X_n}(x_n) \text{ using independence} \\ &= f_X(x_1) f_X(x_2) \times ... \times f_X(x_n) \text{ using identical} \end{aligned}$$

More compactly, we write the joint pdf of the random sample as $f_{\mathbf{X}}(x_1, x_2, .., x_n) = f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i)$. If we want to think about how the joint distribution of the random sample depends on some parameter of the distribution of X, denoted θ, we use the notation:

$$f_{\mathbf{X}}(\mathbf{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)$$

EXAMPLE 2.3.11. For example, suppose that $X \sim N(\mu, \mu)$, a very special normal distribution. Then, the pdf of X is

$$f_X(x;\mu) = \frac{1}{\sqrt{2\pi\mu}} \exp\left(-\frac{1}{2\mu}(x-\mu)^2\right)$$

In this example, the joint pdf of a random sample of size n equals

$$f_{\mathbf{X}}(\mathbf{x};\mu) = \left(\frac{1}{2\pi\mu}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\mu}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Based on this example, we can make two immediate observations.

(1) It is natural to extend the notation to accomodate dependence on a parameter vector θ such as $\left(\mu, \sigma^2\right)$. For example, if $X \sim N\left(\mu, \sigma^2\right)$, then the joint density of the random sample would have been:

$$f_{\mathbf{X}}\left(\mathbf{x}; \mu, \sigma^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right)$$

(2) Second, for a parametric distribution like the normal distribution, it is plain to see that the joint distribution of the random sample can be thought of as a function of the parameter vector. This leads to our definition of likelihood.

DEFINITION 2.3.12. The **Likelihood** of a sample – which we denote as $L\left(\theta; \mathbf{x}\right)$ or $L\left(\theta; \mathbf{X}\right)$ depending on whether we condition on a random sample $\mathbf{X}$ or a realization from that random sample $\mathbf{x}$ – is defined to be the joint density of the sample viewed as a function of the parameters of the population distribution.

Given this definition of likelihood, a maximum likelihood estimator is natural. Pick $\hat{\theta}$ to be the parameter vector that maximizes $L\left(\theta; \mathbf{x}\right)$, given a realization $\mathbf{x}$ of our random sample.

DEFINITION 2.3.13. Formally, the **maximum likelihood estimator (MLE)** of θ given a realization of the random sample $\mathbf{x} = \left(x_1, x_2, ..., x_n\right)$ is defined to be

$$\hat{\theta} = \arg\max_{\tilde{\theta} \in \Theta} L\left(\tilde{\theta}; \mathbf{X}\right)$$

where Θ is the parameter space or the set of allowable parameter values.

Because it is a positive transformation of a positive function, taking the natural log of the likelihood does not change the MLE.[6] For many types of population distributions, this will be an attractive technique to simplify the process of finding the MLE. First, take the log of the likelihood, $l\left(\theta; \mathbf{X}\right) = \log L\left(\tilde{\theta}; \mathbf{X}\right)$, then start taking first order conditions.

EXAMPLE 2.3.14. Suppose the population is $X \sim N\left(\mu, \sigma^2\right)$, we have a random sample $\mathbf{X} = \left(X_1, ..., X_n\right)$ and we would like to use the random sample to construct an estimator for μ. In this setting, the log likelihood is

$$l\left(\mu, \sigma^2; \mathbf{X}\right) = -\frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - \frac{n}{2}\log\left(2\pi\right)$$

Take the first order condition with respect to μ: $\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right) = 0 \Rightarrow \sum_{i=1}^{n} X_i = n\hat{\mu} \Rightarrow \hat{\mu} = \bar{X}$. This is a nice result for motivating $\bar{X}$ as an estimator for μ.

EXERCISE 2.3.15. Write each of the steps in this example in matrix notation. That is, convert sums to inner products and define matrices and vectors appropriately so that the inner products you define are equivalent to the sums in this example.

[6] This is true for the same reason that a positive monotonic transformation of utility does not affect consumer demand functions.

2.3.4.2. *Back to Regression.* Consider a slightly more complicated setting for MLE: linear regression where $(\mathbf{X}, Y)$ is distributed as a multivariate normal random vector. To transform this model into a maximum likelihood estimation problem, we apply a trick to simplify the problem. In the regression model,

$$Y = \mathbf{X}'\beta + U$$

where $(\mathbf{X}, Y)$ is multivariate normal, the error term $U \equiv Y - \mathbf{X}'\beta$ is just a linear combination of normal random variables. Hence, U is normal. Given this observation, we specify the linear regression model in a way that allows us to work with U (one dimensional), rather than the many-dimensional $(\mathbf{X}, Y)$. Denote the variance of U given the regression model as σ_u^2 and exploit the fact that the error term is mean zero. Then, the pdf of U equals:

$$f\left(u; \sigma_u^2\right) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{1}{2\sigma_u^2} u^2\right)$$

The likelihood is defined similarly to the previous section

$$L\left(\sigma_u^2; \mathbf{U}\right) = \left(\frac{1}{2\pi\sigma_u^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_u^2} \sum_{i=1}^{n} U_i^2\right)$$

Notice that there is no explicit dependence on the regression parameter vector in this expression for the likelihood. That is because all of this dependence is wrapped up in U. To reintroduce this dependence, recognize that $U_i = Y_i - \mathbf{X}_i'\beta$. Hence, the likelihood becomes

$$L\left(\beta, \sigma_u^2; \mathbf{X}, Y\right) = \left(\frac{1}{2\pi\sigma_u^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_u^2} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i'\beta\right)^2\right)$$

Taking the log of this likelihood, we can see that the maximum likelihood estimator solves:

$$\left(\hat{\beta}_{MLE}, \hat{\sigma}_{u,MLS}^2\right) = \arg\max_{b, s^2} \left(-\frac{1}{2s^2} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i'b\right)^2 - \frac{n}{2} \log\left(2\pi s^2\right)\right)$$

where $b = (b_0, b_1, ..., b_k)$ is a candidate for the vector of coefficient estimates. The second term does not depend on b and one can show that in the system of first order conditions, s^2 cancels out of the first order conditions for b. Because $-\frac{1}{2s^2}$ is negative, we can drop this part and repose the subproblem of finding the MLE as the minimization problem:

$$\hat{\beta}_{MLE} = \arg\min_{b} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i'b\right)^2$$

This is precisely the least squares problem from which we obtained the OLS estimator for β. In other words, when the data come from a multivariate normal distribution, we have shown that $\hat{\beta}_{MLE}$ solves the ordinary least squares problem. Hence, $\hat{\beta}_{MLE} = \hat{\beta}_{OLS}$.

To obtain the equivalence between MLE and OLS, we needed to assume that we had taken a random sample from a multivariate normal population. MLE is a nice framework because we do not need to make this assumption (and we can consider others). The simplest MLE extension is to assume

some other parametric distribution than multivariate normal, which will produce different FOCs and a different solution for the coefficient estimates.

A more complicated example where MLE gives us more is to use MLE and some knowledge of the setting to relax the assumption of independence among the observations. For example, when we constructed the joint likelihood from the pdfs of each observation, we could have specified a form of dependence between observations. This technique can allow for estimation in a setting where we do not have a random sample, but we know something about the dependence across observations. More on this in Chapter 3.

For the rest of this chapter, we will stick with the OLS estimators because they provide the intuition for econometric analysis, but you should recognize that other methods are available. One reason to stick with OLS estimators is that they generally have nice properties, a topic we explore in the next section.

2.4. Properties of OLS Estimators

We will study three properties of OLS (unbiasedness, consistency and asymptotic normality) and the conditions necessary to guarantee them. More importantly, the hypotheses needed for the derivations in this section are imporant guidelines for thinking about how well OLS estimation does in practical application.

2.4.1. Unbiasedness of OLS Estimators. Before showing the proof in the multiple-regressor context, it is simplest to ground our intuition in single-variable regression.

2.4.1.1. *Unbiasedness of OLS in Single Regression.* Let $\{(X_i, Y_i)\}_{i=1}^n$ be a random sample from the population (X, Y), where $Y = \beta_0 + \beta_1 X + U$. Suppose that we have a statistical model, which implies the orthogonality conditions $E[XU] = 0$. The question of whether $\hat{\beta}_1$ is unbiased amounts to computing its expectation and determining whether it $E\left[\hat{\beta}_1\right] = \beta_1$.

Before taking expectations, a preliminary result is useful to know:

LEMMA 2.4.1. ***The Proof Form of*** $\hat{\beta}_1$. *An alternative equation for the OLS estimator* $\hat{\beta}_1$ *in single linear regression,* $Y = \beta_0 + \beta_1 X + U$ *is*

$$\hat{\beta}_1 \quad = \quad \beta_1 + \frac{\sum_{i=1}^n \left(X_i - \bar{X}_n\right) U_i}{\sum_{i=1}^n \left(X_i - \bar{X}_n\right) X_i}$$

where $U_i = Y_i - \beta_0 - \beta_1 X_i$. *To prove this, substitute* $Y_i = \beta_0 + \beta_1 X_i + U_i$ *into* $\hat{\beta}_1 = \frac{S_{X,Y}}{S_X^2}$ *and perform some simple algebraic manipulations.*

PROOF. In class. □

Returning to unbiasedness, it is clear from Lemma 2.4.1 that $\hat{\beta}_1$ is unbiased for β_1 when

$$E\left[\frac{\sum_{i=1}^n \left(X_i - \bar{X}_n\right) U_i}{\sum_{i=1}^n \left(X_i - \bar{X}_n\right) X_i}\right] \quad = \quad 0$$

This is a condition on our random sample, but we desire a condition on our population that implies $\hat{\beta}_1$ is unbiased for β_1. We can obtain this result is to assume that U is mean independent of X. In other words, $E[U|X] = 0$. Some texts call this the assumption of **zero conditional mean** for the error term (Wooldridge, 2003).

Under this assumption, we can apply the Law of Iterated Expectations as follows:

$$
\begin{aligned}
E\left[\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)U_i}{\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)X_i}\right] &= E_{\mathbf{X}}\left[E_{U|\mathbf{X}}\left[\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)U_i}{\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)X_i}|X_1,X_2,...,X_n\right]\right] \\
&= E_{\mathbf{X}}\left[\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)\underbrace{E_{U_i|\mathbf{X}}\left[U_i|X_1,X_2,...,X_n\right]}_{E_{U_i|X_i}[U_i|X_i] \text{ by iid}}}{\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)X_i}\right] \\
&= E_{\mathbf{X}}\left[\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)\underbrace{E_{U_i|X_i}\left[U_i|X_i\right]}_{=0}}{\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)X_i}\right]=0
\end{aligned}
$$

The proof does not work if we are only allowed to use the orthogonality conditions $E[XU]=0$. Hence, for $\hat{\beta}_1$ to be unbiased for β_1, it must be the case that our error term has zero conditional mean.

CLAIM 2.4.2. The OLS intercept estimate is unbiased for the true intercept under the same assumption. This is simple to verify given the result on the unbiasedness of the slope.

REMARK 2.4.3. Throughout this section, we have assumed that our regressors are stochastic. That is, the elements of $\mathbf{X}$ are random variables and our random sample draws from a joint distribution $(\mathbf{X},Y)$. In economic applications, this is a natural assumption and we will retain **stochastic regressors** throughout the course.

In a statistics course, you may have seen the alternative assumption of **fixed regressors**, where $\mathbf{X}$ is not stochastic. In this case, there is no problem passing the expectation through the expression above as in $E\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)U_i}{\sum_{i=1}^{n}(X_i-\bar{X}_n)X_i}\right]=\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)E[U_i]}{\sum_{i=1}^{n}(X_i-\bar{X}_n)X_i}=0$ because the expectation is over U and not X. Hence, *in the case of fixed regressors*, all we need for unbiasedness of $\hat{\beta}_1$ for β_1 is a mean zero error term.

For our purposes, fixed regressors is a bad assumption because it does not match with economic reality. Therefore, we require U to be mean independent of X for $\hat{\beta}_1$ to be unbiased for β_1.

Next, we turn to multiple regression where we show that the $\hat{\beta}$ is unbiased for β under a similar zero conditional mean assumption. That is, orthogonality conditions are still not enough.

2.4.1.2. *Unbiasedness of $\hat{\beta}$ in Multiple Linear Regression.* Let $\{(\mathbf{X}_i,Y_i)\}_{i=1}^{n}$ be a random sample from the population $(\mathbf{X},Y)$, where $Y=\mathbf{X}'\beta+U$. Suppose that we have a statistical model, which implies the orthogonality conditions $E[\mathbf{X}U]=0$. The question of whether $\hat{\beta}$ is unbiased for β amounts to computing its expectation of the vector and determining whether $E\left[\hat{\beta}\right]=\beta$.

Before taking expectations, a preliminary result is useful to know:

LEMMA 2.4.4. ***The Proof Form of*** $\hat{\beta}$. *An alternative equation for the OLS estimator $\hat{\beta}$ in multiple linear regression, $Y=\mathbf{X}'\beta+U$ is*

$$\hat{\beta} = \beta+\left(\mathbb{X}'\mathbb{X}\right)^{-1}\left(\mathbb{X}'\mathbf{U}\right)$$

where $\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}$ *with* $U_i = Y_i - \mathbf{X}_i'\beta$.

PROOF. Recall the matrix form for our OLS estimator:

$$\hat{\beta} = \left(\mathbb{X}'\mathbb{X}\right)^{-1}\left(\mathbb{X}'\mathbf{Y}\right)$$

and recognize that if we substitute $Y_i = \mathbf{X}_i'\beta + U_i$ into each element of $\mathbf{Y}$, we obtain:

$$\begin{aligned} \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} &= \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix}\beta + \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix} \\ &= \mathbb{X}\beta + \mathbf{U} \end{aligned}$$

Use this expression to simplify our matrix form of the coefficient estimate vector

$$\begin{aligned} \hat{\beta} &= \left(\mathbb{X}'\mathbb{X}\right)^{-1}\left(\mathbb{X}'\left(\mathbb{X}\beta + \mathbf{U}\right)\right) \\ &= \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{X}\beta + \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{U} \\ &= \beta + \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{U} \end{aligned}$$

□

Returning to unbiasedness, it is clear from Lemma 2.4.4 that $\hat{\beta}$ is unbiased for β when

$$E\left[\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{U}\right] = \mathbf{0}$$

For this to be true, we (again) need zero conditional mean of the error term. In multiple regression, the mean independence assumption is $E\left[U|\mathbf{X}\right] = 0$. That is, conditional on all of the regressors, the error term has mean zero.

Under this assumption, we can apply the Law of Iterated Expectations as follows:

$$\begin{aligned} E\left[\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{U}\right] &= E_{\mathbb{X}}\left[E_{U|\mathbb{X}}\left[\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{U}|\mathbb{X}\right]\right] \\ &= E_{\mathbb{X}}\left[\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'E_{U|\mathbb{X}}\left[\mathbf{U}|\mathbb{X}\right]\right] \end{aligned} \tag{2.4.1}$$

We are not quite there yet. Examine the inner expectation $E_{U|\mathbb{X}}\left[\mathbf{U}|\mathbb{X}\right]$ (which is a vector) and consider an arbitrary element j of this expectation

$$\begin{aligned} E_{U|\mathbb{X}}\left[U_j|\mathbb{X}\right] &= E_{U|\mathbf{X}}\left[U_j|\mathbf{X}_j'\right] \text{ by iid} \\ &= E_{U|\mathbf{X}}\left[U|\mathbf{X}'\right] \text{ by identical} \\ &= 0 \end{aligned}$$

This argument applies for each element of $E_{U|X}\left[\mathbf{U}|\mathbb{X}\right]$, hence $E_{U|X}\left[\mathbf{U}|\mathbb{X}\right] = \mathbf{0}$, the zero vector. Plugging this fact into (2.4.1), we have completed the proof that

$$E\left[\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{U}\right] = \mathbf{0}$$

Hence, $\hat{\beta}$ is unbiased for β under mean independence of the error term from the set of regressors.

FACT 2.4.5. *The proof does not work if we are only allowed to use the orthogonality conditions* $E[\mathbf{X}U] = \mathbf{0}$. *Hence, for* $\hat{\beta}$ *to be unbiased for* β, *it must be the case that our error term has zero conditional mean.*

2.4.2. Consistency of OLS Estimators. As we saw in the previous section, unbiasedness requires the strong assumption of mean independence of the error term from the set of regressors. This mean independence condition only holds when the conditional expectation of Y given the set of regressors $\mathbf{X}$ can be expressed as a linear combination of those regressors (that is, $E[Y|\mathbf{X}] = \mathbf{X}'\beta$). Otherwise, OLS estimators $\hat{\beta}$ are biased for β.

In this section, we appeal to asymptotic theory for a related criterion for our estimator, consistency. As before, we will ground our intuition in a proof from single regression and adapt the proof into the multiple regression framework.

2.4.2.1. *Consistency of OLS in Single Regression.* Let $\{(X_i, Y_i)\}_{i=1}^n$ be a random sample from the population (X, Y), where $Y = \beta_0 + \beta_1 X + U$. Suppose that we have a statistical model, which implies the orthogonality conditions ($E[XU] = 0$ and $E[U]$) and assume the **no fat tails conditions** ($E\left[X^4\right], E\left[Y^4\right] < \infty$). The question of whether $\hat{\beta}_1$ is consistent for β_1 amounts to evaluating its probability limit. That is, to what value of c (if any), does $\hat{\beta}_1$ converge? No suspense here: $\hat{\beta}_1 \xrightarrow{P} \beta_1$.

REMARK 2.4.6. At this point, there is no need to use δ, ϵ methodology to evaluate probability limits. In fact, such an approach would be foolish given the power of our prior results. We will obtain this result by applying the weak law of large numbers (WLLN: $\bar{X} \xrightarrow{P} \mu$) and the continuous mapping theorem (CMT), and using some algebra.

In fact, we already proved the result in full detail in Exercise 1.6.3 of Chapter 1. Here are the steps of the proof:

(1) Show that $S_X^2 \xrightarrow{P} Var[X]$ and $S_{X,Y} \xrightarrow{P} Cov[X, Y]$ as long as $Var\left[X^2\right]$ and $Var\left[Y^2\right]$ are finite.
(2) Recognize that $\hat{\beta}_1 = \frac{S_{X,Y}}{S_X^2}$, which is a continuous function of $S_{X,Y}$ and S_X^2 except for when the denominator is zero.
(3) Apply the CMT to argue that $\hat{\beta}_1 \xrightarrow{P} \frac{Cov[X,Y]}{Var[X]} = \beta_1$ as long as $Var[X] \neq 0$, completing the proof.

CLAIM 2.4.7. Using similar methodology (and the result $\hat{\beta}_1 \xrightarrow{P} \beta_1$), it is straightforward to show that $\hat{\beta}_0 \xrightarrow{P} \beta_0$.

2.4.2.2. *Consistency of* $\hat{\beta}$ *for* β *in Multiple Regression.* Recall the analogy principle estimator form of the OLS estimator in multiple regression:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i Y_i\right)$$

Let's examine the inside of the $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'$ matrix to gather some intution about the consistency result. Remember that the set of regressors in multiple regression is $\mathbf{X}_i' = (1, X_{i1}, X_{i2}, ..., X_{ik})$. Hence, each term in the sum $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'$ is an outer product of the following form.

$$\mathbf{X}_i\mathbf{X}_i' = \begin{pmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix} (1, X_{i1}, X_{i2}, ..., X_{ik}) \quad = \quad \begin{pmatrix} 1 & X_{i1} & X_{i2} & ... & X_{ik} \\ X_{i1} & X_{i1}^2 & X_{i1}X_{i2} & ... & X_{i1}X_{ik} \\ X_{i2} & X_{i1}X_{i2} & & & X_{i2}X_{ik} \\ \vdots & \vdots & & \ddots & \\ X_{ik} & X_{ik}X_{i1} & & ... & X_{ik}^2 \end{pmatrix}$$

Each element of the matrix is just a random variable. Given this expression, the term $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'$ equals:

$$\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i' \quad = \quad \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n 1 & \frac{1}{n}\sum_{i=1}^n X_{i1} & \frac{1}{n}\sum_{i=1}^n X_{i2} & ... & \frac{1}{n}\sum_{i=1}^n X_{ik} \\ \frac{1}{n}\sum_{i=1}^n X_{i1} & \frac{1}{n}\sum_{i=1}^n X_{i1}^2 & \frac{1}{n}\sum_{i=1}^n X_{i1}X_{i2} & ... & \frac{1}{n}\sum_{i=1}^n X_{i1}X_{ik} \\ \frac{1}{n}\sum_{i=1}^n X_{i2} & \frac{1}{n}\sum_{i=1}^n X_{i1}X_{i2} & & & \frac{1}{n}\sum_{i=1}^n X_{i2}X_{ik} \\ \vdots & \vdots & & \ddots & \\ \frac{1}{n}\sum_{i=1}^n X_{ik} & \frac{1}{n}\sum_{i=1}^n X_{ik}X_{i1} & & ... & \frac{1}{n}\sum_{i=1}^n X_{ik}^2 \end{pmatrix}$$

By the WLLN, each element of this matrix converges in probability to its population counterpart (as long as the variance of X_iX_j is finite; invoke no fat tails), which we can write as follows:

$$\begin{aligned} \xrightarrow{P} \begin{pmatrix} 1 & E[X_1] & E[X_2] & ... & E[X_k] \\ E[X_1] & E\left[X_1^2\right] & E[X_1X_2] & ... & E[X_1X_k] \\ E[X_2] & E[X_1X_2] & & & E[X_2X_k] \\ \vdots & \vdots & & \ddots & \\ E[X_k] & E[X_kX_1] & & ... & E\left[X_k^2\right] \end{pmatrix} &= E\left[\begin{pmatrix} 1 & X_{i1} & X_{i2} & ... & X_{ik} \\ X_{i1} & X_{i1}^2 & X_{i1}X_{i2} & ... & X_{i1}X_{ik} \\ X_{i2} & X_{i1}X_{i2} & & & X_{i2}X_{ik} \\ \vdots & \vdots & & \ddots & \\ X_{ik} & X_{ik}X_{i1} & & ... & X_{ik}^2 \end{pmatrix}\right] \\ &= E\left[\mathbf{X}\mathbf{X}'\right] \end{aligned}$$

By a similar argument $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_iY_i \xrightarrow{P} E[\mathbf{X}Y]$ (although this is a vector so fewer terms need to be written out).

Note that inversion and matrix multiplication are well-defined continuous operations. As long as the inverse of $E[\mathbf{X}\mathbf{X}']$ exists (invoke no collinearity in $\mathbf{X}$), we can apply the continuous mapping theorem to argue that

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_iY_i\right) \quad \xrightarrow{P} \quad \left(E\left[\mathbf{X}\mathbf{X}'\right]\right)^{-1}E[\mathbf{X}Y] = \beta$$

under the relatively weak conditions we needed to assume throughout the proof.

2.4.2.3. *A Preview on the Relationship to the Causal Model.* This is to say that consistency of $\hat{\beta}$ for β is easy to show relative to showing unbiasedness. As a prelude to the causal interpretation of regression, suppose that $Cov[X, V] \neq 0$, where V is the error term. If this is the case, what goes wrong?

$$\begin{aligned}\hat{\beta}_1 &\xrightarrow{P} \frac{Cov[X,Y]}{Var[X]} = \frac{Cov\left[X, \beta_0 + \beta_1^{causal} X + V\right]}{Var[X]} \\ &= \frac{Cov\left[X, \beta_1^{causal} X\right] + Cov[X,V]}{Var[X]} \\ &= \beta_1^{causal} \frac{Cov[X,X]}{Var[X]} + \frac{Cov[X,V]}{Var[X]} \\ &= \beta_1^{causal} + \frac{Cov[X,V]}{Var[X]}\end{aligned}$$

Hence, even though $\hat{\beta}_1 \xrightarrow{P} \beta_1^{reduced.form}$, our estimator $\hat{\beta}_1$ might not be consistent for the β_1 we care about if $Cov[X,V] \neq 0$. At this point, you may want to revisit a question from earlier in this chapter. In the statistical interpretation of linear regression, what guarantees $Cov[X,U] = 0$?

We get an analogous result in multiple regression. Under the identification assumption, we know:

$$\hat{\beta} \xrightarrow{P} \left(E\left[\mathbf{X}\mathbf{X}'\right]\right)^{-1} E[\mathbf{X}Y]$$

but if the causal error term is correlated with the set of regressors ($E[\mathbf{X}V] \neq \mathbf{0}$), then

$$\begin{aligned}\left(E\left[\mathbf{X}\mathbf{X}'\right]\right)^{-1} E[\mathbf{X}Y] &= \left(E\left[\mathbf{X}\mathbf{X}'\right]\right)^{-1} E\left[\mathbf{X}\left(\mathbf{X}'\beta + V\right)\right] \\ &= \beta + \left(E\left[\mathbf{X}\mathbf{X}'\right]\right)^{-1} E[\mathbf{X}V]\end{aligned}$$

We will explore the problem of causal inference in more detail. This expression will feature prominently in our future discussions on this topic.

2.4.3. Asymptotic Normality. Start with single regression in a statistical interpretation of regression.

2.4.3.1. *Normality OLS Estimators in Simple Linear Regression.* In this setting, suppose we have a random sample $\{(X_i, Y_i)\}_{i=1}^n$ from (X, Y) with

$$Y = \beta_0 + \beta_1 X + U$$

Along the way, we will invoke the assumptions. no fat tails ($E\left[X^4\right] < \infty,\ E\left[Y^4\right] < \infty$) and positive finite variance ($0 < Var[X] < \infty$) as needed. The result on the asymptotic normality of the OLS estimators is summarized in the following theorem.

THEOREM 2.4.8. ***Asymptotic Normality of OLS Estimators*** *(Single-Regressor Case). Let* $\{(X_i, Y_i)\}_{i=1}^n$ *be a random sample from a population* (X, Y) *satisfying the no fat tails and positive, finite variance assumptions. Then, the test statistic* Z_n *has an asymptotically normal distribution. Formally,*

$$Z_n = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2_{\hat{\beta}_i}}} \quad \xrightarrow{d} \quad N(0,1)$$

where

$$\begin{aligned} \sigma^2_{\hat{\beta}_1} &= \frac{1}{n}\frac{Var\left[(X - E[X])U\right]}{\left(Var[X]\right)^2} \\ \sigma^2_{\hat{\beta}_0} &= \frac{1}{n}\frac{Var[HU]}{\left(E\left[H^2\right]\right)^2} \end{aligned}$$

with $H = 1 - \frac{E[X]}{E[X^2]}X$.

PROOF. *(Sketch).* Recall the proof form of the OLS estimator (Lemma 4.2), which implies

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right) U_i}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}$$

The term in the numerator is $\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right) U_i$, which is just a sample mean of $W_i = \left(X_i - \bar{X}_n\right) U_i$. This is approximately $W_i \approx (X_i - \mu_X) U_i$, which has iid terms. If we divide both sides by the population analog of the standard deviation of this term's mean, $\sqrt{\frac{1}{n}Var\left[(X - \mu_X)U\right]}$, our expression becomes:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n}Var\left[(X - \mu_X)U\right]}} \approx \left[\frac{\frac{1}{n}\sum_{i=1}^n (X_i - \mu_X) U_i}{\sqrt{\frac{1}{n}Var\left[(X - \mu_X)U\right]}}\right] \frac{1}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}$$

Note: The term in brackets $\xrightarrow{d} N(0,1)$. The second term (by previous arguments) converges in probability to $Var[X]$. By Slutsky, the term on the RHS, converges in distribution to $N(0,1) \times \frac{1}{Var[X]}$. If the approximation error represented by $\approx$ converges in probability to zero, the LHS converges in distribution as:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n}Var\left[(X - \mu_X)U\right]}} \quad \xrightarrow{d} \quad N\left(0, \frac{1}{\left(Var[X]\right)^2}\right)$$

Alternatively, we could factor out $Var[X]$ and group it with the standard deviation terms on the LHS:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n}\left(Var[(X-\mu_X)U] / (Var[X])^2\right)}} \quad \xrightarrow{d} \quad N(0,1)$$

We denote the term under the square root $\sigma^2_{\hat{\beta}_1}$, which is the asymptotic (approximate) variance of $\hat{\beta}_1$:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2_{\hat{\beta}_1}}} \quad \xrightarrow{d} \quad N(0,1)$$

□

It takes some work to show the analogous result for $\hat{\beta}_0$. As we have seen with the other properties of regression, single regression is nested nicely as a special case of multiple regression. Rather than spend time on a tedious special case, we will appeal to our derivation of asymptotic normality in multiple regression to obtain the result most generally.

In practice, we will use this approximation for "large" finite samples as an approximation to the sampling distribution of $\hat{\beta}_1$:

$$\hat{\beta}_1 \approx N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$$

Claim 2.4.9. In practice, we do not know $\sigma^2_{\hat{\beta}_1}$ because it is a function of unknown parameters of the distribution. Plugging in a consistent estimator $\hat{\sigma}^2_{\hat{\beta}_1}$ for $\sigma^2_{\hat{\beta}_1}$, the convergence in distribution argument we outlined above also holds for:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}} \xrightarrow{d} N(0,1)$$

Moreover, one such consistent estimator for $\sigma^2_{\hat{\beta}_1}$ is

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \frac{\frac{1}{n}\sum_{i=1}^n \left(\left(X_i - \bar{X}_n\right)\hat{U}_i\right)^2}{\left(\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2\right)^2}$$

where $\hat{U}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ is the i^{th} residual.

This is a clunky formula and it turns out that standard output in software packages doesn't produce standard errors based on this formula. Rather, it uses a simplified version of the formula based on a homoskedasticity assumption:

Definition 2.4.10. Homoskedasticity (Constant Variance of the Error Term). The single-regressor regression model satisfies the **homoskedasticity assumption** if

$$Var[U|X] = \sigma^2 = E\left[U^2|X\right]$$

In practice, we can assess the homoskedasticity assumption on our data set by plotting residuals $\hat{U}_i$ versus fitted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. If the resulting plot clearly indicates a pattern in the amount of variability, the model can be said to be heteroskedastic (and the following simplification to the estimator of $\sigma^2_{\hat{\beta}_1}$ is invalid).

Exercise 2.4.11. Under the homoskedasticity assumption, use the law of iterated expectations to show that our expression for $\sigma^2_{\hat{\beta}_1}$ reduces to

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{n}\left(\frac{1}{Var[X]}\right)$$

Plugging in a consistent estimator for $Var[X]$, we get a simpler formula than before for our estimator of the variance of $\hat{\beta}_1$. This is the formula that computer software uses:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\sum_{i=1}^n \hat{U}_i^2}{\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}$$

Why does computer software use this formula by default? It turns out that under the assumption of homoskedasticity (and some other assumptions we have made), the estimator $\hat{\beta}_1$ is the best linear unbiased estimator for β_1. By default, computer software assumes we are in this case. In the multivariate setting, we'll prove this theorem in all of its generality.

Why not always use the more complicated formula? Because the consistent estimators of the robust standard errors make approximation error, we will want to have reasonable suspicion that the data are heteroskedastic before proceeding to use the robust formula for the standard errors.

2.4.3.2. *Asymptotic Normality of $\hat{\beta}$ in Multiple Regression.* Now, let's translate the intuition of the single-regressor proof to the multiple-regressor setting. suppose we have a random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from $(\mathbf{X}, Y)$ with

$$Y = \mathbf{X}'\beta + U$$

Along the way, we will invoke the assumptions. no fat tails ($E\left[X^4\right] < \infty$, $E\left[Y^4\right] < \infty$) and no collinearity in $\mathbf{X}$ as needed. The result on the asymptotic normality of the OLS estimators is summarized in the following theorem.

THEOREM 2.4.12. ***Asymptotic Normality of OLS Estimators*** *(Multiple-Regressor Case). Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be a random sample from a population $(\mathbf{X}, Y)$ satisfying the no fat tails and no multicollinearity. Then,*

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where $\Sigma = E[\mathbf{X}\mathbf{X}']^{-1} Var[XU] \left(E[\mathbf{X}\mathbf{X}']^{-1}\right)'$

PROOF. *(Sketch).* Recall the proof form of the OLS estimator:

$$\hat{\beta} = \beta + \left(\mathbb{X}'\mathbb{X}\right)^{-1}\left(\mathbb{X}'\mathbf{U}\right)$$

Reorganize this expression by subtracting β and multiplying both sides by $\sqrt{n}$:

$$\sqrt{n}\left(\hat{\beta} - \beta\right) = \left(\mathbb{X}'\mathbb{X}\right)^{-1}\sqrt{n}\left(\mathbb{X}'\mathbf{U}\right)$$

In this expression, we can convert the inner products from the matrix multiplication back to summations as in:

$$\begin{aligned}\sqrt{n}\left(\hat{\beta} - \beta\right) &= \left(\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'\right)^{-1}\sqrt{n}\left(\sum_{i=1}^n \mathbf{X}_i U_i\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'\right)^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i U_i\right)\end{aligned}$$

We can apply the weak law of large numbers to argue that $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i' \xrightarrow{P} E[\mathbf{X}\mathbf{X}']$.

Now, consider the term $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i U_i\right)$, which is a vector of sample of vector means premultiplied by $\sqrt{n}$. In this term, the part in the brackets converges in probability to $E[\mathbf{X}U]$, which equals zero by the orthogonality conditions. By the multivariate version of the Central Limit Theorem, we have:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i U_i\right) \xrightarrow{d} N(\mathbf{0}, Var[\mathbf{X}U])$$

Putting these two arguments together using Slutsky's Theorem, we obtain the result of the theorem

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \quad \xrightarrow{d} \quad N\left(\mathbf{0}, \underbrace{E\left[\mathbf{X}\mathbf{X}'\right]^{-1} Var\left[\mathbf{X}U\right]\left(E\left[\mathbf{X}\mathbf{X}'\right]^{-1}\right)'}_{=\Sigma}\right)$$

□

Note: $E\left[\mathbf{X}\mathbf{X}'\right]$ is symmetric so we can drop the transpose on the second $\left(E\left[\mathbf{X}\mathbf{X}'\right]\right)^{-1}$ term.

In practice, we use a consistent estimator $\hat{\Sigma}$ of Σ to construct the variance-covariance matrix of the coefficient estimates $\hat{\beta}$. In this notation, the estimated variance-covariance matrix equals $\frac{\hat{\Sigma}}{n}$. We can – by analogy – construct a consistent estimator for Σ using the sample analogs of the population parameters in Σ.

Just as with single regression, invoking a homoskedasticity assumption simplifies the expression for Σ considerably:

Definition 2.4.13. The error term in the multiple regression model is **homoskedastic** if $Var\left[U|\mathbf{X}\right] = Var\left[U\right]$.

Applying homoskedasticity, we can compute

$$\begin{aligned} Var\left[\mathbf{X}U\right] &= E\left[\left(\mathbf{X}U - E\left[\mathbf{X}U\right]\right)\left(\mathbf{X}U - E\left[\mathbf{X}U\right]\right)'\right] \\ &= E\left[\left(\mathbf{X}U\right)\left(\mathbf{X}U\right)'\right] \\ &= E\left[U^2\mathbf{X}\mathbf{X}'\right] \end{aligned}$$

which simplifies when we apply an iterated expectations argument

$$\begin{aligned} Var\left[\mathbf{X}U\right] &= E\left[\underbrace{E\left[U^2|\mathbf{X}\right]}_{=Var[U]}\mathbf{X}\mathbf{X}'\right] \\ &= Var\left[U\right]E\left[\mathbf{X}\mathbf{X}'\right] \end{aligned}$$

Plugging this expression into our equation for Σ, we cancel one of the $E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$ terms to obtain the much simpler expression:

$$\Sigma \quad = \quad Var\left[U\right]E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$$

In practice, computer software obtains the variance covariance matrix of coefficient estimates given by:

$$\begin{aligned} \frac{\hat{\Sigma}}{n} &= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i'\hat{\beta}\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1} \\ &= \left(\frac{\hat{\mathbf{U}}'\hat{\mathbf{U}}}{n}\right)\left(\mathbb{X}'\mathbb{X}\right)^{-1} \end{aligned}$$

which is a straightforward formula to apply to obtain the variance-covariance matrix. Keep in mind that this formula is only correct if we have homoskedasticity. If there is heteroskedasticity, we need to use a robust method.

Why do we make the assumption of homoskedasticity? It turns out that under this assumption and mean independence, the OLS estimator is the **best linear unbiased estimator** (BLUE) for β. This is a result known as the Gauss-Markov Theorem.

THEOREM 2.4.14. ***Gauss-Markov Theorem.*** *As long as the homoskedasticity assumption* $Var[U|\mathbf{X}] = Var[U] = \sigma^2$ *(in matrix form, this is* $Var[\mathbf{U}] = \sigma^2 I$*) is true, the OLS estimator* $\hat{\beta}^{ols}$ *is the best linear unbiased estimator for* β*. That is, let* $\tilde{\beta} = A\mathbf{Y}$ *be an alternative estimator for* β *such that* $E\left[\tilde{\beta}|\mathbb{X}\right] = \beta$*. Then,* $c'\hat{\beta}^{ols}$ *has a lower variance than* $c'\tilde{\beta}$*. More formally, if we define* $\Sigma^{ols} = Var\left[\hat{\beta}^{ols}|\mathbb{X}\right]$ *and* $\Sigma^a = Var\left[\tilde{\beta}|\mathbb{X}\right]$*,*

$$c'\Sigma^{ols}c \quad \leq \quad c'\Sigma^a c$$

PROOF. *Outline.* First, establish a lemma that the unbiasedness condition on $\tilde{\beta} = A\mathbf{Y}$ implies that $A\mathbb{X} = I$. Next, recognize that $Var\left[\hat{\beta}^{ols} - \tilde{\beta}|\mathbb{X}\right]$ is positive semi-definite because it is a variance covariance matrix and compute it to show:

$$\Sigma^{diff} = Var\left[\tilde{\beta} - \hat{\beta}|\mathbb{X}\right] \quad = \quad Var\left[\tilde{\beta}|\mathbb{X}\right] - Var\left[\hat{\beta}|\mathbb{X}\right] = \Sigma^a - \Sigma^{ols}$$

Apply the same quadratic form using the vector of constants c to both sides of the equation $\Sigma^{diff} = \Sigma^a - \Sigma^{ols}$ to obtain:

$$0 \leq c'\Sigma^{diff}c \quad = \quad c'\Sigma^a c - c'\Sigma^{ols}c$$

which implies $c'\Sigma^{ols}c \leq c'\Sigma^a c$, the result we sought to show. □

2.5. Causal Interpretation of Regression

For simplicity, start with the case of one regressor.

DEFINITION 2.5.1. **The Causal Population**. Let (Y, X, V) be a random vector like the one we studied under the statistical interpretation, except that we will interpret V as a causal error term. In reality, X and V *cause (or determine)* Y through some function g.

$$Y \quad = \quad g(X, V)$$

X is an **observed determinant** of Y, V is the sum of the effects of the **unobserved determinants** of Y. We think of $g(\cdot, \cdot)$ as characterizing the true relationship of Y and its underlying determinants.

The causal function $g(\cdot, \cdot)$ may come from an economic model or it may just represent our ideas about what causes Y. From this causal model, we can generate predictions regarding how Y changes in response to changes in X by taking the partial derivative of g with respect to X.

$$\frac{\partial Y}{\partial X} \quad = \quad \frac{\partial g(X, V)}{\partial X}$$

The partial derivative holds the value of V constant.[7] That is, holding all of the unobserved determinants of Y (captured in V) constant, $\frac{\partial Y}{\partial X}$ is **the effect** of X on Y. In economics classes, we call such partial derivatives **comparative statics**. The econometrics of causal models seeks to estimate and test the comparative statics from economic theory.

DEFINITION 2.5.2. **The Causal Regression Model (Linear Causal Relationship)**. Suppose that the true causal relationship is linear. In this case, the effect of X on Y does not depend on the value of X. This fact allows us to express the causal regression model as:

$$Y = \beta_0 + \beta_1 X + V$$

where $\beta_1 = \frac{\partial g(X,V)}{\partial X}$ is the effect of X on Y and $\beta_0 + V$ the sum of the effects of unobserved determinants, where $E[V] = 0$.

Let's consider an example of how theory can motivate our choice of causal model.

EXAMPLE 2.5.3. **An Example of a Causal Model**. Take an example from consumer theory. Let a consumer have preferences described by a quasilinear utility

$$U(X, Y) = Y + \alpha \log X$$

where P_x is the price of good X, P_y is the price of Y. Denote the consumer's income as M. We know from consumer theory that the Marshallian demand curve for good X is given by $MRS = \frac{\alpha}{X} = \frac{P_x}{P_y}$

$$X^m(P_x, P_y, M) = \frac{\alpha}{P_x/P_y}$$

Plugging in, we can obtain the Marshallian demand curve for good Y

$$Y^m(P_x, P_y, M) = \frac{M}{P_y} - \alpha$$

In this model of demand, these causal determinants of of quantity demanded are prices and income and the derived demand curves yield a causal function that we can study.

As this example illustrates, it is not difficult to imagine a causal model that is not linear. This is not surprising. The world is not linear! In the same way as we used the best linear approximation to the nonlinear statistical $E[Y|X]$, we can use a similar approximation to nonlinear causal models.

To make this comparison more formal, consider the ***augmented statistical* model**:[8]

$$Y = \beta_0 + \beta_1 X + \underbrace{\alpha W + U}_{V}$$

where U is a statistical error term from the best linear predictor problem of Y given X and W, X is an observed predictor of Y, W is a random variable that contains information on how all unobserved predictors of Y are potentially correlated with X as in:

[7] I have adopted the convention that V is a causal error term while U is a statistical error term. I do this because, in practice, it is important to interpretations of the error term.

[8] Technically, this expression cannot be a statistical model because W is unobserved. This is why I refer to the hypothetical regression equation as an augmented statistical model. This description draws on Wooldridge's graduate text and its treatment of regression. Angrist and Pischke introduce the potential outcomes notation (which is an excellent framework) before integrating it into one similar to what is presented in these notes.

$$W = \gamma_1 W_1 + \gamma_2 W_2 + ... + \gamma_M W_M$$

In this expression, $V = \alpha W + U$ is the causal error term that represents the sum of the effects of unobserved factors on Y. The logic of causation recognizes that the regression model contains fewer regressors than exist conceptually, and because of this simplification, we cannot guarantee that $Cov[X, V] = 0$ (which was required to identify the slope coefficient β_1). Our model of causation implies that we pick $(\beta_0, \beta_1, \alpha)$ to minimize the mean squared error distance from $\hat{g}(X, W, U) = g(X, V)$ where $V = \alpha W + U$.

DEFINITION 2.5.4. **Linear Approximation to the Causal Population**. Suppose that the true causal relationship is given by $g(X, V)$, a potentially nonlinear function of the observed and unobserved determinants. It is always possible to express a linear approximation to the causal population relationship:

$$Y = \beta_0 + \beta_1 X + V$$

where β_1 is the slope on the best linear approximation to the true causal relationship between Y and X and β_0 is such that the causal error term V has mean zero, $E[V] = 0$.
This definition extends naturally to a multiple-regressor framework.
Multiple Regressors. Suppose that the true causal relationship is given by $g(\mathbf{X}, V)$, a potentially nonlinear function of the observed and unobserved determinants. It is always possible to express a linear approximation to the causal population relationship:

$$Y = \mathbf{X}'\beta + V$$

where β is the coefficient vector that defines the best linear approximation to the true causal relationship between Y and $\mathbf{X}$. By construction of this approximation, $E[V] = 0$.

FACT 2.5.5. ***Regressors may be Correlated with the Error Term in a causal interpretation.*** *Suppose that we have a well-specified causal model that is actually linear.*

$$Y = \mathbf{X}'\beta + V$$

This equation is the same as in the statistical linear regression model, but the interpretation of the error term and the vector of coefficients is is *vastly* different. In this model, V is the sum of the ununobserved determinants of Y. In contrast to the purely statistical interpretation (where U is merely a statistical discrepancy), V has a "life of its own." We cannot force a causal error term to be uncorrelated with the regressors because these unobserved determinants could be correlated with the observed determinants. For example, Y could be wages, X could be eductational attainment, V could cointain IQ, among other predictors of wages. IQ and educational attainment are clearly correlated with one another. As a practical matter, this implies $Cov[X, V] \neq 0$.

FACT 2.5.6. ***The Causal β is Different from the Statistical β.*** *In a simple linear regression $(Y = \beta_0 + \beta_1 X + V)$, suppose that $Cov[X, V] \neq 0$, where V is the causal error term. In the previous set of notes, we argued that:*

$$\hat{\beta}_1 \xrightarrow{P} \frac{Cov[X, Y]}{Var[X]} = \beta_1^{stat}$$

By plugging in for Y using the causal regression and applying linearity of covariances, we obtain the relationship between the causal effect and the regression coefficient in a statistical model:

$$\beta_1^{stat} = \beta_1^{causal} + \frac{Cov[X,V]}{Var[X]}$$

Hence, even though $\hat{\beta}_1 \xrightarrow{P} \beta_1^{stat}$, our estimator $\hat{\beta}_1$ is no longer consistent for the β_1 we care about. As we discussed in the previous set of notes, multiple regressors have a similar formula:

$$\beta^{stat} = \beta^{causal} + \left(E\left[\mathbf{X}\mathbf{X}'\right]\right)^{-1} E[\mathbf{X}V]$$

2.5.1. Omitted Variable Bias. In both of these expressions, the orthogonality conditions applied to the causal error term are essential to identify the causal effect using an OLS regression model. To put it another way, even if we had infinite data on $(\mathbf{X}, Y)$, a correlation of V with $\mathbf{X}$ makes it impossible to recover the vector of causal effects. We will have two approaches to circumvent the problem of this correlation with the causal error term.

(1) **Include Omitted Variables**. The observed regressors are correlated with the causal error term because omitted variables (presently unobserved variables) are related to the observed regressors. Given this motivation, we may be able to solve the problem by including additional regressors until we can argue that $E[\mathbf{X}V] = 0$. Once we have included all of the variables in $\mathbf{W}$, we can make this argument.[9]
(2) **Use variation in X that is uncorrelated with V instead of using X itself**. This is the motivation behind instrumental variables regression, which we will cover in detail in another set of notes. As a preview for the instrumental variables logic, we will look for a set of variables $\mathbf{Z}$ that is correlated with $\mathbf{X}$ such that $E[\mathbf{Z}V] = 0$. We will use this set of orthogonality conditions to identify β. More on this in Chapter 4.

For now, we have our hands full in understanding multiple regression, which serves as our baseline for understanding how $\mathbf{X}$ determines Y. In addition to helping us identify causal effects, the omitted variables logic is especially useful in understanding our multiple regression estimates and diagnosing potential problems with them. For these reasons, we now study omitted variables in detail.

REMARK 2.5.7. **Omitted Variable Bias**. Suppose that the true causal regression is given by the **long regression**:

$$Y = \beta_0^L + \beta_1^L X_1 + \beta_2^L X_2 + U_L$$

whereas we estimate the **short regression**:

$$Y = \beta_0^S + \beta_1^S X_1 + U_S$$

We know that we are wrong in omitting X_2, but we would like to investigate the extent to which our estimates are biased by this omission. Under the identification assumption and the statistical interpretation of the short regression, we know that

$$\beta_1^S = \frac{Cov[X_1, Y]}{Var[X_1]}$$

[9]This sounds simple, but in practice, there are two problems. First, the dimensionality of m can be quite large. Second, some of the variables in $\mathbf{W}$ are impossible to observe (or very costly to observe). For this reason, it may be impractical to apply this algorithm to recover causal effects. That said, decomposing the omitted variable bias formula is useful to understand the nature of the bias from omitting explanatory variables from the specification.

Plugging in the expression for Y from long regression and using linearity of covariances, we obtain the **omitted variable bias formula**:

$$\beta_1^S = \beta_1^L + \beta_2^L \frac{Cov[X_1, X_2]}{Var[X_1]}$$

Note that we can obtain $\delta_1 = \frac{Cov[X_1,X_2]}{Var[X_1]}$ as the coefficient from a statistical regression of X_1 on X_2. Because it guides our intuition on identifying causal effects, this is one of the most powerful formulas in econometrics.

PROPOSITION 2.5.8. *Suppose that the long regression is given by:*

$$Y = \mathbf{X}'\beta^L + \mathbf{W}'\gamma + U_L$$

where $\mathbf{X} = \begin{pmatrix} 1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$ $\mathbf{W} = \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_m \end{pmatrix}$ *and* $\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{pmatrix}$. *The short regression is given by*

$$Y = \mathbf{X}'\beta^S + U_S$$

Then, the β^L *from the long regression is related to* β^L *from the short regression as in the* ***multiple omitted variables bias formula****.*

$$\beta^S = \beta^L + E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{X}\mathbf{W}'\right]\gamma$$

PROOF. Note that the short regression coefficient estimate has the form

$$\beta^S = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{X}Y\right]$$

To complete the proof, plug in for Y using the long regression:

$$\begin{aligned} \beta^S &= E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{X}\left(\mathbf{X}'\beta^L + \mathbf{W}'\gamma + U_L\right)\right] \\ &= \beta^L + E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{X}\mathbf{W}'\right]\gamma \end{aligned}$$

□

EXERCISE 2.5.9. Suppose there are k regressors in $\mathbf{X}$ and m omitted variables in $\mathbf{W}$. What is the dimension of β^S, β^L, $E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$, $E\left[\mathbf{X}\mathbf{W}'\right]$ and γ. Verify that the dimensions of the above expression are conformable.

In this expression $\Delta = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{X}\mathbf{W}'\right]$ is a $(k+1) \times m$ matrix where each column is of the form:

$$\delta_j = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{X}W_j\right]$$

That is, each column j is a vector of regression coefficients from the statistical regression

$$W_j = \mathbf{X}'\delta_j + U_d$$

EXERCISE 2.5.10. **Single Included Variable. Multiple Omitted Variables**. Suppose that the true causal relationship is described by the long regression:

$$Y = \beta_0^L + \beta_1^L X_1 + \mathbf{W}'\gamma + U_L$$

where $\mathbf{W} = \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_m \end{pmatrix}$ and $\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{pmatrix}$. Derive an expression for the omitted variables bias formula that relates β_1^L to β_1^S in this setting. Be very precise. Is the vector of coefficients that Angrist and Pischke describe from the same regression? If so, what is the specification? If not, explain precisely how you could obtain these coefficients.

2.5.2. Measurement Error. The preceeding discussion suggests that if we want to obtain the *right* estimates (in the causal sense), we should collect information on every variable that is correlated with our regressor of interest X_1 and possibly has predictive power for our response variable Y. This is a good idea, provided that our additional regressors can be observed at low cost and without measurement error. What if we can obtain a noisy measurement of our regressors? The next section develops this idea.

2.5.2.1. *Attenuation Bias in Simple Linear Regression.* Suppose that the true causal relationship is described by

$$Y = \beta_0 + \beta_1 X_1 + U$$

but that we observe $X_1^* = X_1 + \xi$ instead of X_1 where ξ is a random variable that represents some measurement error process. Clearly, if our measurement error ξ is correlated with the true regressor X_1, the mismeasurement will distort our estimator in a perverse way. For this reason, let's focus on the case where X_1 and ξ are uncorrelated, $Cov[X_1, \xi] = 0$.

When we estimate the regression

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X_1^* + \tilde{U}$$

using OLS, the OLS estimator $\hat{\beta}_1^{ols}$ is consistent for $\tilde{\beta}_1$, which can be expressed in terms of the mismeasured regressor and the response variable as

$$\hat{\beta}_1^{ols} \xrightarrow{P} \tilde{\beta}_1 = \frac{Cov[X_1^*, Y]}{Var[X_1^*]}$$

We can plug into this expression for $X_1^* = X_1 + \xi$ and Y using the true regression to obtain:

$$\begin{aligned} \tilde{\beta}_1 &= \frac{Cov[X_1 + \xi, \beta_0 + \beta_1 X_1 + U]}{Var[X_1 + \xi]} \\ &= \frac{\beta_1 Var[X_1]}{Var[X_1] + Var[\xi]} \end{aligned}$$

which is smaller in magnitude than β_1. That is, random measurement error in the case of simple linear regression implies that the OLS estimator $\hat{\beta}_1^{ols}$ is inconsistent for β_1 from the true regression model. This inconsistency is called **attenuation bias** because in the case of simple linear regression the measurement error attenuates the estimated effect, leading to an underestimate of the true parameter value.

2.5.2.2. *Measurement Bias in Multiple Regression.* The result that random measurement error in a regressor attenuates the estimated effects in simple linear regression is hopeful, but it turns out that the measurement error bias can go in either direction once we introduce it to multiple regression.

Suppose that $X_1^* = X_1 + W$ is the only variable that is mismeasured. That is, we would like to estimate the regression model:

$$Y = \mathbf{X}'\beta + U$$

with $\mathbf{X} = \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$, but we use a mismeasured regressor X_1^* in place of X_1. Denote $\mathbf{X}^* = \mathbf{X} + \begin{pmatrix} 0 \\ W \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{X} + \mathbf{W}$. That is, $\mathbf{X}' = (\mathbf{X}^*)' - \mathbf{W}'$. Plug this expression into the true regression relationship to obtain:

$$Y = (\mathbf{X}^*)'\beta + \underbrace{U - \mathbf{W}'\beta}_{U^*}$$

The OLS estimator consistently estimates:

$$\begin{aligned} \beta^* &= E\left[\mathbf{X}^*(\mathbf{X}^*)'\right]^{-1} E\left[\mathbf{X}^* Y\right] \\ &= \beta + E\left[\mathbf{X}^*(\mathbf{X}^*)'\right]^{-1} E\left[\mathbf{X}^* U^*\right] \\ &= \beta + E\left[\mathbf{X}^*(\mathbf{X}^*)'\right]^{-1} E\left[\mathbf{X}^*\left(U - \mathbf{W}'\beta\right)\right] \\ &= \beta - E\left[\mathbf{X}^*(\mathbf{X}^*)'\right]^{-1} \underbrace{E\left[\mathbf{X}^*\mathbf{W}'\beta\right]}_{=E[(\mathbf{X}+\mathbf{W})\mathbf{W}']\beta} \\ &= \beta - \underbrace{E\left[\mathbf{X}^*(\mathbf{X}^*)'\right]^{-1} E\left[\mathbf{W}\mathbf{W}'\right]\beta}_{measurement\ bias} \end{aligned}$$

which does not equal the parameter vector β we would like to estimate.

An Advanced Derivation (for the adventurous student). Explore the measurement bias term in this expression. First, note that $E\left[\mathbf{W}\mathbf{W}'\right]\beta = \begin{pmatrix} 0 \\ E\left[W^2\right]\beta_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. In the matrix multiplication, this expression will pick off the second column in $E\left[\mathbf{X}^*(\mathbf{X}^*)'\right]^{-1}$. Second, apply Woodbury's Theorem for inverting sums of matrices to split out the measurement error from the original regressor.

$$E\left[\mathbf{X}^*(\mathbf{X}^*)'\right]^{-1} = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} - \frac{1}{1+h_{22}} E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{W}\mathbf{W}'\right] E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$$

where $h_{22} = E\left[W^2\right] g_{22}$ where g_{22} is the $(2,2)$ element of the matrix $E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$. Now, notice that the $E\left[\mathbf{X}\mathbf{X}'\right]^{-1} E\left[\mathbf{W}\mathbf{W}'\right] E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$ term is a matrix of zeroes except for the $(2,2)$ element, which equals $g_{22}^2 E\left[W^2\right]$. Denote the second column of the

matrix $E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$ as $g_2 = \begin{pmatrix} g_{21} \\ g_{22} \\ g_{23} \\ \vdots \\ g_{2(k+1)} \end{pmatrix}$. Then, we can express the measurement error term as

$$E\left[\mathbf{X}^*\left(\mathbf{X}^*\right)'\right]^{-1} E\left[\mathbf{W}\mathbf{W}'\right] = E\left[W^2\right]\beta_1 \begin{pmatrix} g_{21} \\ g_{22} - \frac{g_{22}^2 E\left[W^2\right]}{1+g_{22}E\left[W^2\right]} \\ g_{23} \\ \vdots \\ g_{2(k+1)} \end{pmatrix}$$

This equation for measurement error leads to two natural observations about the role of measurement error in multiple regression.

(1) With measurement error in *one* of the regressors (without loss of generality, assume X_1 is the mismeasured regressor), the OLS estimator $\hat{\beta}_j^{ols}$ is (in general) inconsistent for β_j, even if X_j is measured accurately.
(2) The measurement error bias can be in either direction. That is, measurement error in X_1 may lead to an under- or over-estimate for β_j.

2.6. Regression and Linearity

The only reason for a discussion of linearity versus nonlinearity is because we wish to use a powerful tool (linear regression) that happens to be linear. The primary tool in the econometrician's toolkit is linear regression. As the results in this section demonstrate, linear regression is more powerful and less restrictive than it first appears.

REMARK 2.6.1. In the linear approximation to a nonlinear causal population, the error term V depends both on the unobserved determinants of Y and the specification error that arises from using a linear function to approximate a nonlinear relationship (Z, for example, could depend on X^2 and $\log(X)$, which are observed, but not used). In practice, it is difficult to distinguish these two types of misspecification. As long as this we have such a well-specified model, the linear causal population allows us to conceptualize the error term as "unobserved determinants of Y."

This remark about nonlinearity appears to stand in contrast to an important result for the linear approximation of a nonlinear regression relationship: the **average derivative property**. This result is cleanest to see in a single regression setup though a version of it holds in multiple regression.[10]

THEOREM 2.6.2. *Consider the case of one regressor. Suppose that a conditional expectation function. $h(x) = E[Y|X = x]$ is a nonlinear function of the value of X (with derivative $h'(x)$). As long as we assume that X is continuously distributed, the slope coefficient in the linear regression $Y = \beta_0 + \beta_1 X + U$ can be expressed as $\beta_1 = \frac{\int h'(x)\mu_x dx}{\int \mu_x dx}$, the weighted average of the slope of the true CEF with weights given by $\mu_x = (E[X|X \geq x] - E[X|X < x])(1 - P[X \geq x])P[X \geq x]$*

This is a powerful result with an important caveat. This average derivative property is informative as long as the nonlinear CEF is approximately linear. That is, as linearity of the conditional expectation becomes a worse model, it means less for describing the ultimate nonlinear relationship. For example, let X be age (ranging from 9 years to 90 years) and Y be some measure of fertility. Conditional on age, expected fertility rises from about zero to some positive number back to about

[10] See Angrist and Pischke for more details, including a proof in the appendix to Chapter 2.

zero. A linear approximation makes little sense in this setting. Fortunately, linear regression can accommodate this setting for a richer set of interpretations.

REMARK 2.6.3. **Linearity and Transformations**. Although it restricts the number of applications, imposing linearity in the causal model is not as restrictive as it first appears.

(1) The linearity requirement is on the parameters of the causal regression model, not necessarily the regressors. We can transform the variables of interest so that the relationship between the transformed variables is linear. For example, it may be that Y and X have a nonlinear relationship, but $\log Y$ and $\log X$ have a linear relationship. In this case, it would be a good idea to estimate

$$\log Y = \beta_0 + \beta_1 \log X + U$$

(2) In general, we can include multiple observed causal determinants of Y. If the relationship is nonlinear due to correlations with these other determinants, we should include them in the model as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

or (in general):

$$Y = \mathbf{X}'\beta + U$$

Still, there may be a fundamental nonlinearity in the causal model. Multiple regression helps us here, too, because we can use predictors and their transformations as regressors. It is perfectly valid if the regressors are functions of one another (as long as the additional regressors do not provide redundant information). For example, if we believe that the true relationship between Y and X is quadratic, we can specify the causal model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U$$

or if we believe that Y not only depends on X and Z, but also the interaction between the two:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + U$$

These are all linear regression models, but including transformed regressors allows us to analyze nonlinear relationships using linear regression.

REMARK 2.6.4. Just as in the linear causal regression case, **the effect** of X on Y is still the partial derivative of Y with respect to X, $\frac{\partial Y}{\partial X}$ in regression models with transformations and interactions. The only difference is that this partial derivative is computed differently.

For example, in the quadratic regression case,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U$$

the effect is $\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$.

In the logged regressor case,

$$Y = \beta_0 + \beta_1 \log X + U$$

the effect is $\frac{\partial Y}{\partial X} = \frac{\beta_1}{X}$. Solving for β_1, we obtain $\beta_1 = \frac{\partial Y}{\partial X/X}$, an expression we call the **semielasticity** of Y with respect to X.[11]

In the log-log regression case,

[11] We call this expression semielasticity because the denominator is the instantaneous percentage change in X while the numerator is an absolute change in Y. We will often abbreviate the $\frac{\partial X}{X}$ term as $\partial \log X$ for obvious reasons.

$$\log Y = \beta_0 + \beta_1 \log X + U$$

can you show that β_1 is the elasticity of Y with respect to X, $\epsilon_{Y,X} = \frac{\partial Y}{\partial X}\frac{X}{Y}$?

2.7. Some Practical Details of OLS Regression

The terminology in this section shows up in the previous regression notes on projection matrices, but you may find a less vectorized description of these methods informative as a bridge from scalar algebra to linear algebra.

DEFINITION 2.7.1. The **fitted value** is a useful object computed by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ for each observation i. For multiple regression, we compute the fitted value as $\hat{Y}_i = \mathbf{X}_i'\hat{\beta}$.

The **residual** is defined in terms of the fitted value $\hat{U}_i = Y_i - \hat{Y}_i$ as how much we overpredict observation i by using the regression line.

Plots of these residuals will be informative diagnostics of assumptions that we'll want to make in estimating a regression model (independence, linearity and constant variance). Residuals and fitted values can be defined for any $\tilde{Y} = b_0 + b_1 X$ candidate regression line (or $\tilde{Y} = \mathbf{X}'\tilde{b}$ for multiple regression). Using an arbitrary candidate regression line, we can think of the OLS problem as choosing $\tilde{b} = (b_0, b_1, ..., b_k)$ to minimize the sum of squared residuals.

$$\min_{\tilde{b}} \quad \sum_{i=1}^n \tilde{U}_i^2$$

Especially for hypothesis testing (and comparing the performance of regression models to one another), it will be useful to think in terms of sums of squared residuals from various fitted models.

DEFINITION 2.7.2. For the best fitting regression line, define **sum of squared residuals (SSR)**:

$$SSResid \equiv \sum_{i=1}^n \hat{U}_i^2$$

Imagine $\tilde{Y} = \bar{Y}$, the sample mean of Y_i. Using this as our estimator of Y_i regardless of the value of X_i, we could form a similar sum of squared term, called the **sum of squares total (SST)**:

$$SSTotal = \sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2$$

DEFINITION 2.7.3. One measure of goodness of fit is called R^2, which we define to be:

$$\begin{aligned} R^2 &= \frac{SSTotal - SSResid}{SSTotal} \\ &= \frac{SSExplained}{SSTotal} \end{aligned}$$

CLAIM 2.7.4. In the previous definition of R^2, we defined $SSExplained \equiv SSTotal - SSResid$. Through some simple algebraic manipulation, we can also show that $SSExplained = \sum_{i=1}^n \left(\hat{Y}_i - \bar{Y}\right)^2$.

You may also wonder why R^2 has its name. The following claim is why.

CLAIM 2.7.5. It also turns out that

$$R^2 = (r_{XY})^2$$

where $r_{XY} = \frac{S_{X,Y}}{S_X S_Y}$ is the sample correlation between X and Y.

Another important statistic that shows up in inference relating to simple linear regression is the **mean squared residual (MSR)**:[12]

$$MSR = \frac{SSResid}{df}$$

where $df = n - 2$. This is an estimator for the variance of U. If we want an estimate for the standard deviation of U, we can take the square root. $SER = \sqrt{MSR}$. This object is often called the root-MSR or the **standard error of regression**.[13] Like R^2, SER can be a useful measure of the performance of the regression model.

As it estimates the standard deviation of our statistical error term, smaller SER means less unexplained variability in our response variables. If all we care about is prediction within the sample, we may prefer models that have lower SER because their predictive accuracy is greater. In econometrics, however, we have a much more important priority: valid inference.

REMARK 2.7.6. ***Warning about*** R^2: Our estimation routines have been defined for statistical models while a key motivating interest in regression is to assign causation. For this reason, we should be interested in when our statistical model can be used for causal inference. This only happens when $Cov[X, V] = 0$.
Notice that this property that ensures the validity of causal inference holds regardless of the value of R^2. In fact, if R^2 is "suspiciously high," it might cause concern about the validity of our causal inference. This is because some of that explained variation in the numerator of R^2 could be coming from a correlation with unobserved factors (and that correlation undermines our ability to line up our causal model with the statistical model). This is because R^2 is higher when X represents – or is correlated with – a bunch of other factors that also cause Y, but are not in the regression.

2.8. Hypothesis Testing in Regression

To this point, we have analyzed the OLS estimator $\hat{\beta}$ and studied its properties. In this section, we discuss how these properties – specifically, consistency and asymptotic normality – lead naturally to inference about the true parameter vector β. Deep down, we would like to make statements about the true relationships among our predictors in $\mathbf{X}$ and the response variable Y. Inference and hypothesis testing in OLS regression is how we make the link from statistics to statements about the true state of the world.

[12]There is a more general concept here called a mean square, which is computed by taking $\frac{SS}{df}$ for each of the types of sums of squares we have defined thus far. For example, we could define the mean squared explained $MSExplained = \frac{SSExplained}{df_{explained}}$, but $df_{explained}$ equals one when we only have one predictor. This will be more interesting to discuss when we have multiple regressors in the regression model, and it will be an essential calculation for multiple inference.

[13]You may also see the terminology "root-MSE" for square root of mean squared "error." We will try to stick with terminology residual to denote sample differences of our estimator from our response variable, but computer software packages often use a different convention.

As is standard in statistics, we ground our methods of inference in the sampling distribution of our estimator. Similarly to our approach in the probability notes, we want to make minimal assumptions about the population regression. For this reason, we use the limiting distribution of our estimator as the approximate sampling distribution for inference. We have already derived the asymptotic sampling distribution of the OLS estimator:

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \quad \xrightarrow{d} \quad N\left(\mathbf{0},\Sigma\right)$$

where $\Sigma = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} Var\left[\mathbf{X}U\right] E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$ is the asymptotic variance of $\sqrt{n}\hat{\beta}$. Under the assumption of homoskedasticity $Var\left[U|\mathbf{X}\right] = Var\left[U\right] = \sigma^2$, we have a simpler form for $\Sigma = \sigma^2 E\left[\mathbf{X}\mathbf{X}'\right]^{-1}$.

Software packages use the homoskedasticity assumption as the default and compute a consistent estimator for Σ, given by

$$\hat{\Sigma} \quad = \quad \left(\mathbf{Y}-\mathbb{X}\hat{\beta}\right)'\left(\mathbf{Y}-\mathbb{X}\hat{\beta}\right)\left[\mathbb{X}'\mathbb{X}\right]^{-1} = SSResid\left[\mathbb{X}'\mathbb{X}\right]^{-1}$$

2.8.1. Putting the Sampling Distribution to Use. In practice, we will use a consistent estimator for Σ, denoted $\hat{\Sigma}$ and isolate $\hat{\beta}$ in the above converence in distribution argument to argue that

$$\hat{\beta} \quad \approx \quad N\left(\beta, \frac{\hat{\Sigma}}{n}\right)$$

DEFINITION 2.8.1. In this expression, we call $\frac{\hat{\Sigma}}{n}$ the **estimated (approximate) variance-covariance matrix** of $\hat{\beta}$.

As long as we are willing to ground our inference in large sample results, this matrix contains all of the information we will need to use a coefficient vector estimate $\hat{\beta}$ to provide inference for the true parameter vector β. More specifically, $\frac{\hat{\Sigma}}{n}$ contains estimated variances of the estimators $\hat{\beta}_i$ along its diagonal and the estimated covariances in the off diagonal elements:

$$\frac{\hat{\Sigma}}{n} \quad = \quad \begin{bmatrix} \hat{Var}\left[\hat{\beta}_0\right] & \hat{Cov}\left[\hat{\beta}_0,\hat{\beta}_1\right] & \dots & \hat{Cov}\left[\hat{\beta}_0,\hat{\beta}_k\right] \\ \hat{Cov}\left[\hat{\beta}_0,\hat{\beta}_1\right] & \hat{Var}\left[\hat{\beta}_1\right] & & \hat{Cov}\left[\hat{\beta}_1,\hat{\beta}_k\right] \\ \vdots & & \ddots & \\ \hat{Cov}\left[\hat{\beta}_0,\hat{\beta}_k\right] & \hat{Cov}\left[\hat{\beta}_1,\hat{\beta}_k\right] & \dots & \hat{Var}\left[\hat{\beta}_k\right] \end{bmatrix} = MSResid \times \left(\mathbb{X}'\mathbb{X}\right)^{-1}$$

where the second equality assumes homoskedasticity.

REMARK 2.8.2. To obtain **the standard error of an estimate** $\hat{\beta}_i$, extract $\hat{Var}\left[\hat{\beta}_i\right]$ from the variance covariance matrix and compute its square root: $SE\left[\hat{\beta}_i\right] = \sqrt{\hat{Var}\left[\hat{\beta}_i\right]}$.

2.8.1.1. *Tests for Single Regression Coefficients.* Standard errors are useful for inference about the true coefficient β_j because they form the denominator of our **test statistic** for tests concerning single regression coefficients:

$$T \quad = \quad \frac{\hat{\beta}_j - \beta_j^0}{SE\left[\hat{\beta}_j\right]}$$

where β_j^0 is the null-hypothesized value for a hypothesis test regarding β_j. Formally, $H_0: \beta_j = \beta_j^0$. On an intuitive level, we use the test statistic to measure the "distance" in number of standard errors that our computed estimate $\hat{\beta}_j$ is from the hypothesized value β_j^0. If this distance is large enough (above the critical value we set to keep Type I error at α probability), we reject the null hypothesis.

EXAMPLE 2.8.3. Suppose that we estimate the regression model $Y = \beta_0 + \beta_1 X_1 + U$ using OLS, and we want to know if the regressor X_1 is **statistically significant**. That is, we want to know if X_1 has a relationship with Y in the population regression. We can only use our regression estimates to make this conclusion.

The null hypothesis for this test is $H_0: \beta_1 = 0$ and it is natural to use the two-sided alternative $H_1: \beta_1 \neq 0$. Imagine that you want to test this hypothesis at the 5 percent level. That is, you want to limit the probability of making a Type I error to be $\alpha = 0.05$. For a two-sided test, we will reject if T is below the 0.025 quantile of the standard normal or above the 0.975 quantile. In this case, our decision rule is simple: reject if $|T| > 1.96$, fail to reject otherwise.

In practice, we could carry out this test on actual data by computing the OLS estimate of $\hat{\beta}_1$ using our sample and using the $SE\left[\hat{\beta}_1\right]$ computed from the diagonal of the estimated variance-covariance matrix (which is only a 2×2 matrix in this example).[14] Suppose $\hat{\beta}_1 = 0.75$ and $SE\left[\hat{\beta}_1\right] = 0.25$, then the computed value of $T = \frac{0.75-0}{0.25} = 4$. If this is the case, we reject the null hypothesis of no relationship between X_1 and Y, and conclude that X_1 has an independent relationship on Y .

Conducting a hypothesis test regarding a single coefficient in multiple regression is similar.

EXAMPLE 2.8.4. Suppose that we estimate the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$ using OLS, and we want to know if the regressor X_1 is **statistically significant**. Statistical significance is slightly more complicated in multiple regression. Informed by the Frisch-Waugh Theorem, we want to know if X_1 has its own independent relationship with Y, after partialing out the effect of X_2.

The null hypothesis for this test is $H_0: \beta_1 = 0$ and it is natural to use the two-sided alternative $H_1: \beta_1 \neq 0$. As in the simple regression context, our decision rule is simple: reject if $|T| > 1.96$, fail to reject otherwise.

In practice, we could carry out this test on actual data by computing the OLS estimate of $\hat{\beta}_1$ using our sample and using the $SE\left[\hat{\beta}_1\right]$ computed from the diagonal of the estimated variance-covariance matrix (now, a 3×3 matrix). Suppose $\hat{\beta}_1 = 0.4$ and $SE\left[\hat{\beta}_1\right] = 0.1$, then the computed value of $T = \frac{0.4-0}{0.1} = 4$. If this is the case, we reject the null hypothesis of no effect, and conclude that X_1 has its an independent relationship on Y that is distinct from X_2.

2.8.1.2. *Tests for a Single Linear Combination of Regression Coefficients.* The method in the previous section for conducting inference naturally extends to testing hypotheses regarding linear combinations of the regression coefficients. To see how we can extend the method, recognize that $SE\left[\hat{\beta}_j\right] = \sqrt{\hat{Var}\left[\hat{\beta}_j\right]}$.

[14] Many treatments of this material write out the estimated variance-covariance matrix in terms of the cumbersome formulas that arise in the details of the matrix multiplication. You won't apply these non-matrix formulas and they provide little intuition about the properties of standard errors, so why bother learning them? In an effort to more clearly see the concept, we stick to a primarily matrix-based treatment of standard errors. In practice, we use matrices to compute them anyway.

DEFINITION 2.8.5. Consider the estimator defined by the linear combination $\hat{\theta} = r'\hat{\beta}$, where r' is a conformable row vector of constants. The **standard error of the linear combination** $\hat{\theta} = r'\hat{\beta}$ is given by

$$SE\left[\hat{\theta}\right] = SE\left[r'\hat{\beta}\right] = \sqrt{\hat{Var}\left[r'\hat{\beta}\right]} = \sqrt{r'\hat{Var}\left[\hat{\beta}\right]r} = \sqrt{\frac{r'\hat{\Sigma}r}{n}}$$

Just as we did before, we can form the t-ratio as our test statistic for tests about the linear combination $\theta_0 = r'\beta$:

$$T = \frac{\hat{\theta} - \theta_0}{SE\left[\hat{\theta}\right]} = \frac{r'\hat{\beta} - \theta_0}{SE\left[\hat{\theta}\right]}$$

where θ_0 is the null-hypothesized value of the linear combination $r'\beta$. Formally, $H_0 : \ r'\beta = \theta_0$. The same intuition applies as for tests for single regression coefficients. We use the test statistic to measure the "distance" in number of standard errors that our computed estimate of the linear combination $r'\hat{\beta}$ is from the hypothesized value θ_0. If this distance is large enough (above the critical value we set to keep Type I error at α probability), we reject the null hypothesis.

EXAMPLE 2.8.6. Suppose we estimate the regression model $\log(Y) = \beta_0 + \beta_1 \log(K) + \beta_2 \log(L) + \beta_3 \log(M) + U$ where Y is output per year, K, L and M are inputs used per year. In this context, an interesting null hypothesis versus the one sided alternative[15] is

$$\begin{aligned} H_0 : \ \beta_1 + \beta_2 + \beta_3 &= 1 \\ H_1 : \ \beta_1 + \beta_2 + \beta_3 &> 1 \end{aligned}$$

This hypothesis test is exactly

$$\begin{aligned} H_0 : \ r'\beta &= \theta_0 \\ H_1 : \ r'\beta &> \theta_0 \end{aligned}$$

where $r = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$ and $\theta_0 = 1$.

Now, because this is a one-sided test, $\alpha = 0.05$ implies that we obtain a critical value of 1.645 from the standard normal distribution (the 95th percentile). Our decision rule is to reject if $T > 1.645$ and fail to reject otherwise.

To carry out the test, we obtain $\hat{\beta}$ and use it to construct $\hat{\theta} = r'\hat{\beta}$ as our estimator for $\theta = r'\beta$. Next, we need to obtain the standard error. Software will produce the estimated variance-covariance matrix $\hat{V} = \frac{\hat{\Sigma}}{n}$. The estimated variance of $\hat{\theta}$ equals $\hat{Var}\left[r'\hat{\beta}\right] = r'\hat{V}r$. Use this fact to compute $SE\left[\hat{\beta}\right] = \sqrt{r'\hat{V}r}$ and construct the test statistic

$$T = \frac{r'\hat{\beta} - 1}{\sqrt{r'\hat{V}r}}$$

[15] This is actually a test for increasing returns to scale in factor inputs (K, L, M).

according to the general formula given in the probability notes.[16] For example, suppose that $\hat{\beta}$ and $\hat{V}$, the estimated variance-covariance matrix, are:

$$\begin{aligned}\hat{\beta} &= \begin{pmatrix} 0.05 \\ 0.3 \\ 0.6 \\ 0.3 \end{pmatrix} \\ \hat{V} &= \begin{bmatrix} 0.01 & 0.004 & 0.004 & -0.003 \\ 0.004 & 0.015 & -0.007 & -0.0025 \\ 0.004 & -0.007 & 0.02 & -0.002 \\ -0.003 & -0.0025 & -0.002 & 0.030 \end{bmatrix}\end{aligned}$$

From this, we can compute $\hat{\theta} = r'\hat{\beta} = 1.2$ and we can verify (using computer software) that the standard error equals $SE\left[r'\hat{\beta}\right] = 0.2049$. Hence, the test statistic equals $T = \frac{1.2-1}{0.2049} = 0.9759$, which is less than our critical value. Hence, we fail to reject the null hypothesis.

2.8.1.3. *Tests for Multiple Restrictions.* To this point, the hypothesis tests we have considered have had only one restriction. That is, we were interested in testing a single restriction on the parameters (even if that restriction involved a linear combination of several coefficients). In addition to these single-restriction tests, we will occasionally wish to test several (or many) restrictions on the parameters simultaneously. We call such a hypothesis test a **joint test**, or alternatively, a test with **multiple restrictions**.

A naive approach would be to apply the testing procedure we used in the previous section for each restriction we would like to test.

Example 2.8.7. Suppose we estimate the regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + U$$

using OLS and we would like to test the joint hypothesis

$$H_0 : \beta_1 = \beta_2 \quad \text{and} \quad \beta_3 = \beta_4$$

against the alternative that at least one of the two restrictions does not hold.

If we were to take the naive approach to deal with these hypotheses separately, the first restriction could be implemented using $r_1' = (0, 1, -1, 0, 0)$ and $\theta_0^{(1)} = 0$ the second restriction could be implemented using $r_2' = (0, 0, 0, 1, -1)$ and $\theta_0^{(2)} = 0$. Then, we would just construct $T_1 = \frac{r_1'\hat{\beta} - \theta_0^{(1)}}{SE\left[r_1'\hat{\beta}\right]}$ and $T_2 = \frac{r_2'\hat{\beta} - \theta_0^{(2)}}{SE\left[r_2'\hat{\beta}\right]}$ and compare to the critical value that guarantees α.

[16] $r'\hat{\beta}$ has an asymptotically normal distribution. Hence, $T = \frac{\hat{\theta} - \theta_0}{SE\left[\hat{\theta}\right]} \xrightarrow{d} N(0, 1)$

REMARK 2.8.8. **A Comment on the Naive Approach to Multiple Restrictions**. The naive approach is the wrong way to test for joint significance, but it can provide some intuition about the right test to conduct. The problem with this naive approach is that the two test statistics are possibly correlated. For this reason, we cannot just combine the two statistics into one by (for example) computing $X^2 = T_1^2 + T_2^2$.
This squaring-and-summing technique would work for a large enough sample size *and if the two test statistics* T_1 *and* T_2 *were independent.* Because squared $N(0,1)$ RVs are χ_1^2 and summing independent χ^2 RVs produces another χ^2 RV (with *df* equal to the sum of the *df* of the individual χ^2 RVs), we know $X^2 \sim \chi_2^2$ *as long as* T_1 *and* T_2 *are independent.* Even though T_1 and T_2 are not independent, we will do something like squaring and summing the individual test statistics when we construct our F statistic.

In the previous example, a more sophisticated way to conduct multiple restriction hypothesis tests is to construct the F statistic as follows:

$$F = \left(\begin{pmatrix} r_1' \\ r_2' \end{pmatrix} \hat{\beta} - \begin{pmatrix} \theta_0^{(1)} \\ \theta_0^{(2)} \end{pmatrix} \right)' \left[\begin{pmatrix} r_1' \\ r_2' \end{pmatrix} \hat{V} \begin{pmatrix} r_1' \\ r_2' \end{pmatrix}' \right]^{-1} \left(\begin{pmatrix} r_1' \\ r_2' \end{pmatrix} \hat{\beta} - \begin{pmatrix} \theta_0^{(1)} \\ \theta_0^{(2)} \end{pmatrix} \right)$$

where $\hat{V}$ is the estimated variance-covariance matrix. We can simplify the expression by defining

$$\begin{aligned} R &= \begin{pmatrix} r_1' \\ r_2' \end{pmatrix} \\ \mathbf{c} &= \begin{pmatrix} \theta_0^{(1)} \\ \theta_0^{(2)} \end{pmatrix} \end{aligned}$$

If we extend the simpler (R and $\mathbf{c}$) notation to the case where R and $\mathbf{c}$ have a number of rows equal to *df* (each row represents a restriction to be tested, so we test *df* restrictions), we obtain the general formula for the F statistic.

DEFINITION 2.8.9. **F-statistic for Linear Restrictions in Multiple Regression**. To test a null hypothesis that can be expressed in matrix-vector form as $H_0: R\beta = \mathbf{c}$.he F statistic equals

$$F = \left(R\hat{\beta} - \mathbf{c} \right)' \left[R\hat{V}R' \right]^{-1} \left(R\hat{\beta} - \mathbf{c} \right)$$

Constructed this way, $F \xrightarrow{d} \chi_{df}^2$ where *df* = number of rows in R.

If you are comfortable with matrix factorizations, the following formalization of the intuition behind the form of the F statistic may be helpful.

REMARK 2.8.10. **An Advanced Note: Intuition for the form of the F statistic**. If we break the F statistic into different pieces, it becomes somewhat easier to understand why it has the form it does.

(1) The vector $R\hat{\beta} - \mathbf{c}$ contains *precisely* the terms in the numerator of the individual T test statistics that we considered using in the naive approach.
(2) The matrix $\mathcal{V} = Var\left[R\hat{\beta}\right] = R\hat{V}R'$ is the variance-covariance matrix for the elements of the vector $R\hat{\beta} - \mathbf{c}$. We can compute the square root of the diagonal elements of $Var\left[R\hat{\beta}\right]$ to obtain the standard errors from the naive approach. The off-diagonal elements in this matrix give us information about how the elements of $R\hat{\beta}$ are correlated with one another.
   - If only there were a way to take the square root of the matrix and bring along the information on how the linear combinations are correlated with one another. It turns out that there is. Being a positive definite symmetric matrix $\mathcal{V} = Var\left[R\hat{\beta}\right]$ has an inverse and its inverse has a square root decomposition (motivated by an eigenvalue-eigenvector decomposition) Hence, we can write $\mathcal{V}^{-1} = \mathcal{V}^{-\frac{1}{2}}\mathcal{V}^{-\frac{1}{2}}$ and $\mathcal{V}^{-\frac{1}{2}}$ is symmetric.
(3) Given (2), we can rewrite the F statistic. Define $\hat{\mathcal{T}} = \left(R\hat{\beta} - \mathbf{c}\right)\mathcal{V}^{-\frac{1}{2}}$, which looks like a vector equivalent of the naive approach (while accounting for the correlations between individual the test statistics). Given all of this work, the F statistic can be computed as

$$F = \hat{\mathcal{T}}'\hat{\mathcal{T}}$$

This is an inner product, which is a sum of the squared elements of $\hat{\mathcal{T}}$.

That the asymptotic distribution of F turns out to be χ^2_{df} is not surprising given our approximate intuition – made formal in the previous box – that the F statistic is the sum of squared T statistics (adjusted appropriately) from each restriction in the joint test.[17]

There are several important examples of F tests in econometrics:

EXAMPLE 2.8.11. **Omnibus F-test**. Suppose we estimate the regression model

$$Y = \beta_0 + \sum_{i=1}^{5} \beta_i X_i + U$$

and we want to conduct a test of the null hypothesis that none of the regressors has a relationship with the response variable, *jointly*. Formally, this null hypothesis is $H_0 : \beta_1 = \beta_2 = ... = \beta_5 = 0$ and the alternative hypothesis H_1 is *at least one of the β_i coefficients is not zero*. To translate this hypothesis into matrix form, $R = [\mathbf{0}_5, I_5]$ where$\mathbf{0}_5$ is a 5×1 vector of zeros and I_5 is the 5×5 identity matrix. Moreover, $\mathbf{c} = \mathbf{0}_5$.

In this case, the F-statistic has an asymptotic distribution of χ^2_5. At $\alpha = 0.05$, we look up the 95th percentile of the χ^2_5 distribution, which is 11.0705.[18] Our decision rule is to reject if $F > 11.0705$ and fail to reject otherwise. From here, we apply the formula in Definition 3.9 to obtain a computed value of F and conclude about the null hypothesis by comparing F to our critical value.

[17] If we were willing to assume that the population is multivariate normal, the F statistic would have an F distribution with df_{num} equal to the number of restrictions and $df_{denom} = n - dim(\beta)$. Not surprisingly, it is true that $F_{df_{num}, df_{denom}} \xrightarrow{d} \chi^2_{df_{num}}$.

[18] Note: we look up the 95th percentile because the events leading to both tails of the standard normal are translated into the upper tail of the χ^2 distribution because we are squaring. Big negative T values and big positive T values enter positively in the F statistic. In R, the command for obtaining this percentile is `qchisq(0.95,5)`.

In practice, the omnibus F-test is provided by software packages in the the standard table of regression output.

EXAMPLE 2.8.12. **Chow Test**. Suppose there are two groups of individuals in our data set and we define the dummy (binary) variable D

$$D = \begin{cases} 1 & \text{if member of group A} \\ 0 & \text{if member of group B} \end{cases}$$

In our study of this population, we model the relationship between Y and $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U^{(1)}$$

To give some more context, imagine that group A is males and group B is females (so D is a dummy variable that could be called *male*), Y is wages, X_1 is number of children in the household, and X_2 is years of education. Given this context, we might be concerned that the regression relationship between wages, children in the household and years of education is *different in some way* for females than it is for males. That is, the coefficient vector $(\beta_0, \beta_1, \beta_2)'$ depends on group membership.

If the relationship is different by group membership, we would like to know. The test for a difference along these lines is a joint hypothesis test called the Chow Test.[19] To motivate this hypothesis test, suppose we estimated a second regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 D + \beta_4 (D \times X_1) + \beta_5 (D \times X_2) + U^{(2)}$$

using OLS. Then, the null hypothesis of the Chow Test is $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ and that alternative that there are differences in the regression relationship by gender means β_3, β_4 and β_5 are not all zero. The F statistic has an asymptotically χ^2_3 distribution, which has a 95th percentile of 7.814. Hence, our rejection rule is to reject the null hypothesis if $F > 7.814$ and fail to reject otherwise.

We can translate into matrix notation to compute F. In this problem, $R = [\mathbf{0}_{3\times 3}, I_{3\times 3}]$ and $\mathbf{c} = \mathbf{0}_{3\times 1}$. From this matrixization of our hypothesis test and the estimated variance-covariance matrix $\hat{V}$, we can compute F using Definition 3.9. Alternatively, we can used canned methods in R and Stata to produce the desired test statistics.

Although the canned methods in R and Stata use the techniques described in this section (and are a perfectly valid way to conduct empirical research), you should be able to recreate the matrix algebra behind these canned methods. Recreating the calculations behind these canned methods demonstrates mastery of hypothesis testing.

2.9. Chapter Exercises

(1) Consider a single linear regression model, $Y = \beta_0 + \beta_1 X + U$ where X is a binary random variable (can take on only two values, usually zero or one). Show that $E[Y|X]$ is linear and solve for β_0 and β_1 in terms of conditional expectations (not just covariances and variances). Does your expression make intuitive sense?

[19] Here is an insightful discussion of the Chow Test and how to implement it in Stata http://www.stata.com/support/faqs/stat/chow3.html. I agree with the author that it is a good idea to not get worked up about the fact that this test has a name. For our purposes (and for any practical application), you should think of it as an F test.

(2) Suppose you have two binary regressors – X_1 and X_2 – that correspond to distinct attributes of the setting you wish to study. For context, imagine that the unit of observation is US County. Let Y equal the per capita income in the county, let

$$X_1 = \begin{cases} 1 & \text{if the county has county sales tax} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if the county is classified as urban} \\ 0 & \text{otherwise} \end{cases}$$

(a) Consider the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$.

(i) What condition on the random variables X_1 and X_2 must be satisfied to identify the parameter vector $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$? Do you expect this condition to be satisfied in this setting?

(ii) Is $E[Y|X_1, X_2]$ necessarily linear in the parameters? Read Angrist and Pischke on saturated models. What modification to the regression model would you suggest to ensure that $E[Y|\mathbf{X}]$ is linear? Why is the linearity / nonlinearity of $E[Y|\mathbf{X}]$ important for the properties of $\hat{\beta}^{ols}$?

(3) Consider the single linear regression model

$$Y = \beta X + U$$

without an intercept (β is a scalar constant; X is a scalar random variable). Imagine that you have taken a random sample from this population. If you estimate a regression using OLS, but you do not include an intercept, do you know if the mean of the residuals is zero? Explain as precisely as you can.

(4) In the same setting as Question 3, consider the **ratio estimator** $\hat{\beta}^R = \frac{\bar{Y}}{\bar{X}}$ and assume that U is mean independent of X with zero conditional mean, $E[U|X] = 0$.

(a) Prove that $\hat{\beta}^R = \beta + \frac{\sum_{i=1}^n U_i}{\sum_{i=1}^n X_i}$. Hint: note that $\frac{\bar{Y}}{\bar{X}} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}$.

(b) Show that $\hat{\beta}^R$ is unbiased for β. Show that $\hat{\beta}^R$ is consistent for β.

(5) Let (Y, X, U) be a random vector where

$$Y = \beta_0 + \beta_1 X + U$$

for parameters β_0 and β_1, and Y is a binary (Bernoulli) random variable. Interpret this regression model as the best linear predictor of Y given X.

(a) State the minimization problem that (β_0, β_1) solves and find the first order conditions. What do these first order conditions imply about U and X?

(b) How (if at all) does the assumption of linear conditional expectation $E[Y|X]$ change the interpretation of U and β_1?

(c) Suppose that the conditional expectation $E[Y|X]$ is linear. Is U mean independent of X? Show why or why not.

(d) You collect an i.i.d. sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ of size n from (Y, X), which you use to estimate the parameters β_0 and β_1 from

$$Y_i = \beta_0 + \beta_1 X_i + U_i.$$

Assume $0 < Var[X] < \infty$, $E[Y^4] < \infty$, and $E[X^4] < \infty$. What are the OLS estimators of β_0 and β_1? These should be expressed as functions of the sample. Use any special aspects of this question to simplify these expressions as much as possible.

(e) Are your estimators $\hat{\beta}_0$, $\hat{\beta}_1$ unbiased? Show why or why not.

(f) Since Y is a binary (Bernoulli) random variable, show how you can express the conditional probability $P[Y = 1|X]$ is

$$P[Y = 1|X] = \beta_0 + \beta_1 X.$$

(g) Are the error terms U_i $i = 1, \ldots, n$ homoskedastic? Explain why or why not. (Hint: $Var[Y_i|X_1, \ldots, X_n] = Var[U_i|X_1, \ldots, X_n]$.)

(6) **Population Frisch-Waugh Theorem**. Let $(\mathbf{X}, Y, U)$ be a random vector where

$$Y = \mathbf{X}'\beta + U \tag{2.9.1}$$

and $E[U] = 0$, $E[\mathbf{X}U] = 0$ and $\mathbf{X} = \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$. Theorem 2.5 asserts that any parameterβ_j from this model could be interpreted as the parameter β_j^* from the (much simpler) statistical model:

$$Y = \beta_0^* + \beta_j^* \tilde{X}_j + U^* \tag{2.9.2}$$

Note: $\tilde{X}_j$ is a *error term* from a regression of X_j on X_{-j} where

$$X_{-j} \equiv \{X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k\} .$$

In this exercise, we prove this statement of the theorem in several steps.

(a) Solve for β_j^* in terms of properties of the joint distribution of $\left(Y, \tilde{X}_j\right)$?

(b) Show that $Cov\left[\tilde{X}_j, X_l\right] = 0$ for all $l = 1, \ldots j-1, j+1, \ldots, k$. (Hint: remember that $\tilde{X}_j$ is the "error term" from the best linear approximation of $E[X_j|X_{-j}]$.)

(c) Show that $Cov\left[\tilde{X}_j, X_j\right] = Var\left[\tilde{X}_j\right]$.

(d) Show that $Cov\left[\tilde{X}_j, U\right] = 0$. (Hint: what is $\tilde{X}_j$ a function of?)

(e) Using your answers from parts (a-d) show that $\beta_j = \beta_j^*$.

(7) **Sample Frisch-Waugh Theorem**. Suppose that $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ is a random sample from $(\mathbf{X}, Y)$ where $\mathbf{X} = \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$. Consider the linear regression model $Y = \mathbf{X}'\beta + U$. The i^{th} observation from the random sample follows the relationship $Y_i = \mathbf{X}_i'\beta + U_i$. Using the definitions in the notes, we can express the regression model in stacked form as:

$$\mathbf{Y} \quad = \quad \mathbb{X}\beta + \mathbf{U}$$

Note that we can partition the data matrix $\mathbb{X}$ by columns to distinguish two sets of regressors in our model as in

$$\mathbb{X} \quad = \quad \begin{bmatrix} \mathbb{X}_1 & \mathbb{X}_2 \end{bmatrix}$$

Suppose further that we decompose $\mathbb{X}$ by putting the column containing observations on the first regressor in the submatrix $\mathbb{X}_1$ and the rest of the columns (regressors and the

column of 1s) in $\mathbb{X}_2$. [20]

$$\mathbb{X} \;=\; \left[\begin{array}{cc} \mathbb{X}_1 & \mathbb{X}_2 \end{array}\right]$$

If we define the vector $\beta_{two} = \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{pmatrix}$, the stacked regression can be expressed as the partitioned regression specification:

$$\begin{aligned} \mathbf{Y} &= \left[\begin{array}{cc} \mathbb{X}_1 & \mathbb{X}_2 \end{array}\right] \left[\begin{array}{c} \beta_1 \\ \beta_{two} \end{array}\right] + \mathbf{U} \\ &= \mathbb{X}_1\beta_1 + \mathbb{X}_2\beta_{two} + \mathbf{U} \end{aligned}$$

(a) Premultiply the stacked regression by $\mathbb{X}' = \left[\begin{array}{c} \mathbb{X}_1' \\ \mathbb{X}_2' \end{array}\right]$ and use the sample moment conditions $\mathbb{X}'\mathbf{U} = \mathbf{0}$ to simplify the expression. At this point, your expression should be of the form:

$$\left[\begin{array}{cc} A & B \\ C & D \end{array}\right] \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_{two} \end{pmatrix} \;=\; \begin{pmatrix} E \\ F \end{pmatrix}$$

Do these sample moment conditions imply that $\mathbb{X}_1'\mathbf{U} = 0$ and $\mathbb{X}_2'\mathbf{U} = 0$? Explain precisely.

(b) *Tip for using less pencil*: Solve this problem using this A–F notation. Once you have obtained a solution for $\hat{\beta}_1$, substitute the expressions you found in (a) into your final expression.

By matrix multiplication, the expression can be reorganized as follows:

$$A\hat{\beta}_1 \;=\; E - B\hat{\beta}_{two} \tag{2.9.3}$$

$$D\hat{\beta}_{two} \;=\; F - C\hat{\beta}_1 \tag{2.9.4}$$

(i) Assuming the regression model has no collinearity, argue that A and D are invertible.[21] Apply these inverses to both sides of equations (2.9.3) and (2.9.4).

(ii) Substitute your solution for $\hat{\beta}_{two}$ into equation (2.9.3). Solve the resulting expression for $\hat{\beta}_1$.[22]

(c) Now, consider a different approach to obtain an estimator for β_1.

(i) Premultiply the partitioned regression specification by the residual maker matrix $M_2 = I - \mathbb{X}_2\left(\mathbb{X}_2'\mathbb{X}_2\right)^{-1}\mathbb{X}_2'$. For convenience, define $\mathbb{X}_1^* = M_2\mathbb{X}_1$, $\mathbf{Y}^* = M_2\mathbf{Y}$ and $\mathbf{U}^* = M_2\mathbf{U}$ and write down an expression for the transformed regression model. Explain the sense in which $\mathbb{X}_1^*$ is a vector of residuals (from what regression model?).

(ii) Solve for an estimator of β_1 using the transformed regression model. Use the definitions for $\mathbb{X}_1^*$ and $\mathbf{Y}^*$ to express the solution in this part in terms of the residual maker matrix and the untransformed vectors.

---

[20] Without loss of generality, we could reshuffle the columns any which way. Instead of the first regressor, we could place the j^{th} regressor in $\mathbb{X}_1$ instead. We could have more than one regressor in $\mathbb{X}_1$, but the intuition from the result we consider is stronger.

[21] Note: A in this example is a scalar, but to retain the generality of the problem, pretend it is a matrix.

[22] Alternatively, you could use the inverse of partitioned matrices theorem.

(8) For this problem, we will use the `edudat2.dta` data from the data website. Perform all calculations in this question with both R and Stata. Consider the following regression model:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \times exper + U$$

(a) Interpret this regression model as the best linear prediction of log wage, given *educ*, *exper*, and $educ \times exper$.
   (i) Interpret the coefficients in the coefficient vector β. Explain precisely how the interaction affects your interpretation.
   (ii) What does the regression model imply about $E[educ \times U]$?

(b) Estimate this regression model using the `edudat2.dta` data. Make sure you obtain the same answers in both packages.

(c) Use the Frisch-Waugh Theorem to motivate an alternative estimation routine for $\hat{\beta}_1$.
Step 1: Estimate the specification:

$$educ = \alpha_0 + \alpha_1 exper + \alpha_2 educ \times exper + V$$

and use your estimates to construct the residuals of that regression. Call the residuals $\tilde{educ}$.
Step 2: Estimate

$$\log(wage) = \beta_0 + \beta_1 \tilde{educ} + U^*$$

(d) Consider what would happen if you estimated the following regression model in Step 2. Don't run the regression yet. Think about it. Do you expect to obtain the same OLS regression estimates as part (a)? (In your writeup, don't answer this question about what you expect. Just do the pre-regression thought exercise)

$$\log(wage) = \beta_0 + \beta_1 \tilde{educ} + \beta_2 exper + \beta_3 educ \times exper + U$$

Estimate the regression using OLS. How does $\hat{\beta}_1$ from this regression compare to parts (a) and (b)? How do your estimates of $\hat{\beta}_2$ and $\hat{\beta}_3$ from this specification compare to part (a)? Explain precisely why these results are they way they are.

(9) Use R for the following illustration of the Sampling Distribution of $\hat{\beta}_1$. Submit your code as an appendix to this assignment.

(a) Create a population of 200,000 simulated individuals described by the population regression

$$Y = \beta_0 + \beta_1 X + U$$

where $X \sim 15 + 2 \times t(df_X)$ and $U = t(df_U)$. Write your code to allow the user to prespecify the values of the parameters. Namely, start your code specifying the following parameters:

$$\begin{aligned} df_X &= 5 \\ df_U &= 3 \\ \beta_0 &= 5 \\ \beta_1 &= 10 \end{aligned}$$

You may find it convenient to bind (X, Y) together in a data frame object using the `as.data.frame()` and `cbind()` commands. After constructing the population, σ_U (the standard deviation of the error term) for future reference.

(b) Verify that X and U depart significantly from a normal distribution, apply the `qqnorm()` function to the vectors X and U. This function plots the theoretical quantiles of a normal RV against the empirical quantiles of the vector in question. If the

data are normal, this should look like the 45 degree line. Comment on the shape of the population distribution.

(c) Organize the following steps using a `for` loop. Before starting your `for` loop, it will be useful to define two storage vectors `beta1stor` and `seb1stor` to contain the realized values of your sample statistics.[23] We will compute $B = 1000$ random samples of size $n = 150$ from this population.

(i) Using the `sample()` command, draw a sample without replacement of size $n = 150$ from the population you constructed in part (a). Does this sample technically qualify as an iid random sample? As a practical matter, does it?

(ii) Compute OLS estimates of the single regression model $Y = \beta_0 + \beta_1 X + U$ on each of these samples. Store $\hat{\beta}_1$ and $SE\left(\hat{\beta}_1\right)$ computed from each sample.[24] You may use the `lm()` function to estimate $\hat{\beta}_1$.

(iii) Report a histogram of your computed values for $\hat{\beta}_1$. Comment on the shape of this histogram and how this relates to the underlying population. Also, apply `qqnorm()` to your vector of computed $\hat{\beta}_1$ values.

(iv) Compute the standard deviation of $\hat{\beta}_1$. Compute the mean of $SE\left(\hat{\beta}_1\right)$. Comment on what this tells you.

(d) Perform the same calculations as in (c) for $n = 15$ instead of $n = 150$.

(10) Let $(Y_1, X_{1,1}, X_{1,2}), \ldots, (Y_n, X_{n,1}, X_{n,2})$ be an i.i.d. sample from (Y, X_1, X_2) satisfying

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \ ,$$

where $(1, X_1, X_2)$ is not perfectly collinear, $E[Y^4] < \infty$ and $E[X_j^4] < \infty$ for $1 \leq j \leq 2$. You the researcher wish to interpret this regression as the best linear predictor of Y given X_1 and X_2.

(a) Interpret U. Is it necessarily true that $E[U] = 0$? What about $E[X_1 U] = 0$ or $E[X_2 U] = 0$?

(b) Suppose the researcher estimates the equation

$$Y = \beta_0^* + \beta_1^* X_1 + U^*$$

by OLS. Show that

$$\hat{\beta}_1^* \xrightarrow{P} \beta_1 + \beta_2 \frac{\text{Cov}[X_1, X_2]}{\text{Var}[X_1]} \ .$$

(c) Under what conditions will it be true that $\hat{\beta}_1^*$ is consistent for β_1?

(d) Suppose you estimate the single regression using OLS. Is it necessarily true that

$$\sum_{i=1}^{n} X_{i,j} \hat{U}^* = 0$$

for $1 \leq j \leq 2$? What about

$$\sum_{i=1}^{n} \hat{U}_i^* = 0 \ ?$$

(e) Express part (d) in matrix notation. Explain your answer to (d) in terms of the column space of $\mathbb{X}$ and the null space of $\mathbb{X}'$.

[23] I recommend using the `rep()` command for this while setting each value in the storage vector to zero.

[24] Finding the standard error in the output is a little tricky. If you have estimated a model called model.lm, you can find the standard error of the slope coefficient estimate using the command `summary(model.lm)$coefficients[2,2]`

(11) Suppose you estimate the regression model

$$Y = \beta_0 + \beta_1 X + U$$

using OLS and obtain $\hat{\beta}_0$ and $\hat{\beta}_1$. For $a > 0$, define $Y^* \equiv aY$ and $X^* = aX$. Now, suppose you estimate the regression model

$$Y^* = \beta_0^* + \beta_1^* X^* + U^*$$

using OLS and obtain $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$.

(a) What is the relationship between $\hat{\beta}_1$ and $\hat{\beta}_1^*$? $\hat{\beta}_0$ and $\hat{\beta}_0^*$?

(b) What is the relationship between $\hat{\sigma}^2_{\hat{\beta}_1}$ in the two regressions?

(c) What is the relationship between R^2 in the two regressions?

(12) Use the `edudat2.dta` data file from the website for this question. Use both R and Stata for this question. Report your estimates in a nicely-formatted table, but also include an appendix with any relevant code and output.

(a) After reading the data into your statistical packages, produce summary statistics on each variable (e.g., for quantitative variables, mean, median, minimum, maximum and standard deviation) and report these in a nicely-formatted table.

(b) Compute the mean wage for male employees and the mean wage for female employees.

(c) Estimate the simple regression model

$$wage = \beta_0 + \beta_1 gender + U_1$$

using OLS. How do the estimates from this regression compare with the summary measures you computed in the previous parts of the problem. Based on these regression results, do you find evidence that females face discrimination in the labor market? Explain precisely.

(d) How would including *educ* and *pareduc* in the model affect your interpretation of β_1? Do you expect the coefficient estimate $\hat{\beta}_1$ to be the same in a simple regression as it is in a multiple regression?

(e) Estimate the multiple regression model

$$wage = \beta_0 + \beta_1 gender + \beta_2 educ + \beta_3 pareduc + U_2$$

using OLS. How does your estimate of β_1 from this multiple regression model compare with the estimate from the simple regression model? In light of the multiple regression results, reevaluate your answer regarding discrimination of females in the labor force from part (c).

(f) Consider also including an interaction $gender \times educ$ in the regression model. Explain precisely how this interaction changes the interpretation of β_2. Does the interaction change the interpretation of β_1?

(g) Estimate the multiple regression model

$$wage = \beta_0 + \beta_1 gender + \beta_2 educ + \beta_3 pareduc + \beta_4 gender \times educ + U_2$$

using OLS. In light of the regression results, reevaluate your answer regarding discrimination of females in the labor force from part (d).

(13) **Blogging Econometrics**. Economist Justin Wolfers recently wrote a blog post at Freakonomics about a striking scatterplot produced by another economist John Taylor on his own blog. Read Taylor's original post,[25] Wolfers' first response,[26] and Wolfers' second

[25] http://johnbtaylorsblog.blogspot.com/2011/01/higher-investment-best-way-to-reduce.html

[26] http://www.freakonomics.com/2011/03/30/how-to-spot-advocacy-science-john-taylor-edition/

response.[27] This question will guide you through some calculations that will help you evaluate the evidence presented by both sides.

(a) The data set `inv.dta` contains data on the seasonally-adjusted unemployment rate `unrate`, the ratio of fixed private investment to gdp `invtogdp` and the timing of these events, denoted both by the quarter-year `DATE` and a factor that indicates the decade of the observation `dec`. Use this data set to reproduce Taylor's original scatterplot and Wolfers' extended scatterplot. [28]

(b) Estimate the regression model

$$unrate = \beta_0 + \beta_1 invtogdp + U_2$$

using OLS on both the Taylor sample and Wolfers (full) sample. Comment on the relationship between the two regressions.

We will use the full sample for the rest of this problem (unless told to do otherwise).

(c) Estimate the regression model

$$unrate = \sum_{i=1}^{6} I_{\{dec=i\}}\alpha_i + \beta_1 invtogdp + U_2$$

using OLS, where $I_{\{dec=i\}}$ is a dummy variable that equals one if an observation is in decade i and zero otherwise. In this specification, what happens to the estimate for β_1 relative to the simple regression specification? Interpret.

(d) Estimate the simple regression model from (b) separately on each of the six decades. For each of these separate regressions, report $\hat{\beta}_1$ and the p-value for the following hypothesis test:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &< 0 \end{aligned}$$

Relative to part (c), what do you conclude about the statistical relationship of *unrate* and *invtogdp*?

(e) OLS estimation has nice properties when the data are drawn independently and identically from the joint distribution of the regressors and the response variable (that is, when we have a random sample). Is this assumption satisfied in this setting?

(14) Use R for the following illustration of Omitted Variable Bias. Submit your code in an appendix to the assignment.

(a) Construct a sample of 500 simulated individuals with the following random variables.

(i) Prior to running any other commands, use the command `set.seed(20900)`.

(ii) $X_1 = t(5)$, $X_3 \sim 6 + N(0,1)$ and $X_4 \sim 4 + 4 \times t(15)$

(iii) $X_2 = 3 + 5X_1 + U_1$ where $U_1 \sim N\left(0, \sigma^2 = 9\right)$.

(iv) **Population regression relationship**. The response variable equals:

$$Y = 10 + 2X_1 + 7X_2 + 17X_3 + W$$

where $W \sim N\left(0, \sigma_W^2 = 100\right)$.

(v) Given these random variables, create a `data.frame` object by binding (Y, X_1, X_2, X_3, X_4) together into a single data frame. Use `cbind()` and `as.data.frame()`.

(vi) Run the `summary()` command on the data frame you constructed and comment *briefly* on how the summary statistics from this simulated population match with the parameter values you used.

[27] http://www.freakonomics.com/2011/04/01/graph-fight-more-on-taylor%E2%80%99s-scatterplot/

[28] Note: The Taylor sample includes all observations from decades 5 and 6.

(b) Use OLS regression to estimate the single linear regression model

$$Y = \beta_0 + \beta_1 X_1 + U$$

Report your estimates in the first column of a nicely-formatted table. Is your estimate for β_1 within a standard error of the true population value you input into this exercise in part (a)?

(c) Use OLS regression to estimate $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$.
   (i) Report your estimates in the second column of a nicely-formatted table.
   (ii) Is your estimate for β_1 from this specification within a standard error of the true population value you input into this exercise in part (a)?
   (iii) Use the omitted variables bias formula to reconcile your estimate of β_1 from (b) with your estimate from this part.

(d) Suppose that you have an imprecise measurement for X_2 given by $X_2^m = X_2 + M$ where $M \sim N(0, 1)$. Construct a variable X_2^m and append it to your data set. With this mismeasured variable, estimate the regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^m + U$.
   (i) Report your estimates in the third column of a nicely-formatted table.
   (ii) Is your estimate for β_1 from this specification within a standard error of the true population value you input into this exercise in part (a)?

(15) **Reverse Causality?** Revisit the Taylor scatter plot data. These data are available on the website in a file named `taylor.csv`. Note: *govtogdp*1 is the variable in the data set that corresponds to the one-period time lag $govtogdp_{t-1}$. The other additional lagged variables in the file are defined analogously.

(a) Use OLS to estimate the regression model for the t^{th} observation on *govtogdp*

$$govtogdp_t = \beta_0 + \beta_1 govtogdp_{t-1} + \beta_2 unrate_{t-1} + U_t$$

on the subsample of observations spanning from 1990Q1 to the end of the data set. Conduct the hypothesis test that $\beta_2 = 0$ against the two-sided alternative hypothesis. Does lagged unemployment have predictive power above and beyond the first-order autoregression of *govtogdp* on itself?

(b) Use OLS to estimate the regression model for the t^{th} observation on *unrate*

$$unrate_t = \alpha_0 + \alpha_1 govtogdp_{t-1} + \alpha_2 unrate_{t-1} + U_t$$

on the subsample of observations spanning from 1990Q1 to the end of the data set. Conduct the hypothesis test that $\alpha_1 = 0$ against the two-sided alternative hypothesis. Does lagged government purchases / GDP have predictive power above and beyond the first-order autoregression of *unrate* on itself?

(c) The hypothesis tests in (a) and (b) are simple tests for what is known as **Granger causality**, after Clive Granger a proponent of the above testing procedure. A time series X_t *Granger causes* another time series Y_t if it has predictive power above and beyond an autoregression of Y_t on itself of order p (in (a) and (b), $p = 1$). Formally, within the regression model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ...\beta_p Y_{t-p} + \alpha_1 X_1 + ... + \alpha_p X_p + V_t$$

we say that X_t Granger causes Y_t if the α_i are not all zero. In the data file you were given, there are lagged variables up to four lags of *unrate* and *govtogdp*. Use this information to formally test (i) whether *unrate* Granger causes *govtogdp* using lags of up to order $p = 4$ and (ii) whether *govtogdp* Granger causes *unrate* using lags of up to order $p = 4$. What do you conclude about "which series comes first"? Does this mean that increases in government spending cause unemployment to rise?

(d) Extend the analysis you performed in the previous part to the entire data set. In particular, determine whether *unrate* Granger causes *govtogdp* or *govtogdp* Granger

causes *unrate* (or both). From this analysis, what do you learn about the relationship between *unrate* and *govtogdp*?

CHAPTER 3

Going Beyond OLS

In this chapter, we consider some classical violations to the assumptions behind OLS regression. Our discussion will lead us to three classes of solutions: standard error corrections, Generalized Least Squares (GLS) and Maximum Likelihood Estimation (MLE). In the last section of these notes, we discuss the application of these techniques to estimating nonlinear models.

3.1. Extending OLS Regression

Suppose our population regression model is $Y = \mathbf{X}'\beta + U$ and we have a sample from this population that we can organize into the stacked form of the regression model:

$$\mathbf{Y} = \mathbb{X}\beta + \mathbf{U}$$

where $\mathbf{Y}$ is a $n \times 1$ vector of observations on the response variable, $\mathbb{X}$ is the data matrix of the explanatory variables with a column of ones in the first column, and $\mathbf{U}$ is a $n \times 1$ vector of error terms.

Recall that the OLS estimator of β is

$$\hat{\beta} = \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{Y}$$

which we showed has an asymptotically normal distribution given by:

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \Sigma\right)$$

where

$$\Sigma = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} Var\left[\mathbf{X}U\right] E\left[\mathbf{X}\mathbf{X}'\right]^{-1} \tag{3.1.1}$$

To get this form of the asymptotic variance-covariance matrix, we made an assumption that our sample was drawn iid from the population regression relationship.

FACT 3.1.1. *We can relax this assumption and still obtain asymptotic normality, but with an asymptotic covariance matrix of the form:*

$$\Sigma = E\left[\mathbf{X}\mathbf{X}'\right]^{-1} W E\left[\mathbf{X}\mathbf{X}'\right]^{-1} \tag{3.1.2}$$

where V is a positive definite matrix that satisfies: $\hat{W} = \frac{1}{n}\mathbb{X}'\Omega\mathbb{X} \xrightarrow{P} W$ where $\Omega = Var\left[\mathbf{U}|\mathbb{X}\right] = E\left[\mathbf{U}\mathbf{U}'|\mathbb{X}\right]$.

We can motivate this fact by computing the small sample variance of $\hat{\beta}^{OLS}$ conditional on the data $\mathbb{X}$:

$$\begin{aligned} Var\left[\hat{\beta}^{OLS}|\mathbb{X}\right] &= E\left[\left(\hat{\beta}^{OLS}-\beta\right)\left(\hat{\beta}^{OLS}-\beta\right)|\mathbb{X}\right] \\ &= E\left[\left(\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{U}\right)\left(\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{U}\right)'|\mathbb{X}\right] \\ &= \frac{1}{n}\left[\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}\underbrace{\frac{1}{n}\mathbb{X}'Var\left[\mathbf{U}|\mathbb{X}\right]\mathbb{X}}_{\hat{W}}\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}\right] \end{aligned}$$

Then, the probability limit of $Var\left[\sqrt{n}\hat{\beta}^{OLS}|\mathbb{X}\right]$ is equal to the probability limit of the term in $[\cdot]$ brackets.

This form of the Ω matrix allows us to analyze a variety of error processes. To see why this is the case, multiply out Ω:

$$\Omega = \begin{bmatrix} E\left[U_1^2|\mathbb{X}\right] & E\left[U_1U_2|\mathbb{X}\right] & \dots & E\left[U_1U_n|\mathbb{X}\right] \\ E\left[U_1U_2|\mathbb{X}\right] & E\left[U_2^2|\mathbb{X}\right] & \dots & E\left[U_2U_n|\mathbb{X}\right] \\ \vdots & \vdots & \ddots & \vdots \\ E\left[U_1U_n|\mathbb{X}\right] & E\left[U_2U_n|\mathbb{X}\right] & \dots & E\left[U_n^2|\mathbb{X}\right] \end{bmatrix}$$

DEFINITION 3.1.2. With correlation of the error terms across observations, the off-diagonal elements of this matrix will be non-zero. Correlation across observations is called **serial correlation.**

Note that serial correlation is a violation of the assumption that the sample is an iid random sample. With an iid random sample, Ω is a diagonal matrix.

DEFINITION 3.1.3. If the diagonal elements of Ω vary systematically with $\mathbb{X}$, we call this condition **heteroskedasticity**. Homoskedastic errors imply $\Omega = \sigma^2 I$.

We have already seen and analyzed the problem of heteroskedasticity, but discussing it again in this context will provide a useful bridge toward understanding solutions to the serial correlation problem. An error term that is heteroskedastic or serial correlated is called **nonspherical**.

FACT 3.1.4. *As long as the orthogonality conditions hold $E\left[\mathbf{X}U\right] = 0$, the OLS estimator $\hat{\beta}^{OLS} \xrightarrow{P} \beta$. As long as mean independence is satisfied $E\left[U|\mathbf{X}\right] = 0$, $E\left[\hat{\beta}^{OLS}\right] = \beta$. These properties regarding our estimator are true regardless of the form of Ω. That is, nonspherical errors are an issue for the validity of standard errors (which is important), but they are not an issue for the validity of the estimator itself.*

3.1.1. Heteroskedasticity. Our assumption that $\frac{1}{n}\mathbb{X}'\Omega\mathbb{X} \xrightarrow{P} W$ immediately gives us a consistent estimator for the inside term of Σ. Using the continuous mapping theorem,

$$\hat{\Sigma} = \left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}\left(\frac{1}{n}\mathbb{X}'\Omega\mathbb{X}\right)\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}$$

is a consistent estimator for Σ.

Let's explore this expression in more detail. Use the formula for Ω and the fact that $\mathbb{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_k' \end{pmatrix}$ to write

$$\frac{1}{n}\mathbb{X}'\Omega\mathbb{X} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{X}_j E\left[U_i U_j|\mathbf{X}_i,\mathbf{X}_j\right]\mathbf{X}_i'$$

We can break this expression into two pieces:

$$\begin{aligned}\frac{1}{n}\mathbb{X}'\Omega\mathbb{X} &= \frac{1}{n}\sum_{i=1}^{n}E\left[U_i^2|\mathbf{X}_i\right]\mathbf{X}_i\mathbf{X}_i' + \\ &\quad \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}E\left[U_i U_j|\mathbf{X}_i,\mathbf{X}_j\right]\left[\mathbf{X}_j\mathbf{X}_{j-i}' + \mathbf{X}_{j-i}\mathbf{X}_j'\right]\end{aligned}$$

If the only problem is heteroskedasticity, we can obtain a consistent estimator of W by using the residuals as in the following expression:

$$\hat{W} = \frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2\mathbf{X}_i\mathbf{X}_i'$$

This is known as the **White form** of $\hat{W}$. Use this expression for $\hat{W}$ in

$$\hat{\Sigma} = \left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}\hat{W}\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}$$

to obtain (White) heteroskedasticity-robust standard errors. Angrist and Pischke discuss several alternative estimators for W that have better small sample properties. You should read their discussion in Chapter 8 on these alternative estimators (called HC1, HC2 and HC3) for W. The main takeaway from this discussion is that HC3 is the most conservative estimator.

3.1.2. Serial Correlation. To obtain a consistent estimator for the variance-covariance matrix in the presence of serial correlation, we can use the Newey-West estimator for W.

$$\hat{W}_{NW} = \frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2\mathbf{X}_i\mathbf{X}_i' + \frac{1}{n}\sum_{i=1}^{m}\sum_{j=i+1}^{n}\omega_i\hat{U}_j\hat{U}_{j-i}\left[\mathbf{X}_j\mathbf{X}_{j-i}' + \mathbf{X}_{j-i}\mathbf{X}_j'\right]$$

where m is the number of lags (user specified) and $\omega_i = 1 - \frac{i}{n+1}$ decreases in i (the lag number) to put less weight on the observations with higher lags.

REMARK 3.1.5. In this setting, we call robust estimation methods (like White and Newey-West corrections) called Heteroskedasticity and Autocorrelation Consistent (HAC) methods. **HAC standard error corrections** continue to use the OLS estimator for β, but this is a less efficient (higher variance) method for dealing with departures from the standard assumption that $Var\left[\mathbf{U}|\mathbb{X}\right] = \sigma^2 I$. Recall that the OLS estimator is BLUE as long as the error term is spherical. As it is not BLUE in this more general setting, we may be interested in an alternative to OLS. With enough knowledge about the error term, it turns out that this alternative (generalized least squares, GLS) is BLUE. We turn to a discussion of GLS now.

3.2. Generalized Least Squares

Consider the regression model in stacked form as in $\mathbf{Y} = \mathbb{X}\beta + \mathbf{U}$ where $Var[\mathbf{U}|\mathbb{X}] = \Omega$ as in the previous section. For now, suppose that the matrix Ω is known.

Because Ω is a positive definite symmetric matrix, we can decompose it as $\Omega = V\Lambda V^{-1}$ with the square root factorization $\Omega^{\frac{1}{2}} = V\Lambda^{\frac{1}{2}}V^{-1}$. $\Omega^{\frac{1}{2}}$ is itself a positive definite matrix. Hence, this square root matrix has an inverse $\Omega^{-\frac{1}{2}}$. Premultiply the regression model in stacked form by this inverse $\Omega^{-\frac{1}{2}}$ to obtain a transformed stacked regression.

$$\begin{aligned} \underbrace{\Omega^{-\frac{1}{2}}\mathbf{Y}}_{\tilde{\mathbf{Y}}} &= \underbrace{\Omega^{-\frac{1}{2}}\mathbb{X}}_{\tilde{\mathbb{X}}}\beta + \underbrace{\Omega^{-\frac{1}{2}}\mathbf{U}}_{\tilde{\mathbf{U}}} \\ \tilde{\mathbf{Y}} &= \tilde{\mathbb{X}}\beta + \tilde{\mathbf{U}} \end{aligned}$$

As with OLS, our estimator will satisfy the sample moment conditions. This time, however, the sample moment conditions are defined in terms of the transformed regression.

$$\begin{aligned} \tilde{\mathbb{X}}'\tilde{\mathbf{U}} &= \mathbf{0} \\ \text{or} & \\ \mathbb{X}'\Omega^{-1}\mathbf{U} &= \mathbf{0} \end{aligned}$$

These sample moment conditions are called the generalized least squares (GLS) sample moment conditions and we can use them to solve for the GLS estimator of β as follows. First, premultiply the transformed regression by $\tilde{\mathbb{X}}'$. Then, impose the sample moment conditions.

$$\begin{aligned} \tilde{\mathbb{X}}'\tilde{\mathbf{Y}} &= \tilde{\mathbb{X}}'\tilde{\mathbb{X}}\beta + \underbrace{\tilde{\mathbb{X}}'\tilde{\mathbf{U}}}_{=\mathbf{0}} \\ \Rightarrow \hat{\beta}^{GLS} &= \left(\mathbb{X}'\Omega^{-1}\mathbb{X}\right)^{-1}\mathbb{X}'\Omega^{-1}\mathbf{Y} \end{aligned}$$

Exercise 3.2.1. Show that

$$\hat{\beta}^{GLS} = \arg\min_{b} \left(\mathbf{Y} - \mathbb{X}b\right)'\Omega^{-1}\left(\mathbf{Y} - \mathbb{X}b\right)$$

where $b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}$.

(1) Why is the estimator called the generalized least squares estimator of β? In what sense does GLS generalize OLS?
(2) Suppose that $\Gamma = c\Omega$ and consider the estimator $\tilde{\beta} = \left(\mathbb{X}'\Gamma^{-1}\mathbb{X}\right)^{-1}\mathbb{X}'\Gamma^{-1}\mathbf{Y}$. How does $\tilde{\beta}$ relate to $\hat{\beta}^{GLS}$?

Theorem 3.2.2. *Under the assumption that U is mean independent of $\mathbf{X}$, the GLS estimator $\hat{\beta}^{GLS}$ is the Best Linear Unbiased Estimator for β among the class of estimators that are a linear function of $\mathbf{Y}$.*

Proof. Sketch. The transformed model satisfies the assumptions of the Gauss-Markov Theorem for OLS. Namely, the error term $\tilde{\mathbf{U}}$ is homoskedastic, $Var\left[\tilde{\mathbf{U}}\tilde{\mathbf{U}}'\right] = I$. Hence, apply Gauss-Markov for OLS with homoskedastic errors to this setting and conclude that $\hat{\beta}^{GLS}$ is BLUE for β among all estimators that are linear in $\tilde{\mathbf{Y}}$. Suppose that $\hat{\beta}^{alt}$ is an arbitrary

$$\hat{\beta}^{alt} = A\mathbf{Y} = A\Omega^{\frac{1}{2}}\Omega^{-\frac{1}{2}}\mathbf{Y} = \tilde{A}\tilde{\mathbf{Y}}$$

That is, any estimator that is linear in $\tilde{\mathbf{Y}}$ is also linear in $\mathbf{Y}$. Hence, $\hat{\beta}^{GLS}$ is BLUE among all estimators that are linear in $\mathbf{Y}$. Two facts remain to be shown: $\hat{\beta}^{GLS}$ is unbiased for β and $\hat{\beta}^{alt}$ is also unbiased for β for arbitrary A (and is therefore the appropriate comparison for $\hat{\beta}^{GLS}$). These facts are left as an exercise. □

The next theorem establishes the asymptotic normality of the GLS estimator. We prove this analogously to how we proved it in the case of OLS. Start by expressing the GLS estimator in proof form:

$$\begin{aligned}\hat{\beta}^{GLS} &= \left(\mathbb{X}'\Omega^{-1}\mathbb{X}\right)^{-1}\mathbb{X}'\Omega^{-1}\underbrace{\left(\mathbb{X}\beta+\mathbf{U}\right)}_{\mathbf{Y}} \\ &= \beta+\left(\mathbb{X}'\Omega^{-1}\mathbb{X}\right)^{-1}\mathbb{X}'\Omega^{-1}\mathbf{U}\end{aligned}$$

Note: We can express this proof form as $\hat{\beta}^{GLS}=\beta+\left(\tilde{\mathbb{X}}'\tilde{\mathbb{X}}\right)^{-1}\tilde{\mathbb{X}}'\tilde{\mathbf{U}}$.

THEOREM 3.2.3. *Asymptotic Normality of the GLS estimator.* $\sqrt{n}\left(\hat{\beta}^{GLS}-\beta\right)\xrightarrow{d}N\left(\mathbf{0},Q^{-1}\right)$ *where* $\frac{1}{n}\mathbb{X}'\Omega^{-1}\mathbb{X}\xrightarrow{P}Q$.

PROOF. Using the proof form of GLS, we obtain the expression

$$\sqrt{n}\left(\hat{\beta}^{GLS}-\beta\right) = \left(\frac{1}{n}\tilde{\mathbb{X}}'\tilde{\mathbb{X}}\right)^{-1}\left[\sqrt{n}\frac{1}{n}\tilde{\mathbb{X}}'\tilde{\mathbf{U}}\right]$$

Assuming that it exists, let $\frac{1}{n}\tilde{\mathbb{X}}'\tilde{\mathbb{X}}\xrightarrow{P}Q$ denote the probability limit of $\frac{1}{n}\mathbb{X}'\Omega^{-1}\mathbb{X}$. By the multivariate CLT the second term converges in distribution to $N\left(\mathbf{0},Q\right)$, where Q is the probability limit of $\frac{1}{n}\mathbb{X}'\Omega^{-1}Var\left[\mathbf{U}\right]\Omega^{-1}\mathbb{X}=\frac{1}{n}\mathbb{X}'\Omega^{-1}\mathbb{X}$. Putting these two statements together, by Slutsky, we have the result

$$\sqrt{n}\left(\hat{\beta}^{GLS}-\beta\right)\xrightarrow{d}N\left(\mathbf{0},Q^{-1}\right)$$

□

From this result, we obtain an approximating distribution for hypothesis tests when we use the GLS estimator. Namely, if we know Ω, we can conduct hypothesis tests using the approximate normality of GLS and plugging in the approximation $\hat{Q}=\left(\mathbb{X}'\Omega^{-1}\mathbb{X}\right)$

$$\hat{\beta}^{GLS}\approx N\left(\beta,\frac{1}{n}\left(\mathbb{X}'\Omega^{-1}\mathbb{X}\right)^{-1}\right)$$

3.2.1. Weighted Least Squares: An Application of GLS. An application of GLS is weighted least squares. Suppose that we are in a setting where there is heteroskedasticity (but no serial correlation), and we know the form of heteroskedasticity depends on some observable variable Z by some known function $\sigma_i^2=\sigma^2 f\left(Z_i\right)$. In this setting, we actually know Ω and can compute it using the formula $\Omega=\sigma_i^2 I=\sigma^2 f\left(Z_i\right)I$.

In this setting, we can premultiply the stacked regression model by $\Omega^{-\frac{1}{2}}$

$$\Omega^{-\frac{1}{2}}\mathbf{Y} = \Omega^{-\frac{1}{2}}\mathbb{X}\beta+\Omega^{-\frac{1}{2}}\mathbf{U}$$

to obtain the transformed regression model. Estimating the transformed regression model amounts to weighting the i^{th} observation by the known factor $\frac{1}{\sqrt{f(Z_i)}}$. That is to say that we can estimate the σ^2 term as we normally would after reweighting the regression.

EXAMPLE 3.2.4. WLS is an appropriate technique when the response variable is an average. For example, suppose that we are interested in the effect of teacher training hours in the past year *teachtrain* on the student performance as measured by the classroom average on a standardized math test score, *mathavg*. To this end, we would like to estimate

$$mathavg = \beta_0 + \beta_1 teachtrain + U$$

Under the assumption of random sampling, this *mathavg* variable is an average and (accordingly) has a variance equal to $\frac{\sigma^2}{N}$ where N is the teacher's class size. Conditional on a level of teacher training, we can say that $Var\left[mathavg|teachtrain\right] = \frac{\sigma^2}{N}$.

Suppose that *teachtrain* is randomly assigned,[1] but that teachers have different class sizes (e.g., ranging from $N = 20$ to $N = 300$). With the assumptions we made along the way, we can run weighted least squares to obtain an estimator $\hat{\beta}^{WLS}$ by estimating the regression model using OLS.

$$\sqrt{N}mathavg = \beta_0 + \beta_1\sqrt{N}teachtrain + \sqrt{N}U$$

This procedure produces estimates and standard errors that are equivalent to GLS for this setting. We call the resulting estimator the weighted least squares estimator $\hat{\beta}^{WLS}$. Compared with OLS, this weighting scheme has the advantage in that $\hat{\beta}_1^{WLS}$ is BLUE for β_1.

As long the assumptions about the form of heteroskedasticity are correct, WLS is superior to OLS with a standard error correction. After all, WLS is the generalized least squares estimator, which is BLUE for β. If we use the wrong weights, however, WLS may provide incorrect estimates $\hat{\beta}^{WLS}$ in addition to the wrong standard errors. We can minimize the risk of using the wrong weights by constructing an estimate for Ω. This leads to a discussion of feasible GLS.

3.2.2. Feasible GLS. This discussion of attractive properties of the GLS estimator assumes that we know precisely the form of the error process Ω. In practice, we do not know Ω. For applications where we know enough to estimate Ω reliably with a consistent estimator $\hat{\Omega}$, an attractive alternative to GLS is Feasible GLS (FGLS) where we substitute our consistent estimator $\hat{\Omega}$ for the unknown Ω in $\hat{\beta}^{GLS}$ to obtain:

$$\hat{\beta}^{FGLS} = \left(\mathbb{X}'\hat{\Omega}^{-1}\mathbb{X}\right)^{-1}\mathbb{X}'\hat{\Omega}^{-1}\mathbf{Y}$$

In practice, the FGLS procedure is a two-step (or sometimes iterative) procedure where

(1) First, we obtain $\hat{\beta}^{OLS}$ from the multiple regression model and constructing residuals $\hat{U}_i$. Based on these residuals, we construct an estimate for Ω, usually imposing some parametric structure.

[1]We are also implicitly assuming that the teacher training has the same effect on student performance regardless of the size of the class. That is, we do not need to control for class size directly (it only affects the variance). In practice, class size might affect how effective the teacher training is in raising class scores (and we would want to account for that in our specification). As this is a different issue than the issue of weighting (and I would like to focus on the discussion on weighting), we assume that this interaction between class size and effectiveness of teacher training is not present.

(2) Second, we estimate $\hat{\beta}^{FGLS}$ by running OLS of $\tilde{Y} = \hat{\Omega}^{-\frac{1}{2}}\mathbf{Y}$ on $\tilde{\mathbb{X}} = \hat{\Omega}^{-\frac{1}{2}}\mathbb{X}$. This two-step procedure is sufficient to generate a consistent estimator for β, but some practitioners (and packages) iterate on this two-step procedure until convergence.

FGLS estimation has intuitive appeal in that it emulates the form of the BLUE for β, $\hat{\beta}^{GLS}$, but using an imprecise estimator for Ω to construct $\hat{\beta}^{FGLS}$ will impart imprecision on the FGLS estimator. For this reason, there is a tradeoff to using FGLS versus robust OLS methods.

As a general rule, as long as serial correlation is a problem in the setting (Ω is not a diagonal matrix), FGLS estimators based on a reasonable assumed parametric structure of Ω will be an improvement over robust OLS, which does not adjust the estimator $\hat{\beta}$ for potential non-independence among the observations.

3.2.2.1. *Structuring* Ω. A commonly-used structure of dependence of the error terms is that the error term is an **autoregressive process of order one**, abbreviated AR(1). That is, we assume that the dependence of observation t and $t+1$ is given by the autoregressive equation:

$$U_t = \rho U_{t-1} + \xi_t$$

where $\xi_t \sim$iid $\left(0, \sigma_u^2\right)$ across time. If this is the case, $Cov\left[U_t, U_{t-1}\right] = \rho Var\left[U_{t-1}\right]$. For the l^{th} lag, the covariance is $Cov\left[U_t, U_{t-l}\right] = \rho^l Var\left[U_{t-l}\right]$. Assuming that $Var\left[U_t\right] = \sigma^2$, we can use this correlation structure to completely fill out the Ω matrix for an AR(1) process with autocorrelation parameter $\rho < 1$:

$$\Omega = \frac{\sigma_u^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{bmatrix}$$

It turns out that the inverse of Ω can be expressed as

$$\Omega^{-1} = \frac{1}{\sigma_u^2}\begin{bmatrix} 1 & -\rho & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & \vdots & \vdots \\ 0 & -\rho & \ddots & -\rho & 0 \\ \vdots & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & \dots & 0 & -\rho & 1 \end{bmatrix}$$

and as $\Omega^{-1} = P'P$ where

$$P = \frac{1}{\sigma_u}\begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 \\ -\rho & 1 & 0 & \vdots & \vdots \\ 0 & -\rho & \ddots & 0 & 0 \\ \vdots & 0 & -\rho & 1 & 0 \\ 0 & \dots & 0 & -\rho & 1 \end{bmatrix}$$

Exercise 3.2.5. Fill in the details of this previous derivation and write out the GLS estimator.

(1) Explain why $Var\left[U_t\right] = \sigma^2 = \frac{\sigma_u^2}{1-\rho^2}$ (Hint: $\sum_{i=0}^{\infty}\rho^{2i} = \frac{1}{1-\rho^2}$ is a useful result about convergence of geometric series.)
(2) Show by matrix multiplication that $\Omega^{-1}\Omega = I$ and $P'P = \Omega^{-1}$.

(3) Use these expressions to write out the GLS estimator in terms of P rather than Ω as it is usually done.
(4) Show that the FGLS estimator has the form $\hat{\beta}^{GLS} = \left(\left(\mathbb{X}^*\right)'\mathbb{X}^*\right)^{-1}\left(\mathbb{X}^*\right)'\mathbf{Y}^*$ where

$$\begin{aligned}\mathbf{Y}^* &= \begin{pmatrix} rY_1 \\ sY_2 - tY_1 \\ \vdots \\ sY_n - tY_{n-1} \end{pmatrix} \\ \mathbb{X}^* &= \begin{pmatrix} r\mathbf{X}_1' \\ s\mathbf{X}_2' - t\mathbf{X}_1' \\ \vdots \\ s\mathbf{X}_n' - t\mathbf{X}_{n-1}' \end{pmatrix}\end{aligned}$$

and give expressions for the constants r, s and t.
(5) Give a procedure to compute $\hat{\beta}^{FGLS}$ in this setting.

Another type of error process that is easy to specify is a **compound symmetric** error structure. Impose a group structure on the data. That is, U_{gi} denotes the error term for the i^{th} observation in group g. Then, $Cov\left[U_{gi}, U_{gj}\right] = \sigma^2_{within}$ for $i \neq j$ while $Var\left[U_{gi}\right] = \sigma^2$, but for group $h \neq g$, $Cov\left[U_{hi}, U_{gj}\right] = 0$. That is, observations have serial correlation within groups, but not across groups. If we organize $\mathbf{U}$ by group, this leads to Ω of the following form:

$$\Omega = \sigma^2 I + \sigma^2_{within}\begin{bmatrix} 1_{g_1} & 0 & \dots & 0 \\ \vdots & 1_{g_2} & & \vdots \\ & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1_{g_k} \end{bmatrix}$$

where 1_{g_l} denotes a square matrix of ones with dimension equal to the number of group members in group g_l.

R has a function in the `nlme` library called `gls()` that makes it straightforward to implement FGLS estimation for a variety of common correlation structures.

3.3. Maximum Likelihood Estimation

Maximum likelihood estimation ties into this discussion of correcting for non-spherical errors (i.e., $\Omega \neq \sigma^2 I$) because it is one method of estimation we can use to get around the non-independence. We introduce MLE for another reason. Aside from giving us a framework to model the correlation structure, MLE allows us to estimate a variety of nonlinear regression models that have important econometric applications.

3.3.1. Some MLE Theory. For grounding our intuition, assume that the observations are drawn iid from a population with pdf parametrized by a parameter vector $f\left(x;\theta\right)$. Let $\mathbf{X}$ denote the vector of observations from the random sample. Given this notation, the joint distribution of the random sample can be written as

$$f\left(\mathbf{X};\theta\right) = \prod_{i=1}^{n} f\left(X_i;\theta\right)$$

In the regression notes, we have already seen that this joint distribution of the random sample is called the likelihood function when viewed as a function of θ: $L(\theta;\mathbf{X}) = f(\mathbf{X};\theta)$. Taking the natural log of both sides, we can express the log likelihood function as

$$\mathcal{L}(\theta;\mathbf{X}) = \sum_{i=1}^{n} \log f(X_i;\theta)$$

The maximum likelihood estimator of θ maximizes this log-likelihood (or the likelihood) as well as $\frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta)$.

$$\hat{\theta}^{MLE} = \arg\max_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log f(X_i;\theta) \right\}$$

EXERCISE 3.3.1. Describe the conditions under which $\frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta) \xrightarrow{P} E[\log f(X;\theta)]$. Justify.

We could write these formulas in terms of a random sample from a random vector $(\mathbf{Y},\mathbb{X})$ as in regression. In this case, the log-likelihood is given by:

$$\mathcal{L}(\theta;\mathbf{Y},\mathbb{X}) = \sum_{i=1}^{n} \log f(Y_i, X_{1i}, ..., X_{ki};\theta)$$

Suppose that θ does not affect the joint distribution of $\{X_j\}_{j=1}^{k}$. That is,

$$f(x_1, x_2, ..., x_k;\theta_j) = f(x_1, x_2, ..., x_k;\theta_i) = f(x_1, ..., x_k)$$

for any θ_i and θ_j in the parameter space Θ. Then, the log-likelihood (as a function of the random sample) can be written as:

$$\mathcal{L}(\theta;\mathbf{Y},\mathbb{X}) = \mathcal{L}(\theta;\mathbf{Y}|\mathbb{X}) + \sum_{i=1}^{n} \log f(X_{1i}, ..., X_{ki})$$

where $\mathcal{L}(\theta;\mathbf{Y}|\mathbb{X}) = \sum_{i=1}^{n} \log f(Y_i|X_{1i}, ..., X_{ki};\theta)$ is called the conditional log-likelihood function. Often, it will be more natural to specify a conditional log-likelihood function than it will be to express the likelihood of the sample. This is especially true because we are primarily interested in relationships among variables in econometrics.

EXERCISE 3.3.2. Fill in the details from this section of the notes. For $\tilde{\theta}$ that maximizes the log-conditional likelihood, show that $\tilde{\theta} = \hat{\theta}^{MLE}$ as long as θ does not affect the joint distribution of $\{X_j\}_{j=1}^{k}$.

3.3.1.1. *Score Function.* Consider the limit of the expression in Exercise 3.3.1 $E[\log f(X;\theta)]$. This expectation is a function of θ. One view of maximum likelihood estimation is that the parameter we seek to estimate with $\hat{\theta}^{MLE}$ is θ_0, the value of θ that maximizes $E[\log f(X;\theta)]$. More formally,

$$\hat{\theta}^{MLE} = \arg\max_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log f(X_i;\theta) \right\}$$

is an estimator for

$$\theta_0 = \arg\max_{\theta} E[\log f(X;\theta)]$$

The latter expression is the population analog to MLE. Let's consider the first order condition to this population analog:

$$\frac{\partial E\left[\log f\left(X;\theta\right)\right]}{\partial\theta} = 0$$

This first order condition is related to – but not the same as – the expected value of the score function $s\left(X;\theta\right)=\frac{\partial \log f(X;\theta)}{\partial\theta}$ at the true parameter value.

PROPOSITION 3.3.3. *As long as $f\left(X;\theta\right)$ satisfies regularity conditions enough to exchange integration and differentiation, the expectation of the score function equals zero when evaluated at the true parameter value θ_0, $E\left[s\left(X;\theta_0\right)\right]=0$.*

PROOF. Start by writing the expectation in its integral form:

$$\begin{aligned} E\left[s\left(X;\theta\right)\right] &= \int \underbrace{\frac{\partial \log f\left(X;\theta\right)}{\partial\theta}}_{=\frac{\partial f(X;\theta)}{\partial\theta}/f(X;\theta)} f\left(X;\theta_0\right)dX \\ &= \int \frac{\partial f\left(X;\theta\right)}{\partial\theta}\frac{f\left(X;\theta_0\right)}{f\left(X;\theta\right)}dX \end{aligned}$$

Once we evaluate the score function at the true parameter value, the ratio of the pdfs inside the integral cancels to reduce the expression to

$$\begin{aligned} E\left[s\left(X;\theta_0\right)\right] &= \int \frac{\partial f\left(X;\theta_0\right)}{\partial\theta}dX \\ &= \frac{\partial}{\partial\theta}\int f\left(X;\theta_0\right)dX = \frac{\partial\left(1\right)}{\partial\theta}=0 \end{aligned}$$

where we used that we could exchange the order of integration and differentiation and the fact that pdfs integrated over their support integrate to one. □

3.3.1.2. *Technical Details Useful for Understanding Asymptotic Properties of MLE.* Consider an estimator $t\left(\mathbf{X}\right)$ where $\mathbf{X}$ is a vector of observations from the random sample and $t\left(\mathbf{X}\right)$ is a scalar. For this estimator, compute $\frac{\partial E[t(\mathbf{X})]}{\partial\theta}$.

$$\frac{\partial E\left[t\left(\mathbf{X}\right)\right]}{\partial\theta} = \int t\left(\mathbf{X}\right)\frac{\partial f\left(\mathbf{X};\theta\right)}{\partial\theta}d\mathbf{X}$$

as long as we can exchange integration and differentiation. Next, multiply by $\frac{f(\mathbf{X};\theta)}{f(\mathbf{X};\theta)}$ and recognize that $\frac{\partial f(\mathbf{X};\theta)}{\partial\theta}/f(\mathbf{X};\theta)=\frac{\partial \log f(\mathbf{X};\theta)}{\partial\theta}$. Putting these two facts together, we arrive at the expression:

$$\begin{aligned} \frac{\partial E\left[t\left(\mathbf{X}\right)\right]}{\partial\theta} &= \int t\left(\mathbf{X}\right)\frac{\partial \log f\left(\mathbf{X};\theta\right)}{\partial\theta}f\left(\mathbf{X};\theta\right)d\mathbf{X} \\ &= E\left[t\left(\mathbf{X}\right)\frac{\partial \log f\left(\mathbf{X};\theta\right)}{\partial\theta}\right] = Cov\left[t\left(\mathbf{X}\right),\frac{\partial \log f\left(\mathbf{X};\theta\right)}{\partial\theta}\right] \end{aligned}$$

where the expectation is a covariance because the expected score is zero (at θ_0) and

$$E\left[\frac{\partial \log f\left(\mathbf{X};\theta\right)}{\partial\theta}\right] = \sum_{i=1}^{n} E\left[\frac{\partial \log f\left(X_i;\theta\right)}{\partial\theta}\right]=0$$

By the Cauchy-Schwartz Inequality, we know that $\left(Cov\left[X,Y\right]\right)^2 \leq Var\left[X\right]Var\left[Y\right]$. Applying this formula to the covariance term above, we obtain:

$$Var\left[t\left(\mathbf{X}\right)\right]Var\left[\frac{\partial \log f\left(\mathbf{X};\theta\right)}{\partial\theta}\right] \geq \left(Cov\left[t\left(\mathbf{X}\right),\frac{\partial \log f\left(\mathbf{X};\theta\right)}{\partial\theta}\right]\right)^2$$

If we define $J = \frac{1}{n} Var\left[\frac{\partial \log f(\mathbf{X};\theta)}{\partial \theta}\right]$ and substitute out for the covariance term, we obtain:

$$nVar\left(t\left(\mathbf{X}\right)\right) J \quad \geq \quad \left(\frac{\partial E\left[t\left(\mathbf{X}\right)\right]}{\partial \theta}\right)^2$$

Invert the J matrix to obtain a bound on the variance of our estimator $t\left(\mathbf{X}\right)$.

$$nVar\left(t\left(\mathbf{X}\right)\right) \quad \geq \quad J^{-1}\left(\frac{\partial E\left[t\left(\mathbf{X}\right)\right]}{\partial \theta}\right)^2$$

If $t\left(\mathbf{X}\right)$ is unbiased for θ, $E\left[t\left(\mathbf{X}\right)\right] = \theta \Rightarrow \frac{\partial E[t(\mathbf{X})]}{\partial \theta} = 1$, which simplifies the formula:

$$nVar\left(t\left(\mathbf{X}\right)\right) \quad \geq \quad J^{-1}$$

This expression is what is known as the **Cramer-Rao Lower Bound (CRLB) for unbiased estimators**. Showing that an estimator has a variance-covariance matrix equal to the CRLB demonstrates that the estimator has the lowest possible variance. That is, estimators with a variance equal to the CRLB are **efficient** estimators.

Another important result in the theory of maximum likelihood is the information matrix equality. Start with the condition that the expected score function equals zero. For an h-dimensional parameter vector, this is a vector equation:

$$E\left[s\left(X;\theta\right)\right] \quad = \quad \mathbf{0}$$

For an h-dimensional parameter vector, this is an h-dimensional vector equation. Now, take the derivative with respect to this vector. The vector derivative of a vector is a matrix, which at the true parameter value θ_0 must also equal zero.

$$\begin{aligned}\frac{\partial}{\partial \theta'} E\left[s\left(X;\theta\right)\right] &= \int \frac{\partial}{\partial \theta'}\left(\frac{\partial \log f\left(X;\theta\right)}{\partial \theta} f\left(X;\theta\right)\right) dX \\ &= \int \frac{\partial^2 \log f\left(X;\theta\right)}{\partial \theta \partial \theta'} f\left(X;\theta\right) dX + \int \frac{\partial \log f\left(X;\theta\right)}{\partial \theta} \frac{\partial \log f\left(X;\theta\right)}{\partial \theta'} f\left(X;\theta\right) dX\end{aligned}$$

[Useful Note: $\frac{\partial f(X;\theta)}{\partial \theta'} \times \frac{1}{f(X;\theta)} = \frac{\partial \log f(X;\theta)}{\partial \theta'}$]

At the true value of the parameter vector θ_0, the expectation of each term of this matrix is zero (because we would be taking the derivative of a bunch of terms that are identically zero). In other words, we can obtain the equality:

$$E\left[\frac{\partial \log f\left(X;\theta\right)}{\partial \theta} \frac{\partial \log f\left(X;\theta\right)}{\partial \theta'}\right] \quad = \quad -E\left[\frac{\partial^2 \log f\left(X;\theta\right)}{\partial \theta \partial \theta'}\right]$$

This equation is called the **information matrix equality**.

$$\begin{aligned}J = Var\left[\frac{\partial \log f\left(X;\theta\right)}{\partial \theta}\right] &= E\left[\frac{\partial \log f\left(X;\theta\right)}{\partial \theta} \frac{\partial \log f\left(X;\theta\right)}{\partial \theta'}\right] \\ &= -E\left[\frac{\partial^2 \log f\left(X;\theta\right)}{\partial \theta \partial \theta'}\right]\end{aligned}$$

The matrix described by this equality is called the (Fisher) **information matrix**. As we will see at the end of this section, the information matrix is important because it is where to look for MLE standard errors. It will often be the case that computing the $h \times h$ matrix of second derivatives of $\log f\left(X;\theta\right)$ will be much simpler than computing an h-dimensional vector of of derivatives of $\log f\left(X;\theta\right)$ and taking the outer product of that vector of derivatives with itself.

EXERCISE 3.3.4. Fill in the details of this information matrix equality derivation.

3.3.2. MLE as a correction for non-spherical Ω. Just like FGLS estimation, we need to impose some structure on the error term to get anywhere with MLE. In fact for MLE, we need more structure. We need to know the distribution of the errors, rather than just the moment conditions that identify GLS. In this section, we consider the example of an AR(1) error process.

Assume that conditional on $\mathbb{X}$, $\mathbf{U}$ has a multivariate normal distribution given by $N\left(\mathbf{0}, \sigma^2\Omega\right)$ where U_t is AR(1) with parameter ρ as in $U_t = \rho u_{t-1} + \xi_t$ and $\xi_t \sim iid\left(0, \sigma_u^2\right)$. Given this setup, Ω is exactly as we described in the previous section on GLS and fully specifying Ω allows us to write out the likelihood function using the joint distribution of $\mathbf{U}$.

A useful formula here is the joint pdf of a multivariate normal distribution. If $\mathbf{W} \sim N(\mathbf{0}, \Sigma)$, its pdf can be written as:

$$f_{\mathbf{W}}(\mathbf{w}) = \frac{1}{\sqrt{(2\pi)\, det\,(\Sigma)}} \exp\left(-\frac{1}{2}\mathbf{w}'\Sigma^{-1}\mathbf{w}\right)$$

where $\mathbf{w}$ is a vector of realizations of the random vector $\mathbf{W}$.

Given this formula for the joint pdf of a multivariate normal random vector, we can write the log-likelihood function as

$$\mathcal{L}(\beta, \Omega|\mathbf{Y}, \mathbb{X}) = K - \frac{1}{2}(\mathbf{Y} - \mathbb{X}\beta)'\left(\sigma^2\Omega\right)^{-1}(\mathbf{Y} - \mathbb{X}\beta)$$

EXERCISE 3.3.5. Show that the log-likelihood can be expressed this way and that maximizing the likelihood for a given σ^2 is equivalent to the GLS problem for AR(1) errors. Use this finding to propose an iterative method for solving this problem.

A final important result on MLE that we will state and use without proof.[2]

THEOREM 3.3.6. ***Asymptotic Normality and Efficiency of MLE***. *Under some regularity conditions, the maximum likelihood estimator $\hat{\theta}^{MLE}$ is consistent for its population counterpart θ_0, asymptotically normal, and asymptotically efficient in the sense that the variance of the limiting distribution attains the Cramer-Rao lower bound for unbiased estimators.*[3] *In our notation, we can summarize all of these results in the following convergence-in-distribution statement:*

$$\sqrt{n}\left(\hat{\theta}^{MLE} - \theta_0\right) \xrightarrow{d} N\left(0, J^{-1}\right)$$

where $\theta_0 = \arg\max_{\theta} E\left[\log f(X;\theta)\right]$ *and* J^{-1} *is the inverse of the information matrix.*

3.4. Estimating Nonlinear Models For Binary Response: Probit and Logit

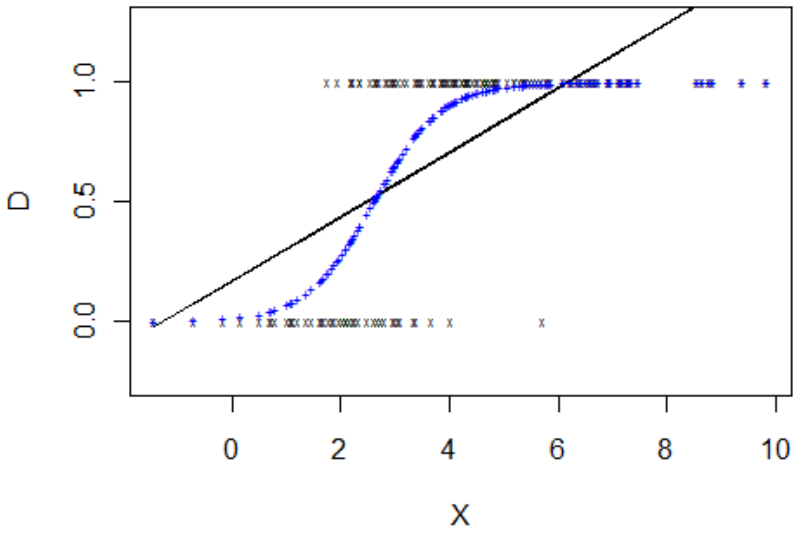
In econometrics, the most important application of nonlinear models is to settings where the response variable D is binary. Our study of OLS regression led us to study binary-response models as a specific application of linear least squares regression. If we could assume that the conditional expectation $E[D|\mathbf{X}]$ is linear, we could argue that the true regression model could be expressed as predicting the probability $D = 1$:

$$D = \mathbf{X}'\beta + U$$

[2]For a full-blown analysis of all of the details of the derivation behind the proof, there is an excellent set of notes available here (http://www2.econ.iastate.edu/classes/econ671/hallam/documents/Asymptotic_Dist.pdf), which uses slightly different notation.

[3]This means that among all consistent, non-super-efficient estimators, MLE has the minimum variance. Of course, there are regularity conditions involved for this to hold.

FIGURE 3.4.1. Fits from Linear Probability Model Versus Fits from Logistic Regression



where $P[D = 1|\mathbf{X}] = \mathbf{X}'\beta$.

Although the regression formulation of binary-response models fits naturally into our framework for understanding OLS regression, there are two undesirable consequences.

(1) The linear probability model (OLS with a binary response variable D) is heteroskedastic. Specifically, the variance is of the form:

$$Var[U|\mathbf{X}] = (\mathbf{X}'\beta)(1 - \mathbf{X}'\beta)$$

(2) The linear probability model often leads to predictions of probabilities greater than one or less than zero. As Figure 3.1 demonstrates, the fitted values from a straight line through the points must go outside of the zero-one interval (except for the rare case of a horizontal fitted line).

Indeed, a better fitting model would have as S shape like the locus of blue points in Figure 3.4.1. In our pursuit of a better fitting model, we turn to the discussion of logit and probit regression. As a starting point, imagine that there is some latent characteristic Y that is linearly related to our vector of regressors $\mathbf{X}$ as in regression:

$$Y = \mathbf{X}'\beta - V$$

In our analysis of the data, we do not have information on Y, but we have coarser information on a related variable D.. We only have observations on a dummy variable D that equals one if Y exceeds some cutoff $Y \geq c$. Otherwise, $D = 0$. Because we have an intercept in our regression model, we can normalize $c = 0$ without loss of generality. Y and D are intimately connected. In indicator function notation, we can express D as:

$$D = I_{\{Y \geq 0\}}$$

Given this setup, what is the probability $D = 1$? It is identically the probability $Y \geq 0$, which depends on the probability distribution of the error term V.

$$\begin{aligned} P\left[D = 1|\mathbf{X}\right] &= P\left[\mathbf{X}'\beta - V \geq 0\right] \\ &= P\left[V \leq \mathbf{X}'\beta\right] = F_V\left(\mathbf{X}'\beta\right) \end{aligned}$$

where $F_V(\cdot)$ is the CDF of V. With knowledge of the form of the CDF, one can write out the likelihood function.

EXERCISE 3.4.1. Imagine that $\{D_i, \mathbf{X}_i\}_{i=1}^n$ is a random sample of observations from the joint distribution $(D, \mathbf{X})$. Show that the likelihood function can be written as:

$$L\left(\beta; \mathbf{D}|\mathbb{X}\right) = \prod_{i=1}^{n} F_V\left(\mathbf{X}_i'\beta\right)^{D_i}\left(1 - F_V\left(\mathbf{X}_i'\beta\right)\right)^{1-D_i}$$

which is equivalent to writing:

$$L\left(\beta; \mathbf{D}|\mathbb{X}\right) = \left(\prod_{\{D_i=1\}} F_V\left(\mathbf{X}_i'\beta\right)\right)\left(\prod_{\{D_i=0\}}\left(1 - F_V\left(\mathbf{X}_i'\beta\right)\right)\right)$$

Given this expression, we can take logs to obtain the log-likelihood:

$$\mathcal{L}\left(\beta; \mathbf{D}|\mathbb{X}\right) = \sum_{i=1}^{n}\left(D_i \log F_V\left(\mathbf{X}_i'\beta\right) + (1 - D_i)\log\left(1 - F_V\left(\mathbf{X}_i'\beta\right)\right)\right)$$

Take the first order conditions from this problem with respect to β to obtain:

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\left(\frac{D_i}{F_V\left(\mathbf{X}_i'\beta\right)} f_V\left(\mathbf{X}_i'\beta\right)\mathbf{X}_i - \frac{(1 - D_i)}{1 - F_V\left(\mathbf{X}_i'\beta\right)} f_V\left(\mathbf{X}_i'\beta\right)\mathbf{X}_i\right) \\ &= \sum_{i=1}^{n}\frac{\left(D_i\left(1 - F_V\left(\mathbf{X}_i'\beta\right)\right) - (1 - D_i)\, F_V\left(\mathbf{X}_i'\beta\right)\right) f_V\left(\mathbf{X}_i'\beta\right)\mathbf{X}_i}{F_V\left(\mathbf{X}_i'\beta\right)\left(1 - F_V\left(\mathbf{X}_i'\beta\right)\right)} \\ &= \sum_{i=1}^{n}\left(D_i - F_V\left(\mathbf{X}_i'\beta\right)\right)\frac{f_V\left(\mathbf{X}_i'\beta\right)\mathbf{X}_i}{F_V\left(\mathbf{X}_i'\beta\right)\left(1 - F_V\left(\mathbf{X}_i'\beta\right)\right)} \end{aligned}$$

This expression mimics the sample moment conditions from OLS. In particular, we can express this set of first order conditions as $\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\hat{U}_i = 0$, but here the residual has the form: $U_i = \left(D_i - F_V\left(\mathbf{X}_i'\beta\right)\right)$ and the $\tilde{\mathbf{X}}_i = \frac{f_V\left(\mathbf{X}_i'\beta\right)\mathbf{X}_i}{F_V\left(\mathbf{X}_i'\beta\right)\left(1 - F_V\left(\mathbf{X}_i'\beta\right)\right)}$.

This derivation is perfectly general, but we are unable to solve the system of equations for an estimator $\hat{\beta}$ until we specify the functional form for the distribution of the errors. There are infinitely many distributions from which we could choose, but two distributional choices are commonly used in practice.

DEFINITION 3.4.2. If we assume that $V \sim N(0, 1)$, the estimator $\hat{\beta}$ that results from solving the maximum likelihood problem is called the **probit estimator** of β, $\hat{\beta}^{probit}$.

More specifically, this assumption of normality means that $F_V(v) = \Phi(v) = \int_{-\infty}^{v}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)dx$. Because the normal CDF (with no closed form solution) shows up in the first order conditions of the maximum likelihood problem, we solve for $\hat{\beta}^{probit}$ using numerical methods. Fortunately, these methods are standard in statistical software.

Although the assumption of normality is popular, practitioners will often depart from the assumption of normality to assume that V has a logistic distribution. This assumption is attractive for a variety of reasons.

DEFINITION 3.4.3. If we assume that $V \sim Logistic$, the estimator $\hat{\beta}$ that results from solving the maximum likelihood problem is called the **logit estimator** of β, $\hat{\beta}^{logit}$. More specifically, this assumption means that $F_V(v) = \frac{\exp(v)}{1+\exp(v)}$. We call this type of regression **logistic regression**.

EXERCISE 3.4.4. If $V \sim Logistic$, the sample moment conditions reduce to

$$\sum_{i=1}^{n} \left(D_i - F_V\left(\mathbf{X}_i'\beta\right)\right)\mathbf{X}_i \quad = \quad 0$$

FACT 3.4.5. *Under an assumption of logistic errors, we can actually invert the expression for* $P[D=1|\mathbf{X}]$ *to obtain an expression that is linear in the parameters:*

$$\log\left(\frac{P[D=1|\mathbf{X}]}{1-P[D=1|\mathbf{X}]}\right) \quad = \quad \mathbf{X}'\beta$$

The left hand side of this expression is the log of the odds of observing $D=1$. Therefore, logistic regression coefficients can be interpreted as the marginal effect on the log of the odds of $D=1$. If we carefully consider the effect of transforming this equation, this expression also implies that the exponentiated coefficient $\exp(\beta_j)$ equals the **ratio of the odds** of $D=1$ when $X_j = \tilde{X}_j + 1$ relative to the odds of $D=1$ when $\tilde{X}_j$. That is, the exponentiated coefficients can be interpreted as having a multiplicative effect on the odds of $D=1$.

REMARK 3.4.6. The logit estimator arises from a generalized linear model for the probability of observing $D=1$. Fundamentally, the model is nonlinear, but under an appropriate transformation of the observed probabilities (in general, this transformation is called a link function), the transformed responses can be modeled as a linear function of the parameters. The function $\log\left(\frac{P}{1-P}\right)$ is called the **logit link function**. In case you are interested, there is a beautifully-elegant theory of generalized linear models that exploits the properties of regular exponential families and natural parametrizations of these regular exponential families to propose natural estimators in a variety of generalized settings. It is nice to have this powerful motivating theory in the background, but we will be satisfied by motivating logistic regression as an application of MLE.

One advantage of logistic regression over probit regression is its relative ease of interpretability through interpreting the exponentiated coefficient estimates as estimated multiplicative effects on the odds ratio. Probit regression coefficients have no such natural interpretation. For this reason, we often transform the estimated regression line $\mathbb{X}\hat{\beta}^{probit}$ back to the scale of predicted probabilities by applying the normal CDF to the fitted values $\hat{P} = \Phi\left(\mathbb{X}\hat{\beta}^{probit}\right)$.

What are the asymptotic properties of this estimator of the probabilities? For one, we can see that $\frac{\partial\Phi(\mathbb{X}\beta)}{\partial\beta} = \varphi(\mathbb{X}\beta) \times \mathbb{X}$ by taking the derivative. We also notice that $Y = \mathbb{X}\beta + V$ is estimated using MLE under an assumption of normality, which we know to be equivalent to OLS estimation. Hence, we know that $\sqrt{n}\left(\hat{\beta}^{probit} - \beta\right) \xrightarrow{d} N(0, \Sigma)$ from before. If we want to know the asymptotic properties of $\Phi\left(\mathbb{X}\hat{\beta}^{probit}\right)$, we can apply the delta method to obtain the asymptotic result.

THEOREM 3.4.7. *Asymptotic Normality of Estimated Probit Probabilities. The probit estimator* $\hat{\beta}^{probit}$ *naturally leads to the result on asymptotic normality:*

$$\sqrt{n}\left(\Phi\left(\mathbb{X}\hat{\beta}^{probit}\right) - \Phi(\mathbb{X}\beta)\right) \quad \xrightarrow{d} \quad N\left(0, \left(\varphi(\mathbb{X}\beta)^2\,\mathbb{X}'\Sigma\mathbb{X}\right)\right)$$

We will often be interested in the change in these probabilities due to a one unit change in a coefficient, called **marginal effects**. This marginal effect can be computed by taking the derivative of these probabilities with respect to the j^{th} regressor $m.effect\left(X_j\right) = \frac{\partial \Phi(\mathbb{X}\beta)}{\partial X_j} = \beta_j \varphi\left(\mathbb{X}\beta\right)$. Under even stronger conditions, an analogous result on the asymptotic normality of the marginal effects. Because statistical software computes these marginal effects with little effort, there is little cost in using probit regression in terms of interpretability of the coefficient estimates. Additionally, it turns out that probit and logit estimates for the marginal effect correspond closely with one another – which technique to use in practice usually boils down to whether you like to interpret odds ratios.

EXERCISE 3.4.8. Let $\Lambda\left(v\right)$ be the logistic CDF and $\Phi\left(v\right)$ be the standard normal CDF.

- Consider the convex combination of these two CDFs $F_V\left(v\right) = \alpha\Lambda\left(v\right) + \left(1-\alpha\right)\Phi\left(v\right)$ parametrized by α. Verify this is a valid CDF.
- Assuming that the CDF of the error term on the latent variable is given by the CDF above, write out the log-likelihood function $\mathcal{L}\left(\alpha, \beta; \mathbf{D}|\mathbb{X}\right)$.
- Compute the first order conditions with respect to α and β (recall: β is a vector) and organize them analogously to the sample moment conditions.
- If our only regressor is the constant, can we separately identify the parameters α and β_0. Be precise.

3.5. Chapter Exercises

(1) Use R to verify Example 3.2.4 numerically in a Monte Carlo experiment.
   (a) Start by constructing a data frame with the following characteristics. Report basic summary statistics on the data frame you create. Append your code to the end of the assignment.
      - At the beginning run the command `set.seed(209TT)`, where TT is the two-digit number you estimate to be Tony's age.
      - Let i denote an individual classroom. Class i has class size of $N_i \in \{8, 10, 12, ..., 122\}$.
      - There are 58 teachers, one in charge of each class.
      - Let T_i denote the teacher training hours. Draw T_i from a Log Normal distribution with location parameter $\mu = 0.07N_i$ and scale parameter (variance of the log) $\sigma^2 = 1$.
         – Suppose teacher training is capped at 2000 hours. That is, your data should be the minimum of the vector you drew. The `pmin()` command is useful here.
         – To make these numbers more "realistic" round them to the nearest integer. Use the rounded numbers as your data on teacher training hours.
      - Conditional on T_i the amount of teacher training in classroom i, the math score of an individual student is drawn iid from $M_{ij} \sim N\left(\mu\left(T_i\right), \sigma^2\right)$, where $\mu\left(T_i\right) = 90 + 0.0015\left(T_i\right)$ and $\sigma^2 = 100$.
      - Construct two data frames.
      - One where the level of the observation is the individual student (and you have student-by-student data)
      - Another where the level of observation is the class (and you only have class average data)
   (b) On the classroom-level data, estimate the regression model

$$class.avg \quad = \quad \beta_0 + \beta_1 Training + U$$

using OLS, OLS with heteroskedastic-robust standard errors and WLS with the appropriate weights. Do you obtain the same estimates for $\hat{\beta}$? Compare the standard errors and comment on their validity.

(c) On the student-level data, estimate the regression model

$$student.score = \beta_0 + \beta_1 Training + U$$

How do your estimates relate to the ones you obtained in part (b)?

(2) **Error Processes**. Consider the regression model in stacked form

$$\mathbf{Y} = \mathbb{X}\beta + \mathbf{U}$$

where the observations in the sample are not necessarily independent of one another. In particular, $\mathbf{U}$ is a n-dimensional vector of observations that follow one of the following processes. Under a homoskedasticity assumption $Var[U_i|\mathbb{X}] = \sigma_U^2$, express $\Omega = Var[\mathbf{U}|\mathbb{X}]$ fully in terms of the parameters:

(a) Moving Average, MA(1)). Let t denote the t^{th} observation

$$U_t = \xi_t - \lambda\xi_{t-1}$$

where we assume that $\xi_t \sim \left(0, \sigma_\xi^2\right)$ independently and identically.

(b) Autoregressive Moving Average, ARMA(1,1). Let t denote the t^{th} observation

$$U_t = \rho U_{t-1} + \xi_t - \lambda\xi_{t-1}$$

where we assume that $\xi_t \sim \left(0, \sigma_\xi^2\right)$ independently and identically.

(3) Maximum Likelihood Practice.

- For each of this problem's parts, (i) Write down the likelihood function and log-likelihood function, (ii) Find $\hat{\theta}_{MLE}$, (iii) Find the information matrix, (iv) Prove whether or not $\hat{\theta}_{MLE}$ is unbiased for θ, (v) Find the asymptotic distribution of $\hat{\theta}_{MLE}$

(a) **Power Density**. Let $\{X_i\}_{i=1}^n$ be an iid sample from the density

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & x \in (0,1) \\ 0 & otherwise \end{cases}$$

Let the parameter space be $\Theta = (0, \infty)$.

(b) **Normal (iid) Density**. Let $\{X_i\}_{i=1}^n$ be an iid sample from a $N(\theta, 1)$ density.

(c) **Special Normal Density**. Let $\{X_i\}_{i=1}^n$ be an iid sample from a $N\left(\theta, \sigma^2 = \theta^2\right)$ density.

(d) **Exponential Density**. Let $\{X_i\}_{i=1}^n$ be an iid sample from the exponential density

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right) & x \in (0,\infty) \\ 0 & otherwise \end{cases}$$

Let the parameter space be $\Theta = (0, \infty)$.

(4) **Consequences of AR(1) errors for OLS.** Use R for this exercise. At the beginning of your code for this problem, use the command `set.seed(209L)` where `L` is the number between 0 and 10 that best describes your love for econometrics.

At date t, national income (Y_t) is determined by human capital stock H_t according to the equation:

$$Y_t = \beta_0 + \beta_1 H_t + U_t$$

where the error term evolves according to an autoregressive process

$$U_t = \rho U_{t-1} + \xi_t$$

where we assume that $\xi_t \sim N\left(0, \sigma_\xi^2\right)$ independently and identically. Assume that $U_0 = 0$.

(a) Suppose that $\sigma_\xi^2 = 1$. For $\rho \in \{-0.6, 0, 0.2, \text{ and } 0.6\}$, generate a sequence of error process realizations of length $T = 25$. For each simulated series of errors, plot these realizations versus date. Comment on your ability to detect the difference visually. In addition, apply the `acf(seriesname)` function to each series of observations.

(b) For practice, generate a synthetic data set for the following values of the parameters: $\beta_0 = 10$, $\beta_1 = 4$, $\rho = 0.6$, $\sigma_\xi = 1$ and for $H_t \sim \chi^2_{df}$ where $df = 12$.

(c) Now, generate $B = 1000$ separate data sets of the form described above (be sure to draw a new vector of ξ errors at each iteration in a loop). This loop may take a little while to run. For each data set, do the following:

(i) Use the `lm()` function to compute $\hat{\beta}_1^{OLS}$ and store its value. Compute the mean of these $B = 1000$ coefficient estimates $\hat{\beta}_1^{OLS}$.

(ii) Use the `gls()` function[4] to compute $\hat{\beta}^{FGLS.1}$ assuming an AR(1) process (but not assuming knowledge of the value of ρ). Store both $\hat{\rho}$ and $\hat{\beta}^{GLS}$ at each iteration in the loop. Compute the mean of these stored values. Do these line up with what we input into this simulation?

(iii) Use the `gls()` function to compute $\hat{\beta}^{FGLS.2}$ assuming an MA(1) process (That is, estimate the model with an incorrectly specified error process).

(iv) On the same scale (using the `xlim = c(low,high)` option), plot the histogram of the calculated values of $\hat{\beta}^{OLS}$, $\hat{\beta}^{FGLS.1}$ and $\hat{\beta}^{FGLS.2}$. Compute the mean and standard deviation of these estimates (use these to construct MSE). What do you conclude about GLS versus OLS?

(d) **Bonus**: Simulate $B = 1000$ data sets using the ARMA(1,1) process described in Question 3. Assume the same parameter values as in parts (a) and (b) with $\rho = 0.6$ and $\lambda = 0.4$. Compare OLS to three alternatives (i) GLS assuming AR(1), (ii) GLS assuming MA(1), (iii) GLS assuming ARMA(1,1). Use the same bases for comparison as in part (c)iv.

(5) **FGLS on Real Data**. For this problem use the `adsensedata.csv` data from the website. Consider the regression specification

$$EarningsPPC = \beta_0 + \beta_1 ClicksPPC + \beta_2 ImpressionsPPC + \beta_3 ClicksPPI + \beta_4 ImpressionsPPI + U$$

(a) Plot the residuals versus the fitted values from an OLS fit. Comment on the homoskedasticity assumption.

(b) Plot the autocorrelation function using the `acf()` command on the residuals from (a). Do you detect evidence of autocorrelation?

(c) Estimate this regresison model using (i) standard OLS, (ii) OLS with robust standard errors, (iii) FGLS assuming an AR(1) correlation structure and (iv) FGLS assuming that the variance of each observation is given by $Var[U|\mathbf{X}] = \frac{\sigma^2}{Impressions}$, where $Impressions = ImpressionsPPC + ImpressionsPPI$. Compare the estimates and standard errors across models.

(d) The Akaike Information Criterion (AIC) is a widely-used methodology to compare the relative closeness to the truth among a set of models for the same set of observations on a response variable. AIC is easy to compute in R, using the AIC function (and this is valid for `lm()` objects and `gls()` objects fit using maximum likelihood; i.e., `gls(formula, data, method=''ML'')`). Use the `AIC()` function in R to compute AIC

[4]This function is in the nlme library. The gls objects are weird in how they store the correlation parameters. You can access the estimated autocorrelation parameter by using the command `intervals(glsobject)$corStruct[2]`. This essentially picks out the center of a confidence interval for the true autocorrelation parameter. It is sad that the package stores its data this way.

for each model you fit in the previous part. For R's formulation of AIC, smaller is better. Based on your computed values of AIC, which model do you prefer?[5]

(6) **Logit and Probit**. Revisit the Yogurt data.

(a) Estimate the basic specification using OLS

$$Y = \mathbf{X}'\beta + \mathbf{P}'\gamma + U \tag{3.5.1}$$

where $\mathbf{X}$ is a vector that contains a constant and a full set of the featured advertisement dummy variables, $\mathbf{P}$ is a vector that contains the prices for each of the brands of yogurt and Y is a dummy variable for whether the individual bought Yoplait.

(b) Estimate the specification in (a) using logit and probit. Compute the estimated probabilities from all three models (OLS, logit and probit). Provide a plot that compares these estimated probabilities across the three specifications (similar to Figure 4.1 in the notes).

(c) Convert your coefficient estimates to marginal effects (at the mean of the other regressors) and compare the results across specifications. Comment on the differences in the results across models.

(7) **Lagged Dependent Variables**. Imagine that you have a time series $\{Y_t\}_{t=1}^T$ of observations and you wish to estimate the simple linear autoregression:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t$$

using OLS.

(a) Using only the data from $t = 2, ..., T$, give an expression for the OLS estimator of β_1. Think of Y_t and Y_{t-1} as different (but related) random variables. In terms of the joint distribution of Y_t and Y_{t-1}, what is the probability limit of the OLS estimator $\hat{\beta}_1^{OLS}$?

(b) Assuming $U_t \sim^{iid} \left(0, \sigma^2\right)$, is $\hat{\beta}_1^{OLS}$ consistent for β_1? Justify your answer.

(c) Now, assume that U_t comes from an autoregressive process $U_t = \rho U_{t-1} + \epsilon_t$ where $\epsilon_t \sim^{iid} \left(0, \sigma_\epsilon^2\right)$. In this setting is $\hat{\beta}_1^{OLS}$ consistent for β_1? Justify your answer.

[5] In the statistics literature, a standard practice is to disregard differences of less than 2 AIC points. If AIC is at least 2 lower for one model than the others, you can reasonably conclude that it is a better fitting model.

CHAPTER 4

Instrumental Variables Methods

In our study of regression, we learned that Ordinary Least Squares is a tremendous method for estimating β from a linear regression model

$$Y = \mathbf{X}'\beta + U$$

where all of the components of the vector of regressors $\mathbf{X} = \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_k \end{pmatrix}$ are known and measured perfectly. In this section of the course, we relax this assumption about $\mathbf{X}$. We have already considered the problem of mismeasurement and omitted variable bias, but by and large, the solutions we proposed in this previous discussion were to (a) include omitted variables in the model and (b) measure mismeasured variables better than we were measuring them before.

The conclusion of that section of the course was unsatisfying because these perfect soltuions are usually unattainable in practice. Now, we turn to a class of methods that are less ambitious about the number of variables to collect, but more ambitious about the properties of the variable we use. The method of instrumental variables was originally proposed as a solution to the mismeasurement problem. For that reason, we start with a more complete discussion of the consequences and fixes of mismeasurement.

4.1. Measurement Bias: Revisited

In simple linear regression, recall from the second set of regression notes that measurement error that is uncorrelated with the regressor ($Cov[X, \xi] = 0$) leads to attenuation bias. In the key part of the derivation, we argued

$$\hat{\beta}_1^{ols} \xrightarrow{P} \tilde{\beta}_1 = \frac{Cov[X^*, Y]}{Var[X^*]}$$

We can plug into this expression for $X^* = X + \xi$ and Y using the true regression to obtain:

$$\begin{aligned} \tilde{\beta}_1 &= \frac{Cov[X + \xi, \beta_0 + \beta_1 X + U]}{Var[X + \xi]} \\ &= \frac{\beta_1 Var[X]}{Var[X] + Var[\xi]} \end{aligned}$$

which is smaller in magnitude than β_1. Our intuition for attenuation bias is that the mismeasured regressor includes two components: (1) The regressor X which has a slope of β_1, (2) the measurement error ξ, which has zero slope in the true population regression. Lumping these two components in one variable means that we can only pick up an average effect between the two parts of the mismeasured regressor. This average is a variance-weighted average of the two effects β_1 and 0.

4.1.1. Proxy Variables. Measuring a variable with error is often confused with the notion of using a proxy variable in place of the true regressor. At this point, it is worth spelling out clearly the difference between a proxy variable and a variable that is mismeasured. If X^p is a proxy variable, it is related to the true regressor X as:

$$X = X^p + V$$

where $Cov[X^p, V] = 0$.

On an intuitive level, one can think of the proxy variable as capturing the essence of the regressor's variability and covariance with Y, but not capturing every element of the variability of Y. An imperfect example is to think about the average education level of a municipality as a proxy for the human capital of the residents who live there. It is true that average educational attainment does not capture every aspect of human capital, but high educational attainment is a good predictor for high human capital. Is the error term that we have to add to the proxy variable unrelated to the value of the proxy X^p? Possibly, and if so, the proxy variable is a good stand in variable for X in a regression.

To see this, notice that the probability limit of the OLS estimator from a regression of Y on X^p is given by

$$\begin{aligned}\hat{\beta}_1^{ols} \xrightarrow{P} \frac{Cov[Y, X^p]}{Var[X^p]} &= \frac{Cov[\beta_0 + \beta_1 X + U, X^p]}{Var[X^p]} \\ &= \beta_1 \frac{Cov[X, X^p]}{Var[X^p]} = \beta_1 \frac{Var[X^p] + Cov[X^p, V]}{Var[X^p]} = \beta_1\end{aligned}$$

That is, using a perfect proxy (in the sense that $Cov[X^p, V] = 0$) in place of the true regressor X implies the OLS estimator $\hat{\beta}_1$ is consistent for β_1. Thus, if we cannot obtain a perfect measurement of an omitted variable, an adequate solution is to find a good proxy, and we would strictly prefer a good proxy to a botched attempt at measuring X.

4.1.2. The case of two mismeasured regressors. Suppose that we have two sets of measurements on the same regressor – $X_1^* = X + \xi_1$ and $X_2^* = X + \xi_2$ – but both contain measurement error that is unrelated to anything else in the problem. That is, $Cov[X, \xi_1] = Cov[X, \xi_2] = 0$ and also $Cov[\xi_1, \xi_2] = 0$ and $Cov[\xi_i, U]$. Because we have two measurements that are uncorrelated with one another, it seems natural that we could improve estimation technique by using both of them.

4.1.2.1. *A Misguided Attempt.* As a first attempt, suppose we use the average of the two measurements in an OLS regression. If we use the formula for the individual mismeasurements, we immediately see that the measurement error problem does not go away.

$$\tilde{X} = \frac{X_1^* + X_2^*}{2} = X + \frac{\xi_1 + \xi_2}{2} = X + \tilde{\xi}$$

In particular, we will still have attenuation bias and we will be estimating

$$\hat{\beta}_1 \xrightarrow{P} \beta_1 \left(\frac{Var[X]}{Var[X] + Var\left[\tilde{\xi}\right]} \right)$$

instead of β_1. Suppose (wlog) that the variance of ξ_1 is less than ξ_2 and we are considering whether to include ξ_2 in an average with ξ_1. Because attenuation bias is worse as the mismeasurement variance increases, we would prefer doing this if $Var\left[\tilde{\xi}\right] < Var[\xi_1]$.

FACT 4.1.1. *Under the assumptions we about the two mismeasurements, we prefer to use the average of the two mismeasurements when $Var[\xi_2] < 3Var[\xi_1]$, but we would rather use a single mismeasured variable.*

4.1.2.2. *A better solution to the measurement error problem.* All of this is to say that averaging the two mismeasurements provides no guarantee of improving the ability of the estimation routine to estimate the correct parameter. If the two mismeasurements are uncorrelated $Cov[\xi_1, \xi_2]$, a better technique is to regress one mismeasured variable on the other, and use the fitted values from that regression in place of X. Suppose we regress X_1^* on X_2^* using the statistical model

$$X_1^* = \underbrace{\gamma_0 + \gamma_1 X_2^*}_{=\tilde{X}} + V$$

If we use this first stage regression to obtain fitted values, and in a second stage, use $\left\{\left(\tilde{X}_i, Y_i\right)\right\}_{i=1}^n$ to obtain the OLS estimator, that estimator has a probability limit given by

$$\begin{aligned}\hat{\beta}^{ols} \xrightarrow{P} \frac{Cov\left[\tilde{X}, Y\right]}{Var\left[\tilde{X}\right]} &= \frac{Cov\left[\tilde{X}, \beta_0 + \beta_1 X + U\right]}{Var\left[\tilde{X}\right]} \\ &= \beta_1 \frac{Cov\left[\tilde{X}, X\right]}{Var\left[\tilde{X}\right]}\end{aligned}$$

Whether $\hat{\beta}^{ols}$ is consistent boils down to whether $\frac{Cov[\tilde{X},X]}{Var[\tilde{X}]} = 1$, which is true because we can substitute out for X using the first mismeasurement and then substitute out for the mismeasurement using the first stage regression:

$$\begin{aligned}\frac{Cov\left[\tilde{X}, X\right]}{Var\left[\tilde{X}\right]} &= \frac{Cov\left[\tilde{X}, X_1^* - \xi_1\right]}{Var\left[\tilde{X}\right]} \\ &= \frac{Cov\left[\tilde{X}, \tilde{X} + V - \xi_1\right]}{Var\left[\tilde{X}\right]} = \frac{Var\left[\tilde{X}\right]}{Var\left[\tilde{X}\right]} = 1\end{aligned}$$

This follows because $\tilde{X}$ is uncorrelated with the error term V by the statistical interpretation applied to the first stage regression and the covariance $Cov\left[\tilde{X}, \xi_1\right] = Cov\left[\gamma_0 + \gamma_1 X_2^*, \xi_1\right] = Cov\left[\gamma + \gamma_1\left(X + \xi_2\right), \xi_1\right] = 0$ as long as $Cov[\xi_2, \xi_1] = 0$. For this reason, $Cov\left[\tilde{X}, \xi_1\right] = 0$. The bottom line from this exercise is that we were able to use two mismeasurements of the same regressor construct a good proxy variable to use in place of the true regressor. It turns out that the intuition of instrumental variables is not much different.

4.2. Omitted Variable Bias: Reconsidered

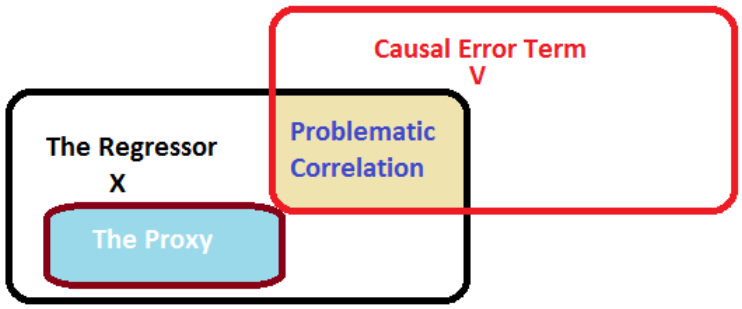
For simplicity, suppose that the best linear approximation to the true causal model of the determinants of Y is given by the long regression

$$Y = \beta_0 + \beta_1^L X + \delta Q + U_L$$

where X is observable, Q is not observable and $Cov[X, Q] \neq 0$. Because we can only observe X, we estimate the short regression

$$Y = \beta_0 + \beta_1^s X + U_S$$

FIGURE 4.2.1. How using a proxy variable can circumvent the omitted variable problem



Our OLS estimator $\hat{\beta}_1$ is consistent for β_1^s, which as we saw in the second set of regression notes, is related to the long regression coefficient we seek to estimate through the omitted variable bias formula

$$\beta_1^s = \beta_1^L + \delta \frac{Cov[X,Q]}{Var[X]}$$

At the time when we analyzed the problem of omitted variable bias, we did not emphasize that the second term in this formula has a related interpretation in terms of the the composite error term in the long regression $V = \delta Q + U_L$

$$\beta_1^s = \frac{Cov[X,Y]}{Var[X]} = \beta_1^L + \frac{Cov[X,V]}{Var[X]}$$

This form of the bias term of the short regression suggests an alternative solution. To see this how this alternative solution works, suppose that we use a proxy variable X^p in place of X. That is, we obtain an X^p that satisfies $X = X^p + V^p$ where $Cov[X^p, V^p] = 0$.

CLAIM 4.2.1. From our discussion of the properties of proxy variables, using X^p instead of X will obtain an estimator that is consistent for the slope of the short regression $\beta_1^s = \frac{Cov[X,Y]}{Var[X]} = \frac{Cov[X^p,Y]}{Var[X^p]} = \beta_1^L + \frac{Cov[X^p,V]}{Var[X^p]}$.

We can think of the proxy variable as picking up on the part (but not all) of the variability in X to identify the effect of X on Y. Moreover, if the proxy variable is correlated with the part of X_1 that is uncorrelated with the causal error term (i.e., $Cov[X^p, V] = 0$), using the proxy variable instead of X can eliminate omitted variable bias. Schematically, this argument is laid out in the Venn Diagram in Figure 4.2.1.

There is a fundamental difference between this approach and our previous solution to the omitted variable bias problem, which was to include the omitted regressors. In this case, we use less variability in an effort to use the right kind of variability to identify the effect of interest. The observation that not all variability is good is critical to the method of instrumental variables.

4.3. What is an Instrumental Variable, Anyway?

Even though we have not explicitly used the terminology, we have been using instrumental variable methods informally in the previous two sections. In this section, we make the discussion more formal.

4.3.1. Simple Regression: One Endogenous Variable, One Instrument. Suppose that our regressor is correlated with the error term ($Cov[X,V] \neq 0$) in the causal regression model

$$Y = \beta_0 + \beta_1^L X + V$$

where $V = \delta Q + U_L$. As we saw in the previous section, our OLS estimator $\hat{\beta}^{ols}$ will be inconsistent for β_1^L. For that reason, we do not want to use OLS, but using what we learned in the previous section, we would like to construct a proxy that is uncorrelated with V. We can do this if we have what is known as a **valid instrument**, Z. An instrument is valid if it satisfies two properties:

(1) **Exogeneity**. $Cov[Z,V] = 0$. Exogeneity is important for ensuring that the variable we construct to use in place of X is uncorrelated with V (which would solve the problem).
(2) **Relevance**. $Cov[Z,X] \neq 0$. Relevance is important for obtaining a variable with enough variability to identify β_1^L. In the background, think of forming fitted values from a regression $X = \gamma_0 + \gamma_1 Z + W$.

Using these properties of a valid instrument, we can solve for the parameter β_1^L in terms of estimable features of the joint distribution of (X,Y,Z). Start by taking the covariance of Y and Z and using linearity of covariances:

$$\begin{aligned} Cov[Y,Z] &= Cov\left[\beta_0 + \beta_1^L X + V, Z\right] \\ &= \beta_1^L Cov[X,Z] + \underbrace{Cov[Z,V]}_{=0} \end{aligned}$$

At this point, we see why it is important to have relevance $Cov[Z,X] \neq 0$. As long as the instrument is relevant, we can solve for β_1^L

$$\beta_1^L = \frac{Cov[Y.Z]}{Cov[X,Z]}$$

That is, in instrumental variables regression, instrument relevance is the assumption we need to identify β_1^L.

4.3.1.1. *A Simultaneous Equations Approach.* The instrumental variable estimator can also be motivated as part of a solution to a system of simultaneous equations. Given our expression for $\beta_1 = \frac{Cov[Z,Y]}{Cov[Z,X]}$, we can transform the coefficient β_1 by dividing top and bottom by $Var[Z]$:

$$\beta_1 = \frac{Cov[Z,Y]/Var[Z]}{Cov[Z,X]/Var[Z]}$$

Both the numerator and denominator are regression slope coefficients – the numerator from a statistical regression of Y on Z, the denominator from a statistical regression of X on Z. If we take this literally, we could obtain the IV estimator by conducting both of these regressions at the same time:

$$\begin{aligned} Y &= \gamma_0 + \gamma_1 Z + U_1 \\ X &= \alpha_0 + \alpha_1 Z + U_2 \end{aligned}$$

In this simultaneous system of equations, we can think about β_1 as the ratio $\frac{\gamma_1}{\alpha_1}$. This way of understanding the relationship of instrumental variables to the the relationship between Y and X makes precise what variables are to be explained (Y, X) and what variables are doing the explaining (Z) within the system. If a variable is on the left hand side of any equation in the system it is called **endogenous** because it is determined as an outcome of the system. If the variable only shows up on the right hand side of the equations in the system, it is called **exogenous** and its values will be taken as given in the system.

4.3.1.2. *Motivating the IV Estimator.* **Analogy Principle**. Using the analogy principle on the population version of the slope coefficient, we can easily obtain an estimator. Given a random sample, a consistent estimator for β_1^L is given by

$$\hat{\beta}_1^{IV} = \frac{S_{Z,Y}}{S_{Z,X}} = \frac{\sum_{i=1}^n \left(Z_i - \bar{Z}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(Z_i - \bar{Z}\right)\left(X_i - \bar{X}\right)}$$

Estimation by Simultaneous Equations. In the simultaneous system of regression equations, one could use an OLS fit on each equation to obtain $\hat{\gamma}_1 = \frac{S_{Z,Y}}{S_Z^2}$ and $\hat{\alpha}_1 = \frac{S_{Z,X}}{S_Z^2}$. Forming the sample ratio, we obtain:

$$\hat{\beta}_1^{SEM} = \frac{\hat{\gamma}_1}{\hat{\alpha}_1} = \frac{S_{Z,Y}}{S_{Z,X}} = \hat{\beta}_1^{IV}$$

which is numerically equivalent to the analogy principle estimator.

Two-Stage Least Squares (2SLS). Alternatively, we could use the motivating intuition that Z enables us to construct a proxy variable X^p to use in place of X. To simplify the derivation, assume that we have demeaned X, Z and Y. This won't affect the slope or the estimator of the slope, but it will allow us to drop the intercept from the model without loss of generality. In this setup, we use the first-stage statistical regression

$$X = \underbrace{\alpha_1 Z}_{X^p} + W$$

to obtain X^p. In a second stage, we use X^p in place of X to predict Y

$$Y = \beta_1 X^p + W$$

Constructed this way, the two-stage least squares estimator has the form

$$\hat{\beta}_1^{2SLS} = \frac{\hat{Cov}\left[Y, X^p\right]}{\hat{Var}\left[X^p\right]} = \frac{\hat{Cov}\left[Y, \frac{S_{X,Z}}{S_Z^2} Z\right]}{\hat{Var}\left[\frac{S_{X,Z}}{S_Z^2} Z\right]}$$

where the second equality comes from substituting $\hat{X}$ from an OLS fit of the first stage in place of X^p. Applying some algebra, we see that the expression for the two-stage estimator is numerically equivalent to the IV estimator we motivated.

$$\hat{\beta}_1^{2SLS} = \frac{\frac{S_{X,Z}}{S_Z^2}\hat{Cov}\left[Y, Z\right]}{\left(\frac{S_{X,Z}}{S_Z^2}\right)^2 \hat{Var}\left[Z\right]} = \frac{\hat{Cov}\left[Y, Z\right]}{\hat{Cov}\left[X, Z\right]} = \hat{\beta}_1^{IV}$$

4.3.1.3. *Properties of the IV Estimator.* Apart from alleviating the problem of omitted variable bias in an effort to identify the causal parameter β_1^L, the instrumental variables estimator is also asymptotically normal.

THEOREM 4.3.1. ***Properties of the IV Estimator.*** *If Z is a valid instrument in a population with finite fourth moments, the IV estimator is consistent and asymptotically normal.*

$$\begin{aligned} \hat{\beta}_1^{IV} &\xrightarrow{P} \beta_1 \\ \sqrt{n}\left(\hat{\beta}_1^{IV} - \beta_1\right) &\xrightarrow{d} N\left(0, \frac{Var\left[\left(Z - E\left[Z\right]\right)U\right]}{Cov\left[Z, X\right]^2}\right) \end{aligned}$$

PROOF. *Outline. We have done a perfectly analogous proof for OLS in the first set of regression notes. As an exercise for yourself, follow the steps of that proof to fill in the details here.*

Consistency. With finite fourth moments, the sample covariances converge in probability to their population counterparts: $S_{Y,Z} \xrightarrow{P} \sigma_{Y,Z}$ and $S_{X,Z} \xrightarrow{P} \sigma_{X,Z}$. Consistency follows from applying the continuous mapping theorem.

Normality. Analogous to our proof of asymptotic normality in OLS, we express the IV estimator in a proof form

$$\hat{\beta}_1^{IV} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n \left(Z_i - \bar{Z}_n\right) U_i}{\frac{1}{n}\sum_{i=1}^n \left(Z_i - \bar{Z}_n\right) X_i}$$

Following exactly the same steps as we did in the previous derivation, we can show

$$\sqrt{n}\left[\hat{\beta}_1^{IV} - \beta_1\right] \xrightarrow{d} \frac{1}{Cov\left[Z, X\right]} N\left(0, Var\left[\left(Z - E\left[Z\right]\right)U\right]\right) = N\left(0, \frac{Var\left[\left(Z - E\left[Z\right]\right)U\right]}{\left(Cov\left[Z, X\right]\right)^2}\right)$$

□

Based on this asymptotic distribution, the approximate variance of the IV estimator equals

$$\sigma^2_{\hat{\beta}_1^{IV}} = \frac{1}{n}\frac{Var\left[\left(Z - E\left[Z\right]\right)U\right]}{\left(Cov\left[Z, X\right]\right)^2}$$

Note: $U = Y - \beta_0 - \beta_1 X$. It is not true that $U = \tilde{U} = Y - \beta_0 - \beta_1 \hat{X}$ as the two-stage least squares intuition suggests. This is to say that the estimator for $\sigma^2_{\hat{\beta}_1^{IV}}$ uses the actual residuals (based on Y and X) rather than the residuals extracted from the second stage regression. In practice, canned instrumental variables methods use a consistent estimator of $\sigma^2_{\hat{\beta}_1^{IV}}$ analogous to what we saw in OLS.

4.3.2. Simple Regression with an endogenous regressor and multiple instruments. Consider the case of two or more instruments for one endogenous regressor. For simplicity of exposition, suppose we have two valid instruments Z_1, Z_2. That is, these instruments sastify both the relevance and exogeneity conditions described in the previous section: $Cov\left[Z_i, X\right] \neq 0$ and $Cov\left[Z_i, U\right] = 0$.

To make the most out of our extra instrument, we form the best linear prediction of X given Z_1 and Z_2 in what is called the **first stage regression**.

$$X = \underbrace{\pi_0 + \pi_1 Z_1 + \pi_2 Z_2}_{=X^*} + W$$

In this formulation, X^* is like a "super-instrument." As long as the vector $(1, Z_1, Z_2)$ is not perfectly collinear, X^* is a more accurate predictor for X than either of the individual instruments. Moreover, because of the way we set things up, X^* inherits the properties of the individual instruments.

$$\begin{aligned} Cov[X^*, U] &= 0 \\ Cov[X^*, X] &\neq 0 \end{aligned}$$

Stated another way, we can write the $X = \Pi' Z + V$ where $\Pi = E[ZZ']^{-1} E[ZX]$. We can be sure that $Cov[X^*, X] \neq 0$, relevance is satisfied as long as $\pi_1 \neq 0$ or $\pi_2 \neq 0$. In practice, this suggests a hypothesis test for relevance in a first stage regression fit by OLS:

$$\begin{aligned} H_0 : \pi_1 = \pi_2 &= 0 \\ H_1 : \pi_i \text{ not all} &= 0 \end{aligned}$$

This test is an F-test of the form that we considered in previous parts of the class. To be sure that instrumental variables regression is a good procedure, we want to reject the null hypothesis that the instruments are not jointly relevant.

REMARK 4.3.2. **More than two instruments?** This test for relevance of the set of instruments extends naturally to l instruments. Just include all of the instruments in the first stage regression and conduct the joint test of the null hypothesis that all l coefficients on the instruments are equal to zero.

Moreover, not only do we want to reject this null hypothesis, but we want to reject it strongly. This is because the degree to which $Cov[X^*, X] \neq 0$ is important for identifying the effect of X on Y. If there is not much covariance between the fitted values in the first stage and the regressor, the IV procedure will be (at minimum) inefficient. In the worst case, a small F-statistic suggests the procedure will suffer from a weak instruments problem (where tiny deviations from exogeneity are exacerbated by the inefficiency of the estimation routine, resulting in a worse estimator than using OLS; more on this later).

4.3.2.1. *Identifying and Estimating* β_1. Just as we did before, we can use the exogeneity condition to solve for the parameter β_1 in terms of estimable features of the joint distribution.

$$\begin{aligned} Cov[X^*, Y] &= Cov[X^*, \beta_0 + \beta_1 X + U] \\ &= \beta_1 Cov[X^*, X] \end{aligned}$$

which leads to the expression $\beta_1 = \frac{Cov[X^*,Y]}{Cov[X^*,X]}$.

CLAIM 4.3.3. Just as in the single-instrument case, we can use this expression for β_1 and the analogy principle to motivate an estimator $\hat{\beta}_1^{IV} = \frac{S_{\hat{X},Y}}{S_{\hat{X},X}}$ where for practical considerations, we substitute the OLS fitted values $\hat{X}$ from the first-stage regression in for the best linear prediction of X, which we denote X^*. As we will see in more detail in the multiple regression setting, this two-stage least squares estimator is a natural extension of our IV regression intuition.

4.3.3. Multiple Regression: One Endogenous Regressor. Consider the multiple regression where U is a composite error term that contains unobserved determinants of Y:

$$Y = \mathbf{X}'\beta + U$$

Suppose that the orthogonality conditions hold for $i \geq 2$, $E[X_i U] = 0$, but not for $i = 1$, $E[X_1 U] \neq 0$. We call X_1 an **endogenous regressor** if it does not satisfy the orthogonality conditions.

Regressors that satisfy the orthogonality conditions are called **exogenous regressors**. Until we develop the idea of two-stage least squares in this setting, assume we only have one instrument. The simultaneous equations methodology suggests that we include a valid instrument to estimate the effect of X_1. In the multiple regression setting, a valid instrument must satisfy two properties:

(1) **Exogeneity**. $Cov[Z, U] = 0$. This condition is the same as in simple regression.
(2) **Relevance**. $Cov\left[\tilde{Z}, X_1\right] \neq 0$, where $\tilde{Z}$ is the residual variation in Z after partialling out all of the exogenous regressor variation.

Why has the relevance condition changed? The short intuition is that multiple regression imposes a Frisch-Waugh interpretation on the regression coefficients. In a phrase that shows up in *Mostly Harmless Econometrics*, multiple regression requires **covariate adjustment**. Hence, when we solve for β_1 in terms of the parameters, we obtain the expression

$$\beta_1 = \frac{Cov\left[\tilde{Z}, Y\right]}{Cov\left[\tilde{Z}, X_1\right]}$$

This expression has the most natural link with an extended version of the simulaneous system of equations we saw in simple regression:

$$\begin{aligned} Y &= \gamma_0 + \gamma_1 Z + \mathbf{X}_-'\gamma + U_1 \\ X_1 &= \alpha_0 + \alpha_1 Z + \mathbf{X}_-'\alpha + U_2 \end{aligned}$$

where $\mathbf{X}_-$ is the vector of regressors after dropping the endogenous regressor X_1. Like before, we can identify β_1 by taking the ratio $\frac{\gamma_1}{\alpha_1}$. The sample analog to this expression is called the covariate-adjusted IV estimator, which is precisely how to think about IV regression for one endogenous variable with one instrument in multiple regression.

4.3.3.1. *Using Orthogonality Conditions to Identify β.* An alternative (and equivalent) approach to obtaining an expression for β is to state the relevance and exogeneity conditions in a way that allows us to use them as we used orthogonality conditions to identify OLS. The instrument validity conditions can also be expressed as:

(1) **Exogeneity**. $E[ZU] = 0$. This implies that $E[\mathbf{W}U] = 0$ for $\mathbf{W} = \begin{pmatrix} 1 \\ Z \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$, which is analogous to the orthogonality conditions from the best linear predictor problem in our study of OLS.
(2) **Relevance**. $\mathbf{W} = \begin{pmatrix} 1 \\ Z \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$ is not perfectly collinear *and* in the regression $X_1 = \mathbf{W}'\alpha + V$, the coefficient on Z is not zero. In a derivation similar to one we used in multiple regression, this implies the identification assumption: $E[\mathbf{W}\mathbf{X}']^{-1}$ exists.

Like we did in single regression, we can use these conditions to characterize β:

$$0 = E[\mathbf{W}U] = E\left[\mathbf{W}\left(Y - \mathbf{X}'\beta\right)\right] = E[\mathbf{W}Y] - E\left[\mathbf{W}\mathbf{X}'\right]\beta$$

Solving and inverting, we obtain the identified parameter:

$$\beta = E\left[\mathbf{W}\mathbf{X}'\right]^{-1} E\left[\mathbf{W}Y\right]$$

As long as the fourth moments of the regressors and instruments are finite, the analogy principle implies that the IV estimator

$$\hat{\beta}^{IV} = \left(\mathbb{W}'\mathbb{X}\right)^{-1}\mathbb{W}'\mathbf{Y}$$

is consistent for β.

4.3.3.2. *Two-Stage Least Squares and IV Estimation.* To make this motivation for IV regression slightly more formal, consider the two-stage least squares approach. Given a random sample, we can consider the first stage regression in stacked form:

$$\mathbf{X}_1 = \mathbb{W}\alpha + \mathbf{V}$$

where $\mathbb{W}$ equals $\mathbb{X}$, but with a column of observations on Z in place of the column of observations on X_1. In this setup, $\hat{\alpha}^{ols} = \left(\mathbb{W}'\mathbb{W}\right)^{-1}\mathbb{W}'\mathbf{X}_1$. We could then obtain fitted values

$$\hat{\mathbf{X}}_1 = \mathbb{W}\left(\mathbb{W}'\mathbb{W}\right)^{-1}\mathbb{W}'\mathbf{X}_1 = P_W\mathbf{X}_1$$

using a projection onto the column space of the vector of exogenous variables (regressors and instrument). Consider the same projection for each column of $\mathbb{X}$ (for example $\hat{\mathbf{X}}_2 = P_W\mathbf{X}_2$) and this projection would perfectly predict the columns relating to exogenous regressors (i.e., $\mathbf{X}_2 = P_W\mathbf{X}_2$). Putting these two ideas together, we can form the second-stage data matrix $\tilde{\mathbb{X}}$ by projecting the entire data matrix $\mathbb{X}$ onto the column space of $\mathbb{W}$, $\tilde{\mathbb{X}} = P_W\mathbb{X}$. The second stage regression is:

$$\mathbf{Y} = \tilde{\mathbb{X}}\beta + \mathbf{U}$$

and we could form the OLS estimator using this modified data matrix, which we call the two-stage least squares estimator for β:

$$\begin{aligned}\hat{\beta}^{2SLS} &= \left(\tilde{\mathbb{X}}'\tilde{\mathbb{X}}\right)^{-1}\tilde{\mathbb{X}}'\mathbf{Y} \\ &= \left(\mathbb{X}'P_W\mathbb{X}\right)^{-1}\mathbb{X}'P_W\mathbf{Y}\end{aligned}$$

Because the number of instruments equals the number of endogenous regressors, we can expand this expression for the two-stage least squares estimator to obtain the common form for the instrumental variables estimator that we derived using the orthogonality conditions

$$\begin{aligned}\hat{\beta}^{2SLS} &= \left(\mathbb{X}'\mathbb{W}\left(\mathbb{W}'\mathbb{W}\right)^{-1}\mathbb{W}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbb{W}\left(\mathbb{W}'\mathbb{W}\right)^{-1}\mathbb{W}'\mathbf{Y} \\ &= \left(\left(\mathbb{W}'\mathbb{X}\right)^{-1}\left(\mathbb{W}'\mathbb{W}\right)\left(\mathbb{X}'\mathbb{W}\right)^{-1}\right)\mathbb{X}'\mathbb{W}\left(\mathbb{W}'\mathbb{W}\right)^{-1}\mathbb{W}'\mathbf{Y} \\ &= \left(\mathbb{W}'\mathbb{X}\right)^{-1}\mathbb{W}'\mathbf{Y} = \hat{\beta}^{IV}\end{aligned}$$

That said, the two-stage least squares estimator is more general than simple IV as we discuss at length in the next section.

4.3.4. Multiple Instruments and Multiple Endogenous Regressors. The derivation and form of the 2SLS estimator $\hat{\beta}^{2SLS}$ presented in the previous section is completely general in the sense that the method works for a setting with multiple instruments *and* multiple endogenous regressors. To generalize our 2SLS procedure, suppose we have h instruments $(Z_1, Z_2, ..., Z_h)$ for m endogenous regressors $(X_1, X_2, ..., X_m)$ among k total regressors with the remaining $k-m$ regressors exogenous $(X_{m+1}, ..., X_k)$.

In line with the previous derivation, create the first-stage matrix $\mathbb{W}$ by deleting the m columns of the usual data matrix corresponding to endogenous regressors (columns 2 through $m+1$), replacing those m columns with h columns that correspond to data on the h instruments.

Then, just as before, the first stage involves projecting $\mathbb{X}$ onto the column space of $\mathbb{W}$. In the second stage, we use these projected

$$\hat{\beta}^{2SLS} = \left(\mathbb{X}' P_W \mathbb{X}\right)^{-1} \mathbb{X}' P_W \mathbf{Y}$$

The only restriction we will place on this setting is that the set of instruments is sufficiently predictive of the set of endogenous regressors. If we do not have as many instruments as endogenous regressors, there is no way to identify a separate effect for each regressor. Hence, two-stage least squares can only be applied successfulluy in two cases: (1) **Just Identified** – when the number of intsruments equals the number of endogenous regressors and (2) **Overidentified** – when the number of instruments exceeds the number of endogenous regressors.

4.3.4.1. *Multiple Instruments for one endogenous regressor (in multiple regression).* As a simple demonstration of how the derivation extends to multiple instruments, suppose we have multiple instruments for one endogenous regressor in multiple regression. For identification of the two-stage least squares estimator, there are two important considerations:

(1) If the first stage regression does not exhibit perfect multicollinearity, $(\mathbb{W}'\mathbb{W})^{-1}$ will exist, allowing us to form fitted values for the endogenous regressor.
(2) If the set of instruments in the first stage regression adds to the predictive power of the exogenous regressors for the endogenous regressor, the fitted values we substitute into $\tilde{\mathbb{X}}$ will not be perfectly collinear with the columns of exogenous in the second stage. That is, the inverse in the second stage regression will exist and the 2SLS procedure makes sense.

This second consideration suggests a natural test for instrument relevance that parallels what we saw with multiple instruments in single regression.

REMARK 4.3.4. **Testing for Relevance with Multiple Instruments**. More formally, we could use an F-test to test the joint significance of the instruments in the first stage to determine if they add to the predictive power of the exogenous regressors enough to identify an independent effect of X_1. That is, for a first stage regression

$$X_1 = \alpha_0 + \alpha_1 Z_1 + ... + \alpha_h Z_h + \alpha_{h+1} X_2 + ... + \alpha_{h+k-1} X_k + V$$

the appropriate null hypothesis to test for whether the set of instruments is relevant is $H_0: \alpha_1 = \alpha_2 = ... = \alpha_h = 0$. As long as we strongly reject this null hypothesis, we are confident that there is enough variability in the fitted values of X_1 to identify an independent effect of X_1 on Y.

4.3.4.2. *Relevance with Multiple Endogenous Regressors.* In the previous subsection, we focused on the case of one endogenous regressor, which simplified testing for relevance. All we needed to do was conduct a joint hypothesis test on whether the set of instruments is significant when controlling for the exogenous regressors.

With multiple endogenous regressors, whether the instruments are relevant for the regressors becomes more complicated. To see why, suppose we have two endogenous regressors. The first stage now involves two regressions:

$$\begin{aligned} X_1 &= \pi_{01} + \pi_{11}Z_1 + \pi_{21}Z_2 + ... + \pi_{h1}Z_h + \pi_{(h+1)1}X_3 + .. + \pi_{(h+k-2)1}X_k + U_1 \\ X_2 &= \pi_{02} + \pi_{12}Z_1 + \pi_{22}Z_2 + ... + \pi_{h2}Z_h + \pi_{(h+1)2}X_3 + .. + \pi_{(h+k-2)2}X_k + U_2 \end{aligned}$$

To emulate the hypothesis test for instrument relevance with one endogenous regressor, we might try conducting a joint test for statistical significance of the instruments within each equation.

The problem with this procedure is that it does not necessarily identify variation in X_1 that is distinct from X_2. For this reason, this procedure is not our test for relevance. On an intuitive level, one could reject both null hypotheses for the same reason – that is, using the same variability in the instruments. Hence, the resulting fitted values $\hat{X}_1$ and $\hat{X}_2$ would not residual variation from one another.

Stata's default test for relevance is called the minimum eigenvalue statistic, which is often called the Cragg-Donald Statistic (`estat first` provides the standard output for this). There are a variety of other tests for relevance, which have been summarized by Stock, Wright and Yogo (2002). The key point underlying all of this discussion is that the first stage should be highly predictive of the set of endogenous regressors. If you are interested in reading more about the underlying details, the Stock, Wright and Yogo paper – though it is advanced and contains terminology we will not have time to study – is worth a read.[1]

4.4. Caveats about IV Regression

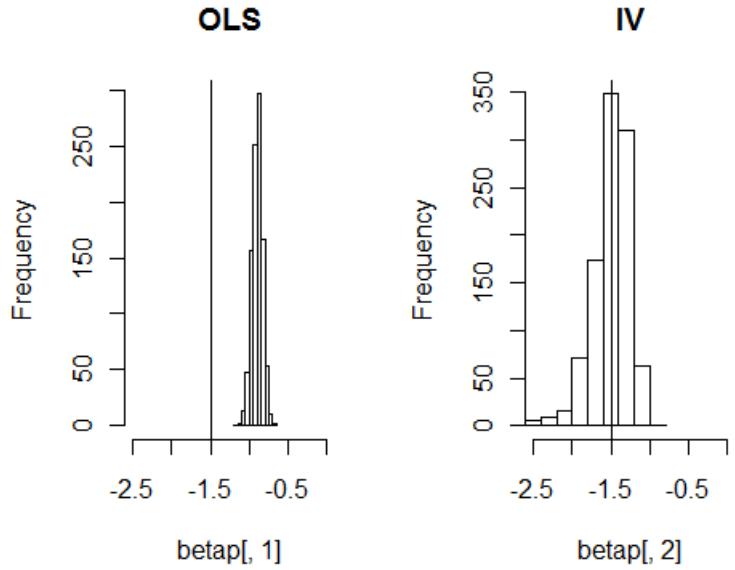
As we saw in the previous section, instrumental variables regression can be a useful tool for learning causal effects when we have no hope for controlling for all omitted variables. Apart from the always difficult task of obtaining a valid instrument, you should be aware of the costs and common pitfalls that arise when using IV regression. We will have time to cover four categories of these costs: (1) Sacrificing efficiency for consistency, (2) weak instruments, (3) forbidden regressions and (4) heterogeneity and the local average treatment effect interpretation of IV regression.

4.4.1. Efficiency versus consistency tradeoff. If you were thinking carefully about how IV regression works, it should come as no surprise to you that IV regression gives up on the goal of obtaining an efficient estimator relative to OLS. After all, the whole point of the first stage is to discard problematic variability (specifically the part of X that is correlated with U) from of the endogenous regressors. In OLS, additional variability in our regressors allows us to obtain estimators of the effect of X on Y that are more precise. Throwing away variability amounts to using less information and an estimator that uses less information is less precise.

Caveat about efficiency aside, as long as the instrument is relevant, the resulting IV estimator will do a better job estimating the fundamental parameter. Figure 4.4.1 displays the results from a Monte Carlo experiment where the underlying slope parameter on an endogenous variable is -1.5. As we can see from the histograms, OLS is biased and inconsistent while being quite precise. On the other hand, IV appears to estimate the slope parameter quite well. In this instance, we may be happy to give up some efficiency to correct the omitted variable bias. In other cases, especially when the first stage is weak and the omitted variable bias is not "so bad," resorting to IV may not be worth it.

[1] `http://www.nber.org/~myogo/papers/published/JBES1002.pdf`

FIGURE 4.4.1. Efficiency Versus Consistency – An Important Tradeoff Between OLS and IV



Histograms computed using a Monte Carlo Simulation (using $B = 1000$ simulated data sets) where the slope parameter on the endogenous variable equals -1.5. Key point: OLS is more efficient, but its sampling distribution is centered on the wrong parameter value.

4.4.2. Weak Instruments and the Bias of 2SLS. We motivated instrumental variables methods by demonstrating that OLS is biased and inconsistent for the causal β. As the 2SLS estimator consistent and asymptotically normal, it seems that 2SLS would be strictly preferred to OLS if we hope to make causal inference. It turns out that 2SLS is biased for finite samples and that it is more biased when the set of instruments is weak (i.e., when the relevance condition is not satisfied in practice). Together with the inefficiency of 2SLS relative to OLS, weak instruments often means that we would be better off using OLS. The catch phrase that describes this best was the title of an important paper on this topic – "The Cure Can Be Worse Than the Disease: A Cautionary Tale Regarding Instrumental Variables" by Bound, Jaeger and Baker.[2]

4.4.2.1. *Why Can the Cure Be Worse Than the Disease?* To see why 2SLS is biased, take the extreme case where the instruments have *no true relationship* with the endogenous regressors. By mere accident, the first stage regression will select some idiosyncratic variability in $\mathbb{X}$ to fit (in practice R^2 is never zero), but when the instruments are not related to the endogenous regressors, this variability is representative of the usual variability in $\mathbb{X}$. On an intuitive level, this representative component of $\mathbb{X}$ has precisely the same small-sample bias properties as OLS, which we already know is biased for the causal parameter. Idiosyncratic fit of the endogenous regressors to the instruments will always contribute this source of bias, but as the systematic fit of the first stage becomes more important, idiosyncratic first-stage fit is weighted less and the bias reduces.

This intuition holds up after applying some rigor. Start with our expression for the 2SLS estimator (and transform it into "proof form" by substituting the stacked regression):

[2] `http://ideas.repec.org/p/nbr/nberte/0137.html`

$$\begin{aligned}\hat{\beta}^{2SLS} &= \left(\mathbb{X}'P_W\mathbb{X}\right)^{-1}\mathbb{X}'P_W\mathbf{Y} \\ &= \beta + \left(\mathbb{X}'P_W\mathbb{X}\right)^{-1}\mathbb{X}'P_W\mathbf{U}\end{aligned}$$

Next, recognize that the first stage regression gives an expression for $\mathbb{X}$:

$$\mathbb{X} = \mathbb{W}\gamma + \mathbf{V}$$

Substitute this expression into the second part of the proof form of $\hat{\beta}^{2SLS}$

$$\hat{\beta}^{2SLS} = \beta + \left(\mathbb{X}'P_W\mathbb{X}\right)^{-1}\left(\mathbb{W}\gamma + \mathbf{V}\right)'P_W\mathbf{U}$$

All of this implies that we can express the bias of 2SLS as

$$E\left[\hat{\beta}^{2SLS} - \beta\right] = E\left[\left(\mathbb{X}'P_W\mathbb{X}\right)^{-1}\gamma'\underbrace{\mathbb{W}'P_W}_{\mathbb{W}'}\mathbf{U}\right] + E\left[\left(\mathbb{X}'P_W\mathbb{X}\right)^{-1}\mathbf{V}'P_W\mathbf{U}\right]$$

At this point, we can apply the *better-than-asymptotic* approximations

$$\begin{aligned}first.term = E\left[\left(\mathbb{X}'P_W\mathbb{X}\right)^{-1}\gamma'\mathbb{W}'\mathbf{U}\right] &\approx E\left[\left(\mathbb{X}'P_W\mathbb{X}\right)\right]^{-1}E\left[\gamma'\mathbb{W}'\mathbf{U}\right] \\ second.term = E\left[\left(\mathbb{X}'P_W\mathbb{X}\right)^{-1}\gamma'\mathbb{W}'\mathbf{U}\right] &\approx E\left[\left(\mathbb{X}'P_W\mathbb{X}\right)\right]^{-1}E\left[\mathbf{V}'P_W\mathbf{U}\right]\end{aligned}$$

See Angrist and Pischke (p 207) for details on this approximation and references to the literature. This result relies on letting the number of instruments and number of observations go to infinity, while preserving the instrument-to-observation ratio at some finite constant. Convergence along this path *feels* more like the problem we are facing with weak instruments. Hence, this approximation provides a decent step toward understanding the bias of 2SLS (even though it is inevitably asymptotic).

The 2SLS bias is the sum of these two terms. The first pertains to the exogeneity condition while the second pertains to relevance. Suppose that the exogeneity condition is satisfied. Then, we know that the first term in this expression for bias equals zero (because $E\left[\gamma'\mathbb{W}'\mathbf{U}\right] = 0$). Under this restriction, the second term

$$E\left[\hat{\beta}^{2SLS} - \beta\right] \approx E\left[\left(\mathbb{X}'P_W\mathbb{X}\right)\right]^{-1}E\left[\mathbf{V}'P_W\mathbf{U}\right]$$

is the bias of 2SLS. From this expression, we see that 2SLS bias is due to there being a correlation between the second stage error term $\mathbf{U}$ and $P_W\mathbf{V}$, a term that represents idiosyncratic fit of $\mathbb{X}$ to the set of instruments $\mathbb{W}$, which we argued at the outset to be representative of the usual variability in $\mathbb{X}$. After some manipulation (see Angrist and Pischke for details), this expression for the bias of 2SLS reduces to

$$E\left[\hat{\beta}^{2SLS} - \beta\right] \approx \frac{Cov\left[V,U\right]}{Var\left[V\right]}\left[\frac{1}{F+1}\right]$$

where F is the F-statistic on the test for relevance. This bias term is the same as the bias of OLS when F is zero and it reduces as we more strongly reject the null hypothesis that the set of instruments is not relevant.

Remark 4.4.1. That the 2SLS estimator is biased in the direction of OLS does not imply that 2SLS estimator is *better* than OLS. Its expectation is closer to the true value of the parameter, but we sacrificed efficiency to make this happen. Moreover, a weaker first stage means that 2SLS has worse bias properties *and* it implies that the 2SLS procedure is less efficient. Putting together these two undesirable aspects of weak instruments, we may obtain an estimator with significantly worse

TABLE 1. Mean, Standard Deviation and MSE of Slope Coefficient Estimates (1000 Replications, $N = 150$)

	$OLS(strong)$	$IV(strong)$	$OLS(weak)$	$IV(weak)$
mean	4.2568	3.9637	4.2765	3.5219
stddev	0.0163	0.2349	0.015	61.1457
MSE	0.0662	0.0565	0.0767	3739.0219

TABLE 2. Mean, Standard Deviation and MSE of Slope Coefficient Estimates (1000 Replications, $N = 15000$)

	$OLS(strong)$	$IV(strong)$	$OLS(weak)$	$IV(weak)$
mean	4.2571	3.9999	4.2766	3.958
stddev	0.0016	0.01	0.0015	0.3052
MSE	0.0661	$1e-04$	0.0765	0.0949

mean squared error. As we see in the simulations in the next section, this property holds even to large sample sizes ($N = 15000$).

4.4.2.2. *An Example: Simulating 2SLS Bias.* The R script file `weakinstruments.R`,[3] constructs 1000 replications of data sets using the data generating process:[4]

Data Generating Process

$$\begin{aligned} Y &= 5 + 4X + 6Q \\ X &= sd_X M + sd_Q Q + sd_Z Z \end{aligned}$$

where each replication takes N draws M, Q, Z, which are each $N(0, 1)$ RVs that are independent of one another. Given this overarching meta-process, the parameter values I picked for relatively strong and relatively weak are as follows:
Relatively Strong: $sd_Z = 5$, $sd_M = 10$, and $sd_Q = 15$
Relatively Weak: $sd_Z = 0.5$, $sd_M = 10$, and $sd_Q = 15$

Tables 1 ($N = 150$) and Table 2 ($N = 15000$) display the results from estimating β_1 from the regression model

$$Y = \beta_0 + \beta_1 X + U$$

on the simulated data from this process using OLS and 2SLS. Within each table, we would rather use IV with strong instruments than with weak instruments. For both of the sample sizes I picked for the simulation, IV on weak instruments has higher MSE than does OLS,[5] but it is *much* worse for the small sample size. Why? The larger sample size increases first stage F statistic for instrument relevance (reducing the true weakness of the instrument).

This feature of the simulation results is difficult to see in the density plots of the F-statistics (Figures 4.4.2 and 4.4.3), but if we compute the mean F statistic within each simulation (Table 3), it becomes apparent. The typical rule of thumb is to proceed as if there is no problem with relevance only

[3]Available on the notes website. Play around with the parameters if you like.

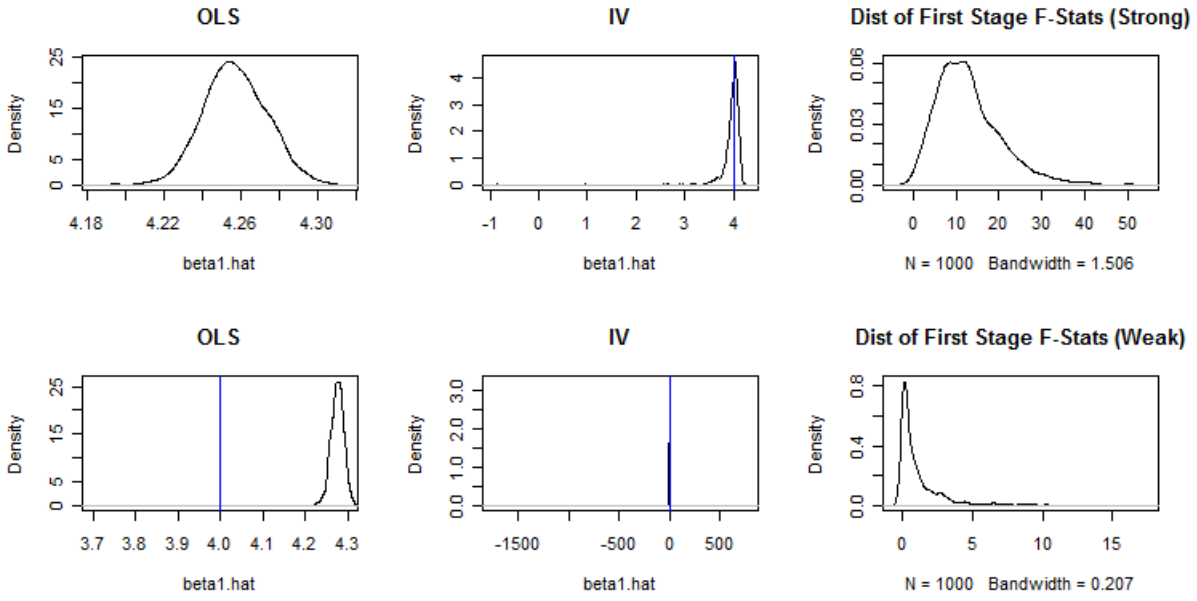
[4]Note: I could have also added an unexplainable systematic component W to this first term to represent the errors in the second stage that have nothing to do with anything, but it wasn't necessary to get the right properties to fall out of the simulation.

[5]Weak instruments versus strong instruments shouldn't matter for OLS and it doesn't in the simulations.

TABLE 3. Average F-statistic in Each Set of Replications

	Relatively Strong	Relatively Weak
$N = 150$	12.823	1.118
$N = 15000$	1156.976	13.889

FIGURE 4.4.2. Density Plots for OLS and IV estimates (1000 Replications, $N = 150$)

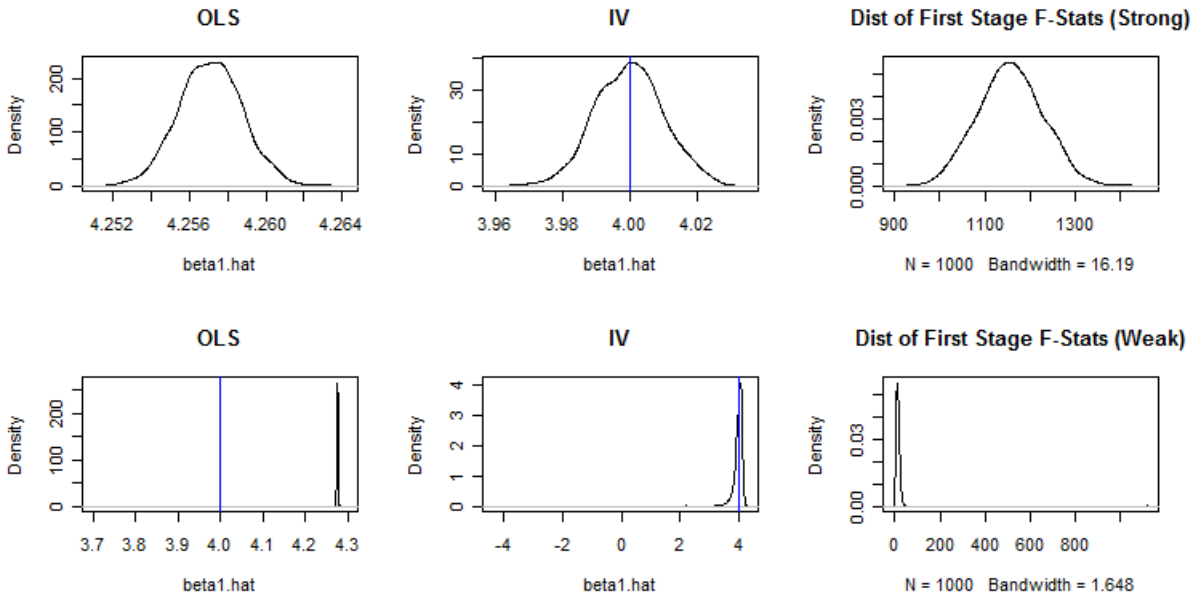


if the F-statistic exceeds 10. This is good advice for small and large samples alike. Although a large sample may make the first stage more relevant, it does not improve the properties of the 2SLS estimator for a given value of the F statistic.

4.4.3. Caution: Implementation Details. There are several points of caution when it comes to implementing instrumental variables methods in practice.

(1) Although motivating IV regression as a two-step process is nice for the intuition of how to construct the estimates (and it constructs the right estimates), it produces the wrong standard errors because the second stage fitted values are – by default – constructed using $\hat{X}$ rather than X. The right thing to do is to use X to compute the fitted values and residuals.
(2) Running 2SLS regression in two stages also bears the risk of user error.
   (a) It is important to note that the first stage regression includes all exogenous variables in the model (including exogenous regressors), not just the instruments.
   (b) When assessing relevance, one should not merely look at the fit of the first stage as measured by R^2, but should only assess whether the instruments significantly predict the endogenous regressors above and beyond the predictive capability of the exogenous regressors. This should concord with your intuition on Frisch-Waugh interpretation of regression coefficients.
(3) **Forbidden Regressions**. If the first stage involves a nonlinear model like probit or logit, it is not valid to plug the fitted values into the second stage to the same degree that it is valid to use OLS in two stages.

FIGURE 4.4.3. Density Plots for OLS and IV estimates (1000 Replications, $N = 15000$)



This is because – even when OLS gives the best linear approximation to a nonlinear causal relationship – OLS is guaranteed to provide residuals that are uncorrelated with the fitted values while nonlinear models like probit and logit are not.

Suppose you have an endogenous dummy variable D in the regression model $Y = \mathbf{X}'\beta + D\alpha + U$. A simple fix is to (1) construct the fitted values $\hat{D}$ using the nonlinear model $D = F(\mathbf{W}'\beta + \xi)$, (2) use the fitted values $\hat{D}$ in a 2SLS procedure to instrument for D. At this step, an important reminder is that we need to construct the first stage estimates in an OLS regression that controls for all exogenous variables in the model:

$$D = \pi_0 + \pi_1\hat{D} + \mathbf{M}'\gamma + V$$

where $\mathbf{M}$ is a vector that contains all exogenous variables in the model (including $\mathbf{X}$ and $\mathbf{W}$).

4.4.4. LATE (Local Average Treatment Effects). All of the previous discussion of IV regression methods has assumed that there is one effect of the regressor on the response variable, called β. It is possible that there is unobserved or unmodeled heterogeneity in the effects across groups.

4.4.4.1. *Intuition in an example.* The effect of increasing education by one year from 11 to 12 years may be quite different than the effect of increasing education by one year from 13 to 14 years. Our instrumental variables methods rely on selective variation, and hence, may only pick up on a subset of the possible effects, which are averaged into the estimate.

For example, compulsory schooling laws are often used as instruments for educational attainment. There is good evidence that laws affect educational attainment where they are binding (perhaps, forcing some individuals to obtain 12 rather than 11 years of education), but the resultant instrument will screen all other variability out of an educational attainment variable.

Depending on our view of the nature of the heterogeneity of the effects of education, we may only be able to cleanly identify the effect of obtaining 12 years of education instead of 11 years using variation in compulsory schooling. In a very precise sense, the instrumental variable technique only allows us to estimate the local average treatment effect of this "experiment."

That said, if we can make a case that the effect of additional educational attainment caused by the instrument is representative of the usual effects of education, the instrumental variables estimates become more interesting. Making such a case is usually more about economic modeling and constructing a compelling arument than rigorous statistical evidence.

4.4.4.2. *A slightly more formal/resctrictive treatment of IV.* Consider the case of an endogenous regressor X as the only explanatory variable for Y. The relationship of interest – second stage – is given by causal model where U contains unobserved determinants of Y.

$$Y_i = \beta_0 + \beta_1^i X_i + U$$

where the notation β_1^i is the effect of X on Y, which differs across individuals i in an unobserved way.

Suppose we have a dummy variable instrument Z. In addition to what we required when there was one effect, we need to distinguish two parts of the relevance condition: the exclusion restriction and the first-stage relevance condition. We also want to strengthen the exogeneity condition to what is known as the independence assumption.

- The **exclusion restriction** is that Z does not have its own causal effect on Y. That is, in the full-blown causal regression model (controlling for all variables that affect Y), the coefficient on Z in the causal regression is zero.
- The **independence assumption** implies that the realized values of Z are as good as randomization if we want to argue that Z causes X and Z causes Y in both reduced form regressions.

$$\begin{aligned} Y &= \pi_0 + \pi_1 Z + U^1 \\ X &= \gamma_0 + \gamma_1 Z + U^2 \end{aligned}$$

- As before, the **relevance condition** implies that Z is correlated with X.

In other words, it would be an ideal setting for instrumental variables if we can make a compelling case that Z causes X, which in turn, causes Y (and Z could not possibly cause Y without affecting X indirectly).

Claim 4.4.2. Using our system-of-equations motivation for instrumental variables, the fact that Z is a dummy variable implies that the heterogeneous effect for persion i is identified by:

$$\hat{\beta}_1^{IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1} \xrightarrow{P} \frac{E[Y_i|Z=1] - E[Y_i|Z=0]}{E[X_i|Z=1] - E[X_i|Z=0]}$$

The problem with converting this expression into an estimator as we have done before is that we rarely observe (X_i, Y_i) for both $Z = 0$ and $Z = 1$. Usually, that would require two observations on the same individual. Fortunately, the conditions on the instrument imply that we treat the differences in these expectations (conditional on Z) as causal.

4.4.4.3. *Understanding the Probability Limit: LATE.* Potential outcomes notation is useful in understanding this probability limit in terms of treatment effects. To bring potential outcomes notation into the problem, let X_i equal $X_i(Z=1)$ if the individual is *treated* and $X_i(Z=0)$ if the individual is *not treated*. These $X_i(\cdot)$ are called **potential outcomes** and their difference $X_i(Z=1) - X_i(Z=0)$ is called the **treatment effect** of Z on X_i.

We can analogously define a treatment effect for Y_i, but like the second stage regression, this treatment effect depends on the value of X_i as in $Y_i(X_i = 1) - Y_i(X_i = 0)$. In general, the treatment effect of Y_i depends on the value of Z as well, but the exclusion restriction allows us to write this dependence on X alone.

For simplicity, suppose that X_i is a dummy variable. In this case, it is easy to relate actual outcomes, potential outcomes and treatment effects:

$$Y_i = Y_i(X_i = 0) + \underbrace{(Y_i(X_i = 1) - Y_i(X_i = 0))}_{=treatment.effect=\beta_1^i} X_i$$

Then, returning to the numerator of the probability limit, we could substitute this potential outcomes notation in for Y_i.

$$\begin{aligned} E[Y_i|Z = 1] &= E\left[Y_i(X_i = 0) + \beta_1^i X_i|Z = 1\right] \\ E[Y_i|Z = 0] &= E\left[Y_i(X_i = 0) + \beta_1^i X_i|Z = 0\right] \end{aligned}$$

The independence condition technically means that potential outcomes are independent of the realizations of the instrument. As a practical matter, applying the independence condition implies that the previous equalities are actually equal to

$$\begin{aligned} E[Y_i|Z = 1] &= E\left[Y_i(X_i = 0) + \beta_1^i X_i(Z = 1)\right] \\ E[Y_i|Z = 0] &= E\left[Y_i(X_i = 0) + \beta_1^i X_i(Z = 0)\right] \end{aligned}$$

Then taking the difference between these to form the numerator, we obtain:

$$E[Y_i|Z = 1] - E[Y_i|Z = 0] = E\left[\beta_1^i(X_i(Z = 1) - X_i(Z = 0))\right]$$

At this point in the derivation, another technical condition comes into play called **monotonicity** (or uniformity). This condition is that the instrument affects everyone's X_i in the same direction. As one case, suppose that $X_i(Z = 1) \geq X_i(Z = 0)$ for all individuals.[6] If this is true, then $X_i(Z = 1) - X_i(Z = 0)$ is either zero or one, so we can rewrite the expression as.

$$E[Y_i|Z = 1] - E[Y_i|Z = 0] = E\left[\beta_1^i|X_i(Z = 1) \geq X_i(Z = 0)\right] P[X_i(Z = 1) \geq X_i(Z = 0)]$$

If we apply similar logic and algebra to the denominator, we obtain:

$$E[X_i|Z = 1] - E[X_i|Z = 0] = P[X_i(Z = 1) \geq X_i(Z = 0)]$$

Hence, this argument shows formally that

$$\hat{\beta}_1^{IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1} \xrightarrow{P} E\left[\beta_1^i|X_i(Z = 1) \geq X_i(Z = 0)\right]$$

In words, the IV estimator – when there are heterogeneous effects – represents the average treatment effect among individuals whose potential outcomes would be changed if they receive the treatment of the instrument. The group of individuals for whom $\{X_i(Z = 1) \geq X_i(Z = 0)\}$ is called the **compliant subpopulation** because changing the value of the instrument changes their choices. This result tells us that the instrumental variables estimator consistently estimates the average treatment effect among the compliant subpopulation, or the **Local Average Treatment Effect** (LATE).

[6]There is only one other case that the monotonicity assumption allows (where $X_i(Z = 1) \leq X_i(Z = 0)$ for all individuals)

4.4.4.4. *Internal and External Validity of IV Estimation.* In a world with heterogeneity of effects, instrumental variables methods will estimate the LATE. If you find yourself in this setting, there are two natural ways to go. First, you could argue that the LATE you estimate is interesting in its own right. In the language of Angrist and Pischke, this is what it means for your IV estimator to be internally valid. In this derivation, we discovered that the LATE interpretation follows from a monotonicity assumption on the nature of the heterogeneity (that the instrument affects choices of the regressor in the same direction). As long as monotonicity holds, LATE is an internally-valid interpretation.

REMARK 4.4.3. Returning to a setting where there is a uniform effect of X on Y called β_1, the LATE equals this single effect β_1. Hence, the results are valid externally in addition to the internal validity afforded by the instrumental variables method. To return to the motivating example, if there is one effect of an additional year of education and we use an instrument that compels some people to get 12 years rather than 11 years of education to estimate the effect of education, the single-effect assumption (feature) allows us to generalize the result to argue that the causal effect of getting 16 years of education rather than 15 years is the same as we estimated using our instrument. That's not necessarily a valid exercise, but this discussion of LATE makes precise that the reason this is not valid is because of heterogeneity in the causal effects.

4.5. Chapter Exercises

(1) **Attenuation Bias Simulation**. Consider the regression model

$$Y = \beta_0 + \beta_1 X + W$$

where we have two measurements of X, measured with error

$$\begin{aligned} X_1 &= X + V_1 \\ X_2 &= X + V_2 \end{aligned}$$

such that $Cov[X, V_1] = Cov[X, V_2] = Cov[V_1, V_2] = 0$. Define the

$$X_{ave} = \frac{X_1 + X_2}{2}$$

(a) Write R code to simulate this data generating process where

$$\begin{aligned} \beta_0 = 5, \quad & \beta_1 = 2 \\ X \sim N\left(0, \sigma_X^2 = 25\right) \quad & W \sim N\left(0, \sigma_W^2 = 4\right) \\ Var[V_1] = 2 \quad & Var[V_2] = 9 \end{aligned}$$

and the sample size $N = 200$.

(b) For $B = 1000$ draws from this data generating process (each a data set containing X, Y, X_1 and X_2), compute the OLS slope estimate $\hat{\beta}_1^{ols}$ and store it for each of the following specifications:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + U \text{ (correct)} \\ Y &= \beta_0 + \beta_1 X_1 + U \text{ (more precise mismeasurement)} \\ Y &= \beta_0 + \beta_1 X_2 + U \text{ (less precise mismeasurement)} \\ Y &= \beta_0 + \beta_1 X_{ave} + U \text{ (average of measurements)} \\ Y &= \beta_0 + \beta_1 \hat{X}_1 + U \text{ (more precise regressed on less precise)} \\ Y &= \beta_0 + \beta_1 \hat{X}_2 + U \text{ (less precise regressed on more precise)} \end{aligned}$$

Note: $\hat{X}_1 = \hat{\gamma}_0^{ols} + \hat{\gamma}_1^{ols} X_2$ and $\hat{X}_2 = \hat{\alpha}_0^{ols} + \hat{\alpha}_1^{ols} X_1$.

(c) Summarize the simulated sampling distributions of each of these methods for computing an estimator for $\hat{\beta}_1$ using the mean estimate, the standard deviation of the estimate and the MSE, where $bias = mean\left(\hat{\beta}_1\right) - 2$.

(d) Plot the histograms of each simulated sampling distribution. If you could not use a perfect measurement of X, what is the next best estimator that you would use in this setting?

(2) **Fertility in Botswana**. The data set `fertility.dta` is available on the website, which contains information on all of the variables you will need for this question. The data provide information on 4361 women in Botswana during 1998.

Of note, *children* is the number of children that the women has, *educ* is years of education, *age* is age of the woman and $frsthalf$ is a dummy variable for whether the woman was born in the first half of the year.

(a) Use the fertility data to estimate the linear regression model

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + U$$

using OLS. Interpret the coefficient estimate on β_1. In a causal interpretation, is *educ* exogenous? Discuss.

(b) Use instrumental variables to estimate the regression model with $frsthalf$ as an instrument for *educ*. How does the coeficient estimate change when we use IV relative to OLS?

(c) Is $frsthalf$ a relevant instrument for *educ*? Provide evidence using output from an appropriate regression.

(d) Evaluate whether the potential endogeneity problem can be solved by adding more variables to the regression model. Specifically, the data set has an interesting set of dummy variables: *usemeth*, *urban*, *evermarr*, *tv*, *radio*, *electric*, *bicycle*, *spirit*, *protest* and *catholic*. How does the inclusion of all of these variables change the coefficient estimate $\hat{\beta}_1$?

(e) Compare the coefficient estimates in (b) and (d). What does this comparison imply about the nature of the omitted variable bias in (d)?

Bibliography

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Casella, G. and R. L. Berger (2002). *Statistical Inference* (Second ed.). Thompson Learning.

Greene, W. H. (2003). *Econometric Analysis* (Fifth ed.). Prentice Hall.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, J. M. (2003). *Introductory Econometrics: A Modern Approach*. Thompson South-Western.