PhD Core Econometrics II

First edition

David M. Kaplan

First edition: Spring 2023
Updated February 21, 2023

However, the second time round, she came upon a low curtain she had not noticed before, and behind it was a little door about fifteen inches high: she tried the little golden key in the lock, and to her great delight it fitted! Alice opened the door and found that it led into a small passage, not much larger than a rat-hole: she knelt down and looked along the passage into the loveliest garden you ever saw. How she longed to get out of that dark hall, and wander about among those beds of bright flowers and those cool fountains, but she could not even get her head through the doorway...

Lewis Carroll, *Alice's Adventures in Wonderland*

Brief Contents

Contents

Preface

This text was prepared for the 15-week 2nd-semester core PhD econometrics course at the University of Missouri. The main focus is identification, from perspectives of both structural models and potential outcomes, using conditional independence, instrumental variables (IV), or panel data. Additionally, the generalized method of moments (GMM) is discussed (after the special case of IV), as well as maximum likelihood. Probit/logit models are also presented, which allows introduction of important concepts for any nonlinear models.

The assumed background is the first-semester core PhD econometrics at the University of Missouri, which uses (roughly) the first nine chapters of Hansen (2020a) and related material from Hansen (2020b).

As with my *Introductory Econometrics* (Kaplan, 2022a) and *Distributional and Nonparametric Econometrics* (Kaplan, 2021), this text's source files are freely available. Instructors may modify them as desired, or copy and paste LaTeX code into their own lecture notes, with usage subject to the Creative Commons license linked on the copyright page. I wrote the text in Overleaf, an online (free) LaTeX environment that includes knitr support. You may see, copy, and download the entire project from my website.[1]

Another unusual feature is the prevalence of in-class discussion questions. I find these very helpful (for more actively engaging students, for gauging how students are tracking, and for breaking up my lecturing), and students seem to appreciate them, too.

Thanks to everyone for their help and support: my past econometrics instructors, my colleagues and collaborators, my students, and my family.

David M. Kaplan
Spring, 2023
Columbia, Missouri, USA

[1] https://kaplandm.github.io/teach.html

Textbook Learning Objectives

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

The textbook learning objectives (TLOs) are the following.

1. Define terms and concepts, both mathematically and intuitively.
2. Develop intuition for fundamental concepts to enable you to understand econometrics papers/books that you need to read later for your own research.
3. Describe various econometric methods both mathematically and intuitively, including their objects of interest and assumptions, and the logical relationship between the assumptions and corresponding theorems and properties.
4. For a given economic question, dataset, and econometric method, judge whether the method is appropriate and assess the economic significance and statistical significance of the results.
5. Use Stata to manipulate and analyze data, interpreting results both economically and statistically.

Notation

Variables

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

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

and a k-dimensional non-random vector is

$$\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{pmatrix}$$

Unless otherwise specified, vectors are column vectors. The transpose of a column vector is a row vector. For example, using the $\boldsymbol{z}$ defined above,

$$\boldsymbol{z}' = (z_1, z_2, \ldots, z_k)$$

Note: displayed math like above should always have appropriate punction (comma, period) at the end! . . . unless you are defining notation and worry about confusing people.

Greek letters like $\boldsymbol{\beta}$ and θ generally denote fixed population parameters.

I sometimes make exceptions to match convention. For example, ϵ is a Greek letter but is conventionally used for a regression error term or white noise.

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

I try to put "hats" on other quantities computed from the sample, too. For example, a t-statistic would be $\hat{t}$ (a random variable computed from the sample) instead of just t (which looks like a non-random scalar). Or, a J-statistic would be $\hat{J}$, even though J is already uppercase, to emphasize that it is computed from data (rather than data itself).

Besides hats, tildes and bars may indicate estimators of parameters, and bars indicate sample averages. For example, there may be multiple alternatives for estimating θ: $\hat{\theta}$, $\tilde{\theta}$, and $\bar{\theta}$. The sample average of $Y_1, \ldots, Y_n$ is $\bar{Y}$.

Estimators and other **statistics** (i.e., things computed from data) may sometimes have a subscript with the sample size n to remind us of the asymptotic perspective of a sequence (indexed by n) of random variables. For example, with n denoting sample size, $\hat{\theta}_n$, $\hat{t}_n$, and $\bar{Y}_n$.

The following is a summary.

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

Symbols

In addition to the following symbols, vocabulary words and abbreviations (like "quantile" or "IVQR") can be looked up in the Index in the very back of the text.

$\Longrightarrow$	implies; see Chapter 2
$\Longleftarrow$	is implied by; see Chapter 2
$\Longleftrightarrow$	if and only if; see Chapter 2
$\lim_{n\to\infty}$	limit
$\operatorname{plim}_{n\to\infty}$	probability limit
$\to$	converges to (deterministic)

$\xrightarrow{p}$	converges in probability to; see Hansen (2020b, §7.3)
$\xrightarrow{\text{a.s.}}$	converges almost surely to; see Hansen (2020b, §7.14)
$\xrightarrow{d}$	converges in distribution to; see Hansen (2020b, §8.2)
$\rightsquigarrow$	converges weakly to
$\equiv$	is defined as
$\approx$	approximately equals
$\doteq$	equals when ignoring smaller-order terms
$\sim$	is distributed as
$\overset{a}{\sim}$	is distributed approximately (or asymptotically) as
$X \perp\!\!\!\perp Y$	X and Y are statistically independent
$\mathrm{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathrm{N}(0, 1)$	standard normal distribution
$\Phi(\cdot)$	cumulative distribution function (CDF) of $\mathrm{N}(0, 1)$
$\phi(\cdot)$	probability density function (PDF) of $\mathrm{N}(0, 1)$
$F_Y(\cdot)$	cumulative distribution function (CDF) of Y
$Q_Y(\cdot)$	quantile function of Y
$f_Y(\cdot)$	probability density function (PDF) of Y (or PMF if discrete)
$\mathbb{1}\{\cdot\}$	indicator function: $\mathbb{1}\{A\} = 1$ if event A occurs, else $\mathbb{1}\{A\} = 0$
$\mathrm{P}(A)$	probability of event A
$\mathrm{P}(A \mid B)$	conditional probability of A given B
$\mathrm{E}(Y)$	expected value of Y
$\widehat{\mathrm{E}}(Y)$	expectation for sample distribution; same as $\frac{1}{n}\sum_{i=1}^{n} Y_i$
$\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$	CEF (function of $\boldsymbol{x}$); see Hansen (2020a, §2.5)
$\mathrm{E}(Y \mid \boldsymbol{X})$	expected value of Y given $\boldsymbol{X}$; this is a random variable
$\mathrm{Var}(Y)$	variance of Y
$\mathrm{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x})$	conditional variance (a non-random value)
$\mathrm{Var}(Y \mid \boldsymbol{X})$	conditional variance (a random variable)
$\mathrm{Cov}(Y, X)$	covariance
$\mathrm{Corr}(Y, X)$	correlation
$b \in \{a, b, c\}$	b is in the set containing a, b, and c
$\mathcal{S}_1 \cup \mathcal{S}_2$	the union of sets $\mathcal{S}_1$ and $\mathcal{S}_2$
$\bigcup_{j=1}^{J} \mathcal{S}_j$	the union of $\mathcal{S}_1, \ldots, \mathcal{S}_J$
$\mathcal{S}_1 \cap \mathcal{S}_2$	the intersection of sets $\mathcal{S}_1$ and $\mathcal{S}_2$
$\bigcap_{j=1}^{J} \mathcal{S}_j$	the intersection of $\mathcal{S}_1, \ldots, \mathcal{S}_J$
$\mathbb{N}$	the set of natural numbers, $\{1, 2, 3, \ldots\}$
$\mathbb{R}$	the set of real numbers (which excludes $\pm\infty$)
$\mathbb{R}_{\geq 0}$	the non-negative real numbers
$\mathbb{R}_{>0}$	the strictly positive real numbers
$\bar{\mathbb{R}}$	the extended real numbers, $\mathbb{R} \cup \{-\infty, \infty\}$

$\mathbb{R}^k$	k-dimensional Euclidean space
$\mathbb{Z}$	the set of integers, $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{Z}_{\geq 0}$, $\mathbb{Z}_{>0}$	analogous to $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$
$\mathrm{SE}(\hat{\theta})$	standard error of estimator $\hat{\theta}$
$\arg\min_g f(g)$	the value of g that minimizes $f(g)$
$\underline{\boldsymbol{I}}_k$	$k \times k$ identity matrix (ones on main diagonal, zeros elsewhere)
$\|\cdot\|$	norm (Euclidean unless otherwise defined)
$\mathrm{tr}(\underline{\boldsymbol{v}})$	trace of matrix $\underline{\boldsymbol{v}}$
$\underline{\boldsymbol{v}}'$	transpose of matrix $\underline{\boldsymbol{v}}$
$\underline{\boldsymbol{v}}^{-1}$	inverse of matrix $\underline{\boldsymbol{v}}$
$\underline{\boldsymbol{v}} > 0$	matrix $\underline{\boldsymbol{v}}$ is positive definite
$\underline{\boldsymbol{v}} \geq 0$	matrix $\underline{\boldsymbol{v}}$ is positive semi-definite

Part I

Foundations

Introduction

This part may be largely review, but it is helpful to have a deeper understanding of "basic" ideas before adding complexity. Eventually the focus narrows to identification of causal effects, specifically how in linear regression "control variables" can help reduce omitted variable bias but usually do not eliminate it.

Chapter 1
Stata

Unit learning objectives for this chapter

1.1. Access Stata and code basic commands for data manipulation and analysis. [TLO 5]

This chapter provides a brief overview of Stata, which you will use for the end-of-part exercises in this book.

Optional resources for this chapter

- Many user-contributed Stata commands can be installed from SSC, including `bcuse` (Baum, 2012), `ivreg2` (Baum, Schaffer, and Stillman, 2002), and `ranktest` (Kleibergen, Schaffer, and Windmeijer, 2007), which are used in this class.
- UCLA resources: `https://stats.idre.ucla.edu/stata`

1.1 Access

As a student at Mizzou, you can use Software Anywhere for free, even from home.[1] It currently has Stata version 15 (StataCorp, 2017), which is a few versions old but sufficient for this class.

The on-campus computing sites also provide a variety of statistical software. You can check which computing sites/labs have your favorite software on the Computing Sites Software web page.[2]

[1] `https://doit.missouri.edu/services/software/software-anywhere/`
[2] `https://doit.missouri.edu/services/computing-sites/sites-software/`

1.2 Pros and Cons

As with econometric paradigms, different statistical software packages have complementary strengths; none is "best" for every case. Stata is commonly used by eocnomists, especially in applied microeconomics. Below are some general strengths and weaknesses.
Strengths:

1. Very intuitive and simple; easy to do most common tasks.
2. Popular among applied economists $\implies$ lots of support, data often available in Stata format, used in jobs, etc.
3. I think the help files within Stata are very helpful (once you know the basic structure and syntax).

Weaknesses:

1. Not as many fancy functions as R, although econometricians are getting better about providing code in Stata (e.g., lots of the new RD methods).
2. Not as easy to code your own functions (vs. R, based on my experiences doing both).
3. Can only have one dataset in memory at a time.
4. Can be slower, but depends on Stata version (some support parallel processing) and the particular computational task.

1.3 General Setup

When you open Stata, you should see one "window" with multiple "panes" inside. The big one is the **console**, which shows the commands you run and the corresponding output. The very short one at the bottom (below the console) is the **command line**, where you can enter commands one at a time. However, it's generally best to keep all your code in a "do-file" (see below), unless you're just opening a help file or browsing the data manually. The other panes you can probably just close and ignore. If you make a graph, it will open in a new window. If you make another graph, it will open in the same (second) window, unless you include an option to name your graph.

Generally, you should write all your code in a **do-file**. These are files with .do file extension. They are not "programs" but **scripts**: a sequence of commands for Stata to run in order. The do-file editor is a separate window. To open it, hit Ctrl+9 (Windows) or use the "Window" file-menu, then go down to "Do-file Editor," then "New Do-file Editor." In this new window, you can start typing a new do-file, or open an old file, or save your current file. I suggest having this window open on the left half of your screen, with the console fully visible on the right half of your screen.

You can run code from the do-file editor without using the mouse or switching to the console. Simply highlight whichever line(s) of code you want to run, and hit Ctrl+D (Windows; probably Cmd+D Mac?). You should see the your highlighted code appear line-by-line in the console, along with any resulting output.

Economists care about **replicability**, meaning somebody else on a different computer should be able to exactly reproduce any result you report in your research. Making sure

everything you do is in a do-file helps ensure replicability. When you (think you) are "done" with your research, you should be able to take (only) the raw dataset(s) and your do-file(s), and run your do-files in order to generate every single number or graph included in your research paper. Saving log files (see below) also helps.

Besides the built-in Stata commands, there are many high-quality (and some low-quality) user-contributed commands. These often have corresponding articles in the *Stata Journal*, like the articles for commands `distcomp` (Kaplan, 2019) and `sivqr` (Kaplan, 2022b). They are often very easy to install, too. For example, you can issue command `net from https://kaplandm.github.io/stata` from the Stata command line, then click a few times to install any Stata command I have made available. These and others are also available on SSC and can be installed from the Stata command line with `ssc install sivqr` (or whatever the command name). These commands' code is usually in an **ado-file** with .ado file extension. Such files are plain-text, so you can see the code yourself (open it in the do-file editor or just in a text editing program). They define a **program** and sometimes refer to other functions designed for internal use.

Stata includes some additional functionality in **Mata**, which is more like a regular programming language. This is helpful if you are writing your own functions, like if you need to do some numerical optimization, but probably nothing you would need to use for empirical work.

1.4 Administrative Commands

There are some commands in Stata that you probably want to use at the beginning of all your do-files, to help get everything set up. This section briefly covers some common such administrative commands.

One thing that is not a command but is very help is **comments**. A comment is ignored by Stata, but lets you describe your do-file's goals (and put your name at the very top of your file), or the reason for a particular line of code. Comments are helpful for working with coauthors, but also just helpful for working with your future self. Research projects usually last at least one year, and often you do not remember why you wrote line 157 of your do-file after one year. Here is an example of using code comments to communicate effectively with your future self: xkcd.com/1421. In Stata, any line starting with an asterisk `*` is treated as a comment. Also, even if not at the beginning of a line, double-slash `//` tells Stata to ignore the rest of the line. You can also use C-style multi-line comments, starting with `/*` and ending with `*/`.

Here are some suggested commands to include at the top of your do-file.

- `clear all`: clears any data in memory; keep in mind that you will need to re-run your code many, many times (fixing bugs, adjusting data prep, etc.), so you should plan for having just run your code but maybe having gotten an error.

- `capture log close`: closes any log file left hanging open (again, if you got an error in the middle of your last run...).

- `cd`: change the current working directory (to a particular directory on your computer, where your data and such are stored, and where output will get saved by default, etc.).
- `log using FILENAME.log , replace`: start saving a log file, so all your commands and output will be saved; the ".log" makes it plain-text (my preference, instead of Stata markup language [like HTML]), and the `replace` option tells it to over-write the currently saved one (because again, this may be your 64th time running this do-file). Important: at the very end of your do-file, put `log close` as your very last line.
- `pwd`: prints the filepath of the current working directory in the console.
- `version`: prints the current version of Stata that you're running.
- `which`: prints the version of any user-contributed command (or even built-in command); for example, `which sivqr` to see the current version of `sivqr` you have installed.
- `set more off`: make sure Stata just runs through your whole do-file without waiting for you to click "more"; very important! (Otherwise you may start your file running, go work on something else for an hour, and come back only to realize it ran for 3 minutes before waiting for you to click "more.")

1.5 Data Input and Examination

Stata can read/input a few different formats of data, using the following commands.

- `use` loads data from .dta files, the proprietary Stata format; just tell it the name of the file, and usually you want to use the `clear` option to clear out the current dataset in memory, like `use my-file.dta , clear` (or you can omit the .dta part).
- `insheet` loads comma-separated values (.csv) data files or tab-separated files (often .txt or .dat); for example, something like `insheet using my-file.csv , comma names` to load a CSV file with a header row (`names`).
- `infile` can handle fixed-format data files (or whitespace-delimited files), although those are less common nowadays.

Stata can also output a few different formats of data, using the following commands.

- `save` generates a Stata-format .dta file, and the `replace` option is usually helpful (to overwrite an existing file with the same name); for example, `save my-new-file , replace` (the .dta is added to the filename automatically).
- `export delimited` can produce comma-delimited or tab-delimited files that are easy to import into other statistical software (like R).

- `export excel` creates an Excel file (which is rarely useful), and other `export` variants can work with databases or SAS.

Once your dataset is loaded into memory, you can examine it a bit.

- `describe`
- `codebook`
- `list in 1/5` to print the first five rows (observations) to the console; or `list x y z in 1/5` to print only variables `x`, `y`, and `z`, etc.
- `browse` to open the data browser.

Usually `edit` is a bad idea because any changes you make will not be replicated by running (only) your do-file; any changes you make to the data should be done "programmatically," with code in your do-file that can be replicated by other users.

1.6 Data Manipulation Commands

There are many ways to manipulate (change) your data, including with the following Stata commands.

- `order`: reorder the columns (variables) in your dataset; for example, `order y x` to put the `y` variable as the first column and `x` as the second (and leave the remaining variables in the same order).
- `keep` and `drop`: retain or delete the specified columns in the data; for example, `drop y` to delete the `y` variable (column), or `keep x` to keep only the `x` variable (and delete everything else).
- `keep if` and `drop if`: retain or delete the rows in the data satisfying the specified condition; for example, `keep if !missing(x)` to keep only observations (rows) with non-missing `x` value, or `drop if missing(x)` to drop observations with missing `x` value.
- `sort` and `gsort`: sort the rows according to the value of some variable(s); for example, `sort x` to sort the rows in ascending order by the value of their `x` variable.
- `generate`, `replace`, and `egen`: generate a new variable (column), or replace certain values; for example, `generate wx=w*x` creates a new column/variable named `wx` that equals the product of existing variables `w` and `x`, or `replace z=1 if x>0` to replace the value of variable `z` in observations with strictly positive `x`.

There are a few more commands that are complicated but useful.

Command `reshape` is mostly useful with panel data that includes multiple observations of the same "unit" (individual, firm, etc.) in different time periods. Sometimes, you might get a dataset where each unit has only one row, but there are variables like `inc2013` and `inc2014` that give the `inc` value observed in years 2013 and 2014. For most Stata commands for panel data, you instead want one row per unit-year, like one row for unit 1 in year 2013 and a separate row for the same unit in 2014. To convert such data, use `reshape long`; for example, `reshape long inc , i(id) j(year)` when variable `id` uniquely identifies the units and the original dataset has variables like `inc2013` and `inc2014`, which get converted into a single `inc` variable (column) plus a `year` column that stores the 2013 or 2014 from the original variable names. To go in the reverse direction (less useful), use `reshape wide`.

Command `collapse` can aggregate or summarize your data. For example, if you have wages for many individuals across many states and years, you could do `collapse (mean) wage , by(state year)` to create a dataset with only one observation (row) per unique state-year combination, containing the mean wage among all the individuals in that state-year.

There are also a couple useful commands for combining the dataset in memory with another data file, the simpler one being `append`. For example, this can be useful if you have separate datasets for separate years of data, but they all have the same variables, so you just need to stack the datasets on top of each other. In that case, you load the earliest file into memory, say year 2013, then `append using data2014`, then `append using data2015`, etc., assuming your files are named like `data2014.dta`.

Command `merge` instead combines datasets "horizontally." Each dataset needs to have identifier variable(s) that can be matched across datasets. For example, if you had one dataset with each person's `id` number and `height`, and another dataset with `id` and the person's `weight`, then you could merge them together to get a single dataset with both height and weight: you'd load the first dataset (say `heightdatafile.dta`), then `merge 1:1 id using weightdatafile` where the `1:1` indicates that `id` is a unique identifier in each dataset (there are never multiple rows with the same `id` value). Instead of such a "one-to-one" merge, sometimes a "many-to-one" merge is useful. For example, if your data in memory (the "master" data) has individuals living in different states, with multiple individuals per state, and you have another data file (the "using" data) with the current sales tax rate in each state, with only one row per state, then you could do a many-to-one merge like `merge m:1 state using taxdatafile` if the second dataset is `taxdatafile.dta`. Conversely, you can also do a one-to-many merge with `merge 1:m`. Finally, `merge` creates a new variable named `_merge` whose values you should check, because they indicate whether each row (in the newly merge dataset) has data from the master data (value `1`), from the using data (value `2`), or both (value `3`). It's not necessarily bad to have values of `1` or `2`, as long as you are expecting them and treat them properly in your subsequent analysis.

1.7 Data Analysis

There are many, many Stata commands for data analysis, most of which you can easily Google; these are just a few examples.

- `summarize`: compute summary statistics; you can also use the `detail` option to get even more statistics.
- `tabulate`: helpful summary for discrete or categorical data; for example, `tabulate region` to see how frequently each value of `region` appears in your dataset.
- `regress`: linear regression. Just specify the dependent variable followed by the regressors, and the type of standard errors you want in the `vce` option; for example, `regress y x1 x2 , vce(robust)` to get heteroskedasticity-robust standard errors. There are some operators that can make it easier for you to include polynomial, interaction, dummy, lagged, and differenced terms in your model, like `reg y c.x##c.x` to include a polynomial in `x`, or `reg y i.region` to generate dummy variables for different values of categorical variable `region`, or `reg y x L.x` to use lagged `x`, etc.
- `histogram` and `scatter`: make graphs.

Some of the end-of-chapter exercises provide additional code for more complex econometric analysis.

Chapter 2
Logic

Unit learning objectives for this chapter

2.1. Define and apply basic logic terms and relationships [TLO 1]

This chapter is mostly taken from Section 6.1 of Kaplan (2022a).

Some basic logic is useful for understanding certain parts of econometrics. First, logic is useful for understanding the relationship among different conditions. Often these conditions are assumptions used in various theorems. Second, logic is useful for understanding what a theorem actually claims. Third, logic is helpful for interpreting results. The following may not be fully technically correct from a philosopher's perspective, like perhaps I conflate logical implication with the material conditional, but it suffices for econometrics.

2.1 Terminology

⟹ Kaplan video: Logic Terms Example

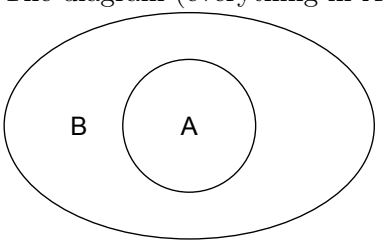
Many words and notations can refer to the same logical relationship. Let A and B be two statements that can be either true or false. For example, maybe A is "$Y \geq 10$" and B is "$Y \geq 0$." Or, A is "this animal is a cat," and B is "this animal is a mammal." The following ways of describing the logical relationship between A and B all have the same meaning.

1. If A is true, then B is true (often shortened: "if A, then B")
2. $A \implies B$
3. A **implies** B
4. $B \impliedby A$
5. B is **implied by** A

6. B is true **if** A is true
7. A is true **only if** B is true
8. A is a sufficient condition for B (shorter: "A is **sufficient** for B")
9. B is a necessary condition for A (shorter: "B is **necessary** for A")
10. A is **stronger** than B
11. B is **weaker** than A
12. It is impossible for B to be false when A is true (but it is fine if both are true, or both are false, or A is false and B is true)
13. The truth table (T=true, F=false):

A	B	$A \implies B$
T	T	T
T	F	F
F	T	T
F	F	T

14. The diagram (everything in A is also in B):



To state equivalence of A and B, opposite statements can be combined. Specifically, any of the following have the same meaning.

1. $A \iff B$ (meaning both $A \implies B$ and $A \impliedby B$)
2. A is true **if and only if** B is true (meaning A is true if B is true *and* A is true only if B is true)
3. B is true if and only if A is true
4. A is necessary and sufficient for B
5. B is necessary and sufficient for A
6. A and B are equivalent
7. It is impossible for A to be false when B is true, and impossible for A to be true when B is false.
8. The truth table (T=true, F=false):

A	B	$A \iff B$
T	T	T
T	F	F
F	T	F
F	F	T

Variations of $A \implies B$ have the following names. Read $\neg A$ as "not A": $\neg A$ is false when A is true, and $\neg A$ is true when A is false.

- $\neg A \implies \neg B$ is the **inverse** of $A \implies B$.
- $B \implies A$ is the **converse** of $A \implies B$.
- $\neg B \implies \neg A$ is the **contrapositive** of $A \implies B$.

The statement $A \implies B$ is logically equivalent to its contrapositive. That is, statements "$A \implies B$" and "$\neg B \implies \neg A$" can be both true or both false, but it's impossible for one to be true and the other false.

The statement $A \implies B$ is not logically equivalent to either its inverse or converse. (The inverse and converse are equivalent to each other because the inverse is the contrapositive of the converse.)

Example 2.1 (Kaplan video)**.** Let A be "$X \leq 0$" and let B be "$X \leq 10$."

- $A \implies B$: any number below 0 is also below 10.
- We could equivalently say "A implies B" or "B is true if A is true" or "A is stronger than B" or "A is sufficient for B."
- The contrapositive is $X > 10 \implies X > 0$, which is also true: any number above 10 is also above 0.
- The inverse is $X > 0 \implies X > 10$, which is false: e.g., if $X = 5$, then $X > 0$ but not $X > 10$.
- The converse is $X \leq 10 \implies X \leq 0$, also false: again if $X = 5$, then $X \leq 10$ but not $X \leq 0$.

2.2 Theorems

Theorems all have the same logical structure: if assumption A is true, then conclusion B is true. Sometimes A and B have multiple parts, like A is really "A_1 and A_2." (Like, "If SUTVA holds and treatment is randomized, then the ATE is identified and equals the mean difference.") The theorem's practical use is: if we can verify that A is true, then we know B is also true.

What if we think A is false? Then, B could be false, or it could be true. This may be seen most readily from the picture version of the A and B relationship in Section 2.1: we could be somewhere inside B but outside A (i.e., B true, A false); or we could be outside both (both false). That is, as in Section 2.1, the theorem $A \implies B$ is not equivalent to its inverse.

Also from Section 2.1, a theorem is equivalent to its contrapositive. That is, if the theorem's conclusion is false, then we know at least one of its assumptions is false.

Example 2.2 (Kaplan video)**.** Consider three line segments, x, y, and z. Let A be "x, y, and z form a triangle"; let B be "the length of z is less than or equal to the sum of the lengths of x and y."

- In Euclidean geometry, if assumption A is true, then conclusion B is true; the theorem "$A \implies B$" is correct (known as the triangle inequality).

- The inverse is, "if A is false, then B is false," or: "if x, y, and z do not form a triangle, then the length of z is greater than the sum of the lengths of x and y." This statement is incorrect: if the segments do not form a triangle, then they can be any lengths.
- The contrapositive is, "if B is false, then A is false," or: "if the length of z is greater than the sum of the lengths of x and y, then the three segments do not form a triangle." The contrapositive is true; if you have three such segments, it's impossible to arrange them into a triangle.

2.3 Comparing Assumptions

To compare assumptions, the terms "stronger" and "weaker" are most commonly used. Let A_1 and A_2 denote different assumptions. Per Section 2.1, "A_1 is stronger than A_2" is equivalent to $A_1 \implies A_2$, which is also equivalent to "A_2 is weaker than A_1."

All else equal, it is more useful to have a theorem with weaker assumptions because it applies to more settings. That is, if $A_1 \implies A_2$, then we prefer a theorem based on A_2, the weaker assumption. A theorem based on A_1 can only be used when A_1 is true. In contrast, a theorem based on A_2 can be used not only when A_1 is true (because $A_1 \implies A_2$), but also sometimes when A_1 is false (but A_2 is still true).

Example 2.3 (Kaplan video)**.** Let assumption A_1 be, "a city is in Missouri," and let assumption A_2 be, "a city is in the United States." Consider the theorems $A_1 \implies B$ and $A_2 \implies B$. (The conclusion is irrelevant here, but to be concrete you could imagine B is "the city is in the northern hemisphere.") Because Missouri is part of the United States, $A_1 \implies A_2$, i.e., A_1 is the stronger assumption and A_2 is the weaker assumption. We prefer the theorem based on the weaker assumption because it applies to more cities. For example, only the theorem $A_2 \implies B$ applies to Houston; A_1 is false, but A_2 is true. (And recall that when A_1 is false, the theorem $A_1 \implies B$ does not conclude that B is false; it just says, "I don't know if B is true or false," i.e., it is useless.)

Discussion Question 2.1 (median theorem logic)**.** Consider the theorem, "If sampling is iid, then the sample median consistently estimates the population median." Hint: draw a picture and/or write it as $A \implies B$.

a) What does this tell us about consistency of the sample median when sampling is not iid?
b) What does this tell us about sampling when the sample median is not consistent?

Discussion Question 2.2 (mean theorem logic)**.** Consider the theorem, "If sampling is iid and the population mean is well-defined, then the sample mean consistently estimates the population mean." Hint: there may be multiple possible pictures that show this relationship among A_1 (iid), A_2 (well-defined), and B (consistency).

a) What does this tell us about consistency of the sample mean when sampling is not iid?

b) What does this tell us about sampling when the sample mean is not consistent?

Discussion Question 2.3 (logic with feathers)**.** Consider two theorems. Theorem 1 says, "If X is an adult eagle, then it has feathers." Theorem 2 says, "If X is an adult bird, then it has feathers."

a) Describe each theorem logically: what's the assumption (A), what's the conclusion (B), what's the relationship?
b) State Theorem 1's contrapositive; is it true?
c) Compare: does Theorem 1 or Theorem 2 have a stronger assumption? Why?
d) Compare: which theorem is more useful? (Which applies to more situations?)

Chapter 3

The Big Picture

Unit learning objectives for this chapter

3.1. Define terms and concepts fundamental to econometrics as a whole and the portion of it on which we will focus, including the interpretation and significance of empirical results. [TLOs 1, 2, and 4]

This chapter provides a view of the wide world of econometrics, including fundamental ideas that will recur throughout the book. The section titles use the word "and" instead of "versus" to emphasize that different paradigms may be helpful in different contexts or even complement each other in the same context; it is not a fight about which is "best," because there is no universal "best." If you end up using econometrics for research, then you will (hopefully) not be using methods directly from this class but more sophisticated methods that you learn about later. I hope this chapter (and book) helps you more readily understand the new methods you encounter later.

Optional resources for this chapter

- Quantifying uncertainty and statistical significance (Masten video)
- Bayesian vs. frequentist cookie inference example (StackExchange)
- Structural vs. reduced form approach (Masten video)
- Structural modeling advantages (Masten video)
- Greater external validity for "structural" results (Masten video)

3.1 Description, Prediction, and Causality

This section draws from Section 4.3 of Kaplan (2022a).

There are three categories of questions that econometric methods can help answer, related to description, prediction, and causality. Description is essentially about features of the joint distribution of observable variables, like correlations and conditional means. Prediction is guessing an unknown value based on other observed values; the "best" guess depends on the consequences of wrong guesses, which are often used to make a decision. Causality is important for making decisions, like should our firm spend more on advertising, or should we raise or lower the minimum wage? Causality is about the effect of such policy changes on other variables.

Example 3.1 (Kaplan video)**.** Consider the relationship between an individual's employment status and mental health, specifically anxiety. A descriptive question is: what's the proportion of employed individuals who have generalized anxiety disorder (GAD), and how much higher or lower is that proportion among unemployed individuals? A predictive question is: given somebody's employment status, what's the "best" guess of their score on the GAD-7 anxiety measure? A causal question is: how does being employed (instead of unemployed) affect an individual's level of anxiety as quantified by the GAD-7?

Discussion Question 3.1 (description, prediction, causality)**.** Which type of question (description, prediction, causality) is each of the following? Explain why. Hint: there's one of each.

a) If you only know whether an individual is from Canada or the U.S., what is your best guess of their income?
b) You are currently working in the U.S. but considering moving to Canada. How will your income change if you do?
c) Which country's population has higher income: Canada or the U.S.?

3.2 Population and Sample

Generally, the population is what we want to learn about, and the sample is the data from which we can learn. There are different ways to mathematically model a population, depending on the object of interest. Often, we can learn about a feature of a population by computing the same feature of the sample. More details about how we interpret the population and sample are in Section 3.4.

In this book, the population is a represented by a joint probability distribution of random variables, and the sample is a set of n observations of values drawn from that distribution. Population features are mirrored by features of the sample. For example, population random variable Y has mean $\mathrm{E}(Y)$, and given $Y_1, Y_2, \ldots, Y_n$, the sample mean is $\widehat{\mathrm{E}}(Y) = (1/n)\sum_{i=1}^{n} Y_i$. The **sample distribution** or **empirical distribution** is a discrete probability distribution with probability $1/n$ on each value of Y_i; the sample mean is thus the mean of the empirical distribution.

3.2.1 Population Types

⟹ Kaplan video: Population Types

This subsection is a shorter version of Section 2.2 of Kaplan (2022a).

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

Beyond our scope...

Recently, there has been some renewed interest in finite-population methods in econometrics; for example, see Abadie, Athey, Imbens, and Wooldridge (2020).

Generally, the finite population perspective cares more about the outcomes of a finite group of individuals, whereas the other two population types care more about properties of the underlying mechanisms that generated the outcomes, often called the **data-generating process** (DGP).

A **finite population** is closest to the regular English word "population," which means all the people living in some area. For example, if we are interested in the outcomes of (only) everyone in Missouri in 2023, then we have a finite population. Other examples of finite populations are (for a given time period) all employees at a particular firm, all firms in a particular industry, all students in a particular school, or all hospitals of a certain size. In a finite population, we care only about the actual outcomes, not underlying reasons; for example, maybe we want to know how many individuals in Missouri actually earned the minimum wage in January 2023, but we do not care about the determinants of wage. Hypothetically, if we could observe every single member of a finite population, then we could fully answer our question, with no uncertainty. That is, our confidence interval would just be a single point, equal to the true value.

Sometimes a finite population is so large compared to the sample size (i.e., the number of population members we observe) that an **infinite population** is a reasonable approximation. For example, if we observe only 600 individuals out of the 6+ million in Missouri, econometric results based on finite and infinite populations are practically identical. Although "infinite" sounds more complex than "finite," it is actually simpler mathematically: instead of needing to track every single member of a finite population, an infinite population is succinctly described by a probability distribution or random variable.

Besides this convenience, sometimes there is no finite population (however large) that answers your question. For example, imagine there's a new manufacturing process for carbon monoxide monitors that should sound an alarm above 50ppm. Most work properly, but some are faulty and never alarm. Specifically, this manufacturing process corresponds to some probability of producing a faulty monitor. Mathematically, the manufacturing

process can be modeled as random variable W with some probability of the value "faulty." If you want to learn this probability (i.e., this property of the manufacturing process), then there is no finite number of monitors that can exactly answer your question; no finite number of realizations exactly determines $\mathrm{P}(W = \text{faulty})$. This is an infinite population question.

One variation of the infinite population is the **superpopulation** (coined by Deming and Stephan, 1941). This imagines (infinitely) many possible universes; our actual universe is just one out of infinity. Thus, even if it appears we have a finite population, we could imagine that our universe's finite population is actually a single sample from an infinite number of universes' finite populations. The term "superpopulation" essentially means "population of populations." Our universe's finite population "is only one of the many possible populations that might have resulted from the same underlying system of social and economic causes" (Deming and Stephan, 1941, p. 45). For example, imagine we want to learn the relationship between U.S. state-level unemployment rates and state minimum wage levels. It may appear we are stuck with a finite population because there are only 50 states, each of which has an observable unemployment rate and minimum wage. However, observing all 50 states still doesn't fully answer our question about the underlying mechanism that relates unemployment and minimum wage, so a finite population seems inappropriate. But we can't just manufacture new states like we can manufacture new carbon monoxide monitors, so an infinite population also seems inappropriate. The superpopulation imagines manufacturing new entire universes, each with 50 states and the same economic and legal systems.

In Sum: Population Type

Hypothetically, could a finite number of observations fully answer your question?
No $\implies$ superpopulation or infinite population, modeled as probability distribution (as in this textbook)
Yes $\implies$ finite population (use different methods unless sample is much smaller than population)

Example 3.2 (employment status)**.** Consider the employment status of individuals in Missouri. A finite population is more appropriate if you want to document the actual percentage of Missouri individuals unemployed last week. A superpopulation is more appropriate if you want to learn about the underlying mechanism that relates education and unemployment. That is, knowing each individual's employment status fully answers the first question, but not the second question.

Example 3.3 (employee productivity)**.** Consider the productivity of employees at your company (you're the CEO). If you want to know each employee's productivity over the past fiscal quarter, then a finite population is more appropriate. If you want to learn how a particular company policy affects productivity, then a superpopulation is more appro-

priate. That is, knowing each employee's productivity fully answers the first question, but not the second question.

Discussion Question 3.2 (student data)**.** Imagine you're a high school principal. You have data on every student, including their standardized test scores from last spring.

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

3.2.2 Before and After Sampling: Two Perspectives

⟹ Kaplan video: "Before" and "After" Perspectives of Data

This subsection is from Section 2.1 of Kaplan (2022a).

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

Let variable W represent the outcome. After the flip, the outcome is known: either $W = h$ or $W = t$. Before the flip, both $W = h$ and $W = t$ are possible. If the coin is "fair," then possible outcome $W = h$ has probability $1/2$, as does $W = t$.

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

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

Other terms for W include a **random draw** (or just **draw**), or more specifically a random draw (or "randomly drawn") from a particular probability distribution. Seeing the population as a probability distribution (see Section 3.2.1), we could say W is randomly sampled from its population distribution, or if there are multiple random variables $W_1, W_2, \ldots$ (e.g., multiple flips of the same coin), we could say they are randomly sampled from the population or that they collectively form a **random sample**; see Section 3.2.3 for more about sampling.

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

Example 3.4 (Kaplan video)**.** Let $R = 1$ if it rains in Columbia, MO on Tuesday and $R = 0$ if not. If today is Monday, then either outcome is possible, so we have

the "before" view: R is a random variable, with some probability of $R = 0$ and some probability of $R = 1$. If instead today is Wednesday, then what happened Tuesday is already determined, so we have the "after" view. If it rained, then $R = 1$; if not, $R = 0$. There is only a single value, not multiple possible values. Even if we don't know the realized value r, we know it's just a single value.

Extending the above are the **before sampling** and **after sampling** perspectives, or "before observation" and "after observation." Similar to above, "before" corresponds to random variables, whereas "after" corresponds to realized values. Before sampling a unit (person, firm, etc.) from a population, we don't know which one we'll get, so there are multiple possible values. After sampling, we can see the specific values we got.

Example 3.5 (age as random variable)**.** Imagine you plan to record the age of one person living in your city. You take a blank piece of paper on which you'll write the age. After you find a person and write their age ("after sampling"), that number can be seen as a realized value, like w. In contrast, before sampling, there are many possible numbers that could end up on your paper. It's not that peoples' ages are undetermined; they each know their own age. But before you "sample" somebody, it's undetermined whose age will end up on your paper. It could be your neighbor DeMarcus, age 88. It could be your kid's friend Lucia, age 7. It could be your colleague Xiaohong, age 35. The random variable W is like your blank paper: it has many possible values, like $W = 88$, $W = 7$, or $W = 35$.

Discussion Question 3.3 (web traffic)**.** Let $Y = 1$ if you're logged into the course website and $Y = 0$ if not.

a) From what perspective is Y a non-random value?
b) From what perspective is Y a random variable?

In Sum: Before & After

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

3.2.3 Sampling Types

$\implies$ Kaplan video: Types of Sampling

This subsection is a shorter version of Section 3.2 of Kaplan (2022a).

Properties of estimators depend on how a sample is drawn from the population. However, this book focuses mostly on identification, so generally iid sampling (see below) is assumed for simplicity. One exception is the discussion of "cluster-robust" confidence intervals when using panel data. There are also problems related to sampling like sample selection bias and missing data; for example, see Section 12.3 ("Threats to Internal Validity") of Kaplan (2022a) or Chapter 21 ("Missing Data") of Kaplan (2021).

Notationally, we observe the values from n **units**, which could be individuals, firms, countries, etc. (I often refer to units as "individuals," too.) Let $i = 1$ refer to the first unit, $i = 2$ to the second, etc., up to $i = n$, where n is the **sample size**. The corresponding values are $Y_1, Y_2, \ldots, Y_n$, with Y_i more generally denoting the observation for unit i. A particular dataset may have specific values like $Y_1 = 5$, $Y_2 = 8$, etc., but to analyze (frequentist) statistical properties, each Y_i is seen as a random variable as in Section 3.2.2. You can imagine n buckets (or pieces of paper), initially empty, that will eventually contain information from n observations. The sampling procedure does not determine the specific numeric values that end up in the buckets, but it determines how the buckets get filled.

In this section, two important sampling properties are considered: "independent" and "identically distributed." If both hold, then the Y_i are called **independent and identically distributed** (iid) random variables (or "sampled iid"), and "sampling is iid." Sometimes the vague phrase **random sampling** refers to iid sampling.

Notationally, iid sampling is indicated by $\overset{iid}{\sim}$. For example, with population CDF $F_Y(\cdot)$,

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

The F_Y can be replaced by another distribution function or name.

Independent

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

Mathematically, independence means

$$Y_i \perp\!\!\!\perp Y_k \text{ for any } i \neq k, \tag{3.2}$$

where $\perp\!\!\!\perp$ denotes statistical independence. That is, $Y_1 \perp\!\!\!\perp Y_2$, $Y_1 \perp\!\!\!\perp Y_8$, $Y_6 \perp\!\!\!\perp Y_4$, etc. For any $i \neq k$, independent sampling implies (but is not implied by), among other properties,

$$\mathrm{Cov}(Y_i, Y_k) = 0, \quad \mathrm{Var}(Y_i + Y_k) = \mathrm{Var}(Y_i) + \mathrm{Var}(Y_k), \quad \mathrm{E}(Y_i \mid Y_k) = \mathrm{E}(Y_i). \tag{3.3}$$

Example 3.6 (Kaplan video)**.** You plan to flip a coin and record $Y_1 = 1$ if heads and $Y_1 = 0$ if tails. You plan flip the same coin again and record $Y_2 = 1$ if heads and $Y_2 = 0$ if tails. These are independent: $Y_1 \perp\!\!\!\perp Y_2$. Although the probabilities are very closely related (actually identical), the realization of the first flip (heads or tails) has no relationship with the second flip. For example, even if we know the first flip is heads, this does not change the probability of heads for the second flip: $\mathrm{P}(Y_2 = 1 \mid Y_1 = 1) = \mathrm{P}(Y_2 = 1)$.

Example 3.7 (Kaplan video)**.** You plan to pick a random person in the world and record how many years of formal education they've had as Y_1. You plan to then pick another

random person and record their years of education in Y_2. The way you sample Y_2 has no relation to the first sampled person or their Y_1 value, so there is independence: $Y_1 \perp\!\!\!\perp Y_2$. Among other implications, this means Y_1 and Y_2 have zero correlation (uncorrelated) and zero covariance, $\text{Cov}(Y_1, Y_2) = 0$.

Identically Distributed

The **identically distributed** property means that from the "before" view, the distribution of Y_i is the same for any i. Qualitatively, all units are sampled from the same population. Mathematically, given shared population CDF $F_Y(\cdot)$, $Y_i \sim F_Y$ for all $i = 1, \ldots, n$; or without specifying F_Y explicitly, identically distributed means $Y_i \overset{d}{=} Y_k$ for any i, k. This further implies equalities like $\text{E}(Y_i) = \text{E}(Y_k)$ and $\text{Var}(Y_i) = \text{Var}(Y_k)$.

Example 3.8 (Kaplan video)**.** The Y_1 and Y_2 in Example 3.6 are identically distributed because they are from the same coin, so the probability of heads is the same each time. (Unless you cheat or flip it differently or something, but those are nuances for physics class, not econometrics.)

Discussion Question 3.4 (i/id sampling)**.** You are planning to sample values Y_1 and Y_2, but you have not yet sampled them. Each of the following four statements implies one of the four sampling properties: 1) independent, 2) not independent (i.e., dependent), 3) identically distributed, 4) not identically distributed. Which is which?

a) You are just as likely to get $Y_1 = 3$ as $Y_2 = 3$, and similarly for any other value besides 3.
b) If you get a negative Y_1, then you'll probably get a negative Y_2; but if you get a positive Y_1, then you'll probably get a positive Y_2.
c) Separately and simultaneously, you will randomly sample Y_1 while your friend samples Y_2.
d) For Y_1 you are going to get the salary of somebody with an economics degree, and Y_2 will be the salary of somebody with an art history degree.

Example 3.9 (Kaplan video)**.** Imagine randomly picking a Mizzou student ID number, then randomly picking a 2nd, then 3rd, then 4th. The corresponding Y_i are both independent and identically distributed (iid). They are independent because each ID number is randomly drawn without any consideration of how the other numbers are drawn, and without any consideration of the other observed Y_i values. They are identically distributed because each ID number is drawn from the same population (anyone who has a Mizzou student ID).

Example 3.10 (Kaplan video)**.** Each Mizzou student is classified as either a resident of Missouri ("in-state") or not ("non-resident"). Imagine buckets 1 and 2 say "in-state," while buckets 3 and 4 say "non-resident": observations Y_1 and Y_2 are from in-state students, while Y_3 and Y_4 are from non-resident students. (This is "stratified sampling": assigning buckets to different strata before sampling.) For most variables, the in-state distribution

differs from the non-resident distribution, so the distribution of Y_1 and Y_2 (in-state) differs from the distribution of Y_3 and Y_4 (non-resident). That is, sampling is not identically distributed. Thus, even if the samples are all independent, sampling is not iid.

Example 3.11 (Kaplan video)**.** Imagine randomly picking a class (like introductory econometrics) at Mizzou, and filling the first two buckets (Y_1 and Y_2) with two random students from that class; then randomly picking another class, and another two students for the other buckets (Y_3 and Y_4). (This is an example of "clustered sampling," where each class is a "cluster"; this differs from "clustering" in cluster analysis.) Observations are identically distributed (because each Y_i has the same probability of getting any particular student) but probably not independent. For example, dependence may come from students in the same class being similarly affected by their shared experience. Here, buckets 1 and 2 are correlated, and 3 and 4 are correlated, but not 1 and 3, nor 2 and 4, etc. Thus, sampling is not iid.

Example 3.12 (Kaplan video)**.** Imagine randomly picking 2 Mizzou students (like with random ID numbers), then observing them this semester and next semester. For example, imagine bucket 1 contains the first student's GPA this semester, bucket 2 contains the same student's GPA next semester, and buckets 3 and 4 contain the other student's GPAs from this semester and next semester. Buckets 1 and 2 (Y_1 and Y_2) are probably both high or both low, rather than one high and one low, and similarly for buckets 3 and 4 (Y_3 and Y_4). That is, buckets 1 and 2 are correlated, and 3 and 4 are correlated. Further, observations may not even be identically distributed if fall GPA and spring GPA do not have the same distribution. Thus, sampling is not iid.

Discussion Question 3.5 (rural household sampling)**.** You want to learn about household consumption in rural Indonesia. In an area with 100 villages, you either i) pick 5 villages at random, then survey every household in each of the 5 villages; or ii) make a list of all households in all 100 villages, then randomly pick 5% of them. Explain why each approach is or isn't iid.

3.3 Frequentist and Bayesian

⟹ Kaplan video: Bayesian and Frequentist Perspectives

This subsection is a shorter version of Section 3.1 of Kaplan (2022a).

The **Bayesian** and **frequentist** (or **classical**) frameworks have both produced valuable econometric methods. This book uses the frequentist framework. Often the practical difference is small, although in some cases it can be large (e.g., Kaplan and Zhuo, 2021).

The goal of this section is to develop a basic understanding of both frameworks, including how sampled data is used to learn about the population, as well as how uncertainty is quantified.

3.3.1 Very Brief Overview: Bayesian Approach

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

Mathematically, "belief" is a probability distribution. For example, let random variable B represent your belief about the population mean. If you think there's a 50% chance the mean is negative, then $\mathrm{P}(B < 0) = 50\%$. If you think there's a 1/4 probability that B is below -1, then $\mathrm{P}(B < -1) = 1/4$. (Elsewhere, you may see this written more confusingly as $\mathrm{P}(\theta < 0)$ and $\mathrm{P}(\theta < -1)$.)

For example, imagine you see a bird flying in your backyard, and you grab your binoculars to try to identify it. Let θ represent the true species, while B is your belief. Imagine (for simplicity) you only ever see three types of bird in your backyard, all woodpeckers: downy, hairy, and red-bellied, written $\theta = d$, $\theta = h$, and $\theta = r$. Based on the location and habitat, you know hairy is somewhat less likely in general, so your prior is $\mathrm{P}(B = d) = \mathrm{P}(B = r) = 0.4$, $\mathrm{P}(B = h) = 0.2$. Looking through your binoculars (looking at the data), you're pretty sure it's not the red-bellied, but it's too far to distinguish downy from hairy, so your updated posterior belief has $\mathrm{P}(B = d) = 0.6$, $\mathrm{P}(B = h) = 0.3$, $\mathrm{P}(B = r) = 0.1$. The low probability of red-bellied comes from the data, whereas the higher probability of downy than hairy comes from your prior.

The posterior distribution is the Bayesian way of quantifying uncertainty. It is relatively intuitive, similar to how people talk about uncertainty in daily life. The posterior distribution is often summarized by a **credible interval**, i.e., a range of values that you're pretty sure (like 90% sure) contains the true θ. Or in the above example with categorical θ, the **credible set** $\{d, h\}$ has 90% posterior belief: you'd say, "I'm 90% sure it's a downy or hair woodpecker, although I think there's a 10% chance I'm wrong and it's a red-bellied woodpecker."

3.3.2 Very Brief Overview: Frequentist Approach

The core of the frequentist approach is the "before" perspective (Section 3.2.2), which can also be described in terms of **repeated sampling**. Instead of the belief probabilities of a Bayesian posterior, frequentist probabilities are from the "before" view of the dataset (and thus value of estimator and such). Equivalently, as a thought experiment, we can imagine many different random samples drawn from the same population; the "before" probabilities are then how often certain values occur in these many random datasets.

For intuition, imagine you could randomly sample 100 datasets from the same population. Then, the frequentist probability of an event is approximately how many times that event occurs among the 100 samples. For example, we could compute the sample mean $\bar{Y}$ in all 100 samples; because the datasets are all different, the sample means $\bar{Y}$ are also all different. If $\bar{Y} \leq 0$ in 50 of the 100 hypothetical samples, then $\mathrm{P}(\bar{Y}_n \leq$

0) $\approx 50/100 = 50\%$. Or, if $\bar{Y}$ is in the interval $[-0.4, 0.4]$ in 70 of 100 samples, then $\mathrm{P}(-0.4 \leq \bar{Y} \leq 0.4) \approx 70\%$.

3.3.3 Bayesian and Frequentist Differences

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

First, the frameworks treat different variables as random or non-random. The frequentist framework treats the population mean and other population features as non-random values, whereas it treats the data as random. For example, the population mean $\mu = \mathrm{E}(Y)$ is a non-random value, whereas an observation Y is a random variable. In contrast, the Bayesian framework treats (beliefs about) population features as random, whereas it treats the data as non-random values (the "after" view).

Second, due to this different treatment, the frameworks answer different types of questions, especially when quantifying uncertainty. The Bayesian framework answers questions about our beliefs after seeing the data. The frequentist framework answers questions about probabilities of seeing different features in the data, given the true population values.

Example 3.13 (Kaplan video)**.** Consider the question, "Given the observed data, what do I believe is the probability that the population mean is above 1/2?" This is a Bayesian question. Mathematically, if y is the "observed data," this question is commonly written as $\mathrm{P}(\mu > 1/2 \mid y)$, noting the conventional but confusing notation where μ represents beliefs. This question makes no sense from the frequentist perspective: either $\mu > 1/2$ or not; it cannot be "maybe," with some probability.

Example 3.14 (Kaplan video)**.** Consider the question, "Given the value of $\mu = \mathrm{E}(Y)$, what's the probability that the sample mean is above 1/2?" This is a frequentist question. Mathematically, this is usually written $\mathrm{P}(\bar{Y} > 1/2)$, or $\mathrm{P}_\mu(\bar{Y} > 1/2)$ to be explicit about the dependence on μ. The sample mean $\bar{Y}$ is a function of data, so it is treated as a random variable. This question makes no sense from the Bayesian perspective: we can see the data, so we can see either $\bar{Y} > 1/2$ or not; it cannot be "maybe," with some probability.

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

Third, frequentist methods use only the data, whereas Bayesian methods can formally incorporate additional knowledge. In practice, though, even frequentist results should be interpreted in light of other knowledge. The difference is that this process is not formalized within the frequentist methodology itself. Unfortunately, many people do not combine

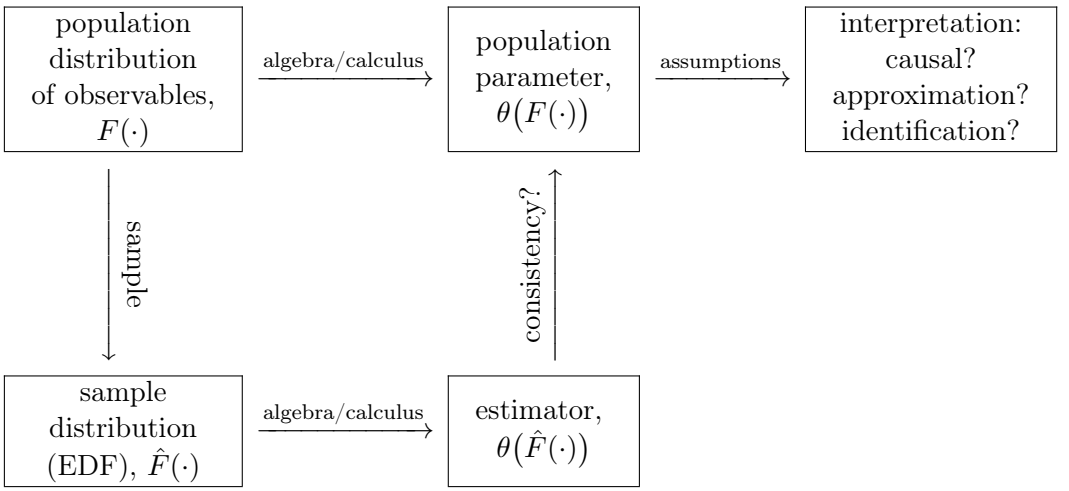


Figure 3.1: Map of (part of) the world of econometrics (Mercator projection).

frequentist results with other knowledge, instead interpreting frequentist results as if one single dataset contains the full, absolute truth of the universe; please do not do this!

In Sum: Bayesian & Frequentist

Frequentist: "before" view of data (random variables); assess methods' performance across repeated random samples from same population
Bayesian: "after" view of data (non-random); model beliefs (about population features) as random variables

Discussion Question 3.6 (frequentist vs. Bayesian inference)**.** Discuss https://xkcd.com/1132. Note the null hypothesis is that the sun has not exploded; the alternative hypothesis is that the sun has exploded.

a) Explain why the p-value is indeed computed correctly.
b) Given the machine's output, do you think the sun exploded? Why/not?

3.4 Identification, Estimation, and Inference

Figure 3.1 shows one perspective of what (some) econometrics is about. Different parts of the "map" corresponds to identification and estimation.

Identification relates to the top-right of Figure 3.1. There are two different ways to think about identification. First, as in the map, imagine population parameter θ is a feature of the joint population distribution of observable variables, $F(\cdot)$. For example,

maybe $\theta = \mathrm{E}(Y)$ is the mean of Y, or maybe $\theta = \mathrm{Cov}(Y, X)/\mathrm{Var}(X)$ is the slope of the linear projection of Y onto $(1, X)$. In some cases, this slope can be interpreted as the causal effect of X on Y. Identification can be understood in terms of the set of assumptions under which the population feature θ has this particular causal interpretation. Alternatively, imagine we define β as the causal effect of X on Y. This β is not a feature of the population distribution of (Y, X). However, β is **identified** if (under a set of **identifying assumptions**) it equals a feature of the population distribution of observables. More specifically, a parameter β is **point identified** $F(\cdot)$ uniquely determines the value of β. That is, if we somehow knew $F(\cdot)$, then we would also know the value of β. This book focuses on point identification of causal parameters.

Beyond our scope...

Parameters can also be **set identified** or **partially identified**, meaning that $F(\cdot)$ does not uniquely map to a single value of the parameter but rather a set of possible parameters. For example, maybe knowing $F(\cdot)$ lets us narrow down the possible values of β to the interval $[a, b]$, but $a < b$. For example, see Part VI of Kaplan (2021).

Estimation relates to the bottom of Figure 3.1. Given identification, our object of interest is a feature of the population distribution of observables. In many cases, to get an estimator, we simply compute the same feature of the sample distribution, also called the empirical distribution, $\hat{F}(\cdot)$. This is called the **analogy principle** or **plug-in principle**. Other population parameters are defined as the solution to a population optimization problem, in which case the estimator solves the sample version of the problem. The OLS estimator can be thought of from both perspectives: it estimates the population parameter $\boldsymbol{\beta} = [\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\mathrm{E}(\boldsymbol{X}Y)$ by $\hat{\boldsymbol{\beta}} = [\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\widehat{\mathrm{E}}(\boldsymbol{X}Y)$, or equivalently it estimates the population parameter $\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}} \mathrm{E}[(Y - \boldsymbol{X}'\boldsymbol{b})^2]$ by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} \widehat{\mathrm{E}}[(Y - \boldsymbol{X}'\boldsymbol{b})^2]$.

Inference is a vague word and also not well represented by Figure 3.1. People use inference in a variety of contexts with different meaning: Bayesian inference, causal inference, statistical inference, etc. In this book, it refers to methods (mostly confidence intervals) that quantify the statistical uncertainty about a certain population feature, which may not be the actual parameter of interest if the identifying assumptions are violated. For example, if you run `reg y x` in Stata, the confidence interval is for the population linear projection coefficients; even if you are interested in the causal effect of `x` on `y`, the confidence interval cannot account for your uncertainty about the identifying assumptions required for the linear projection slope to have a causal interpretation. However, sometimes there are ways to empirically assess certain identifying assumptions, as we will see.

3.5 General Equilibrium and Partial Equilibrium

This subsection is Section 4.3.3 of Kaplan (2022a).

Another econometric dichotomy is between **general equilibrium** (GE) and **partial equilibrium** (PE) analysis. GE more ambitiously tries to model entire markets, sometimes multiple markets, whereas PE takes current market equilibria as given. The tradeoff is that the GE framework can analyze policies that change equilibria (i.e., that have **general equilibrium effects**), but it requires stronger assumptions to do so.

Example 3.15 (Kaplan video)**.** Imagine you were analyzing the impact of free public childcare on mothers' employment. A PE analysis would consider how mothers might respond to different childcare policies given the current prices of private childcare, current wages, etc. A GE analysis might further model the childcare and labor markets, to allow for the possible general equilibrium effects of public childcare policy on the prices in those markets. If there is a big expansion of free public childcare, then private childcares may indeed change their prices. If the expansion allows many mothers to enter the workforce, then the labor supply curve shifts out, which could lower wages. However, if the proposed changes to childcare policy are relatively small, then such GE effects may be negligible, and PE analysis may suffice.

The famous Lucas critique (Lucas, 1976) argues in part that macroeconomic policy analysis requires structural, GE models. Lucas writes (p. 41), "Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models." That is, he says that if we want to guess how people and firms will behave in the future, under new macroeconomic policies, then we need to account for GE effects, which requires deeper, structural understanding and modeling of economic behavior.

In Sum: General & Partial Equilibrium Models

Partial equilibrium models treat prices and other market equilibria as fixed, whereas **general equilibrium models** allow markets to change.

3.6 Structural and Reduced-Form Approaches

This subsection is a shorter version of Section 4.3.2 of Kaplan (2022a).

There are two general approaches to learning about causality: the reduced-form approach, and the structural approach. Confusingly, the reduced-form approach is sometimes called **causal inference** even though the structural approach also aims to learn about causality. (Also confusingly, "reduced form" can refer to other related but different concepts.)

Both approaches consider **counterfactual** analysis, but in different ways. Broadly, a counterfactual is a universe that's different than our actual universe. Usually, the

counterfactual universe is nearly identical to our actual universe except for one particular policy whose effect we want to learn.

The **reduced-form** approach tries to isolate causal effects by using comparisons that are either randomized or "as good as randomized." For example, **randomized** treatment means that individuals are randomly assigned to be treated or not, without regard to their characteristics. Hopefully, it is then appropriate to interpret the mean difference as the average effect of the treatment. "As good as randomized" means that although we did not explicitly randomly assign treatment, the actual assignment mechanism did not depend on individuals' characteristics anyway. More often this is (hoped to be) true after some other adjustment is made.

In contrast, the **structural approach** tries to more explicitly model the inner workings of causal systems. Structural models often come from economic theory, like decision-making or market equilibria models. The goal is to estimate such models' parameters, like elasticities, discount factors, risk aversion, and demand curves. There are different ways people define "structural," but I think a helpful definition is: a model that is invariant to a set of policies under consideration. If we are considering very large, macro-level policy changes, then we would need a relatively complex model, otherwise the policy changes could change the model (for example, through general equilibrium effects). If we are considering relatively small, micro-level policy changes, then a simpler model may suffice. Either way, the hope is that we can estimate the structural model and use it to guess the causal effect of each possible policy change.

The structural and reduced-form approaches have complementary advantages, and often both are helpful; for example, see the survey by Lewbel (2019). Structural models often require stronger (less realistic) assumptions, but in return they can analyze a wider variety of possible policies. Also, there can be relatively vague "structural" models (like in this book!), or relatively complex reduced-form models.

Example 3.16 (Kaplan video)**.** Imagine trying to learn how a retirement pension formula (i.e., how much money somebody gets paid after retiring, based on their years of experience, age, and salary history) affects the age at which a teacher decides to retire. A reduced-form analysis might compare the mean retirement age of teachers who joined a school in the year 1998 with the mean retirement age of teachers who joined in 1999, just after the formula was changed, hoping that the two groups of teachers are otherwise "as good as randomized." A structural analysis might explicitly model a teacher's retirement decision within an expected utility framework that "discounts" the value of future periods (like net present value). The structural analysis requires strong (maybe unrealistic) assumptions about things like the utility function and the distribution of unobserved variables. However, it can then evaluate the effect of hypothetical pension changes that may have never been implemented before, rather than only estimating the effect of the historical 1999 pension change.

Example 3.17. Imagine trying to learn about the effect of free public childcare on how much mothers work in the formal sector. A reduced-form analysis might estimate how

much mothers work in cities that just opened such childcare centers last year compared to mothers in cities that plan to open them next year. The hope is that whether a city opens the childcare centers last year or next year is "as good as randomized," so that the mean difference in hours worked can be interpreted as the effect of the childcare (rather than the effect of something else that's different). A structural analysis might try to estimate an economic model of a mother's decision to work in the formal sector, including variables like the price of childcare, wages, and utility from different activities. Such a model requires strong assumptions (although "as good as randomized" may also be unrealistic!), but can then be used to evaluate the effects of a wide variety of hypothetical policies, not only the effect of the childcare centers that opened last year.

In Sum: Structural & Reduced-Form Approaches

Reduced-form: randomized or "as good as randomized" comparisons to isolate causality
Structural: more explicit economic models of causal relationships

3.7 Linear Regression

Discussion Question 3.7 (Model interpretation)**.** Interpret $Y = \beta_0 + \beta_1 X + U$. (As a concrete example: Y is wage, X is years of education.) In particular,

a) what does β_1 mean?
b) what does U mean?

The method of **ordinary least squares** (OLS) can be defined in multiple ways that each help illustrate a more general point.

The following notation is used in later chapters, too. Let Y be a scalar random variable that is the outcome of interest, like an individual's earnings or a state's traffic fatality rate. Let $\boldsymbol{X}$ be a column vector containing all the **regressors**, also known as **covariates** or **predictors** or **right-hand-side variables** or **independent variables**, usually with 1 as the first element, like $\boldsymbol{X} = (1, X_2, X_3, \ldots)'$. (Note that Wooldridge (2010) defines $\boldsymbol{X}$ as a row vector to avoid needing as many transpose symbols, but at the expense of needing to remember which vectors are columns vs. rows.) In this book, usually U is an unobserved scalar structural error term (with some causal meaning), whereas V is a statistical error term defined with respect to a linear projection or conditional mean.

3.7.1 Linear Projection

This subsection draws from Sections 7.3–7.5 of Kaplan (2022a).

Fundamentally, OLS estimates the coefficients of the **linear projection** (LP) of Y onto $\boldsymbol{X}$. The LP is a population object:

$$\text{LP}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\beta}, \tag{3.4}$$

where vector

$$\boldsymbol{\beta} = [\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{X}Y) \tag{3.5}$$

contains the **linear projection coefficients** (LPCs). This $\boldsymbol{\beta}$ is a feature of the population, i.e., it is a summary of the joint distribution of $(Y, \boldsymbol{X}')$. Assuming $\boldsymbol{\beta}$ is well-defined, iid sampling is sufficient for the OLS estimator $\hat{\boldsymbol{\beta}}$ to be consistent for $\boldsymbol{\beta}$, written $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$; details are below.

The **best linear predictor** (BLP) is another interpretation of the population LP. Often we think of "prediction" in terms of data, but here it is meant in the sense of trying to guess Y given $\boldsymbol{X}$ in the population. The "best" guess depends on the consequences of a wrong guess. For mathematical convenience, this is often quantified by squaring the difference between the guessed value and the true value Y, which is called **quadratic loss** (or L_2 loss). In BLP, "linear predictor" means a guess of the form $\boldsymbol{X}'\boldsymbol{g}$ (a linear combination of the predictors $\boldsymbol{X}$, where $\boldsymbol{g}$ is a non-random vector). It turns out that the LPC $\boldsymbol{\beta}$ solves

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{g}} \mathrm{E}[(Y - \boldsymbol{X}'\boldsymbol{g})^2]. \tag{3.6}$$

That is, among all possible predictions of the form $\boldsymbol{X}'\boldsymbol{g}$, the mean squared prediction error $\mathrm{E}[(\text{true} - \text{guess})^2]$ is minimized by setting $\boldsymbol{g} = \boldsymbol{\beta}$. Besides the caveat that quadratic loss may not reflect the actual consequences of our incorrect predictions, another caveat is that "best" does not mean "good": it could be that all predictions of the form $\boldsymbol{X}'\boldsymbol{g}$ are awful, and the BLP is merely the least bad.

Beyond our scope...

What if we replace quadratic loss with another loss function? (For binary Y, see Section 14.3.) If instead of squaring the error we take the absolute value, then we get the so-called "median regression" estimator. If more generally we use a tilted version of the absolute value function that allows positive errors to be worse (or better) than negative errors, then we get "quantile regression." For example, see Part II of Kaplan (2021).

The **best linear approximation** (BLA) is yet another interpretation of the LP. "Best" again refers to minimizing mean squared error (and again does not mean "good"!), and "linear" again refers to the functional form $\boldsymbol{X}'\boldsymbol{g}$ that takes a linear combination of $\boldsymbol{X}$. "Approximation" refers to approximation of the conditional mean $\mathrm{E}(Y \mid \boldsymbol{X})$. That is, the LPC $\boldsymbol{\beta}$ also solves

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{g}} \mathrm{E}\{[\mathrm{E}(Y \mid \boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{g}]^2\}. \tag{3.7}$$

This could also be written in terms of the **conditional mean function** (CMF), also called the **conditional expectation function** (CEF),

$$m(\boldsymbol{x}) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}). \tag{3.8}$$

Note that $m(\cdot)$ is a non-random function: it maps each possible non-random value $\boldsymbol{x}$ (lowercase) to the corresponding non-random scalar $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$, i.e., the mean Y among individuals in the subpopulation with $\boldsymbol{X} = \boldsymbol{x}$. Given this $m(\cdot)$, (3.7) can be written

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{g}} \mathrm{E}\{[m(\boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{g}]^2.\} \tag{3.9}$$

As noted above, OLS most fundamentally estimates the LP/BLP/BLA. The OLS estimator can be written as the sample analog of (3.5),

$$\hat{\boldsymbol{\beta}} = [\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{X}Y). \tag{3.10}$$

Given iid sampling, sample moments converge in probability to the corresponding population moments by the weak law of large numbers (WLLN), so

$$\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}') \xrightarrow{p} \mathrm{E}(\boldsymbol{X}\boldsymbol{X}'), \quad \widehat{\mathrm{E}}(\boldsymbol{X}Y) \xrightarrow{p} \mathrm{E}(\boldsymbol{X}Y),$$

which can then be combined by the continuous mapping theorem (again assuming everything is well-defined):

$$\hat{\boldsymbol{\beta}} = [\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{X}Y) \xrightarrow{p} [\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{X}Y) = \boldsymbol{\beta}.$$

The above **sample analog** form of the OLS estimator can be derived from the "least squares" definition that mirrors (3.6),

$$\hat{\boldsymbol{\beta}} \equiv \arg\min_{\boldsymbol{b}} \widehat{\mathrm{E}}[(Y - \boldsymbol{X}'\boldsymbol{b})^2] = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i'\boldsymbol{b})^2. \tag{3.11}$$

The objective function is clearly convex (satisfying the second-order condition), so the (unique global) minimizer solves the first-order condition

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{b}}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i'\boldsymbol{b})^2\bigg|_{\boldsymbol{b}=\hat{\boldsymbol{\beta}}} = \frac{1}{n}\sum_{i=1}^{n}2\boldsymbol{X}_i(Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}) = \frac{2}{n}\sum_{i=1}^{n}(\boldsymbol{X}_iY_i - \boldsymbol{X}_i\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}).$$

Dividing by 2 and solving for $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_iY_i = [\widehat{\mathrm{E}}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{X}Y).$$

Just as we can interpret the population $\boldsymbol{\beta}$ in terms of LP, BLP, or BLA, we can also interpret the estimator in terms of the sample minimization problem (3.11) that parallels the BLP population minimization problem, or as the sample analog (3.10) of the population LPC expression in (3.5).

Often the linear projection model is written in **error form**, but note that the above does not fundamentally require any such "error term." Nonetheless, sometimes it is convenient to define the linear projection error V as the difference between the true Y and the linear projection:

$$V \equiv Y - \text{LP}(Y \mid \boldsymbol{X}) = Y - \boldsymbol{X}'\boldsymbol{\beta}. \tag{3.12}$$

Given this definition of V, it follows automatically that

$$\text{E}(\boldsymbol{X}V) = \boldsymbol{0},$$

meaning every element of the vector $\text{E}(\boldsymbol{X}V)$ is zero. This is not an assumption: it is a property that follows from the definition of V as the linear projection error.

3.7.2 Conditional Mean Function

Linear CMF is LP/BLP/BLA

While OLS most fundamentally estimates the LP/BLP/BLA, under stronger assumptions it can estimate the CMF defined in (3.8). From (3.9), if the true CMF happens to have the linear functional form $m(\boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\gamma}$ for some non-random vector $\boldsymbol{\gamma}$, then

$$m(\boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{\gamma} - \boldsymbol{X}'\boldsymbol{b}$$

can be set to zero by setting $\boldsymbol{b} = \boldsymbol{\gamma}$, in which case the entire RHS is zero and thus the global minimum because the RHS is non-negative (due to the squaring). Thus, the RHS is $\boldsymbol{\gamma}$ (defined above to be the CMF coefficient vector), and the LHS is the LPC $\boldsymbol{\beta}$, so $\boldsymbol{\beta} = \boldsymbol{\gamma}$ and the LP and CMF are equal. That is, if the CMF is "linear" in the sense of having the same functional form as the LP, then the LP and CMF are equal, so we can interpret the OLS estimand (which fundamentally is the LP) as the CMF. In some cases, this is always true, like if $\boldsymbol{X} = (1, X)'$ where X is a dummy variable (only takes value 0 or 1).

> **Beyond our scope...**
>
> There are ways to estimate the CMF without requiring that you guess the exact functional form ahead of time, called **nonparametric regression**; for example, see Part V of Kaplan (2021).

CMF in Error Form

Like the LP, the CMF can also be written in error form. Parallel to the LP error defined in (3.12), the CMF error is defined as

$$V \equiv Y - m(\boldsymbol{X}), \tag{3.13}$$

where $m(\cdot)$ is the CMF defined in (3.8). The property $\mathrm{E}(V \mid \boldsymbol{X}) = 0$ follows from the definition; it is not an additional assumption. That is, if we write

$$Y = m(\boldsymbol{X}) + V \tag{3.14}$$

with $m(\cdot)$ the CMF from (3.8), then it automatically follows that

$$\mathrm{E}(V \mid \boldsymbol{X}) = \mathrm{E}(Y - m(\boldsymbol{X}) \mid \boldsymbol{X}) = \mathrm{E}(Y \mid \boldsymbol{X}) - m(\boldsymbol{X}) = 0, \tag{3.15}$$

where the first equality plugs in the definition in (3.13), the second uses the linearity property of $\mathrm{E}(\cdot)$, and the third uses the definition of $m(\cdot)$.

CMF vs. Conditional Mean

One common confusion is the difference between $m(\boldsymbol{x})$ and $m(\boldsymbol{X})$. The former is a non-random function evaluated at a non-random value, hence $m(\boldsymbol{x})$ is a non-random value, like 7 or -1.1. The latter is a non-random function evaluated at a random value $\boldsymbol{X}$, hence $m(\boldsymbol{X})$ is also a random variable. For example, imagine scalar X with $\mathrm{P}(X = 0) = \mathrm{P}(X = 1) = 0.5$, and $m(x) = x + 2$; then $\mathrm{P}(m(X) = 2) = \mathrm{P}(X = 0) = 0.5$ and $\mathrm{P}(m(X) = 3) = \mathrm{P}(X = 1) = 0.5$, showing that $m(X)$ is a random variable, whereas $m(0) = 2$ and $m(1) = 3$ are non-random values.

Similarly, $\mathrm{E}(V \mid \boldsymbol{X})$ is a random variable, whereas $\mathrm{E}(V \mid \boldsymbol{X} = \boldsymbol{x})$ is a non-random value. The expression $\mathrm{E}(V \mid \boldsymbol{X}) = 0$ means that zero is the only possible value the random variable takes. Equivalently, we could also write $\mathrm{E}(V \mid \boldsymbol{X} = \boldsymbol{x}) = 0$ for all possible values $\boldsymbol{x}$, which may be easier to understand.

CMF with Binary Regressor

Consider the special case with binary X. Let $m(x) = \beta_0 + \beta_1 x$. Then we can solve for the parameters from $m(0) = \beta_0$ and $m(1) = \beta_0 + \beta_1$, implying $\beta_1 = m(1) - m(0)$. That is, the intercept β_0 is the conditional mean of Y when $X = 0$, and the slope β_1 is the difference in conditional means of Y between $X = 1$ and $X = 0$.

3.7.3 Causal Interpretation

Under additional assumptions, the LP or CMF can have a causal interpretation. This is left to later chapters.

3.8 Economic Significance

3.8.1 Basic Idea

The term **economic significance** refers to the magnitude of an estimated parameter. An estimate is not economically significant if it is "economically" negligible (not meaningfully

different than zero). "Economically" just means "for real-world purposes," like whether it is important to consider for policy purposes. One way to think about this is: would you personally care about the difference? For example, imagine $\hat{\theta}$ estimates the effect on your final exam score of studying an additional hour per week. Would you care about having a final exam score that's $\hat{\theta}$ percentage points higher? If $\hat{\theta} = 0.01$, then no; if $\hat{\theta} = 50$, then yes. Of course, it's a continuum, so somewhere between "yes" and "no" are varying degrees of "maybe," corresponding to varying degrees of moderate economic significance.

Example 3.18. Would you care if you had $\hat{\theta} = 2$ additional years of education? This is a lot, like an entire master's degree, so presumably you would indeed care.

3.8.2 Units of Measure

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

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

3.8.3 Log Models

In addition to units of measure, coefficient interpretations depend on whether a variable enters the model in levels or in logs. In economics, "log" always refers to the natural log.

The interpretations of different log models are detailed in Section 8.1 of Kaplan (2022a). Here is a summary. If control variables are added to the model, the interpretations do not change, unless there are interaction terms involving the regressor of interest X. The approximations below come from a linearization of the (natural) log function $\log(w)$ around $w = 1$: $\log(w) \approx w - 1$. Such an approximation is pretty good if $|w - 1| \leq 0.1$ or so, with the approximation error increasing in $|w - 1|$.

A **log-linear** model has the form

$$\log(Y) = \beta_0 + X\beta_1 + U. \tag{3.16}$$

A one-unit increase in X is associated with a $100(e^{\beta_1} - 1)\%$ change in Y. If β_1 is near zero, then this is approximately $100\beta_1\%$. A d-unit increase in X is associated with a $100(e^{d\beta_1} - 1)\%$ change in Y, or approximately $100d\beta_1\%$ for small enough $d\beta_1$.

A **linear-log** model has the form

$$Y = \beta_0 + \log(X)\beta_1 + U. \tag{3.17}$$

A 1% increase in X is associated with a $\beta_1 \log(1.01)$-unit change in Y, which is approximately a $\beta_1/100$-unit change. A $100p\%$ increase in X is associated with a $\beta_1 \log(1+p)$-unit change in Y. For small p, this is approximately a $p\beta_1$-unit change.

A **log-log** model has the form

$$\log(Y) = \beta_0 + \log(X)\beta_1 + U. \tag{3.18}$$

A 1% increase in X is associated with a $100(1.01^{\beta_1} - 1)\%$ change in Y, which is approximately a $\beta_1\%$ change in Y (i.e., an elasticity). A $100p\%$ increase in X is associated with a $100((1+p)^{\beta_1} - 1)\%$ change in Y. For small $p\beta_1$, this is approximately $100p\beta_1\%$.

3.9 Quantifying Uncertainty

See also Section 3.7 of Kaplan (2022a), as well as Section 3.8 ("Quantifying Uncertainty: Misinterpretation and Misuse").

While an estimator provides a best guess about the true population value given the data (roughly speaking), we usually also want a sense of our uncertainty about the true value. The most common frequentist methods to quantify uncertainty are confidence intervals and p-values from hypothesis tests. This book focuses on confidence intervals because most econometricians agree they are less likely to be misinterpreted or misused (than p-values). Additionally, in many settings, with large n the frequentist confidence interval is very similar to the Bayesian credible interval. Hopefully you have already learned the basics of hypothesis testing and p-values (because they are still reported, and occasionally still useful), like that an estimate $\hat{\theta}$ is **statistically significant** at level α if the p-value for $H_0\colon \theta = 0$ is below α, or that failing to reject such a null hypothesis should not be interpreted as our best guess being $\theta = 0$. For example, if $\hat{\theta} = 0.1$ and $p = 0.33$, then we would not be surprised to see such a dataset if indeed $\theta = 0$, but we would also not be surprised to see such a dataset if $\theta = 0.2$, etc.

A **confidence interval** (CI) only quantifies uncertainty due to random sampling, not uncertainty about identifying assumptions. For example, a CI for the linear projection slope accounts only for the uncertainty due to having finite sample size n, not due to uncertainty about the true CMF being linear, nor due to uncertainty about the CMF slope having a causal interpretation. This can be misleading. If we have very large n, then our CI will be very short (because we have very little uncertainty about the LPC), even if we are very uncertain about the CMF being linear or having a causal interpretation. For this reason, it is useful to know the fundamental population parameter that a particular estimator is consistent for, like how OLS is fundamentally consistent for the LPC.

A CI provides a range of values that should contain the true value with high probability. Recall that from the frequentist perspective, "probability" is from the before-sampling perspective, and the CI is random (because it depends on the observations, which are modeled as random), whereas the true population value is non-random. That is, a CI can be seen as a procedure such that (before sampling) we have a high probability of randomly sampling a dataset for which the CI contains the true value. This probability is called

the **coverage probability** (CP). That is, given population value θ and CI $[\hat{L}, \hat{U}]$, where the lower and upper endpoints $\hat{L}$ and $\hat{U}$ are computed from data (thus random variables), the coverage probability is

$$\text{CP} \equiv \text{P}(\hat{L} \leq \theta \leq \hat{U}) = \text{P}(\theta \in [\hat{L}, \hat{U}]). \tag{3.19}$$

The **confidence level** or **nominal coverage probability** is the desired coverage probability. Usually a CI is justified by an asymptotic argument such that asymptotically, its coverage probability equals the nominal level. (Sometimes the asymptotic coverage probability is only shown to be greater than or equal to the nominal level.) However, for finite n, the coverage probability may be higher or lower than desired. If it is higher than desired, then the CI is "too conservative": hypothetically, it could be shortened and still achieve the desired coverage probability. Lower than desired CP is usually considered even worse: we are over-confident about how precise our estimates are. Of course, even with a 95% CI, our CI fails to include the true value 5% of the time, so we should never be too confident anyway. This is why replication is an important part of any science (although that begs the question of whether economics is truly a science!).

Coverage probability can also be interpreted in terms of repeated sampling. For example, if we have a 90% probability of randomly sampling a dataset for which the CI contains the true value, and we randomly sample 100 datasets, then in roughly 90 of the 100 datasets the CI should contain the true value.

Instead of a binary label of "statistically significant at a 5% level" whenever $p < 0.05$, it is more helpful to look at the full range of possible population values included in the CI when quantifying uncertainty. At minimum: consider the economic significance of the lowest value in the CI, the estimated value, and the highest value in the CI. If a confidence level $100(1-\alpha)\%$ CI does not contain zero, then it is "statistically significant at level α," but that is usually not the most helpful statement to make. For example, if a 95% CI is $[0.1, 1.1]$, then there is statistical significance at a 5% level.

Discussion Question 3.8 (salary increase significance)**.** Imagine you compute a 95% CI of $[4.1, 5.9]$ around your estimated annual salary effect of $\hat{\theta} = 5$ dollars per year. Are these results statistically significant? Are they economically significant? Explain. Hint: would you care if your annual salary increased by $\hat{\theta} = 5$ dollars per year?

Discussion Question 3.9 (significance: distance and education)**.** Let Y be years of education, and let X be distance from someone's childhood home to the nearest college or university, measured in kilometers ($1\,\text{km} = 0.6\,\text{mi}$). Let θ be the causal effect of X on Y. You think you found an "as good as randomized" natural experiment, from which you estimate $\hat{\theta} = -0.03$. You compute a 95% CI of $[-0.05, -0.01]$.

a) How economically significant is the point estimate of -0.03? Hint: consider the units.
b) Is this statistically significant at a 5% level?
c) More generally, discuss your CI and uncertainty.

There are many possible ways to misinterpret or misuse confidence intervals (or p-values), including the following (not exhaustive!).

- Multiple testing: if you take enough random samples, or test enough different hypotheses in the same sample, you will eventually get a "statistically signficant" result; for example, see this insightful comic (xkcd.com/882) that illustrates the **multiple testing problem** (or **multiple comparisons problem**), or this video.
- Publication bias: if statistically significant results are more likely to be published, then it's similar to the multiple testing problem in the linked comic, where we only read about the one significant result but not the 19 not-significant results.
- Assumptions: a CI may not be valid if it is based on iid sampling but the actual data were not sampled iid; and the CI does not account for additional interpretations of the population value based on identifying assumptions.
- Frequentist results may be misinterpreted as Bayesian, like a p-value being misinterpreted as the probability that the null hypothesis is true.
- Unlikely events happen: even if you only run one test on one dataset with confidence level 99%, your dataset may be in the unlucky 1% for which the true value is outside the CI.

Example 3.19 (Kaplan video)**.** Your friend claims to have magical powers. You have a deck of playing cards; you repeatedly draw a card (without showing it) and ask your friend to guess whether the card is black or red. You record the data and compute a 90% CI for your friend's probability p of guessing correctly. Random guessing would yield $p = 0.5$, but your CI is $[0.52, 0.61]$, all values above 0.5. Your friend's interpretation is that statistics have now proved true the claim of magical powers. However, you think it was just luck and ask to gather more data. Indeed, the new dataset's 90% CI is $[0.44, 0.51]$. You try another few datasets, and those CIs also contain 0.5. It seems the first result was simply luck, not magic.

Discussion Question 3.10 (frequentist or Bayesian?)**.** For each of the following, say whether it is a frequentist question, Bayesian question, neither, or both; if both, explain the two possible interpretations. Hint: use Section 3.3 as well as Section 3.9.

a) What's the probability that the current natural unemployment rate in the U.S. is between 4.5% and 7.5%?
b) Can we create a diagnostic tool for our company's daily website traffic data to identify whether it's normal or has been hacked, limiting the rate of falsely reporting "hacked" on normal days to only 1% of normal days?
c) What is the probability that the true unemployment rate is within 1 percentage point of the estimated unemployment rate?
d) Is the positive estimate $\hat{\theta} > 0$ primarily due to the income effect or substitution effect?

3.10 Quantifying Accuracy of an Estimator

This section is mostly from Section 3.6 of Kaplan (2022a).

From the frequentist perspective, an estimator's accuracy can be quantified by comparing features of its sampling distribution to the true population value. The **sampling distribution** views the estimator from the before-sampling perspective; for intuition, you can imagine taking 1000 random samples and plotting a histogram of the estimated values. Bias is an important, commonly mentioned property, but it is not sufficient to quantify accuracy. Mean squared error better quantifies accuracy.

Throughout, let θ be the population parameter estimated by $\hat{\theta}_n$; for example, $\theta = \mathrm{E}(Y)$ and $\hat{\theta}_n = \bar{Y}_n$.

3.10.1 Bias

Definitions

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

$$\mathrm{Bias}(\hat{\theta}_n) \equiv \mathrm{E}(\hat{\theta}_n) - \theta. \tag{3.20}$$

The bias captures if the estimator systematically differs from θ in a particular direction, i.e., how wrong the average $\hat{\theta}_n$ is.

There are four types of bias:

$$\begin{aligned}
\textbf{upward bias (positive bias)}:&\ \mathrm{E}(\hat{\theta}_n) > \theta,\\
\textbf{downward bias (negative bias)}:&\ \mathrm{E}(\hat{\theta}_n) < \theta,\\
\textbf{attenuation bias (bias toward zero)}:&\ 0 < \frac{\mathrm{E}(\hat{\theta}_n)}{\theta} < 1, \text{so } |\mathrm{E}(\hat{\theta}_n)| < |\theta|,\\
\text{bias away from zero}:&\ \frac{\mathrm{E}(\hat{\theta}_n)}{\theta} > 1, \text{so } |\mathrm{E}(\hat{\theta}_n)| > |\theta|.
\end{aligned}$$

An estimator is **unbiased** if its bias is zero. Using (3.20),

$$\mathrm{Bias}(\hat{\theta}) = 0 \iff \mathrm{E}(\hat{\theta}) = \theta, \tag{3.21}$$

where symbol $\iff$ can be read as "is equivalent to" (see Chapter 2).

Example 3.20 (Kaplan video)**.** With iid sampling, the sample mean is an unbiased estimator of the population mean. The estimator is $\hat{\theta}_n = \bar{Y}_n$, and the population parameter is $\theta = \mathrm{E}(Y)$. With $n = 1$, $\bar{Y}_1 = Y_1$, so $\mathrm{E}(\bar{Y}_1) = \mathrm{E}(Y_1) = \mathrm{E}(Y)$. With $n = 2$,

$$\mathrm{E}[\bar{Y}_2] = \mathrm{E}[(1/2)Y_1 + (1/2)Y_2] = \overbrace{(1/2)\,\mathrm{E}(Y_1)}^{\mathrm{E}(Y)/2} + \overbrace{(1/2)\,\mathrm{E}(Y_2)}^{\mathrm{E}(Y)/2} = \mathrm{E}(Y), \tag{3.22}$$

using the linearity property of $\mathrm{E}(\cdot)$. Similar derivations hold for any n, so $\mathrm{E}(\bar{Y}_n) = \mathrm{E}(Y)$, thus the bias is zero given (3.21).

Example 3.21 (Kaplan video)**.** The estimator $\hat{\theta}_n = \bar{Y}_n + 1$ has positive bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(\bar{Y}_n + 1) = \mathrm{E}(\bar{Y}_n) + 1 = \mathrm{E}(Y) + 1 > \mathrm{E}(Y)$. The estimator $\hat{\theta}_n = \bar{Y}_n - 2$ has negative bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(\bar{Y}_n - 2) = \mathrm{E}(\bar{Y}_n) - 2 = \mathrm{E}(Y) - 2 < \mathrm{E}(Y)$. The estimator $\hat{\theta}_n = 0.5\bar{Y}_n$ has attenuation bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(0.5\bar{Y}_n) = 0.5\,\mathrm{E}(\bar{Y}_n) = 0.5\,\mathrm{E}(Y)$, so $0 < [\mathrm{E}(\hat{\theta}_n)/\,\mathrm{E}(Y)] = 0.5 < 1$.

Insufficiency of Bias to Quantify Accuracy

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

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

$$\mathrm{P}(\hat{\theta}_1 = -100) = \mathrm{P}(\hat{\theta}_1 = 100) = 1/2. \tag{3.23}$$

The second estimator's distribution is

$$\mathrm{P}(\hat{\theta}_2 = 1) = 1. \tag{3.24}$$

The first estimator has smaller bias. The estimators' means are

$$\mathrm{E}(\hat{\theta}_1) = (1/2)(-100) + (1/2)(100) = 0, \quad \mathrm{E}(\hat{\theta}_2) = (1)(1) = 1. \tag{3.25}$$

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

$$\mathrm{Bias}(\hat{\theta}_1) = \mathrm{E}(\hat{\theta}_1) - \theta = 0 - 0 = 0, \quad \mathrm{Bias}(\hat{\theta}_2) = \mathrm{E}(\hat{\theta}_2) - \theta = 1 - 0 = 1. \tag{3.26}$$

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

But intuitively, $\hat{\theta}_2$ is much better. It always differs from the true θ by only 1, whereas $\hat{\theta}_1$ always differs by 100, which is much worse. That is, regardless of the dataset, $\hat{\theta}_2$ is always 100 times closer than θ_1 to the true $\theta = 0$. This illustrates how bias alone does not properly quantify our preferences: it tells us to prefer $\hat{\theta}_1$ (lower bias) when in fact we strongly prefer $\hat{\theta}_2$ (always much closer to θ).

3.10.2 Mean Squared Error

⟹ Kaplan video: MSE Examples

The **mean squared error** (MSE) is a more complete measure of "how bad" an estimator is. The idea is analogous to using quadratic loss for prediction as in (3.6). Among other possible loss functions, this is most common and generally reasonable. MSE is mean quadratic loss:

$$\mathrm{MSE}(\hat{\theta}) \equiv \mathrm{E}[L_2(\hat{\theta}, \theta)] = \mathrm{E}[(\hat{\theta} - \theta)^2]. \tag{3.27}$$

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

$$\begin{aligned}\text{MSE}(\hat{\theta}_1) &= \text{E}[(\hat{\theta}_1 - \theta)^2] = (1/2)(-100-0)^2 + (1/2)(100-0)^2 = 10{,}000,\\ \text{MSE}(\hat{\theta}_2) &= \text{E}[(\hat{\theta}_2 - \theta)^2] = (1)(1-0)^2 = 1.\end{aligned}$$

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

MSE can also be decomposed into variance plus squared bias. The variance is

$$\text{Var}(\hat{\theta}) \equiv \text{E}[(\hat{\theta} - \text{E}(\hat{\theta}))^2]. \tag{3.28}$$

(The square root of this is the standard deviation, also called the "standard error" of the estimator $\hat{\theta}$.) Skipping the math, using the bias and variance definitions in (3.20) and (3.28),

$$\text{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2. \tag{3.29}$$

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

Example 3.22 (Kaplan video)**.** Continue the previous example, but instead of assuming $\theta = 0$, let

$$\text{P}(\hat{\theta}_1 = \theta - 100) = \text{P}(\hat{\theta}_1 = \theta + 100) = 1/2, \quad \text{P}(\hat{\theta}_2 = \theta + 1) = 1. \tag{3.30}$$

The MSEs are the same as before because the θ cancels out:

$$\begin{aligned}\text{MSE}(\hat{\theta}_1) &= \text{E}[(\hat{\theta}_1 - \theta)^2] = (1/2)(\theta-100-\theta)^2 + (1/2)(\theta+100-\theta)^2 = 10{,}000,\\ \text{MSE}(\hat{\theta}_2) &= \text{E}[(\hat{\theta}_2 - \theta)^2] = (1)(\theta+1-\theta)^2 = 1.\end{aligned} \tag{3.31}$$

Example 3.23 (Kaplan video)**.** Imagine we know the bias and variance of two estimators, but not the full sampling distributions. This is still sufficient to compute MSE using (3.29). For example, let

$$\text{Bias}(\hat{\beta}_1) = 1, \text{Var}(\hat{\beta}_1) = 16, \quad \text{Bias}(\hat{\beta}_2) = 10, \text{Var}(\hat{\beta}_2) = 9. \tag{3.32}$$

Plugging these into (3.29),

$$\text{MSE}(\hat{\beta}_1) = 1^2 + 16 = 17, \quad \text{MSE}(\hat{\beta}_2) = 10^2 + 9 = 109. \tag{3.33}$$

According to MSE, $\hat{\beta}_1$ is better because it has lower MSE ("less bad") than $\hat{\beta}_2$. In this case, although $\hat{\beta}_1$ has larger variance, its bias is enough smaller than its overall MSE is also smaller.

Discussion Question 3.11 (estimator MSE)**.** Consider three estimators of the population mean $\mu = \text{E}(Y)$, and their three sampling distributions: $\hat{\mu}_1 \sim \text{N}(\mu, 25)$, $\hat{\mu}_2 \sim \text{N}(\mu+3, 16)$, and $\hat{\mu}_3 \sim \text{N}(\mu+2, 9)$, i.e., the sampling distributions of the three estimators are all normal distributions with respective means μ, $\mu+3$, and $\mu+2$, and respective variances 25, 16, and 9.

a) Compute the MSE of each estimator.

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

3.10.3 Consistency and Asymptotic MSE

At the intuitive level, an estimator is **consistent** if in "large" samples (large n), there is a "high" probability of the estimator being "close" to the true value. This is similar to the idea of "probably approximately correct" in computer science: estimator $\hat{\theta}_n$ is "consistent" if with large n it is "probably approximately correct." Unfortunately, there are usually no precise quantitative definitions of "large," "high," and "close."

If $\hat{\theta}_n$ is not consistent, then it has **asymptotic bias**: even with infinite data, the estimator would still be biased. One way to formally define asymptotic bias is

$$\text{AsyBias}(\hat{\theta}_n) \equiv \operatorname*{plim}_{n\to\infty} \hat{\theta}_n - \theta. \tag{3.34}$$

Analogous to "unbiasedness" being "zero bias," here "consistency" is "zero asymptotic bias": roughly speaking, with a large dataset, there is very little bias. There are the same four types of asymptotic bias as bias: upward/positive, downward/negative, attenuation, and away from zero.

It is also possible to compare approximate (asymptotic) mean squared error by comparing asymptotic distributions. Again, lower is better, and it depends on both bias and variance components. For two consistent estimators, this reduces to comparing asymptotic variance. For example, if $\sqrt{n}(\hat{\theta}_1 - \theta) \xrightarrow{d} \text{N}(0, \sigma_1^2)$ and $\sqrt{n}(\hat{\theta}_2 - \theta) \xrightarrow{d} \text{N}(0, \sigma_2^2)$, then we prefer estimator $\hat{\theta}_1$ (and call it more **efficient** than $\hat{\theta}_2$) iff $\sigma_1 < \sigma_2$.

Beyond our scope...

In contexts like nonparametric regression, there is also an important bias term, even asymptotically, and procedures are designed to try to minimize the asymptotic MSE; for example, see Chapter 18 ("Model Selection") of Kaplan (2021).

Chapter 4
Identification by Independence

Unit learning objectives for this chapter

4.1. Explain mathematically and verbally how an independence condition can achieve identification, in both structural and potential outcomes models. [TLOs 2 and 3]

4.2. In real-world examples, provide reasons why the key identifying assumption probably does (not) hold. [TLO 4]

To develop intuition and vocabulary, this chapter explains identification in the simplest structural and potential outcomes models.

Some material is from Chapters 4 and 6 of Kaplan (2022a). Some of the same topics as in Section 4.1 are covered in Sections 21.1–3 of Wooldridge (2010).

Optional resources for this chapter

- Causal inference intro (Masten video)
- Correlation vs. causation (Masten video)
- Potential outcomes and SUTVA (Wikipedia)
- Potential outcomes example (Masten video)
- SUTVA and spillovers (Masten video)
- Individual causal effects (Masten video)
- ATE (Masten video)
- ATT (Masten video)
- Assumptions for randomized experiment validity (Masten video)

4.1 Average Treatment Effect

First, the potential outcomes framework and notation are introduced. Then, the average treatment effect is defined, after which identification results are given.

4.1.1 Potential Outcomes

⟹ Kaplan video: Potential Outcomes and the ATE

This subsection is a shorter version of Section 4.4.1 of Kaplan (2022a).

The **potential outcomes framework** is also called the **Neyman–Rubin causal model** after its two earliest contributors (although sometimes Neyman's name is dropped). It is popular not only in economics, but statistics, medicine, political science, and other fields.

The terms **treatment** and **treatment effect** just refer to any variable and its causal effect on another variable. In English, usually "treatment" makes us think narrowly about medicine (or lumber. . . and facials?), but it can be anything. For example, the "treatment" could be a job training program, and the "treatment effect" is the causal effect of the program on a person's wage. Or, a treatment could be going to a charter school (instead of public school). Another treatment could be a policy or law, like a higher sales tax, or a certain labor law.

As throughout this book, "individual" can mean a firm, county, school, etc.

Imagine two parallel universes. The universes are identical except for one difference: whether or not an individual is treated. The individual's outcome in the universe without treatment is their **untreated potential outcome**, and the individual's outcome in the universe with treatment is their **treated potential outcome**.

Notationally, Y^t represents the treated potential outcome and Y^u the untreated potential outcome. Elsewhere, often Y_1 and Y_0 represent the treated and untreated potential outcomes, or $Y(1)$ and $Y(0)$.

Potential outcomes Y^u and Y^t are not always observable. Often, if an individual is untreated in our universe, then we can observe her untreated potential outcome Y^u, but not her Y^t; conversely, if she is treated, then we observe Y^t but not Y^u. This partial observability makes causal inference more difficult than description or prediction.

Example 4.1 (Kaplan video)**.** Imagine one universe where a student wins the lottery to enter a popular charter school, and another universe where the student remains in the conventional public school. Potential outcomes Y^t and Y^u are dummy (binary) variables for whether or not the student eventually graduated from college in each respective universe. Again, in our universe, we can observe Y^t if the student wins the lottery and Y^u if not, but we cannot observe both.

4.1.2 Treatment Effects

This subsection is a shorter version of Section 4.4.2 of Kaplan (2022a).

The difference $Y^t - Y^u$ between an individual's two potential outcomes is that individual's **treatment effect**. Just as different individuals can have different (Y^u, Y^t), individuals can have different treatment effects $Y^t - Y^u$; i.e., individuals can be affected differently by the same treatment. The fancy term for people being different is **heterogeneity**, more specifically here "treatment effect heterogeneity."

Example 4.2 (Kaplan video)**.** In the charter school example (Example 4.1), $Y^t - Y^u$ is the treatment effect of the charter school on college graduation. That is, it is the difference between the college graduation outcomes in the charter school universe and the public school universe. Because the outcome is binary (1 if graduate college, 0 if don't), there are only four possible values of (Y^u, Y^t) (student types) and only three possible treatment effect values: $Y^t - Y^u = 1$ if the student graduates in the charter school universe $(Y^t = 1)$ but not the public school universe $(Y^u = 0)$; $Y^t - Y^u = -1$ if they only graduate in the public school universe $(Y^u = 1)$ but not the charter school universe $(Y^t = 0)$; and $Y^t - Y^u = 0$ if they graduate either in both universes $(Y^t = Y^u = 1)$ or neither $(Y^t = Y^u = 0)$. This is seen in the later example of Table 4.1.

In economics, where many systems are interrelated, sometimes it's difficult merely to specify which "effect" we care about. For example, consider racial differences in salary. In the parallel universe that's "identical" except for the individual's race, does "identical" include having the same job at the same firm? Or does it allow for an effect of race on hiring? Does it allow for an effect on educational opportunities, or an effect on family background (parents' education, wealth, etc.)? There is no "right" or "wrong" specification, but each answers a different question.

In Sum: Causality in Potential Outcomes Framework

Treatment effect: the difference in outcomes between parallel universes identical except for treatment

4.1.3 Average Treatment Effect

⟹ Kaplan video: Potential Outcomes and the ATE (again)

This subsection is a shorter version of Section 4.5 of Kaplan (2022a).

Although the full distribution of potential outcomes (Y^u, Y^t) contains the most information, usually only certain summary features are studied; here, we focus on the mean.

The **average treatment effect** (ATE) is $\mathrm{E}(Y^t - Y^u)$. "Average" refers to the population mean, while "treatment effect" refers to $Y^t - Y^u$. Thus, the ATE may be interpreted as the probability-weighted average (mean) of all possible individual treatment effects in

the population. Another name for the ATE is the **average causal effect** (ACE), but I use ATE to emphasize that this concept is from the potential outcomes framework.

The ATE has another interpretation. Using the linearity of the expectation operator,

$$\text{ATE} \equiv \text{E}(Y^t - Y^u) = \text{E}(Y^t) - \text{E}(Y^u). \tag{4.1}$$

Here, $\text{E}(Y^t)$ is the mean treated potential outcome, and $\text{E}(Y^u)$ is the mean untreated potential outcome. This could be interpreted as "the treatment effect on the mean outcome": treatment causes the mean outcome to change from $\text{E}(Y^u)$ to $\text{E}(Y^t)$.

Example 4.3 (Kaplan video)**.** Table 4.1 shows a numerical version of the charter school example. The four student "types" refer to the four possible values of (Y^u, Y^t), and each type has its own probability. Given the probabilities, the mean untreated outcome $\text{E}(Y^u)$, mean treated outcome $\text{E}(Y^t)$, and ATE $\text{E}(Y^t - Y^u)$ are

$$\text{E}(Y^u) = (0.3)(0) + (0.3)(0) + (0.1)(1) + (0.3)(1) = 0.4, \tag{4.2}$$
$$\text{E}(Y^t) = (0.3)(0) + (0.3)(1) + (0.1)(0) + (0.3)(1) = 0.6, \tag{4.3}$$
$$\text{E}(Y^t - Y^u) = (0.3)(0) + (0.3)(1) + (0.1)(-1) + (0.3)(0) = 0.2. \tag{4.4}$$

To verify (4.1),

$$\text{E}(Y^t - Y^u) = 0.2 = 0.6 - 0.4 = \text{E}(Y^t) - \text{E}(Y^u). \tag{4.5}$$

Table 4.1: Charter school example population of potential outcomes and ATE.

Student type	Probability	Y^u	Y^t	$Y^t - Y^u$
1	0.3	0	0	0
2	0.3	0	1	1
3	0.1	1	0	−1
4	0.3	1	1	0
Mean		0.4	0.6	0.2

There are some important limitations of the ATE, including the following.

- Zero ATE does not mean zero effect (e.g., it could affect variance).
- ATE compares a universe where everybody is treated to a universe where nobody is treated, which may be unrealistic; often we are interested in more marginal policy changes.

See Section 4.5.2 of Kaplan (2022a) for details and examples.

4.1.4 ATE Identification

Besides their potential outcomes, each individual has a treatment dummy X such that their observed outcome Y is

$$Y = (1 - X)Y^u + XY^t. \tag{4.6}$$

That is, if $X = 0$, then $Y = Y^u$, whereas if $X = 1$, then $Y = Y^t$.

Assumption A4.1 (SUTVA)**.** Everyone with $X = 1$ receives the same treatment, and one individual's treatment does not affect any other individual's potential outcomes.

Assumption A4.1 is usually just called SUTVA, but the main part of it is often called **no interference** (or **non-interference**).

Assumption A4.2 (independence)**.** Treatment is independent of the potential outcomes: $X \perp\!\!\!\perp (Y^u, Y^t)$.

Assumption A4.2 has many names: **independence**, **ignorability**, or **unconfoundedness**. The combination of A4.2 and A4.3 is sometimes called **strong ignorability**. For more detail, history, and discussion, see Imbens and Wooldridge (2007).

Assumption A4.3 (overlap)**.** There is strictly positive probability of both treatment and non-treatment: $0 < \mathrm{P}(X = 1) < 1$.

Assumption A4.3 is intuitive: if everybody (or nobody) is treated, then it's impossible to compare treated and untreated outcomes. For example, if $\mathrm{P}(X = 1) = 0$, then nobody is treated, so it's impossible to learn about $\mathrm{E}(Y^t)$ because Y^t is never observed. Although trivial in this simple context, overlap becomes more important to consider when other conditioning variables are included.

Theorem 4.1 formally states the ATE identification result. Intuitively, the key is that A4.2 allows us to observe representative samples of both Y^u and Y^t; treatment cannot be chosen or assigned based on an individual's potential outcomes. Mathematically, A4.2 implies that the means of the potential outcomes do not statistically depend on the treatment X:

$$\mathrm{E}(Y^t) = \mathrm{E}(Y^t \mid X = 1), \quad \mathrm{E}(Y^u) = \mathrm{E}(Y^u \mid X = 0). \tag{4.7}$$

This condition is called **mean independence**: conditioning on X does not affect the mean of Y^t or of Y^u. Independence implies mean independence; mean independence is weaker than independence. We observe $Y = Y^t$ when $X = 1$ and $Y = Y^u$ when $X = 0$, so

$$\mathrm{E}(Y^t \mid X = 1) = \mathrm{E}(Y \mid X = 1), \quad \mathrm{E}(Y^u \mid X = 1) = \mathrm{E}(Y \mid X = 0). \tag{4.8}$$

Combining (4.7) and (4.8), this says that the population mean of the treated *potential* outcome, $\mathrm{E}(Y^t)$, equals the mean of the *observed* outcome in the treated population, $\mathrm{E}(Y \mid X = 1)$. Thus, $\mathrm{E}(Y^t) = \mathrm{E}(Y \mid X = 1)$ is identified. Similarly, $\mathrm{E}(Y^u) = \mathrm{E}(Y \mid X = 0)$ is identified, so $\mathrm{E}(Y^t) - \mathrm{E}(Y^u)$ is identified.

Theorem 4.1 (ATE identification)**.** *Under A4.1–A4.3, the ATE is identified:*

$$\mathrm{E}(Y^t - Y^u) = \mathrm{E}(Y^t) - \mathrm{E}(Y^u) = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0),$$

which is also the slope in the linear CMF model $\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x$*. More generally, A4.2 can be replaced by the mean independence condition in* (4.7)*.*

Proof. Using the above,

$$\begin{aligned}\text{ATE} \equiv \overbrace{\mathrm{E}(Y^t - Y^u)}^{\text{use linearity, (4.1)}} &= \overbrace{\mathrm{E}(Y^t) - \mathrm{E}(Y^u)}^{\text{use (4.7)}} \\ &= \overbrace{\mathrm{E}(Y^t \mid X = 1)}^{\text{use (4.8)}} - \overbrace{\mathrm{E}(Y^u \mid X = 0)}^{\text{use (4.8)}} \\ &= \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0).\end{aligned}$$

□

Beyond our scope...

We can also learn about "quantile treatment effects" like the median treatment effect, defined as the difference between the medians of the treated and untreated potential outcome distributions. The same identification argument goes through if we assume median independence instead of mean independence; with full independence, all quantile treatment effects (and the ATE) are identified. For example, see Chapter 6 of Kaplan (2021).

Discussion Question 4.1 (college and wage)**.** Let $X = 1$ if an individual has a college degree (the "treatment") and otherwise $X = 0$. Let Y be the individual's wage at age 45, with Y^u and Y^t the potential outcomes. Explain specifically why A4.2 is violated.

Discussion Question 4.2. In which direction do you think self-selection would bias the ATE estimator in the following cases? (Hint: draw pictures.) (Hint: imagine the true ATE is just zero for simplicity; is the sample mean difference positive or negative?)

a) Everyone has the same Y^u.
b) Everyone has the same Y^t.
c) The treatment effect $Y^t - Y^u$ is decreasing in Y^u (i.e., larger Y^u corresponds to lower $Y^t - Y^u$).

4.1.5 SUTVA Violations

As alluded to above, SUTVA can be violated in many ways, especially in economics. This is not about sampling, or randomization, or data; it is about the potential outcomes framework itself. Without SUTVA, it's unclear what "treatment effect" even means.

One common violation of SUTVA is from **spillover effects** that benefit even untreated individuals. That is, the treatment's benefit "spills over" into untreated individuals. Perhaps the treated individuals can share the treatment itself with others, or perhaps others benefit from the improved outcomes of treated individuals.

Example 4.4. Consider a treatment that provides treated individuals with helpful information about financial planning. Treated individuals might share such information with their untreated friends and family. Thus, an untreated individual's outcome may depend on whether or not their friend is treated. This spillover effect violates the "no interference" part of SUTVA.

Example 4.5. Consider a "treatment" that leads to less binge drinking among treated individuals. Even if the treatment itself is not shared, the reduction in binge drinking may reduce social pressure and result in less binge drinking among untreated individuals. Here, untreated individuals are affected by the treatment through the changed behavior of treated individuals. This spillover effect violates SUTVA.

Another common violation of SUTVA is from **general equilibrium effects** (Section 3.5), such as changing market prices.

Example 4.6 (Kaplan video)**.** Consider a new agricultural technology hoping to increase cacao farmers' earnings (through increased productivity). If only one farmer gets this treatment (technology), then she benefits from increased production, selling more cacao at the current global price. But if all farmers in the world get the technology, then the global cacao supply curve shifts and the price drops. Thus, each farmer's untreated and treated potential outcomes (earnings) are affected by all other farmers' treatment status, which affects the market equilibrium price. This violates SUTVA.

Example 4.7. Consider the "treatment" that provides a subsidy for buying a house. This increases demand, which increases prices. This general equilibrium effect violates SUTVA.

Discussion Question 4.3 (cash transfer spillovers)**.** Consider the effect of income on food consumption (Y) in a rural village. Consider an "unconditional cash transfer" program (like GiveDirectly) that gives the equivalent of $1000 to a treated individual. Describe different possible spillover effects that would violate SUTVA.

Beyond our scope...

Check your intuition at https://doi.org/10.3982/ECTA17945 that reports estimates of such spillover effects.

4.1.6 ATT Identification

A common variant of the ATE is the **average treatment effect on the treated** (ATT), less commonly abbreviated ATET. The definition is

$$\text{ATT} \equiv \text{E}(Y^t - Y^u \mid X = 1). \tag{4.9}$$

The ATT is the ATE for the subpopulation of individuals who are actually treated in our universe ($X = 1$).

The ATT identifying assumptions are similar but slightly weaker than for the ATE. Specifically, Assumption A4.4 requires mean-independence of Y^u but not Y^t like before. Intuitively, for the actually-treated ($X = 1$) subpopulation, we always observe $Y = Y^t$, so we only need identifying assumptions to learn about the unobserved Y^u. The precise argument is seen in the proof of Theorem 4.2.

Assumption A4.4 (untreated mean independence)**.** The untreated potential outcome is mean-independent of the treatment: $\mathrm{E}(Y^u \mid X) = \mathrm{E}(Y^u)$.

Theorem 4.2 (ATT identification by independence)**.** *Under A4.1, A4.3, and A4.4, the ATT is identified:* $\mathrm{ATT} = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$.

Proof. Starting from the definition of ATT in (4.9),

$$\begin{aligned}
\mathrm{ATT} &= \overbrace{\mathrm{E}(Y^t - Y^u \mid X = 1)}^{\text{use linearity of } \mathrm{E}(\cdot)} \\
&= \overbrace{\mathrm{E}(Y^t \mid X = 1)}^{\text{use } Y = Y^t \text{ when } X = 1} - \overbrace{\mathrm{E}(Y^u \mid X = 1)}^{\text{use A4.4}} \\
&= \mathrm{E}(Y \mid X = 1) - \overbrace{\mathrm{E}(Y^u \mid X = 0)}^{\text{use } Y = Y^u \text{ when } X = 0} \\
&= \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0).
\end{aligned}$$

□

In economics, often the ATT does not equal the ATE. Mathematically, they can be equal; note that in both Theorems 4.1 and 4.2, the right-hand side is $\mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$, and the assumptions of Theorem 4.2 are strictly weaker than (i.e., are implied by) those of Theorem 4.1, so the assumptions of Theorem 4.1 imply ATE = ATT. However, more generally, often economic agents "select into" treatment (i.e., choose $X = 1$) if they benefit more from it. In such cases, we should generally guess that the ATT makes the treatment appear more beneficial than does the ATE.

Example 4.8 (Kaplan video)**.** Consider the "treatment" of a small business receiving a loan, and the outcome of monthly sales revenue. Although there are certainly other factors, economic theory suggests that the small businesses applying for loans tend to be the ones that would most benefit from loans. Thus, we'd guess that the ATT (the effect on small businesses who in reality got a loan) is probably higher than the overall ATE that includes businesses who did not apply for loans. That is, because small businesses can (partially) self-select into treatment depending on their benefit from the treatment, the benefit is probably higher among the actually-treated businesses.

Table 4.2: Potential outcomes example: ATE vs. ATT.

Y^u	Y^t	X	Probability
0	8	1	0.25
4	6	1	0.25
3	1	0	0.25
1	1	0	0.25

Discussion Question 4.4 (ATE vs. ATT)**.** Consider Table 4.2.

a) Explain why the different types of individuals (different rows in the table) choose the X value shown, based on their potential outcomes.
b) Compute the ATE.
c) Compute the ATT.
d) Compute $\mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$, the observed treated-untreated mean difference.
e) Explain why the ATT is identified but the ATE is not, both mathematically and intuitively.
f) Explain why the ATT is larger than the ATE in this example.

4.1.7 Estimation

Under an additional assumption about sampling (like iid) and assuming all relevant moments are well-defined and finite, consistent estimation follows:

$$\widehat{\mathrm{E}}(Y \mid X = 1) \xrightarrow{p} \mathrm{E}(Y \mid X = 1), \quad \widehat{\mathrm{E}}(Y \mid X = 0) \xrightarrow{p} \mathrm{E}(Y \mid X = 0), \tag{4.10}$$

so (by the continuous mapping theorem) the sample treated-untreated mean difference is consistent for the population mean difference. Asymptotic normality can also be derived.

For this class, rather than deriving asymptotic properties through the precise applications of the weak law of large numbers, continuous mapping theorem, and central limit theorem, we will focus on the interpretation of results. The estimator $\widehat{\mathrm{E}}(Y \mid X = 1) - \widehat{\mathrm{E}}(Y \mid X = 0)$ can always be computed, but its interpretation requires critical thought specific to each empirical setting. Most fundamentally, we can interpret it as simply an estimate of the population mean difference. This may still be valuable for description. If the identifying assumptions of Theorem 4.2 hold, then we can additionally interpret it as an estimate of the population ATT. If we further assume the ATT and ATE are equal (as implied by A4.2), then we can additionally interpret it as an estimate of the population ATE.

For dissertation-level research, this critical thought requires deep familiarity with your empirical setting. For example, if you are looking at a therapy program for prisoners, you would need to know how individuals get assigned to the program: do they freely choose? Are they assigned based on some observable characteristics $\boldsymbol{W}$? If a counselor chooses who participates, are counselors randomly assigned to prisoners, do different counselors have different probabilities of assigning individuals to therapy? Is it a group therapy where spillover effects may be important? Are policy-makers considering a program expansion, or ending it? We will practice thinking critically about assumptions, but for research you will also need to acquire the extensive knowledge as the input to your critical thinking.

4.2 Linear Structural Model

To identify parameters in a structural model, generally we need some sort of exogeneity condition that says X is unrelated to other causal determinants of Y. Below are some examples.

4.2.1 Fixed Coefficients

Consider the linear structural model

$$Y = \beta_0 + \beta_1 X + U, \tag{4.11}$$

where the unobserved scalar U captures the combined effect on Y of everything besides the common linear effect β_1 of X (and β_0 and β_1 are non-random parameter values). That is, U contains heterogeneity (if some individuals' effect of X is above β_1, or below), as well as nonlinearity in X (like if Y also depends on X^2), as well as omitted variables (like if some other Q has an effect on Y). If all these other effects are "unrelated" to X, then X is called **exogenous** and β_1 is identified; if not, then X is called **endogenous**. Mathematically, here "exogenous" means uncorrelated. In other contexts, "exogenous" may required mean independence or independence (which here are sufficient but not necessary).

Theorem 4.3 (linear structural identification)**.** *Given* (4.11)*, if* $\text{Cov}(X, U) = 0$*, then* β_1 *is identified and equals the slope of the linear projection* $\text{LP}(Y \mid 1, X)$*.*

Proof. The slope of $\text{LP}(Y \mid 1, X)$ is

$$\frac{\text{Cov}(Y, X)}{\text{Var}(X)} = \frac{\text{Cov}(\beta_0 + \beta_1 X + U, X)}{\text{Var}(X)} = \frac{\beta_1 \text{Cov}(X, X) + \text{Cov}(U, X)}{\text{Var}(X)} = \frac{\beta_1 \text{Var}(X) + 0}{\text{Var}(X)}$$
$$= \beta_1.$$

□

Discussion Question 4.5 (college and wage: endogeneity)**.** As in DQ 4.1, let $X = 1$ if an individual has a college degree and otherwise $X = 0$. Let Y be the individual's wage at age 45. Explain one real-world reason why $\text{Cov}(X, U) \neq 0$, including the sign (positive or negative).

4.2.2 Random Coefficients

Consider the linear structural **random coefficients** model

$$Y = U_0 + U_1 X, \tag{4.12}$$

where U_0 and U_1 are unobserved random variables. That is, each individual is represented by (Y, U_0, U_1, X); you can think of "random" as just meaning "individual-specific." This model more explicitly shows the heterogeneity in the intercept and slope. If X is binary, then the potential outcomes model in (4.6) can be rewritten as

$$Y = Y^u + (Y^t - Y^u)X, \tag{4.13}$$

which is (4.12) with $U_0 = Y^u$ and $U_1 = Y^t - Y^u$ (the individual's treatment effect).

Discussion Question 4.6 (college and wage: random coefficients)**.** Let $X = 1$ if an individual has a college degree and otherwise $X = 0$. Let Y be the individual's wage at age 45.

a) How do you interpret the economic meaning of U_0?
b) How do you interpret the economic meaning of U_1?
c) Why do you think individuals have different U_0?
d) Why do you think individuals have different U_1?

Theorem 4.4 (linear random coefficient identification)**.** *Given* (4.12)*, if* U_0 *and* U_1 *are mean-independent of* X*, then* $\mathrm{E}(U_0)$ *and* $\mathrm{E}(U_1)$ *are identified and equal to the linear CMF intercept and slope, respectively.*

Proof. Take the conditional mean of (4.12):

$$\mathrm{E}(Y \mid X) = \mathrm{E}(U_0 + U_1 X \mid X) = \mathrm{E}(U_0 \mid X) + \mathrm{E}(U_1 \mid X)X = \mathrm{E}(U_0) + \mathrm{E}(U_1)X,$$

where the first equality is from (4.12), the second equality is by the linearity property of expectation, and the third equality is from the assumed mean independence. Altogether, this shows that the conditional mean of Y given X is linear (affine) in X, with non-random intercept $\mathrm{E}(U_0)$ and non-random slope $\mathrm{E}(U_1)$. □

We can connect the random coefficients model to the fixed coefficients model in (4.11). Rewrite (4.12) as

$$\begin{aligned} Y &= U_0 + U_1 X \\ &= U_0 + U_1 X + \overbrace{\mathrm{E}(U_0) - \mathrm{E}(U_0)}^{=0} + \overbrace{[\mathrm{E}(U_1) - \mathrm{E}(U_1)]X}^{=0} \\ &= \overbrace{\mathrm{E}(U_0)}^{\beta_0} + \overbrace{\mathrm{E}(U_1)}^{\beta_1} X + \overbrace{U_0 - \mathrm{E}(U_0) + [U_1 - \mathrm{E}(U_1)]X}^{U}. \end{aligned}$$

4.3 Nonseparable Structural Model

Consider the **all-causes model**

$$Y = h(X, \boldsymbol{U}) \tag{4.14}$$

that shows how Y is fully determined by observable scalar X and unobserved vector $\boldsymbol{U}$, through the function $h(\cdot)$. That is, $\boldsymbol{U}$ contains all determinants of Y besides X, and thus is very large. This model is called **nonseparable** because the X and unobservables enter $h(\cdot)$ together, not additively separable like $Y = f(X) + g(\boldsymbol{U})$. The nonseparable model is more general because it still allows for additive separability but does not require it.

The **average structural function** (ASF) is a common object of interest, defined as

$$\mathrm{ASF}(x) \equiv \mathrm{E}[h(x, \boldsymbol{U})], \tag{4.15}$$

where the expectation is with respect to the unconditional distribution of $\boldsymbol{U}$. Like the CMF, the ASF is a non-random function (because it plugs in a non-random x and then averages out the $\boldsymbol{U}$).

The **average structural effect** (ASE) is the partial derivative of the ASF:

$$\text{ASE}(x) \equiv \frac{\partial}{\partial x}\text{ASF}(x) = \text{E}[\frac{\partial}{\partial x}h(x, \boldsymbol{U})]. \tag{4.16}$$

If x is discrete instead of continuous, then as usual the partial derivative can be replaced by a discrete difference like

$$\text{ASE} = \text{ASF}(1) - \text{ASF}(0) = \text{E}[h(1, \boldsymbol{U}) - h(0, \boldsymbol{U})].$$

As with the ATE, due to linearity of expectation, we can either think of the ASE as the difference between two points on the ASF (or derivative), or the mean of the individual-level causal effects. For example, in the binary X case, the causal effect of changing $X = 0$ to $X = 1$ is

$$C(\boldsymbol{U}) \equiv h(1, \boldsymbol{U}) - h(0, \boldsymbol{U}), \tag{4.17}$$

which depends on an individual's $\boldsymbol{U}$ (some individuals' Y may be more responsive to X changes than others'). The average such causal effect is

$$\text{E}[C(\boldsymbol{U})] = \text{E}[h(1, \boldsymbol{U}) - h(0, \boldsymbol{U})] = \text{E}[h(1, \boldsymbol{U})] - \text{E}[h(0, \boldsymbol{U})] = \text{ASF}(1) - \text{ASF}(0) = \text{ASE}\,. \tag{4.18}$$

To connect back with the ATE, first consider a binary X. An individual with unobserved $\boldsymbol{U}$ has potential outcomes

$$Y^u = h(0, \boldsymbol{U}), \quad Y^t = h(1, \boldsymbol{U}). \tag{4.19}$$

That is, the model says if we change the individual's $X = 0$ to $X = 1$, their Y will change from $h(0, \boldsymbol{U})$ to $h(1, \boldsymbol{U})$.

Theorem 4.5. *Given the structural all-causes model in (4.14) with binary $X \in \{0, 1\}$ and the potential outcomes in (4.19), if $X \perp\!\!\!\perp \boldsymbol{U}$, then the slope coefficient β_1 of the CMF $\text{E}(Y \mid X = x) = \beta_0 + \beta_1 x$ equals the ATE and equals the ASF.*

Proof. Given $X \perp\!\!\!\perp \boldsymbol{U}$, (4.14), and (4.15),

$$\text{E}(Y \mid X = 1) = \text{E}[h(X, \boldsymbol{U}) \mid X = 1] = \text{E}[h(1, \boldsymbol{U}) \mid X = 1] = \text{E}[h(1, \boldsymbol{U})] = \text{ASF}(1),$$

and similarly

$$\text{E}(Y \mid X = 0) = \text{E}[h(0, \boldsymbol{U})] = \text{ASF}(0).$$

The CMF slope equals $\text{E}(Y \mid X = 1) - \text{E}(Y \mid X = 0)$, which thus equals $\text{ASF}(1) - \text{ASF}(0)$, which is the average structural effect of X on Y. Taking expectations of (4.19), $\text{E}(Y^u) = \text{ASF}(0)$ and $\text{E}(Y^t) = \text{ASF}(1)$, so additionally the ASE equals $\text{E}(Y^t) - \text{E}(Y^u)$, which equals the ATE. □

Theorem 4.6. *Given the structural all-causes model in (4.14), if $X \perp\!\!\!\perp \boldsymbol{U}$, then the ASF is identified and equals the CMF (and thus the ASE is the derivative or difference of the CMF).*

Proof. For any x,

$$\mathrm{E}(Y \mid X = x) = \mathrm{E}[h(X, \boldsymbol{U}) \mid X = x] = \mathrm{E}[h(x, \boldsymbol{U}) \mid X = x] = \mathrm{E}[h(x, \boldsymbol{U})] = \mathrm{ASF}(x),$$

where the first equality is from plugging in (4.14), the second is because it conditions on $X = x$, the third is from $X \perp\!\!\!\perp \boldsymbol{U}$, and the fourth is the definition of ASF in (4.15). □

Note: although the nonseparable model may seem fancy, it's still essentially asking: what if we changed every single individual in the population from $X = x$ to $X = x + 1$? (Or $X = x + \mathrm{d}x$.) Often a policy only affects certain individuals' X values (and such "marginal" individuals may differ in important ways from the population as a whole).

Discussion Question 4.7 (wage and education: nonseparable)**.** Imagine an "audit study" where fake resumes are posted to apply to jobs online, and years of experience $X \in \{0, 1, 2, \ldots, 15\}$ is randomized while holding other applicant characteristics constant (or varying them independently of X). Let $Y = 1$ if the employer requests a follow-up interview, otherwise $Y = 0$.

a) Interpret the identification result in Theorem 4.6 in terms of this example.
b) How/does the identification result help us estimate the effect of interest from our audit study data? Explain.

Chapter 5

Identification by Conditional Independence

Unit learning objectives for this chapter

5.1. Explain how the intuition of causal identification from independence extends to a conditional setting. [TLO 2]

The intuition for identification by independence can be extended to conditional independence, for both treatment effects and structural models.

Optional resources for this chapter

- Potential outcomes and CATE (Masten video)
- conditional independence/unconfoundedness (Masten video)
- ATE/conditional independence example (Masten video)
- Overlap assumption (Masten video)

5.1 Conditional Average Treatment Effect

Consider the ATE for the subpopulation of individuals with $\boldsymbol{W} = \boldsymbol{w}$. For example, this could be the subpopulation of individuals who are 40 years old, have 16 years of education, and are married. This is called the **conditional average treatment effect** (CATE), here denoted

$$\text{CATE}(\boldsymbol{w}) \equiv \text{E}(Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w}) = \text{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}) - \text{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}), \tag{5.1}$$

where as in (4.1) the equality is due to the linearity of the expectation operator.

By the law of iterated expectations, the ATE can be written as

$$\text{ATE} = \text{E}[\text{CATE}(\boldsymbol{W})], \tag{5.2}$$

where the expectation is with respect to the population distribution of $\boldsymbol{W}$.

Imagine we run an experiment where we randomize treatment, but the treatment probability is higher for unemployed individuals, and the outcome Y is wage one year later. Let $X = 1$ if treated and $X = 0$ if untreated; let $W = 1$ if unemployed and $W = 0$ otherwise. We randomize with $\text{P}(X = 1 \mid W = 1) = 0.8$ and $\text{P}(X = 1 \mid W = 0) = 0.1$. Our earlier independence assumption is likely violated. For simplicity, imagine the treatment is useless, so $Y^t = Y^u = Y$ for everyone. Assuming the unemployed individuals tend to have lower wages, then a simple comparison of treated and untreated wages misleadingly suggests the treatment has a negative effect. For example, simplifying further, imagine all unemployed individuals have $Y = 15$ and all employed individuals have $Y = 25$, and there are 10 unemployed individuals and 20 employed individuals. Given the treatment probabilities, the treatment group consists of 8 of the 10 unemployed individuals with $Y = 15$, plus 2 of the 20 employed individuals with $Y = 25$; altogether, 10 individuals, with average wage 17. The remaining individuals are untreated, with average wage $[(2)(15) + (18)(25)]/(2 + 18) = 24$, much higher than the treated group! This is because the independence assumption A4.2 fails: X is not independent of (Y^u, Y^t).

However, we can still identify the true ATE by using conditional independence. Given W, X is independent of potential outcomes. For example, if we only look at unemployed individuals, then we have a randomized experiment where A4.2 holds; and similarly if we only look at employed individuals. Thus, the conditional ATEs are identified, and the ATE is identified by taking a weighted average of the CATEs (rather than pooling the data like above).

Assumption A5.1 (conditional independence)**.** Treatment X is conditionally (on $\boldsymbol{W}$) independent of the potential outcomes: $(Y^u, Y^t) \perp\!\!\!\perp X \mid \boldsymbol{W}$.

Assumption A5.2 (overlap)**.** There is strictly positive probability of both treatment and non-treatment for every subpopulation: $0 < \text{P}(X = 1 \mid \boldsymbol{W} = \boldsymbol{w}) < 1$ for all $\boldsymbol{w}$.

The key argument again relies on the (conditional) independence assumption. Assumption A5.1 implies

$$\begin{aligned} \text{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}) &= \text{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) \\ \text{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}) &= \text{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 0). \end{aligned} \tag{5.3}$$

This condition is called **conditional mean independence**: after conditioning on $\boldsymbol{W} = \boldsymbol{w}$, further conditioning on X does not affect the conditional mean of Y^t or Y^u. Given (4.6),

$$\begin{aligned} \text{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) &= \text{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) \\ \text{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 0) &= \text{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 0). \end{aligned} \tag{5.4}$$

This is an example of **nonparametric identification** because we have not restricted the CMF $\mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = x)$ to be linear or quadratic or have any other specific functional form (parameterization). Even if we end up estimating the CMF parametrically, it is still reassuring that our underlying identification argument does not rely on our knowing the true functional form.

Theorem 5.1 (CATE identification)**.** *Under A4.1, A5.1, and A5.2, each CATE is identified:*

$$\mathrm{CATE}(\boldsymbol{w}) = \mathrm{E}[Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w}] = \mathrm{E}(Y \mid X = 1, \boldsymbol{W} = \boldsymbol{w}) - \mathrm{E}(Y \mid X = 0, \boldsymbol{W} = \boldsymbol{w}).$$

Thus, the ATE is also identified. More generally, A5.1 can be replaced by conditional mean independence as in (5.3).

Proof. Using the above ingredients,

$$\begin{aligned}
\mathrm{CATE}(\boldsymbol{w}) &\equiv \overbrace{\mathrm{E}(Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w})}^{\text{use linearity, (5.1)}} \\
&= \overbrace{\mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}) - \mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w})}^{\text{use (5.3)}} \\
&= \overbrace{\mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}, X = 1)}^{\text{use (5.4)}} - \overbrace{\mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 0)}^{\text{use (5.4)}} \\
&= \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) - \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 0),
\end{aligned}$$

which is a feature of (only) the joint population distribution of observables $(Y, \boldsymbol{W}, X)$. By (5.2), the ATE is thus also identified. □

Beyond our scope...

How can we estimate the CATE using the identification result in Theorem 5.1? In principle, given iid data (or otherwise restricted dependence), we can consistently estimate any feature of the joint distribution of $(Y, X, \boldsymbol{W})$. For example, if X and W are both binary, then $\mathrm{E}(Y \mid W = 0, X = 0)$ can be estimated by $\widehat{\mathrm{E}}(Y \mid W = 0, X = 0)$, the sample mean of the Y_i for observations with $W_i = 0$ and $X_i = 0$. However, if W is continuous, then $\mathrm{P}(W_i = w) = 0$ for any $w \in \mathbb{R}$, so this estimation approach fails. We need to either assume the CMF is linear or quadratic (or some other specific functional form), or else use nonparameteric regression, as introduced in Part V of Kaplan (2021).

Discussion Question 5.1 (doctor certification)**.** Let $W \in \mathbb{R}$ be a continuous scalar random variable representing a doctor's quality, and let $X = \mathbb{1}\{W \geq 0\}$ be a dummy variable for whether the doctor receives a publicly visible certification as being high-quality. Let $Y^t \in [0, 10]$ be the doctor's patient satisfaction rating in the world where

the doctor is certified, and $Y^u \in [0, 10]$ the rating in the world where the doctor is not certified (but everything else is identical, including true quality W).

a) Give a specific, real-world reason why probably $(Y^u, Y^t) \perp\!\!\!\perp X$ fails; explain both intuitively and mathematically. (If it helps: try graphing the functions $\mu_t(w) = \mathrm{E}(Y^t \mid W = w)$ and $\mu_u(w) = \mathrm{E}(Y^u \mid W = w)$.)

b) Explain how it is possible to satisfy Assumption A5.1 even if patient satisfaction increases with true quality W.

c) Explain why Assumption A5.2 (overlap) fails, and intuitively why thus we cannot estimate $\mathrm{E}(Y \mid W = w, X = 1) - \mathrm{E}(Y \mid W = w, X = 0)$.

Beyond our scope...

The overlap problem in DQ 5.1 can be addressed by (roughly speaking) comparing doctors who are just barely above zero to doctors who are just barely below: their W is very similar, but the former have $X = 1$ while the latter have $X = 0$. This approach is called **regression discontinuity** and is covered in the ECON 9446/9447 sequence.

5.2 CATT

Analogous to the ATE and ATT, the **conditional average treatment effect on the treated** (CATT) is identified under somewhat weaker assumptions than the CATE. Specifically, we only need an identifying assumption about Y^u, not Y^t. The CATT is defined as

$$\begin{aligned} \mathrm{CATT}(\boldsymbol{w}) &\equiv \mathrm{E}(Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) \\ &= \mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) - \mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 1), \end{aligned} \tag{5.5}$$

and parallel to (5.2) the unconditional ATT is

$$\mathrm{ATT} = \mathrm{E}[\mathrm{CATT}(\boldsymbol{W}), X = 1] \tag{5.6}$$

by the law of iterated expectations, where the (outer) expectation is with respect to the population distribution of $\boldsymbol{W}$ conditional on $X = 1$.

Assumption A5.3 (untreated conditional mean independence)**.** The untreated potential outcome is conditionally mean-independent of the treatment: $\mathrm{E}(Y^u \mid \boldsymbol{W}, X) = \mathrm{E}(Y^u \mid \boldsymbol{W})$.

Theorem 5.2 (CATT identification)**.** *Under A4.1, A5.2, and A5.3, each CATT is identified:*

$$\begin{aligned} \mathrm{CATT}(\boldsymbol{w}) &= \mathrm{E}[Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 1] \\ &= \mathrm{E}(Y \mid X = 1, \boldsymbol{W} = \boldsymbol{w}) - \mathrm{E}(Y \mid X = 0, \boldsymbol{W} = \boldsymbol{w}). \end{aligned}$$

Thus, the ATT is also identified.

Proof. Starting from the definition of CATT, similar to the arguments in the proof of Theorem 5.1,

$$\begin{aligned}\mathrm{CATT}(\boldsymbol{w}) &\equiv \overbrace{\mathrm{E}(Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 1)}^{\text{use linearity, (5.5)}} \\ &= \overbrace{\mathrm{E}(Y^t \mid \boldsymbol{W} = \boldsymbol{w}, X = 1)}^{Y^t = Y \text{ because } X=1} - \overbrace{\mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 1)}^{\text{use A5.3}} \\ &= \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) - \overbrace{\mathrm{E}(Y^u \mid \boldsymbol{W} = \boldsymbol{w}, X = 0)}^{Y^u = Y \text{ because } X=0} \\ &= \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 1) - \mathrm{E}(Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 0),\end{aligned}$$

which is a feature of (only) the joint population distribution of observables $(Y, \boldsymbol{W}, X)$. By (5.6), the ATT is thus also identified. □

Discussion Question 5.2. Consider expanding Medicaid (health insurance for low-income people) in Missouri by increasing the income threshold (below which somebody is eligible) for individuals from \$18,754/yr to \$24,000/yr.

a) Which (sub)population do we care about for the purpose of assessing the possible benefit of this particular policy change? That is, do we care about the average benefit for everybody in Missouri (the ATE)? For current Medicaid recipients (the ATT)? Some other subpopulation (conditional on other $\boldsymbol{W}$)? Explain.
b) Do you think the benefits of Medicaid are higher or lower or similar for the subpopulation you described compared to the full Missouri population? Explain.
c) Do you think the benefits of Medicaid are higher or lower or similar for the subpopulation you described compared to current Missouri Medicaid recipients? Explain.

5.3 Linear Structural Model

Consider the linear structural model

$$Y = X\beta_1 + \boldsymbol{W}'\boldsymbol{\beta}_2 + U, \tag{5.7}$$

where $\boldsymbol{W}$ includes an intercept as well as control variables. The following assumptions and arguments are similar to Appendix 7.2 of Stock and Watson (2015).

Assumption A5.4 (conditional mean independence)**.** Given (5.7), $\mathrm{E}(U \mid \boldsymbol{W}, X) = \mathrm{E}(U \mid \boldsymbol{W})$: given $\boldsymbol{W}$, U is mean-independent of X.

Assumption A5.5 (linear error expectation)**.** The conditional mean of U is linear in $\boldsymbol{W}$: $\mathrm{E}(U \mid \boldsymbol{W}) = \boldsymbol{W}'\boldsymbol{\delta}$.

Theorem 5.3. *Given (5.7), under Assumptions A5.4 and A5.5, the structural slope coefficient β_1 is identified by the corresponding CMF slope coefficient.*

Proof. Taking the conditional mean of (5.7),

$$\begin{aligned} m(x, \boldsymbol{w}) &\equiv \mathrm{E}(Y \mid X = x, \boldsymbol{W} = \boldsymbol{w}) \\ &= \mathrm{E}(X\beta_1 + \boldsymbol{W}'\boldsymbol{\beta}_2 + U \mid X = x, \boldsymbol{W} = \boldsymbol{w}) \\ &= x\beta_1 + \boldsymbol{w}'\boldsymbol{\beta}_2 + \mathrm{E}(U \mid X = x, \boldsymbol{W} = \boldsymbol{w}) \\ &= x\beta_1 + \boldsymbol{w}'(\boldsymbol{\beta}_2 + \boldsymbol{\delta}). \end{aligned}$$

Thus, the (linear) CMF coefficient on x is β_1, identical to the structural coefficient on x. (However, the CMF coefficient vector for $\boldsymbol{w}$ does not match the structural $\boldsymbol{\beta}_2$, unless $\boldsymbol{\delta} = \mathbf{0}$.) □

Discussion Question 5.3 (student-teacher ratio)**.** The following is similar to a running example in Stock and Watson (2015). Let Y be the average math test score of an elementary school in Missouri. Let X be the school's student-teacher ratio; for example, if the school has 500 students and 25 teachers, then $X = 500/25 = 20$. Interest is in the causal effect of X on Y.

a) Explain how family income could be a source of omitted bias.
b) Let $\boldsymbol{W}$ include the percentage of students who qualify for free lunch (due to low family income). Explain mathematically and verbally what it would mean for Assumption A5.4 to hold.
c) Explain one potential reason A5.4 does not hold.

The linearity assumptions can be relaxed, as in Section 5.A, but the main point is that sometimes "conditional exogeneity" (like A5.4) lets us learn about a structural relationship between Y and X if we have sufficiently helpful control variables in $\boldsymbol{W}$, although the structural coefficients on $\boldsymbol{W}$ are not identified. That is, the estimated coefficients on control variables cannot be interpreted causally. This means we should not necessarily worry if such estimated coefficients have the opposite sign of our intuition. For example, with scalar W for simplicity, maybe we think in the real world $\beta_2 > 0$ (W has a positive causal effect on Y), but possibly $\delta < 0$ and moreover $\beta_2 + \delta < 0$, in which case the population CMF coefficient is negative even though the population structural coefficient is positive.

Appendix to Chapter 5

5.A Nonseparable Structural Model

The intuition here is the same as in Section 5.1: if X is "as good as randomized" conditional on $\boldsymbol{W}$, then we should be able to learn about the causal effect of X on Y. The mathematical details differ.

Extending (4.14), consider the nonseparable all-causes model

$$Y = h(X, \boldsymbol{W}, \boldsymbol{U}), \tag{5.8}$$

where $(Y, X, \boldsymbol{W})$ is observable but not $\boldsymbol{U}$.

First consider binary X. For an individual with $(\boldsymbol{w}, \boldsymbol{u})$, the causal effect (structural effect) of X on Y is

$$C(\boldsymbol{w}, \boldsymbol{u}) \equiv h(1, \boldsymbol{w}, \boldsymbol{u}) - h(0, \boldsymbol{w}, \boldsymbol{u}). \tag{5.9}$$

If we condition on $\boldsymbol{w}$ but average out $\boldsymbol{u}$,

$$\text{CASE}(\boldsymbol{w}) = \text{E}[C(\boldsymbol{w}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}], \tag{5.10}$$

where the expectation is taken with respect to the conditional distribution of $\boldsymbol{U}$ given $\boldsymbol{W} = \boldsymbol{w}$. To connect back to the CATE, the potential outcomes are

$$Y^u = h(0, \boldsymbol{W}, \boldsymbol{U}), \quad Y^t = h(1, \boldsymbol{W}, \boldsymbol{U}), \tag{5.11}$$

so the CATE is

$$\begin{aligned} \text{CATE}(\boldsymbol{w}) &= \text{E}(Y^t - Y^u \mid \boldsymbol{W} = \boldsymbol{w}) \\ &= \text{E}[h(1, \boldsymbol{W}, \boldsymbol{U}) - h(0, \boldsymbol{W}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}] \\ &= \text{E}[C(\boldsymbol{w}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}] \\ &= \text{CASE}(\boldsymbol{w}). \end{aligned}$$

Similar to Theorem 4.5, the CASE with binary X is identified following an argument parallel to CATE. As with CATE, the key is a conditional independence assumption, parallel to A5.1.

Assumption A5.6 (conditional independence)**.** Regressor of interest X is conditionally (on $\boldsymbol{W}$) independent of the unobserved determinants of Y: $\boldsymbol{U} \perp\!\!\!\perp X \mid \boldsymbol{W}$.

Theorem 5.4 (CASE identification, binary)**.** *Given* (5.8)*, under Assumptions A5.2 and A5.6, the CASE is identified by the difference in conditional means (the "slope" of the CMF):* $\text{CASE}(\boldsymbol{w}) = \text{E}(Y \mid X = 1, \boldsymbol{W} = \boldsymbol{w}) - \text{E}(Y \mid X = 0, \boldsymbol{W} = \boldsymbol{w})$.

Proof. The proof follows the same logic as that of Theorem 5.1:

$$\begin{aligned}
\text{CASE}(\boldsymbol{w}) &\equiv \text{E}[h(1, \boldsymbol{w}, \boldsymbol{U}) - h(0, \boldsymbol{w}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}] \\
&= \text{E}[h(1, \boldsymbol{w}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}] - \text{E}[h(0, \boldsymbol{w}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}] \\
&= \text{E}[h(1, \boldsymbol{W}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}] - \text{E}[h(0, \boldsymbol{W}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}] \\
&= \text{E}[h(1, \boldsymbol{W}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}, X = 1] - \text{E}[h(0, \boldsymbol{W}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}, X = 0] \\
&= \text{E}[h(X, \boldsymbol{W}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}, X = 1] - \text{E}[h(X, \boldsymbol{W}, \boldsymbol{U}) \mid \boldsymbol{W} = \boldsymbol{w}, X = 0] \\
&= \text{E}[Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 1] - \text{E}[Y \mid \boldsymbol{W} = \boldsymbol{w}, X = 0],
\end{aligned}$$

which is a feature of (only) the joint population distribution of observables $(Y, \boldsymbol{W}, X)$. □

The CASE is defined and identified more generally with non-binary X, including continuous X. The discrete X identification follows the same definition and proof as binary X, merely replacing 0 and 1 with general values a and b. For continuous X, similar to (4.16), let

$$\text{CASE}(x, \boldsymbol{w}) = \text{E}[\frac{\partial}{\partial x} h(x, \boldsymbol{w}, \boldsymbol{U}) \mid X = x, \boldsymbol{W} = \boldsymbol{w}]. \tag{5.12}$$

For notational simplicity, assume $\boldsymbol{U}$ is also continuous, even conditional on any subset of $(X, \boldsymbol{W})$, so the expectation can be written as an integral against the conditional PDF of $\boldsymbol{U}$. Writing $\mathcal{U}$ as the support of $\boldsymbol{U}$,

$$\text{E}[\frac{\partial}{\partial x} h(x, \boldsymbol{w}, \boldsymbol{U}) \mid X = x, \boldsymbol{W} = \boldsymbol{w}] = \int_{\mathcal{U}} \frac{\partial}{\partial x} h(x, \boldsymbol{w}, \boldsymbol{u}) f_{\boldsymbol{U}|X,\boldsymbol{W}}(\boldsymbol{u} \mid X = x, \boldsymbol{W} = \boldsymbol{w}) \, \mathrm{d}\boldsymbol{u}. \tag{5.13}$$

The partial derivative of the CMF is

$$\begin{aligned}
\frac{\partial}{\partial x} m(x, \boldsymbol{w}) &= \frac{\partial}{\partial x} \text{E}[Y \mid X = x, \boldsymbol{W} = \boldsymbol{w}] \\
&= \frac{\partial}{\partial x} \text{E}[h(X, \boldsymbol{W}, \boldsymbol{U}) \mid X = x, \boldsymbol{W} = \boldsymbol{w}] \\
&= \frac{\partial}{\partial x} \text{E}[h(x, \boldsymbol{w}, \boldsymbol{U}) \mid X = x, \boldsymbol{W} = \boldsymbol{w}] \\
&= \frac{\partial}{\partial x} \int_{\mathcal{U}} h(x, \boldsymbol{w}, \boldsymbol{u}) f_{\boldsymbol{U}|X,\boldsymbol{W}}(\boldsymbol{u} \mid X = x, \boldsymbol{W} = \boldsymbol{w}) \, \mathrm{d}\boldsymbol{u}.
\end{aligned}$$

Under a relatively weak technical condition, the derivative and integral can be interchanged; doing that and then applying the product rule,

$$\begin{aligned}\frac{\partial}{\partial x}m(x,\boldsymbol{w}) &= \int_{\mathcal{U}} \frac{\partial}{\partial x}[h(x,\boldsymbol{w},\boldsymbol{u})f_{\boldsymbol{U}|X,\boldsymbol{W}}(\boldsymbol{u}\mid X=x,\boldsymbol{W}=\boldsymbol{w})]\,\mathrm{d}\boldsymbol{u}\\ &= \int_{\mathcal{U}} \big\{[\frac{\partial}{\partial x}h(x,\boldsymbol{w},\boldsymbol{u})f_{\boldsymbol{U}|X,\boldsymbol{W}}(\boldsymbol{u}\mid X=x,\boldsymbol{W}=\boldsymbol{w})]\\ &\qquad + [h(x,\boldsymbol{w},\boldsymbol{u})\frac{\partial}{\partial x}f_{\boldsymbol{U}|X,\boldsymbol{W}}(\boldsymbol{u}\mid X=x,\boldsymbol{W}=\boldsymbol{w})]\big\}\,\mathrm{d}\boldsymbol{u}\\ &= \mathrm{CASE}(x,\boldsymbol{w}) + \int_{\mathcal{U}} h(x,\boldsymbol{w},\boldsymbol{u})\frac{\partial}{\partial x}f_{\boldsymbol{U}|X,\boldsymbol{W}}(\boldsymbol{u}\mid X=x,\boldsymbol{W}=\boldsymbol{w})\,\mathrm{d}\boldsymbol{u}.\end{aligned}$$

Under Assumption A5.6, after conditioning on $\boldsymbol{W}=\boldsymbol{w}$, the distribution (or equivalently here PDF) of $\boldsymbol{U}$ does not depend on $X=x$, so

$$f_{\boldsymbol{U}|X,\boldsymbol{W}}(\boldsymbol{u}\mid X=x,\boldsymbol{W}=\boldsymbol{w}) = f_{\boldsymbol{U}|\boldsymbol{W}}(\boldsymbol{u}\mid \boldsymbol{W}=\boldsymbol{w}),$$

with no dependence on $X=x$. Thus, taking a derivative with respect to x yields zero, which zeroes out the second term in the expression above, so the CMF partial derivative equals the CASE.

Theorem 5.5 (CASE identification)**.** *Given* (5.8)*, under Assumption A5.6 (and an overlap condition), the CASE is identified by the corresponding partial derivative of the CMF:* $\mathrm{CASE}(x,\boldsymbol{w}) = \frac{\partial}{\partial x}\,\mathrm{E}(Y\mid X=x,\boldsymbol{W}=\boldsymbol{w})$.

Proof. See above. □

Beyond our scope...

Like the CATE identification, this is another example of nonparametric identification: we do not assume that either the structural model or the CMF is linear, or quadratic, or any other specific functional form. To estimate the CASE, we would need to either specify a function form or use nonparametric regression. We could also try to reduce our statistical uncertainty by doing further averaging (in both the population object and estimator), like averaging the CASE over the distribution of $\boldsymbol{W}$ and/or X.

Chapter 6
OVB and Proxy Variables

Unit learning objectives for this chapter

6.1. Define terms and concepts related to omitted variable bias and proxy variables. [TLO 1]

6.2. Describe how proxy variables can help reduce omitted variable bias, both intuitively and mathematically. [TLO 3]

This chapter first describes and quantifies the problem known as omitted variable bias, for linear structural models. Then, it shows how "proxy variables" can help reduce this bias.

This problem only really applies to estimating causal effects. For description, for example, if we want to estimate the linear projection slope of $\text{LP}(Y \mid 1, X)$, then it doesn't matter what other variables there are; the linear projection depends only on Y and X. (Mathematically, you could ask about bias in the estimated coefficient on X in $\text{LP}(Y \mid 1, X, Q)$ if Q is omitted and instead $\text{LP}(Y \mid 1, X)$ is estimated, but that's not usually a situation faced in practice.) For prediction, we do not care about coefficients, only prediction accuracy; omitting a predictor may make our accuracy worse, but we wouldn't say it's "biased."

Optional resources for this chapter

- OVB/confounders (Masten video)
- Collider bias examples: https://doi.org/10.1093/ije/dyp334
- Collider bias review (very detailed): https://doi.org/10.1146/annurev-soc-071913-043455

6.1 Omitted Variable Bias

6.1.1 Allegory for Intuition

The following allegory is from Kaplan (2022a). Imagine a ghost (Q) that often accompanies a child (X), i.e., the ghost and child are often in the same place at the same time. The ghost always makes a huge mess (Y): spilling flour, knocking over chairs, drawing on walls, etc. The child's parents only observe the child and the mess; they do not observe the ghost. The parents note that when the child is in the kitchen, then there is often a mess in the kitchen, and when the child is in the bathroom, then there is often a mess in the bathroom, etc. Thus, they infer that the child (X) causes the mess (Y). However, we know that it only appears that way because

GHOST.1 the ghost (Q) often accompanies the child (X) and

GHOST.2 the ghost (Q) causes a mess (Y).

The child is the regressor. The ghost is the omitted variable. The parents are economists who over-estimate how much mess the child causes. This phenomenon is **omitted variable bias** (OVB).

6.1.2 Formal Characterization of OVB

Mathematically, consider the linear structural model

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + Q\gamma + V, \tag{6.1}$$

where $\boldsymbol{X}$ includes an intercept. Assume the structural error V is "well-behaved" in the sense of satisfying the linear projection error property:

$$\mathrm{E}(\boldsymbol{X}V) = \mathbf{0}, \quad \mathrm{E}(QV) = 0. \tag{6.2}$$

Thus, if we could observe Q, then we could consistently estimate the structural coefficients $\boldsymbol{\beta}$ and γ by OLS because they are also linear projection coefficients. However, if Q is not observed, then

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + U, \quad U \equiv Q\gamma + V. \tag{6.3}$$

If $Q\gamma$ and $\boldsymbol{X}$ are related, then U is not an LP error, so $\boldsymbol{\beta}$ is not the LPC and thus not the OLS estimand.

To precisely characterize the OVB, let

$$\mathrm{LP}(Q \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\delta}, \quad R \equiv Q - \boldsymbol{X}'\boldsymbol{\delta}. \tag{6.4}$$

Plug these into (6.1):

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{\delta} + R)\gamma + V = \boldsymbol{X}'(\boldsymbol{\beta} + \gamma\boldsymbol{\delta}) + (V + R\gamma). \tag{6.5}$$

Note $V + R\gamma$ satisfies the LP error property:

$$\mathrm{E}[\boldsymbol{X}(V + R\gamma)] = \mathrm{E}[\boldsymbol{X}V + \boldsymbol{X}R\gamma] = \overbrace{\mathrm{E}(\boldsymbol{X}V)}^{=0 \text{ by } (6.2)} + \overbrace{\mathrm{E}(\boldsymbol{X}R)}^{=0 \text{ by } (6.4)} \gamma = \mathbf{0}. \tag{6.6}$$

Thus, (6.5) is a linear projection of Y onto $\boldsymbol{X}$ in error form, so OLS is consistent for $\boldsymbol{\beta} + \gamma\boldsymbol{\delta}$ (the LPC), meaning the asymptotic bias is $\gamma\boldsymbol{\delta}$.

Theorem 6.1 (OVB)**.** *Given the structural model in* (6.1)*, with the structural error satisfying* (6.2)*, and given the definitions in* (6.4)*, the linear projection of Y onto $\boldsymbol{X}$ is* $\mathrm{LP}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'(\boldsymbol{\beta} + \gamma\boldsymbol{\delta})$.

Proof. See above. □

Theorem 6.1 shows why both Conditions GHOST.1 and GHOST.2 are required for OVB. Condition GHOST.1 is about $\boldsymbol{\delta}$, which is the vector of "partial correlations" of Q with $\boldsymbol{X}$. If $\boldsymbol{\delta} = \mathbf{0}$, then the OVB term becomes zero. Condition GHOST.2 says $\gamma \neq 0$, recalling γ is the coefficient on Q in the structural model.

Corollary 6.2 (OVB)**.** *Assume conditions such that OLS estimator consistently estimates the linear projection coefficients. Let $\hat{\boldsymbol{\beta}}$ be the OLS estimator from regressing Y onto $\boldsymbol{X}$. Given Theorem 6.1, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} + \gamma\boldsymbol{\delta}$, also written $\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \gamma\boldsymbol{\delta}$. The "OVB" (or "asymptotic bias") is thus*

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \gamma\boldsymbol{\delta}.$$

For the jth vector element $\hat{\beta}_j$, the OVB is $\gamma\delta_j$, which equals zero if either $\gamma = 0$ or $\delta_j = 0$ (or both). If $\gamma \neq 0$ and $\delta_j \neq 0$, then the sign (positive or negative) of OVB depends on the signs of γ and δ_j.

Note if all $\delta_j = 0$ except $\delta_k \neq 0$, then $\delta_k = \operatorname{Cov}(Q, X_k)/\operatorname{Var}(X_k)$, which is easier to think about than a general partial correlation (LP coefficient).

Example 6.1 (kindergarten effect)**.** Let Y be annual earnings of an individual at age 30, and let $X = 1$ if (as a child) the individual had more than 24 students in their kindergarten class, otherwise $X = 0$.

a. Let Q be the size of the individual's first-grade class. If class size affects students at all (in terms of long-term earnings), then probably larger class size (less attention from teacher) causes lower earnings, so $\gamma < 0$. Because most students stay at the same school for kindergarten and first grade, the class sizes are probably positively correlated, $\delta_1 > 0$. Thus, there is negative OVB because $\gamma\delta_1 < 0$, so $\operatorname{plim}_{n\to\infty} \hat{\beta}_1 < \beta_1$. Most likely the true kindergarten effect is negative (larger class size causes smaller earnings), though possibly very small, so negative OVB actually makes the magnitude appear larger than it really is.

b. Let Q be the number of cubbies (places to put clothes, backpacks, etc.) in the kindergarten classroom. Naturally, this is correlated with the number of students, so $\delta_1 > 0$. However, the number of cubbies probably does not affect future earnings; for example, if I sneak into my child's classroom and add a cubby, it will not cause higher future earnings. Thus, $\gamma = 0$, so $\gamma\delta_1 = 0$ and this is not a source of OVB.

c. Let $Q = 1$ if the kindergarten is in a high-income neighborhood, otherwise $Q = 0$. Because higher-income neighborhoods tend to have better-funded schools who can afford to hire more teachers, $\delta_1 < 0$ (we are more likely to see $Q = 1 - X$ than $Q = X$). Further, being in a higher-income neighborhood itself causes higher future earnings, so $\gamma > 0$. Thus, again we have negative OVB because $\gamma\delta_1 < 0$.

d. As a sanity check: repeat the previous example but defining $Q = 1$ for low-income and $Q = 0$ for high-income. In that case, $\delta_1 > 0$ (instead of < 0), and $\gamma < 0$ (instead of > 0), but the resulting OVB is still negative because again $\gamma\delta_1 < 0$. That is, the way we define the variable Q does not affect the OVB.

Discussion Question 6.1 (assessing OVB)**.** Among public elementary schools (students mostly 5–11 years old) in California, let Y be the average standardized math test score among a school's 5th-graders, and let X be the school's student-teacher ratio for 5th-graders (like average number of students per class). Consider a simple regression of Y on X. For *any two of the following* variables, assess each OVB condition separately, and then decide whether you think it's a source of OVB.

a) School's parking lot area per student. (Remember 5–11-year-olds don't have cars to park.)
b) Time of day of the test.
c) School's total spending per student (including books, facilities, etc.).
d) Percentage of English learners (non-native speakers) among a school's 5th-grade students.

Discussion Question 6.2 (wage OVB)**.** Let Y be log wage, X_1 experience, X_2 years of education, and Q unobserved "ability," and assume that the structural model $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + Q\gamma + V$ has structural error term V satisfy $\mathrm{E}[(1, X_1, X_2, Q)V] = \mathbf{0}'$. Also assume for simplicity $\mathrm{LP}(Q \mid 1, X_1, X_2) = \delta_0 + X_1\delta_1 + X_2\delta_2$ has $\delta_1 = 0$.

a) Do you think $\delta_2 < 0$, $\delta_2 > 0$, or $\delta_2 = 0$? Explain why.
b) Do you think β_2 is < 0, > 0, or $= 0$? Explain.
c) Do you think γ is < 0, > 0, or $= 0$? Explain.
d) Given the above, would $\mathrm{plim}_{n\to\infty} \hat{\beta}_2$ be $< \beta_2$, $> \beta_2$, or $= \beta_2$? Explain.

Discussion Question 6.3 (OVB: ES habits)**.** For my introductory econometrics class, let Y be a student's final semester score ($0 \leq Y \leq 100$), and $X = 1$ if the student starts the exercise sets well ahead of the deadline (otherwise $X = 0$).

a) What's one variable that might cause OVB? Explain why you think both OVB conditions are satisfied.
b) Which direction of asymptotic bias would your omitted variable cause? Explain both mathematically and intuitively.

6.1.3 Measurement Error

One special case of OVB is due to measurement error. The observed Y and/or X variable can be written in terms of the true value plus an error. The error is then an omitted variable, which may (or may not) cause OVB depending on its properties.

Measurement Error in Outcome Variable

See also Section 12.3.2 of Kaplan (2022a) and Section 4.4.1 of Wooldridge (2010).

Consider measurement error in Y. The true but unobserved (**latent**) value is Y^*, but we observe Y, which has **measurement error**

$$M \equiv Y - Y^*. \tag{6.7}$$

In the simplest case, imagine we want to learn the population $\mathrm{E}(Y^*)$. For estimation, if we observed Y^*, then the sample mean (generally) is consistent for the population mean. However, we instead observe Y and can consistently estimate $\mathrm{E}(Y)$. The identification question here is: does $\mathrm{E}(Y^*) = \mathrm{E}(Y)$? That is, can we equate the parameter we care about (mean of Y^*) with a feature of the population distribution of the observed Y?

The following formally states a simple identification result.

Proposition 6.3. *Give $Y = Y^* + M$, under the identifying assumption $\mathrm{E}(M) = 0$, the true population mean $\mathrm{E}(Y^*)$ is identified and equal to the observable mean $\mathrm{E}(Y)$.*

Proof. Using linearity of expectation,

$$\mathrm{E}(Y) = \mathrm{E}(Y^* + M) = \mathrm{E}(Y^*) + \mathrm{E}(M) = \mathrm{E}(Y^*) + 0,$$

where the final equality relies on the assumption $\mathrm{E}(M) = 0$. □

However, if $\mathrm{E}(M) \neq 0$, then the true mean is not identified. If $\mathrm{E}(M) > 0$, then $\mathrm{E}(Y) > \mathrm{E}(Y^*)$, so there is upward (positive) bias. If $\mathrm{E}(M) < 0$, then $\mathrm{E}(Y) < \mathrm{E}(Y^*)$, so there is downward (negative) bias.

Discussion Question 6.4 (exercise)**.** Imagine you ask people how many minutes they exercised last week, to try to learn how much exercise people do each week.

a) Define each variable's meaning. What's Y? What's Y^*? What's M?
b) Explain a reason we might see $\mathrm{E}(M) > 0$.
c) Explain a reason we might see $\mathrm{E}(M) < 0$.

Extending to regression, consider the LP of interest

$$Y^* = \boldsymbol{X}'\boldsymbol{\beta} + V, \quad \mathrm{E}(\boldsymbol{X}V) = \mathbf{0}. \tag{6.8}$$

Substituting $Y^* = Y - M$ using (6.7),

$$\begin{aligned} Y - M &= \boldsymbol{X}'\boldsymbol{\beta} + V, \\ Y &= \boldsymbol{X}'\boldsymbol{\beta} + M + V = \boldsymbol{X}'\boldsymbol{\beta} + U, \quad U \equiv M + V, \end{aligned} \tag{6.9}$$

so $\boldsymbol{\beta}$ remains the LPC if and only if $M + V$ satisfies the LP error property

$$\mathbf{0} = \mathrm{E}[\boldsymbol{X}(M+V)] = \mathrm{E}(\boldsymbol{X}M) + \overbrace{\mathrm{E}(\boldsymbol{X}V)}^{=\mathbf{0} \text{ by } (6.8)} = \mathrm{E}(\boldsymbol{X}M). \tag{6.10}$$

If we do not care about the intercept term (the coefficient on $X_1 = 1$), then this says we need $\mathrm{Corr}(X_j, M) = 0$ for all $j = 2, \ldots, \dim(\boldsymbol{X})$. For example, this is implied by $M \perp\!\!\!\perp \boldsymbol{X}$: the measurement error is independent of the regressors.

Conversely, if the measurement error is correlated with some regressors, then it is a source of omitted variable bias. In fact, (6.9) is a special case of (6.3) with $\gamma = 1$ and $Q = M$. Thus, the following is a corollary of Theorem 6.1.

Corollary 6.4 (measurement error in outcome)**.** *Given the setup of* (6.8) *and* (6.9)*, applying the results of Theorem 6.1, the linear projection of the observable Y onto $\boldsymbol{X}$ is* $\mathrm{LP}(Y \mid \boldsymbol{X}) = \boldsymbol{X}'(\boldsymbol{\beta} + \boldsymbol{\delta})$*, where $\boldsymbol{\delta}$ is the LPC in* $\mathrm{LP}(M \mid \boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\delta}$*. If* $\mathrm{Corr}(M, X_j) = 0$ *for all $j = 2, \ldots, \dim(\boldsymbol{X})$, and assuming $X_1 = 1$ is the intercept term like usual, then the slope coefficients of* $\mathrm{LP}(Y^* \mid \boldsymbol{X})$ *are identified and equal to the slope coefficients of* $\mathrm{LP}(Y \mid \boldsymbol{X})$*.*

Proof. Apply Theorem 6.1 with $\gamma = 1$ and $Q = M$. If $\mathrm{E}(M) \neq 0$, and assuming $X_1 = 1$ is an intercept term like usual, then we can rewrite

$$Y = [\boldsymbol{X}'\boldsymbol{\beta} + \mathrm{E}(M)] + [M - \mathrm{E}(M)],$$

where the intercept term is now $\beta_1 + \mathrm{E}(M)$, and $M - \mathrm{E}(M)$ satisfies $\mathrm{E}[X_j(M - \mathrm{E}(M))] = \mathrm{Cov}(X_j, M) = 0$ under the identifying assumption $\mathrm{Corr}(X_j, M) = 0$ stated in Corollary 6.4. Thus, $M - \mathrm{E}(M)$ satisfies the LP error property $\mathrm{E}[\boldsymbol{X}(M - \mathrm{E}(M))] = \mathbf{0}$, so the LPC is $(\beta_1 + \mathrm{E}(M), \beta_2, \beta_3, \ldots)'$, i.e., $\boldsymbol{\beta}$ but with the intercept adjusted by $\mathrm{E}(M)$. □

Discussion Question 6.5 (exercise and gym membership)**.** Let Y^* be an individual's true minutes of exercise per week, and Y is their self-reported value (i.e., how much exercise the say they did when asked on a survey). Let $X = 1$ if somebody is a gym member, and $X = 0$ otherwise.

a) Explain a reason we might see $\mathrm{Cov}(X, M) \neq 0$, and whether this would make the covariance $>$ or $<$.
b) Compare the LP slope of Y on $(1, X)$, which is $\mathrm{Cov}(X, Y)/\mathrm{Var}(X)$, with the LP slope of Y^* on $(1, X)$, which is $\beta_1 = \mathrm{Cov}(X, Y^*)/\mathrm{Var}(X)$. What's the direction of asymptotic bias?

Instead of additive measurement error, we could write it as a multiplicative error, $Y = MY^*$. Then, $\log(Y) = \log(M) + \log(Y^*)$. Thus, for a regression with a logged outcome and multiplicative measurement error, there's just a $\log(M)$ term floating around, so we check if $\mathrm{Cov}(X_j, \log(M)) = 0$ or not.

Discussion Question 6.6 (self-reported scrap rate)**.** (See also DQ 6.9.) Let Y^* be the true "scrap rate" of a manufacturing firm: how many products (out of 100) need to be "scrapped" (put in trash) because their quality is too low, so high scrap rate is bad. For example, $Y^* = 0.04$ means a 4% scrap rate. Consider a government program that provides grant money to manufacturing firms to lower their scrap rate. The government randomly assigns firms to a control group and treatment group, to run an experiment. On January 1, the treated firms receive grant money, which they are supposed to use to improve efficiency. All firms self-report their scrap rates on December 31; this is $Y = Y^* + M$.

a) Describe a reason why treated firms might want to systematically over-report ($M > 0$) or under-report ($M < 0$) their scrap rates.
b) In that case, and assuming untreated firms report accurately ($M = 0$), would we overestimate or underestimate the treatment effect of a grant? (The estimator is the slope coefficient in the regression of Y on $(1, X)$.) Why? (To get started: consider if the true effect/slope is zero; how does the measurement error make it appear as if there is a non-zero effect?)
c) If the government uses these incorrect estimates to decide whether or not to continue the program, what incorrect decision might they make? Why? (To get started: again imagine there is zero true effect.)

Measurement Error in Regressor

The following is mostly from Section 12.3.3 of Kaplan (2022a); see also Section 4.4.2 of Wooldridge (2010).

Now consider measurement error in a regressor. This is sometimes called **errors-in-variables**.

The following uses a simple regression (one non-constant regressor) to show how measurement error can cause asymptotic bias. The true LP with latent X^* is

$$Y = \beta_0 + \beta_1 X^* + R, \quad \mathrm{E}(R) = \mathrm{Cov}(X^*, R) = 0. \tag{6.11}$$

Because the observed X is $X = X^* + M$, substituting in $X^* = X - M$,

$$Y = \beta_0 + \beta_1 (X - M) + R = \beta_0 + \beta_1 X + (R - \beta_1 M). \tag{6.12}$$

The asymptotic bias is

$$\mathrm{AsyBias}(\hat{\beta}_1) = \frac{\mathrm{Cov}(X, R - \beta_1 M)}{\mathrm{Var}(X)},$$

so the asymptotic bias is zero if and only if $\mathrm{Cov}(X, R - \beta_1 M) = 0$, i.e., if the observed X is uncorrelated with the unobserved "error term" $R - \beta_1 M$. Using (6.11) and linearity,

$$\begin{aligned}
\mathrm{Cov}(X, R - \beta_1 M) &= \mathrm{Cov}(X, R) - \mathrm{Cov}(X, \beta_1 M) \\
&= \mathrm{Cov}(X^* + M, R) - \beta_1 \,\mathrm{Cov}(X, M) \\
&= \overbrace{\mathrm{Cov}(X^*, R)}^{=0} + \mathrm{Cov}(M, R) - \beta_1 \,\mathrm{Cov}(X, M).
\end{aligned}$$

If M is uncorrelated with the LP error $R = Y - \beta_0 - \beta_1 X^*$, and if $\beta_1 = 0$ (which means Y and the true X^* are not correlated), then this is zero. Otherwise, there is almost certainly asymptotic bias, in particular when $\text{Cov}(X, M) \neq 0$.

Unfortunately, $\text{Cov}(X, M) = 0$ is very unlikely. Consider what seems to be the best-case scenario: M is just random noise unrelated to the true value X^*, so $\text{Cov}(X^*, M) = 0$. This is sometimes called **classical measurement error**, or more specifically the **classical errors-in-variables** assumption. Unfortunately, using $\text{Cov}(X^*, M) = 0$,

$$\text{Cov}(X, M) = \text{Cov}(X^* + M, M) = \overbrace{\text{Cov}(X^*, M)}^{=0} + \overbrace{\text{Cov}(M, M)}^{=\text{Var}(M)} = \text{Var}(M). \tag{6.13}$$

Assuming $\text{P}(M = 0) < 1$, then $\text{Var}(M) > 0$, so $\text{Cov}(X, M) > 0$. Thus, even if $\text{Cov}(M, R) = 0$, the asymptotic bias is not zero because $-\beta_1 \text{Cov}(X, M) \neq 0$.

In this case with $\text{Cov}(X, M) > 0$ and $\text{Cov}(M, R) = 0$, the resulting bias is called **attenuation bias**. That is, the estimates are systematically pushed closer to zero by the measurement error. Putting together the above equations, the asymptotic bias is

$$-\beta_1 \frac{\text{Cov}(X, M)}{\text{Var}(X)} = -\beta_1 \frac{\text{Var}(M)}{\text{Var}(X)}, \tag{6.14}$$

which has the opposite sign of β_1 because variances are positive. That is, if $\beta_1 > 0$ then the bias is negative, whereas if $\beta_1 < 0$ then the bias is positive, so the bias always pushes toward zero. Further, the magnitude of the bias is never larger than β_1, so it can never "overshoot" zero, because $\text{Var}(M) < \text{Var}(X)$:

$$\text{Var}(X) = \text{Var}(X^* + M) = \text{Var}(X^*) + \text{Var}(M) + 2\overbrace{\text{Cov}(X^*, M)}^{=0 \text{ by CEV}} = \text{Var}(X^*) + \text{Var}(M).$$

Even if we cannot fix the attenuation bias, it is helpful to know the direction of the bias. For example, if we estimated $\hat{\beta}_1 = 7$, and we suspect attenuation bias, then our best guess is that β_1 might be even larger than 7. Similarly, if the corresponding 95% CI is $[4, 10]$, then we may feel even more confident about the lower endpoint, though we may not think 10 is the upper bound.

Discussion Question 6.7 (EIV example)**.** Let $X^* \in \{1, 2, 3\}$. Assume if $X^* = 1$ or $X^* = 3$, then $X = X^*$ and $M = 0$. But if $X^* = 2$, then $\text{P}(M = -1) = \text{P}(M = 1) = 0.5$, i.e., $\text{P}(X = 1 \mid X^* = 2) = \text{P}(X = 3 \mid X^* = 2) = 0.5$. So, $\text{E}(M \mid X^* = x) = 0$ for $x = 1, 2, 3$, which sounds nicely behaved.

a) Is $\text{Corr}(X, M) = 0$, > 0, or < 0? Why? (Hint: graph possible values of (X, M).)
b) Let $Y = \beta_0 + \beta_1 X^* + V$, where for simplicity $V = 0$ (always), $\beta_0 = 0$, and $\beta_1 = 2$. Graph all possible values of (X^*, Y), and then (on the same axes but with a different plot symbol shape or color) graph all possible values of (X, Y). Draw a best-fit/OLS line through each set of points.
c) What type of bias does this measurement error cause?

Unfortunately, outside the very special case of classical errors-in-variables with a linear model, the direction of bias may differ. It is not necessarily attenuation bias. In particular, if $\text{Cov}(M, R) \neq 0$ and $|\text{Cov}(M, R)| > |\beta_1 \text{Cov}(X, M)|$, then the sign of the bias is the sign of $\text{Cov}(M, R)$, i.e., positive bias if $\text{Cov}(M, R) > 0$ or negative bias if $\text{Cov}(M, R) < 0$. So, generally, any type of asymptotic bias is possible, depending how the measurement error is related to other variables.

One way to address measurement error is with instrumental variables, as in Chapter 8.

6.2 Proxy Variables

Continue from the model in (6.1) and (6.2), where if we could observe Q then the linear structural model coefficients can be estimated by OLS.

Consider a **proxy variable** Z that we can observe and use as a control variable, hoping that it captures enough about Q to reduce OVB. Usually Z is assumed **redundant** in the structural model, meaning it does not appear in (6.1). (With other mathematical setups, the definition differs, but the qualitative idea is the same.) This is essentially implied by (6.2): if Z were part of V, and Z is correlated with Q (see below), then V would not satisfy the LP error property, so the original model (6.1) even including Q would suffer from OVB.

Assumption A6.1 (proxy redundancy)**.** The proxy variable Z is not part of the structural model; mathematically, $\text{E}(ZV) = 0$.

Similar to the derivation of Theorem 6.1, the Q in (6.1) can be replaced by its linear projection in error form. Similar to (6.4) but now with Z,

$$\text{LP}(Q \mid \boldsymbol{X}, Z) = \boldsymbol{X}'\boldsymbol{\rho} + Z\theta_1, \quad R \equiv Q - \text{LP}(Q \mid \boldsymbol{X}, Z). \tag{6.15}$$

Plugging in,

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{\rho} + Z\theta_1 + R)\gamma + V = \boldsymbol{X}'(\boldsymbol{\beta} + \gamma\boldsymbol{\rho}) + \gamma\theta_1 Z + (V + R\gamma). \tag{6.16}$$

Note $V + R\gamma$ satisfies the LP error property:

$$\begin{aligned} \text{E}[(\boldsymbol{X}', Z)(V + R\gamma)] &= \text{E}[(\boldsymbol{X}', Z)V + (\boldsymbol{X}', Z)R\gamma] \\ &= \overbrace{\text{E}[(\boldsymbol{X}', Z)V]}^{=\mathbf{0} \text{ by (6.2) and A6.1}} + \overbrace{\text{E}[(\boldsymbol{X}', Z)R]}^{=\mathbf{0} \text{ by (6.15)}}\gamma = \mathbf{0}. \end{aligned}$$

Thus, (6.5) is a linear projection of Y onto $\boldsymbol{X}$ in error form, so OLS is consistent for $\boldsymbol{\beta} + \gamma\boldsymbol{\delta}$ (the LPC), meaning the asymptotic bias is $\gamma\boldsymbol{\delta}$.

The difference with Theorem 6.1 is having $\boldsymbol{\rho}$ instead of $\boldsymbol{\delta}$. Thus, the key to success of a proxy variable is how much of the variation in Q it "soaks up" in the linear projection, reducing the linear projection coefficients on $\boldsymbol{X}$.

Theorem 6.5 (proxy OVB)**.** *Given the structural model in* (6.1)*, with the structural error satisfying A6.1 and* (6.2)*, and given the definitions in* (6.15)*, the linear projection of Y onto $\boldsymbol{X}$ and Z is* $\text{LP}(Y \mid \boldsymbol{X}, Z) = \boldsymbol{X}'(\boldsymbol{\beta} + \gamma\boldsymbol{\delta}) + Z\theta_1$.

Proof. See above. □

Corollary 6.6 (perfect proxy)**.** *A **perfect proxy** for Q satisfies $\boldsymbol{\delta} = \mathbf{0}$ (or allowing $\delta_1 \neq 0$, which only affects the intercept), meaning* $\text{LP}(Q \mid \boldsymbol{X}, Z) = \rho_1 + Z\theta_1$ *with no dependence on $\boldsymbol{X}$. Then, the structural slope coefficients are all linear projection coefficients, for which the OLS estimator is consistent under relatively weak sampling and finite-moment conditions.*

Proof. From Theorem 6.5, the linear projection coefficients are $\boldsymbol{\beta} + \gamma\boldsymbol{\delta}$, so regardless of γ, if any $\delta_j = 0$ then the corresponding LPC is β_j, the structural coefficient. □

Discussion Question 6.8 (recidivism and therapy)**.** Consider the causal effect of a particular cognitive behavioral therapy (CBT) program on the future criminal activity of current prison inmates. Specifically, there is a group of individuals who were previously in prison, with $X = 1$ if they participated in CBT and $X = 0$ if not, and who were then tracked for five years after their release from prison, with Y the number of additional days spent in prison during that five-year window.

a) Prisoners are more likely to be assigned to CBT if they committed a more severe crime; prisoners are also more likely to commit future crimes (and more severe future crimes) if their initial crime is more severe. Explain which direction of OVB this generates.
b) We also observe Z, the length (in days) of the prisoner's initial prison sentence. Explain with words and equations the conditions under which Z would be a perfect proxy.

Discussion Question 6.9 (scrap rate and grants)**.** This example is similar to Example 4.4 of Wooldridge (2010); see also III.5 later. Let Y be the log "scrap rate" of a manufacturing firm: how many products (out of 100) need to be "scrapped" (put in trash) because their quality is too low, so high scrap rate is bad. Consider a government program that provides grant money to manufacturing firms to lower their scrap rate, but the grants are not randomized. Specifically, grants target firms who have more room for improvement of their scrap rate.

a) Explain which direction of OVB the above setup generates, both mathematically and intuitively. Say explicitly whether this makes the grants look more or less helpful than they really are.
b) Let Z be a firm's lagged (last year's) scrap rate. Explain with words and equations the conditions under which Z would be a perfect proxy.

However, if a variable is a really bad proxy, then it can actually worsen OVB, as well as increasing standard errors. Thus, you need to think carefully about whether a variable actually should proxy for a particular omitted variable, rather than simply including all

available variables in the data. For example, see the simple example at the very end of Chapter 4 (page 72) of Wooldridge (2010).

6.3 Collider Bias

This section is taken from Section 9.6 of Kaplan (2022a).

Although OVB shows the risk of omitting certain types of variables, other types of variables actually *should* be omitted, otherwise they cause a different type of (asymptotic) bias.

A **collider** or **common outcome** is a variable on which both X and Y have a causal effect. For example, imagine you want to learn the effect of a firm's ownership structure (say $X = 1$ for family-owned, $X = 0$ otherwise) on its research and development expenditure Y. Both X and Y affect the firm's performance Z, so Z is a collider.

Including a collider as a regressor causes **collider bias** when estimating a causal relationship. This is not as intuitive as OVB, but it can be just as problematic. Section 6.A provides a detailed example.

Appendix to Chapter 6

6.A Collider Bias: Example

The following example is a modification of https://doi.org/10.1093/ije/dyp334.

Imagine you're interested in the causal effect of eating falafel or salad on having the flu (which is zero effect), and you have a sample of 200 individuals. You randomly assigned 100 people to eat falafel for lunch, and 100 salad; a few hours later, you test each for flu (assume there is no testing error). Let $Y = 1$ if somebody has the flu (otherwise $Y = 0$), and $X = 1$ if somebody ate falafel for lunch ($X = 0$ if salad). Let $Z = 1$ if the individual has a fever (otherwise $Z = 0$). Sadly, the salad had some romaine contaminated with E. coli, so 40% of those who ate salad got a fever from the E. coli, unrelated to whether or not they had the flu. Among individuals with flu, 90% have a fever, but 10% don't.

Table 6.1: Counts in falafel/salad/flu example.

			Fever		No fever	
	Flu	No flu	Flu	No flu	Flu	No flu
Falafel	50	50	45	0	5	50
Salad	50	50	47	20	3	30

Table 6.1 shows the number of individuals in different categories. Overall, there is no relationship between lunch and flu, so the flu rate is the same in the falafel and salad groups. To make the numbers easier, the overall flu rate is 50% (100/200 overall, 50/100 in each group). Because nobody who ate falafel got E. coli, the only reason for fever is the flu, which has a 90% fever rate. Thus, among the 50 with flu who at falafel, $(50)(0.9) = 45$ have a fever and 5 do not. This entirely explains the Falafel row. In the salad row, given the statistical independence of flu (probability 0.5) and E. coli (probability 0.4), the

probability of having neither is

$$\begin{aligned}\mathrm{P}(\text{not flu and not E. coli}) &= \mathrm{P}(\text{not flu})\,\mathrm{P}(\text{not E. coli}) = [1-\overbrace{\mathrm{P}(\text{flu})}^{0.5}][1-\overbrace{\mathrm{P}(\text{E. coli})}^{0.4}] \\ &= (0.5)(0.6) = 0.3,\end{aligned}$$

hence $(100)(0.3) = 30$ salad-eaters who have neither flu nor E. coli, and thus no fever. This explains the No fever / No flu entry of 30 in the Salad row. Similarly,

$$\begin{aligned}\mathrm{P}(\text{flu, not E. coli}) &= (0.5)(0.6) = 0.3 \quad (30 \text{ people}),\\ \mathrm{P}(\text{flu, E. coli}) &= (0.5)(0.4) = 0.2 \quad (20 \text{ people}),\\ \mathrm{P}(\text{not flu, E. coli}) &= (0.5)(0.4) = 0.2 \quad (20 \text{ people}).\end{aligned}$$

The "not flu and E. coli" are the 20 individuals who have a fever (from the E. coli) but not flu. The 20 with both flu and E. coli all have a fever, due to E. coli. Among the 30 with flu but not E. coli, 90% have a fever, i.e., $(30)(0.9) = 27$ have a fever, so 3 do not. This 3 is the No fever / Flu entry in the Salad row. The 27 combine with the 20 who had both illnesses to make 47 who have both flu and a fever in the Salad row.

If we regress Y (flu) on X (food), then we correctly estimate zero effect, but if we also use Z (fever), then we incorrectly estimate a non-zero effect. If we only look at the "no fever" group, then there is (appropriately) zero difference: the flu rate for the falafel eaters is $5/55 = 1/11$, identical to the $3/33 = 1/11$ for the salad eaters. Mathematically, these "rates" are estimates of the conditional mean of the binary Y flu variable; e.g., $5/55 = \widehat{\mathrm{E}}(Y \mid \text{falafel, no fever})$, recalling $\mathrm{E}(Y) = \mathrm{P}(Y = 1)$ for binary Y. However, if we also look at the "fever" group, the flu rate is much higher in the falafel group. In fact, the falafel group's flu rate is $45/45 = 100\%$, whereas the salad group's flu rate is only $47/(47 + 20) = 70\%$, substantially lower. Mathematically,

$$\begin{aligned}&\overbrace{\widehat{\mathrm{E}}(Y \mid X = 1, Z = 0)}^{5/55} - \overbrace{\widehat{\mathrm{E}}(Y \mid X = 0, Z = 0)}^{3/33} = 0,\\ &\underbrace{\widehat{\mathrm{E}}(Y \mid X = 1, Z = 1)}_{45/45} - \underbrace{\widehat{\mathrm{E}}(Y \mid X = 0, Z = 1)}_{47/67} = 0.30.\end{aligned} \tag{6.17}$$

This suggests eating falafel causes flu, but this incorrect conclusion is entirely collider bias.

Exercises

Exercise I.1. Write a Stata do-file as follows. In general, each step corresponds to one line of code, except where otherwise noted. The data files are available at:
https://drive.google.com/file/d/0B-_LUSJVBv20SjBYd2pwYkYtcnc/view?resourcekey=0-DMCuTq__SV1c0PxaQ-wTOQ
https://drive.google.com/file/d/0B-_LUSJVBv20U2E2R2tBWnItaU0/view?resourcekey=0-dkguDqOtIoH5VWJgm-4T4g

a. Include the usual top-of-file items:

   i. Make the first line a "comment" (starting with an asterisk) with your name, the class name, and today's date.
   ii. Clear all variables in memory with `clear all`
   iii. Close any log file that may currently be open, without displaying an error if none is open, with `capture log close`
   iv. Issue the command `set more off` so that Stata doesn't wait for your input if there's more than one screen of output.
   v. Change the current directory to the one where you have downloaded the raw data and have saved this do-file, using the `cd` command.
   vi. Start writing a plaintext log to a file with suffix ".log", replacing the existing file if applicable (with the `replace` option).

b. Read into memory the data in file `Kaplan_Stata1_fake_data_grades.csv` using the command `insheet`:
   `insheet using "Kaplan_Stata1_fake_data_grades.csv" , clear`

c. Using a `keep if` statement, keep only rows for undergraduates, who are identified by their student type being `UG`.

d. Create a new variable named `cl_grade_num` that translates the string variable `cl_grade` into the corresponding numeric values. For example, if row 7 contains `cl_grade` equal to `D`, then the new variable should equal one; A=4, B=3, C=2, D=1, F=0. Note: this step requires multiple lines of code; the first is a `generate` command, and the rest are `replace` commands.

e. Create (with command `generate`) another new variable: name it `cl_grade_pts`, and store the product of `cl_units` and `cl_grade_num`.

f. Collapse (with command `collapse`) the data to one row per student, calculating the sum of `cl_units` and `cl_grade_pts`.

g. Create a new variable named `s_GPA` as the quotient of the summed `cl_grade_pts` and the summed `cl_units`; this is the grade point average (GPA), the average of the grades weighted by the units per class.

h. Drop (with `drop`) the variables containing the summed `cl_grade_pts` and summed `cl_units`.

i. Sort by `s_id`.

j. Check whether `s_id` is a unique identifier: `isid s_id`.

k. Save the dataset to a new .dta file, replacing the existing version (if applicable): `save "Kaplan_Stata1_fake_data_GPA.dta" , replace`

l. Load the data in `Kaplan_Stata1_fake_data_parents.csv` using `insheet`.

m. Rename (command `rename`) the variable `student_id` to `s_id` to match the other file's convention.

n. Convert the variable `s_id` from string to numeric, ignoring the leading `A` in each: `destring s_id, replace ignore("A")`

o. Reshape the data to have only one row per student, with variable `p_edu1` containing parent 1's education and `p_edu2` for parent 2: `reshape wide p_edu,i(s_id) j(parent)`

p. Make a new variable named `p_edu_max` that is the maximum of all variables with prefix `p_edu`, using the `egen` command with `rowmax` (ignoring missing values): `egen p_edu_max = rowmax(p_edu*)`

q. Sort by `s_id`

r. Check that `s_id` is a unique identifier with `isid`

s. Merge the data currently in memory 1:1 by `s_id` with the temporary file with GPA that you saved earlier.

t. Drop observations containing data only from the parent dataset and not from the GPA dataset, i.e., when the generated variable `_merge` equals one.

u. Order the columns in the dataset so that `s_id` is first, then `s_GPA`, then other variables: `order s_id s_GPA`

v. Print the dataset to the console/log using the `list` command (with no arguments or options). (Note: in reality, you would rarely print an entire dataset since they are usually much bigger than this artifical example.)

w. Save your dataset to a new .dta file.

x. Close your log file.

After thoroughly debugging, run your file all the way through all at once, and submit your .log file and .do file electronically through Canvas.

Exercise I.2. Write a Stata do-file as follows. The data are from a New York Times article on December 28, 1994.

a. Do the usual top-of-file items from Exercise I.1(a).

b. Run `ssc install bcuse` to ensure command `bcuse` is installed, and then load the dataset with `bcuse wine, clear`

c. View basic dataset info with Stata command `describe`

d. View the first few rows of the dataset with Stata command `list if _n<=5`

e. Rename the `alcohol` column, which measures liters of alcohol from wine (consumed per capita per year): `rename alcohol wine`

f. Add a column named `id` whose value is just 1, 2, 3, 4, 5, etc.: `generate id = _n`

g. Display the countries with fewer than 100 heart disease deaths per 100,000 people: `list country if heart<100`

h. Display the rows for the countries with the 5 lowest death rates, sorted by death rate: `sort deaths` followed by (next line) `list if _n<=5`

i. Add a column with the sum of heart and liver disease deaths per 100,000: `generate heart_plus_liver = heart + liver`

j. Generate a variable with the squared death rate: `gen deaths_sq = deaths^2`

k. Display the sorted death rates: `sort deaths` followed by `list deaths`

l. Add a column with the proportion of heart deaths to total deaths with command `generate heart_prop = heart / deaths`

m. Create a histogram of liver deaths: `histogram liver`

n. Create a scatterplot of liver death rates (vertical axis) against wine consumption (horizontal axis): `scatter liver wine`

Exercise I.3. Consider the effect of being assigned to a job training program, where assignment was randomized. The specific program was the National Supported Work Demonstration in the 1970s in the U.S. Data are originally from LaLonde (1986), via Wooldridge (2020). You will look at effects on earnings. The `train` variable indicates (randomized) assignment to job training if it equals 1, and it equals 0 otherwise.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse jtrain2 , clear`

c. Run `describe re78 train` and read the variable labels to understand the meaning and units of measure.

d. Run `ttest re78 , by(train) unequal` and explain in words briefly (1 sentence) what that code does.

e. Run `reg re78 train , vce(robust)` and explain in words briefly (1 sentence) what that code does.

f. Rounding to three significant figures (and including units of measure), report the estimated average effect of being assigned to training, and discuss the estimate's economic significance (magnitude).

g. Rounding to three significant figures (and including units of measure), report the corresponding 95% confidence interval, and discuss what this tells us about uncertainty (be precise).

h. Describe the "potential outcomes" in this example, and explain why the average treatment effect of assignment to job training seems to be identified.

i. If this job training program were scaled up and offered to every individual in the country, would you guess the average effect would be higher or lower (due to general equilibrium effects)? Explain in 1–2 sentences.

Exercise I.4. The data are originally from Card (1995), with individual-level observations of (log) wages, years of education, and other variables. Note the dataset lacks variable labels, but they can be found online.[1]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse card , clear`

c. Create a dummy to compare high-school (12 years education) and college (16 years education):

```
gen d_coll = .
replace d_coll=0 if educ==12
replace d_coll=1 if educ==16
```

d. Regress log wage on years of education `reg lwage educ , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction.

e. Run `reg lwage d_coll , vce(robust)` and re-phrase your above concern (about OVB) in terms of why the average treatment effect is not identified (make sure to define the potential outcomes first).

f. Run `reg lwage educ IQ , vce(robust)`

   i. Explain the conditions under which IQ would be a perfect proxy for unobserved "ability."

[1] http://fmwww.bc.edu/ec-p/data/wooldridge/card.des

ii. Briefly describe one type of "ability" that IQ does not capture (so is not a perfect proxy). Given this, do you think it's better or worse (or neither) to use IQ as a proxy for ability?

iii. Does the estimated slope change in the direction that suggests reduced OVB? Explain briefly.

iv. Discuss the economic significance of the estimated slope on `educ`.

v. Explain what the confidence interval tells us about our uncertainty; be precise and explicit about whichever population value(s) you refer to, and about sources of uncertainty, etc.

g. Run `reg lwage educ IQ exper expersq black smsa south , vce(robust)` and then briefly compare with previous results, focusing on the returns to education.

Exercise I.5. Go through the analysis in I.4 but with the `nls80` dataset, noting that now `iq` is lowercase.

Exercise I.6. Consider the causal effect of being an athlete on a college student's grades (GPA). Note the dataset lacks variable labels, but they can be found online.[2]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse gpa2 , clear`

c. Regress GPA on the athlete dummy: `reg colgpa athlete , vce(robust)`

i. Interpret the estimated coefficient on `athlete` in terms of a conditional mean model.

ii. In terms of structural model $Y = \beta_0 + \beta_1 X + U$ (Y is GPA, X is the athlete dummy, U is the combined effect of unobserved determinants of Y), explain one reason why β_1 is not identified, and in which direction there is omitted variable bias. (Feel free to "cheat" and do the next parts first to get an idea!)

iii. Repeat your argument about identification failure, but in terms of a potential outcomes model and treatment effect.

d. Run `reg colgpa athlete female , vce(robust)` and explain why this seems to help (slightly) the omitted variable bias; try `tab athlete female` too.

e. Run `reg colgpa athlete female sat , vce(robust)` and explain what `sat` helps proxy for and why this helps reduce omitted variable bias.

f. Run `reg colgpa athlete female sat verbmath hsperc hsize hsizesq black white , vce(robust)`

i. Discuss the economic significance of the estimated coefficient on `athlete` and briefly compare with the original estimate from the simple regression in part (c).

[2] http://fmwww.bc.edu/ec-p/data/wooldridge/gpa2.des

ii. Explain what the confidence interval tells us about our uncertainty; be precise and explicit about whichever population value(s) you refer to, and about sources of uncertainty, etc.

Exercise I.7. Consider the causal effect of using a 401(k) retirement plan on net total financial assets. Variable descriptions are included in the dataset's variable labels.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse 401ksubs , clear`

c. Regress (net total financial) assets on the 401(k) participation dummy: `reg nettfa p401k , vce(robust)`

   i. Interpret the estimated coefficient on `p401k` in terms of a conditional mean model.

   ii. In terms of structural model $Y = \beta_0 + \beta_1 X + U$ (Y is assets, X is the 401(k) dummy, U is the combined effect of unobserved determinants of Y), explain one reason why β_1 is not identified, and in which direction there is omitted variable bias. (Feel free to "cheat" and do the next parts first to get an idea!)

   iii. Repeat your argument about identification failure, but in terms of a potential outcomes model and treatment effect.

d. Run `reg nettfa p401k inc , vce(robust)` and explain why this seems to help the omitted variable bias; try `reg p401k inc` too.

e. Run `reg nettfa p401k inc marr male age fsize , vce(robust)`

   i. From the potential outcomes perspective: what is the name and interpretation of the population object we hope to estimate by the coefficient on `p401k`?

   ii. Discuss the economic significance of the estimated coefficient on `p401k` and briefly compare with the original estimate from the simple regression in part (c).

   iii. Explain what the confidence interval tells us about our uncertainty; be precise and explicit about whichever population value(s) you refer to, and about sources of uncertainty, etc.

Exercise I.8. Consider the relationship between an infant's birthweight (which when too low is associated with other negative health outcomes) and the mother's cigarette smoking. Note the dataset lacks variable labels, but they can be found online.[3]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse bwght , clear`

c. Create a dummy to compare no smoking to any smoking: `gen d_smk = (cigs>0)`

[3] http://fmwww.bc.edu/ec-p/data/wooldridge/bwght.des

d. Regress log birthweight on the amount of smoking `reg lbwght cigs , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction. (Feel free to "cheat" and look below to get ideas.)

e. Run `reg lbwght d_smk , vce(robust)` and re-phrase your above concern (about OVB) in terms of why the average treatment effect is not identified (make sure to define the potential outcomes first).

f. Run `reg lbwght cigs motheduc , vce(robust)`

   i. Explain mathematically how OVB can be reduced by using `motheduc` as a proxy for unobserved mother's knowledge about prenatal health, even if it is not a perfect proxy.

   ii. Does the estimated slope change (when adding `motheduc` as a control variable) in the direction that suggests reduced OVB? Explain briefly.

   iii. Discuss the economic significance of the estimated slope on `cigs`.

   iv. Explain what the confidence interval tells us about our uncertainty; be precise and explicit about whichever population value(s) you refer to, and about sources of uncertainty, etc.

g. Provide a reason/argument why even conditional on `motheduc`, `d_smk` is not (mean) independent of the potential outcomes.

Part II

Instrumental Variables

Introduction

Part II concerns the instrumental variables approach to learn about causal effects. Both structural and treatment effect models are developed. The topics are similar to Chapter 5 and Section 21.4.3 of Wooldridge (2010).

Beyond being a potential solution to omitted variable bias, instrumental variables can also address endogeneity due to **simultaneity**, meaning both X and Y are determined at the same time through some "economic" process. The classic example is supply and demand, which was the original motivation for instrumental variable regression, developed in the early 1900s by Philip and Sewall Wright (father and son) to estimate supply and demand curves for products like butter. The observe market prices and quantities are equilibrium values, at the intersection of the supply and demand curves, so if both supply and demand curves move around, we just see a cloud of different equilibrium points. (Try drawing a lot of different supply and demand curves on the same graph, and then make a dot at each crossing point.) Thus, if we simply regress quantity on price, we cannot estimate the demand curve (or supply curve). However, if we could find a source of variation that moves the supply curve a lot (but not the demand curve), then it could help "trace out" the demand curve. (Try graphing a single demand curve and lots of supply curves, and again draw a dot at each intersection; now connecting the dots (the observed equilibria) recovers the demand curve.)

Discussion Question 6.10 (crime and police and crime)**.** Let Y be a city's crime rate (per capita) in a given year, and X its number of police officers per capita.

a) Do you think cities consider Y when choosing X? Do you think larger Y would cause a city to choose larger or smaller X?

b) In which direction would this bias our estimator of the causal effect of X on Y if we simply regress Y on X and look at the estimated slope coefficient? (Drawing a scatterplot of (X, Y) may help.)

Chapter 7
Local Average Treatment Effect

Unit learning objectives for this chapter

7.1. Describe the identification and estimation of the local average treatment effect, both mathematically and intuitively. [TLOs 2 and 3]

7.2. Interpret IV results and judge validity of an instrument in real-world examples. [TLO 4]

This section considers a binary treatment X and binary instrument Z. This setting is simple, yet rich enough to develop concepts and intuition.

7.1 Wald Estimator and Estimand

The intuition for an **instrumental variable** (IV) is one that generates "as good as random" variation in regressor of interest X, without affecting Y through other economic channels. Thus, we can see how Y varies with Z, and see how X varies with Z, and attribute the changes in Y to the causal effect of the changes in X. As in Chapters 4 and 5, we will focus on mean effects, which are also effects on the mean (due to the linearity of expectation).

Beyond our scope...

Analogous to the quantile treatment effect variation on the average treatment effect idea, there is a **local quantile treatment effect** (LQTE) variation of LATE, to help us learn about treatment effects across the full outcome distribution (not just the mean). For example, see Section 7.2.2 ("Local Quantile Treatment Effect") of Kaplan (2021) and references therein.

Putting the intuition into a formula, recalling that both X and Z are binary, the so-called **Wald estimator** is

$$\hat{\theta}_{\text{Wald}} = \frac{\widehat{\mathrm{E}}(Y \mid Z = 1) - \widehat{\mathrm{E}}(Y \mid Z = 0)}{\widehat{\mathrm{E}}(X \mid Z = 1) - \widehat{\mathrm{E}}(X \mid Z = 0)}. \tag{7.1}$$

Because X is binary, the denominator is equivalent to $\widehat{\mathrm{P}}(X = 1 \mid Z = 1) - \widehat{\mathrm{P}}(X = 1 \mid Z = 0)$. The Wald estimator is equivalent to the IV regression estimator in this binary setting.

The fundamental population estimand of (7.1) is

$$\frac{\mathrm{E}(Y \mid Z = 1) - \mathrm{E}(Y \mid Z = 0)}{\mathrm{E}(X \mid Z = 1) - \mathrm{E}(X \mid Z = 0)}. \tag{7.2}$$

Consistency follows by applying the appropriate weak law of large numbers and the continuous mapping theorem. Of course, the denominator cannot be zero; see Section 9.1. In fact, there are problems even if the denominator is merely "close" to zero; see Section 9.2.

7.2 Types of Individuals

To connect (7.2) with treatment effects, we need both potential outcomes and "potential treatments." To fix ideas, imagine $Z = 1$ means an individual is assigned to be treated, otherwise $Z = 0$. However, actual treatment can differ from the assignment: assigned individuals may refuse the treatment, or unassigned individuals may still get treated. Parallel to potential outcomes, the potential treatments are the values of X (actual treatment) in the parallel universes where the individual is unassigned ($Z = 0$) or assigned ($Z = 1$), respectively. Notationally, let

$$\begin{aligned} X^u &\equiv \text{treatment status when "unassigned" } (Z = 0), \\ X^a &\equiv \text{treatment status when "assigned" } (Z = 1). \end{aligned} \tag{7.3}$$

The observed actual treatment is thus

$$X = (1 - Z)X^u + ZX^a. \tag{7.4}$$

This implicitly defines four types of individuals based on the pair (X^u, X^a).

A Always-takers: $(X^u, X^a) = (1, 1)$, always treated regardless of Z.

N Never-takers: $(X^u, X^a) = (0, 0)$, never treated regardless of Z.

D Defiers: $(X^u, X^a) = (1, 0)$, always "defy" the assignment Z and do the opposite $(X = 1 - Z)$.

C Compliers: $(X^u, X^a) = (0, 1)$, always "comply" with the assignment and do what it says, getting treated if $Z = 1$ but not if $Z = 0$.

Table 7.1: Potential treatments and outcomes example.

Type	Probability	X^u	X^a	Y^u	Y^t
N	1/3	0	0	10	0
A	1/3	1	1	0	10
D	0	1	0	6	0
C	1/3	0	1	1	7

Table 7.1 shows an example of potential treatments for the four types, along with mean potential outcomes within each type. We could replace Y^u with $\mathrm{E}(Y^u \mid \text{type})$, and replace Y^t with $\mathrm{E}(Y^t \mid \text{type})$, but the intuition is the same. Note that defiers are assumed not to exist in this population (zero probability); this turns out to be a critical identifying assumption.

Discussion Question 7.1. Using Table 7.1, and assuming $\mathrm{P}(Z = 1 \mid \text{type}) = 0.5$ for each type, compute and interpret the following.

a) $\mathrm{E}(Y \mid Z = 0)$
b) $\mathrm{E}(Y \mid Z = 1)$
c) $\mathrm{P}(X = 1 \mid Z = 0)$
d) $\mathrm{P}(X = 1 \mid Z = 1)$

7.3 LATE Identification

The **local average treatment effect** (LATE) is

$$\text{LATE} \equiv \mathrm{E}(Y^t - Y^u \mid X^a - X^u = 1) = \mathrm{E}(Y^t \mid X^a - X^u = 1) - \mathrm{E}(Y^u \mid X^a - X^u = 1), \quad (7.5)$$

where $X^a - X^u = 1$ refers to (only) the compliers defined in Section 7.2. That is, LATE is the ATE for the subpopulation of compliers.

There is a long-running debate about the merits of the LATE. It has clear limitations: like ATE/ATT, it may not refer to the "marginal" population that would be affected by a particular policy change, and moreover we do not even know who a "complier" is in the real-world, and "complier" depends on the instrument (so even if two instruments both satisfy the identifying assumptions, the corresponding IV estimators have different LATE estimands). However, the identifying assumptions are weaker than those of certain other causal parameters, and LATE provides clarity and transparency about what the IV estimator is estimating. As with most econometric debates, it seems LATE has strengths and weaknesses that complement other approaches to causal identification.

To identify the LATE, in addition to SUTVA and overlap (sort of), we now replace the treatment independence assumption with instrument independence, as well as assuming no defiers and assuming the treatment is related to the instrument so that the denominator is non-zero.

Assumption A7.1 ("overlap")**.** The population contains some compliers.

Assumption A7.2 (instrument independence)**.** The instrument is independent of potential outcomes and potential treatments: $Z \perp\!\!\!\perp (Y^u, Y^t, X^u, X^a)$.

Assumption A7.3 (monotonicity / no defiers)**.** There are no defiers, so the potential treatments are monotonic in the instrument: $X^a \geq X^u$. (If $X^a \geq X^u$, then just redefine the instrument as $1 - Z$.)

Assumption A7.4 (relevance)**.** The instrument is **relevant**: $\mathrm{E}(X \mid Z = 1) - \mathrm{E}(X \mid Z = 0) \neq 0$.

Discussion Question 7.2. Using your calculations from DQ 7.1, compute and interpret

$$\frac{\mathrm{E}(Y \mid Z = 1) - \mathrm{E}(Y \mid Z = 0)}{\mathrm{E}(X \mid Z = 1) - \mathrm{E}(X \mid Z = 0)}.$$

Theorem 7.1 (LATE identification)**.** *Under Assumptions A4.1 and A7.1–A7.4, the LATE is identified and equal to* (7.2)*.*

Proof. See Section 7.A if you are interested. □

Discussion Question 7.3 (Vietnam draft)**.** Consider a version of the famous Vietnam War draft lottery, which Angrist (1990) used to estimate the causal effect of having served in the military on earnings later in life as a civilian (outside the military). During the war, every American male of a certain age is assigned a random number (based on date of birth), and if the number is below a certain threshold, military service is required; if not, military service is optional. In this case, $Z = 1$ if service is required, otherwise $Z = 0$; and $X = 1$ if the individual actually serves in the military, otherwise $X = 0$. Let Y be earnings 15 years after the potential military service.

a) Describe a "defier" (in the IV sense) in this example. Is somebody a "defier" if their number was below the threshold and yet still refused military service? What portion of the population would you guess are defiers, and why?
b) Describe a "complier" in this example, as well as never-taker and always-taker.
c) Describe the LATE in this example. How do you think it compares to the ATE for never-takers? Why?
d) Hypothetically, if Vietnam military service had a negative effect on earnings (as estimated in the paper), then being drafted (required service) should cause some individuals to actually serve in the military and then have lower earnings; but then how can the assumption of instrument independence (A7.2) hold? Explain.

Appendix to Chapter 7

7.A Proof of LATE Identification

Proof of Theorem 7.1. Starting from the statistical object in (7.2), the identifying assumptions are used to work toward a causal interpretation.

First, we can write the observed Y in terms of potential outcomes and potential treatments. Rearranging (4.6),

$$Y = Y^u + X(Y^t - Y^u), \tag{7.6}$$

and similarly rearranging (7.4),

$$X = X^u + Z(X^a - X^u). \tag{7.7}$$

Substituting (7.7) into (7.6),

$$Y = Y^u + [X^u + Z(X^a - X^u)](Y^t - Y^u) = Y^u + X^u(Y^t - Y^u) + Z(X^a - X^u)(Y^t - Y^u). \tag{7.8}$$

For the terms in the numerator of (7.2), plugging in (7.8) and then using A7.2 along with the linearity of $\mathrm{E}(\cdot)$,

$$\begin{aligned}
\mathrm{E}(Y \mid Z=1) &= \mathrm{E}[Y^u + X^u(Y^t - Y^u) + Z(X^a - X^u)(Y^t - Y^u) \mid Z=1] \\
&= \mathrm{E}(Y^u \mid Z=1) + \mathrm{E}[X^u(Y^t - Y^u) \mid Z=1] \\
&\quad + \mathrm{E}[Z(X^a - X^u)(Y^t - Y^u) \mid Z=1] \\
&= \mathrm{E}(Y^u) + \mathrm{E}[X^u(Y^t - Y^u)] + \mathrm{E}[(X^a - X^u)(Y^t - Y^u)], \\
\mathrm{E}(Y \mid Z=0) &= \mathrm{E}[Y^u + X^u(Y^t - Y^u) + Z(X^a - X^u)(Y^t - Y^u) \mid Z=0] \\
&= \mathrm{E}(Y^u \mid Z=0) + \mathrm{E}[X^u(Y^t - Y^u) \mid Z=0] \\
&\quad + \mathrm{E}[Z(X^a - X^u)(Y^t - Y^u) \mid Z=0] \\
&= \mathrm{E}(Y^u) + \mathrm{E}[X^u(Y^t - Y^u)].
\end{aligned}$$

Subtracting,

$$\mathrm{E}(Y \mid Z=1) - \mathrm{E}(Y \mid Z=0) = \mathrm{E}[(X^a - X^u)(Y^t - Y^u)].$$

Now there are three possible values of $X^a - X^u$, so we can apply the law of total expectation:

$$\begin{aligned}
\mathrm{E}[(X^a - X^u)(Y^t - Y^u)] &= \mathrm{P}(X^t - X^u = 1)\,\mathrm{E}[(1)(Y^t - Y^u) \mid X^t - X^u = 1] \\
&\quad + \overbrace{\mathrm{P}(X^t - X^u = 0)\,\mathrm{E}[(0)(Y^t - Y^u) \mid X^t - X^u = 0]}^{=0} \\
&\quad + \overbrace{\mathrm{P}(X^t - X^u = -1)}^{=0 \text{ by A7.3}}\,\mathrm{E}[(-1)(Y^t - Y^u) \mid X^t - X^u = -1] \\
&= \overbrace{\mathrm{P}(X^t - X^u = 1)}^{=\mathrm{P}(\text{complier})}\,\overbrace{\mathrm{E}(Y^t - Y^u \mid X^t - X^u = 1)}^{\text{LATE}}.
\end{aligned}$$

For the denominator,

$$\begin{aligned}
\mathrm{E}(X \mid Z = 1) - \mathrm{E}(X \mid Z = 0) &= \mathrm{E}(X^a \mid Z = 1) - \mathrm{E}(X^u \mid Z = 0) \\
&= \mathrm{E}(X^a) - \mathrm{E}(X^u) \\
&= \mathrm{P}(X^a = 1) - \mathrm{P}(X^u = 1) \\
&= \mathrm{P}(\text{A or C}) - \mathrm{P}(\text{A or D}) \\
&= [\mathrm{P}(\text{A}) + \mathrm{P}(\text{C})] - [\mathrm{P}(\text{A}) + \overbrace{\mathrm{P}(\text{D})}^{=0 \text{ by A7.3}}\ ] \\
&= \mathrm{P}(\text{C}) = \mathrm{P}(X^a - X^u = 1).
\end{aligned}$$

Finally, taking the quotient,

$$\begin{aligned}
\frac{\mathrm{E}(Y \mid Z = 1) - \mathrm{E}(Y \mid Z = 0)}{\mathrm{E}(X \mid Z = 1) - \mathrm{E}(X \mid Z = 0)} &= \frac{\mathrm{P}(X^a - X^u = 1)\,\mathrm{E}(Y^t - Y^u \mid X^t - X^u = 1)}{\mathrm{P}(X^a - X^u = 1)} \\
&= \mathrm{E}(Y^t - Y^u \mid X^t - X^u = 1),
\end{aligned}$$

which is the LATE. □

Chapter 8
IV Regression

Unit learning objectives for this chapter

8.1. Define terms and concepts related to instrumental variables identification and estimation from the structural perspective. [TLO 1]

8.2. Describe IV regression estimators, including their estimands and assumptions, from the structural perspective, both mathematically and intuitively. [TLOs 2 and 3]

8.3. Judge whether an instrument is valid in real-world examples. [TLO 4]

Optional resources for this chapter

- Reverse causality and simultaneity (Masten video)

This chapter consider the instrumental variables approach from a structural regression perspective.

8.1 Simple IV Regression

Consider a simple setting with structural model

$$Y = \beta_0 + \beta_1 X + U \tag{8.1}$$

and scalar instrument Z. Sometimes Z is called an **excluded instrument** because it does not appear in the structural model (nor is it related to the structural error, as formalized below). Imagine X is endogenous, here meaning correlated with U. OLS is

consistent for the linear projection slope coefficient $\text{Cov}(Y,X)/\text{Var}(X)$, but

$$\frac{\text{Cov}(Y,X)}{\text{Var}(X)} = \frac{\text{Cov}(\beta_0+\beta_1 X+U,X)}{\text{Var}(X)} = \frac{\beta_1\text{Var}(X)+\text{Cov}(U,X)}{\text{Var}(X)} = \beta_1 + \frac{\text{Cov}(U,X)}{\text{Var}(X)}. \tag{8.2}$$

If X is exogenous in the sense of $\text{Cov}(U,X)=0$, then the structural slope β_1 is identified and equals the LP slope, but if $\text{Cov}(U,X)\neq 0$ then there is asymptotic bias.

Generally, the fixed (non-random) structural coefficient β_1 in (8.1) does not easily generalize to an interpretation as the mean of a random coefficient. This is a limitation. In contrast, the LATE framework allows arbitrary individual-level treatment effects (the equivalent of random β_1). However, there are ways to extend IV regression to allow random coefficients, although they are beyond our scope.

Beyond our scope...

One way to use the IV approach in a random coefficients model is to model the coefficients as non-random functions of a random scalar "rank variable." The seminal work of Chernozhukov and Hansen (2005) provides identification results for an IV quantile regression model that allows the slope coefficient to vary with the individual's rank variable value. For example, see Section 7.1 of Kaplan (2021) and references therein, or try the `sivqr` Stata command introduced by Kaplan (2022b), based on Kaplan and Sun (2017).

The following subsections establish identification of β_1 using different approaches. Some approaches help develop intuition, and others generalize better to more complex models.

The two critical IV regression assumptions are qualitatively similar to the independence and relevance conditions from Assumptions A7.2 and A7.4.

Assumption A8.1 (exogeneity)**.** The instrument Z is exogenous in the sense of uncorrelated with the structural error U from (8.1): $\text{Cov}(Z,U)=0$.

Assumption A8.2 (relevance)**.** The instrument Z is relevant in the sense of correlated with the regressor X: $\text{Cov}(Z,X)\neq 0$.

8.1.1 Ratio of Covariances

One way to think of the IV strategy is to separate the "endogenous part" of X from the "exogenous part": the instrument Z should vary with X but not U. The intuition is the same as in Section 7.1: see how Y varies with Z, then see how X varies with Z, and infer how much variation in Y is caused by X by dividing. This is formalized in Theorem 8.1.

Theorem 8.1 (simple IV identification)**.** *Given structural model* (8.1)*, under Assumptions A8.1 and A8.2, the structural slope* β_1 *is identified and equals* $\text{Cov}(Z,Y)/\text{Cov}(Z,X)$.

Proof. Starting from the statistical object, whose denominator is non-zero due to Assumption A8.2, and plugging in for Y from (8.1) and using the linearity of $\text{Cov}(\cdot)$,

$$\frac{\text{Cov}(Z,Y)}{\text{Cov}(Z,X)} = \frac{\text{Cov}(Z,\beta_0+\beta_1 X+U)}{\text{Cov}(Z,X)} = \frac{\beta_1\,\text{Cov}(Z,X)+\text{Cov}(Z,U)}{\text{Cov}(Z,X)} = \beta_1 + \frac{\text{Cov}(Z,U)}{\text{Cov}(Z,X)}, \tag{8.3}$$

and the numerator of the second term is zero by Assumption A8.1. □

Discussion Question 8.1 (IVs for education)**.** This example is similar to Example 5.1 of Wooldridge (2010). Let Y be log wage and X years of education, with structural model $Y = \beta_0 + \beta_1 X + U$. For each of the following, discuss why you think it does or does not satisfy each of Assumptions A8.1 and A8.2.

a) Z: years of education of the individual's mother
b) Z: last digit of the individual's Social Security number
c) Z: quarter of birth ($Z = 1$ if January through March, $Z = 2$ if April through June, $Z = 3$ if July through Sept., else $Z = 4$); note that many U.S. states require you to attend school until a certain age (say 16), but the corresponding grade level depends on which month you were born in

8.1.2 Ratio of LP Slopes

To develop intuition, consider another derivation of the same IV estimator. Let

$$\text{LP}(X \mid 1, Z) = \delta_0 + \theta Z, \quad R \equiv X - \text{LP}(X \mid 1, Z), \tag{8.4}$$

and plug the LP in error form into the structural model (8.1):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + U \\ &= \beta_0 + \beta_1(\delta_0 + \theta Z + R) + U \\ &= (\beta_0 + \beta_1\delta_0) + \beta_1\theta Z + (\beta_1 R + U). \end{aligned} \tag{8.5}$$

The term $\beta_1\theta$ is the slope of $\text{LP}(Y \mid 1, Z)$ because $\beta_1 R + U$ is uncorrelated with Z: using linearity,

$$\text{Cov}(Z, \beta_1 R + U) = \beta_1 \overbrace{\text{Cov}(Z,R)}^{=0 \text{ by } (8.4)} + \overbrace{\text{Cov}(Z,U)}^{=0 \text{ by A8.1}} = 0. \tag{8.6}$$

Thus, the structural slope β_1 is the ratio of the slope in $\text{LP}(Y \mid 1, Z)$ and the slope in $\text{LP}(X \mid 1, Z)$: $\beta_1\theta/\theta = \beta_1$.

The LP slopes are sometimes called **reduced form parameters**, meaning they are just "statistical" parameters (that can usually be estimated consistently, here by OLS). The ratio-of-slopes estimator is sometimes called the **Wald estimator**. Note that with binary X and Z, the ratio of LP slopes equals the expression in (7.1).

Corollary 8.2 (simple IV identification: LP slope ratio)**.** *Given structural model* (8.1)*, under Assumptions A8.1 and A8.2, the structural slope* β_1 *is identified and equals the ratio of the slope in* $\text{LP}(Y \mid 1, Z) = \rho_0 + Z\rho_1$ *and the slope in* $\text{LP}(X \mid 1, Z) = \delta_0 + Z\theta$*:* $\beta_1 = \rho_1/\theta$.

Proof. The LP slopes can be written in terms of covariances and variances like usual: $\rho_1 = \mathrm{Cov}(Z,Y)/\mathrm{Var}(Z)$ and $\theta = \mathrm{Cov}(Z,X)/\mathrm{Var}(Z)$. The ratio is thus

$$\frac{\mathrm{Cov}(Z,Y)/\mathrm{Var}(Z)}{\mathrm{Cov}(Z,X)/\mathrm{Var}(Z)} = \mathrm{Cov}(Z,Y)/\mathrm{Cov}(Z,X),$$

which is the expression proved to equal β_1 (under these assumption) in Theorem 8.1. □

8.1.3 Isolating Exogenous Part of Regressor

Another approach explicitly finds the "exogenous part" of X by projecting it onto the exogenous instrument Z. More specifically, let

$$X^* \equiv \mathrm{LP}(X \mid 1, Z) = \delta_0 + \theta Z, \tag{8.7}$$

using notation from (8.4). If we run OLS with X^* instead of X, then we can estimate the slope of

$$\mathrm{LP}(Y \mid 1, X^*) = \gamma_0 + \gamma_1 X^*.$$

With $V \equiv Y - \mathrm{LP}(Y \mid 1, X^*)$,

$$Y = \gamma_0 + \gamma_1 X^* + V = \gamma_0 + \gamma_1(\delta_0 + \theta Z) + V = (\gamma_0 + \gamma_1\delta_0) + (\gamma_1\theta)Z + V. \tag{8.8}$$

By the LP error property,

$$0 = \mathrm{Cov}(V, X^*) = \mathrm{Cov}(V, \delta_0 + \theta Z) = \theta\,\mathrm{Cov}(V, Z), \tag{8.9}$$

and by Assumption A8.2 (relevance), $\theta \neq 0$, so it must be that $\mathrm{Cov}(V,Z) = 0$. That is, the RHS of (8.8) is also a linear projection in error form, specifically $\mathrm{LP}(Y \mid 1, Z)$. Finally, recall from (8.5) that the slope of $\mathrm{LP}(Y \mid 1, Z)$ is $\beta_1\theta$, which must equal the slope on the RHS of (8.8). Thus, $\beta_1\theta = \gamma_1\theta$, which along with A8.2 implies $\gamma_1 = \beta_1$. That is, the structural parameter β_1 is identified and equal to the slope of $\mathrm{LP}(Y \mid 1, X^*)$, where X^* was constructed to be the "exogenous part" of X.

The estimator corresponding to this identification strategy would be first to run OLS to estimate $\mathrm{LP}(X \mid 1, Z)$, and second to run OLS to estimate $\mathrm{LP}(Y \mid 1, \hat{X})$, where $\hat{X} = \hat{\delta}_0 + \hat{\theta}Z$. This is the origin of the name **two-stage least squares** (2SLS), also sometimes abbreviated TSLS. However, this is not how any modern statistical software computes the IV regression estimator, nor does it provide intuition that generalizes well to other settings (like IV quantile regression).

8.1.4 Method of Moments

The most general perspective of IV regression is in terms of moment conditions. Let $\boldsymbol{Z} = (1, Z)'$ be the **full instrument vector**. In this simple model, Z is the only **excluded instrument** (because it does not appear in the structural model), and 1 is the only

included instrument (because it does appear in the structural model, implicitly in the intercept term; i.e., it's an exogenous regressor). The exogeneity assumption is that

$$\mathrm{E}(\boldsymbol{Z}U) = \mathbf{0}, \tag{8.10}$$

which means $\mathrm{E}(U) = 0$ (first element) and $\mathrm{E}(ZU) = 0$ (second element), the latter of which is equivalent to $\mathrm{Cov}(Z, U) = 0$ given that $\mathrm{E}(U) = 0$. As usual, if $\mathrm{E}(U) \neq 0$, then the intercept will be biased by $\mathrm{E}(U)$, but the slope coefficients are unaffected.

From (8.10), we can plug in for the structural error U from the structural model (8.1), and then solve for the corresponding population coefficient vector $\boldsymbol{\beta} \equiv (\beta_0, \beta_1)'$. Also defining $\boldsymbol{X} \equiv (1, X)'$,

$$\mathbf{0} = \mathrm{E}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{\beta})] = \mathrm{E}(\boldsymbol{Z}Y) - \mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')\boldsymbol{\beta} \tag{8.11}$$

using the linearity of expectation. The $\boldsymbol{\beta}$ can be isolated by moving that term to the other side and pre-multiplying by the inverse of the matrix $\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')$, yielding

$$\boldsymbol{\beta} = [\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{Z}Y). \tag{8.12}$$

This formula generalizes the covariance ratio from Theorem 8.1. It provides "identification" in that the LHS is a structural parameter (causal interpretation), whereas the RHS is a feature of the joint distribution of observable variables (Y, X, Z).

The first equality in (8.11) is an example of a **moment condition**. That is, the expected value of some function of observable variables (here Y, $\boldsymbol{X}$, and $\boldsymbol{Z}$) and parameters (here $\boldsymbol{\beta}$) equals zero. This restricts the possible values of the parameter that are consistent with the population. With enough restrictions, the parameter is uniquely determined, i.e., identified. If there are not enough restrictions to determine the parameter's value, then the parameter is **underidentified** (or "unidentified" or just "not identified"). If there are even more restrictions than we need, the parameter is **overidentified**. If there are just enough restrictions for identification, then the parameter is called **just-identified** or **exactly identified**.

> **Beyond our scope...**
>
> In some cases, there are enough restrictions to narrow down the parameter to a set of possible values, but not a single value, in which case the parameter is called **partially identified** or **set identified**. For example, see Part VI of Kaplan (2021) and references therein.

The RHS of (8.12) requires that the matrix inverse indeed exists. This is the required instrument relevance condition. It is also called a **rank condition** because a matrix is invertible if and only if it is full rank.

The following assumptions are equivalent to A8.1 and A8.2 in this simple model, but they generalize more readily.

Assumption A8.3 (exogeneity)**.** Given structural error term U from (8.1), the instrument vector $\boldsymbol{Z} = (1, Z)'$ satisfies $\mathrm{E}(\boldsymbol{Z}U) = \boldsymbol{0}$.

Assumption A8.4 (rank condition)**.** The matrix $\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')$ is invertible (or equivalently, full rank).

Theorem 8.3 (simple IV identification by moments)**.** *Given structural model* (8.1) *under Assumptions A8.3 and A8.4 (and assuming the moments* $\mathrm{E}(\boldsymbol{Z}Y)$ *and* $\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')$ *exist and are finite), the structural parameter vector* $\boldsymbol{\beta}$ *is identified and equals* $[\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{Z}Y)$.

Proof. Repeating the arguments in the text above: combining (8.1) with A8.3 yields $\boldsymbol{0} = \mathrm{E}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{\beta})]$, and solving for $\boldsymbol{\beta}$ yields the formula given, which is well-defined given the finite moments and invertibility of A8.4. □

The formula in Theorem 8.3 suggests the sample analog estimator

$$\hat{\boldsymbol{\beta}} = [\widehat{\mathrm{E}}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{Z}Y). \tag{8.13}$$

Indeed, assuming the sampling type is such that a weak law of large number holds (and again assuming the population moments are well-defined and finite), the sample means converge to population means, and they can be combined by the continuous mapping theorem.

Asymptotic normality can also be established with an argument very similar to that for OLS. Plugging in for Y in (8.13),

$$\hat{\boldsymbol{\beta}} = [\widehat{\mathrm{E}}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}[\boldsymbol{Z}(\boldsymbol{X}'\boldsymbol{\beta} + U)] = \boldsymbol{\beta} + [\widehat{\mathrm{E}}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\boldsymbol{Z}U). \tag{8.14}$$

Centering and scaling like usual, and writing the last term in summation notation,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = [\widehat{\mathrm{E}}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i U_i. \tag{8.15}$$

We get $[\widehat{\mathrm{E}}(\boldsymbol{Z}\boldsymbol{X}')]^{-1} \xrightarrow{p} [\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}$ by a WLLN, and then the other term is asymptotically mean-zero normal by a central limit theorem, because $\mathrm{E}(\boldsymbol{Z}_i U_i) = \boldsymbol{0}$ by A8.3. Again, the specific WLLN/CLT depends on the type of sampling; iid sampling is sufficient, but not necessary.

8.2 IV with One Instrument

The first generalization of Section 8.1 is to allow exogenous regressors. Here are a few equations, with details saved for Section 8.3.

The structural model is now

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + U, \tag{8.16}$$

where $\boldsymbol{X}' = (X_1, X_2, \ldots, X_k)$ with $X_1 = 1$ (intercept term) and

$$\mathrm{E}(X_j U) = 0, \quad j = 1, \ldots, k-1. \tag{8.17}$$

As usual, if $\mathrm{E}(U) \neq 0$, then only the intercept term is affected; because the intercept usually does not have much economic importance, we do not worry about this.

The full instrument vector is $\boldsymbol{Z} = (X_1, \ldots, X_{k-1}, Z)'$, including all the exogenous regressors (including the constant $X_1 = 1$) and the excluded instrument Z.

The excluded instrument Z is assumed to be **relevant** in the sense that it has a non-zero coefficient in $\mathrm{LP}(X_k \mid X_1, \ldots, X_{k-1}, Z)$. Although not obvious, this is equivalent to $\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')$ being invertible (full rank).

The excluded instrument Z must also be exogenous in the sense of $\mathrm{Cov}(Z, U) = 0$, which again is equivalent to $\mathrm{E}(ZU) = 0$ given $\mathrm{E}(U) = 0$.

The population coefficient vector can be solved for from the moment condition generated by the exogeneity assumptions. The corresponding moment condition is $\mathbf{0} = \mathrm{E}(\boldsymbol{Z}U) = \mathrm{E}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{\beta})]$. Using the linearity of expectation and algebra,

$$\boldsymbol{\beta} = [\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')]^{-1}\,\mathrm{E}(\boldsymbol{Z}Y), \tag{8.18}$$

which is identical to the formula in Theorem 8.3.

Alternatively, extending Section 8.1.3, let

$$X_k^* \equiv \mathrm{LP}(X_k \mid X_1, \ldots, X_{k-1}, Z) = (X_1, \ldots, X_{k-1})\boldsymbol{\delta} + Z\theta, \quad R \equiv X_k - X_k^*. \tag{8.19}$$

Again, relevance requires $\theta \neq 0$. The coefficient on X_k^* in $\mathrm{LP}(Y \mid X_1, \ldots, X_{k-1}, X_k^*)$ is the structural coefficient β_k.

8.3 IV with Multiple Instruments

Generalizing further, consider the same structural model where only X_k is endogenous, but now there are multiple excluded instruments, $(Z_1, \ldots, Z_m)$.

8.3.1 Some Intuition

It is not obvious how to proceed. The formula in Theorem 8.3 no longer works because $\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')$ is not even a square matrix, so it cannot be invertible.

If we really felt stuck, then we could just ignore $(Z_2, \ldots, Z_m)$ and only use Z_1. However, if we really believe we have multiple valid IVs, then this feels like we are throwing away information (because we are), which intuitively should make our estimator less "efficient" (i.e., higher variance of sampling distribution / higher standard errors / more uncertainty). We could also use something like $(Z_1 + \cdots + Z_m)/m$ as our single instrument. This feels better, but also feels like an arbitrary way to combine our IVs. However, both Z_1 and $(Z_1 + \cdots + Z_m)/m$ are valid instruments (assuming each Z_j is valid), in

which case both should yield a consistent IV estimator. If we only cared about consistency, then we would not care which we used. (This is not fully true: recalling the LATE interpretation, we may worry that different instruments identify different causal estimands; but we will wait to worry about that more formally in Chapter 9.) But, we also want the most precise (lowest standard error / uncertainty) estimator possible. The question is how to combine the m instruments optimally to minimize the (asymptotic) variance. This question is not fully addressed until Chapter 10.

Extending (8.7) and (8.19), consider the LP

$$X_k^* \equiv \text{LP}(X_k \mid X_1, \ldots, X_{k-1}, Z_1, \ldots, Z_m) = (X_1, \ldots, X_{k-1})\boldsymbol{\delta} + (Z_1, \ldots, Z_m)\boldsymbol{\theta}. \quad (8.20)$$

Again, this X_k^* is a linear combination of exogenous variables (or "instruments"), including both exogenous regressors ("included instruments") and excluded instruments. Thus, X_k^* itself is exogenous (again here meaning uncorrelated with U). That is, the scalar X_k^* is a valid instrument for the endogenous scalar regressor X_k, so we have reduced the problem to IV regression with a single instrument, and we can use previous results. For example, letting $\tilde{\boldsymbol{Z}} \equiv (X_1, \ldots, X_{k-1}, X_k^*)'$,

$$\boldsymbol{\beta} = [\text{E}(\tilde{\boldsymbol{Z}}\boldsymbol{X}')]^{-1}\,\text{E}(\tilde{\boldsymbol{Z}}Y), \quad (8.21)$$

and the sample analog provides a consistent estimator (which can be proved by more formal arguments).

Although fundamentally the idea is to use X_k^* as an IV (which generalizes to other contexts like IV quantile regression), in this case there is an equivalence with using X_k^* as a regressor in a second-stage linear projection. Because each X_j for $j = 1, \ldots, X_{k-1}$ is the projection of itself onto $\boldsymbol{Z}$, we can write

$$\tilde{\boldsymbol{Z}}' = \boldsymbol{Z}'[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\,\text{E}(\boldsymbol{Z}\boldsymbol{X}'), \quad \tilde{\boldsymbol{Z}} = \text{E}(\boldsymbol{X}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\boldsymbol{Z}, \quad (8.22)$$

so

$$\begin{aligned}
\text{E}\{\tilde{\boldsymbol{Z}}\boldsymbol{X}'\} &= \text{E}\{\overbrace{\text{E}(\boldsymbol{X}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\boldsymbol{Z}}^{\tilde{\boldsymbol{Z}}}\boldsymbol{X}'\} \\
&= \text{E}(\boldsymbol{X}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\,\text{E}(\boldsymbol{Z}\boldsymbol{X}') \\
&= \text{E}(\boldsymbol{X}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\overbrace{\text{E}(\boldsymbol{Z}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}}^{\text{identity matrix}}\text{E}(\boldsymbol{Z}\boldsymbol{X}') \\
&= \text{E}\{\overbrace{\text{E}(\boldsymbol{X}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\boldsymbol{Z}}^{\tilde{\boldsymbol{Z}}}\overbrace{\boldsymbol{Z}'[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\,\text{E}(\boldsymbol{Z}\boldsymbol{X}')}^{\tilde{\boldsymbol{Z}}'}\} \\
&= \text{E}\{\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}'\}.
\end{aligned}$$

Thus, plugging into (8.21),

$$\boldsymbol{\beta} = [\text{E}(\tilde{\boldsymbol{Z}}\boldsymbol{X}')]^{-1}\,\text{E}(\tilde{\boldsymbol{Z}}Y) = [\text{E}(\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}')]^{-1}\,\text{E}(\tilde{\boldsymbol{Z}}Y),$$

which is the familiar formula for the LPC of $\text{LP}(Y \mid \tilde{\boldsymbol{Z}})$, which is what OLS regression of Y on $\tilde{\boldsymbol{Z}}$ estimates. This is the origin of the name **two-stage least squares**, where the first stage is the linear projection of X_k onto $(X_1, \ldots, X_{k-1}, Z_1, \ldots, Z_m)$, and the second stage is the linear projection of Y onto $\tilde{\boldsymbol{Z}}$, although again this is sort of a coincidence rather than a fundamental concept, so it does not generalize to other contexts like IV quantile regression.

8.3.2 Identification

The following formalizes the above results.

Assumption A8.5 (exogeneity)**.** The regressors $(X_1, \ldots, X_{k-1})$ (with $X_1 = 1$ to include an intercept) and the excluded instruments $(Z_1, \ldots, Z_m)$ are exogenous in the sense that $\text{E}(\boldsymbol{Z}U) = \boldsymbol{0}$, where $\boldsymbol{Z} \equiv (X_1, \ldots, X_{k-1}, Z_1, \ldots, Z_m)'$ and U is the structural error term from structural model $Y = \boldsymbol{X}'\boldsymbol{\beta} + U$ in (8.16).

Assumption A8.6 (rank condition)**.** The matrix $\text{E}(\boldsymbol{Z}\boldsymbol{X}')$ has full column rank; or equivalently, at least one component of $\boldsymbol{\theta}$ is non-zero in the LP in (8.20).

Theorem 8.4 (IV identification)**.** *Given structural model* (8.16)*, under Assumptions A8.5 and A8.6 and assuming all elements of* $\text{E}(\boldsymbol{Z}\boldsymbol{X}')$ *and* $\text{E}(\boldsymbol{Z}Y)$ *are well-defined and finite, the structural parameter vector* $\boldsymbol{\beta}$ *is identified and equals* $[\text{E}(\tilde{\boldsymbol{Z}}\boldsymbol{X}')]^{-1}\,\text{E}(\tilde{\boldsymbol{Z}}Y)$ *where* $\tilde{\boldsymbol{Z}} \equiv \text{E}(\boldsymbol{X}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\boldsymbol{Z}$*.*

Proof. See Section 8.3.1. To summarize: first,

$$\text{E}(\tilde{\boldsymbol{Z}}U) = \text{E}\{\text{E}(\boldsymbol{X}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\boldsymbol{Z}U\} = \text{E}(\boldsymbol{X}\boldsymbol{Z}')[\text{E}(\boldsymbol{Z}\boldsymbol{Z}')]^{-1}\overbrace{\text{E}\{\boldsymbol{Z}U\}}^{=\boldsymbol{0}\text{ by A8.5}} = \boldsymbol{0}.$$

Thus, plugging in $U = Y - \boldsymbol{X}'\boldsymbol{\beta}$ from (8.16),

$$\boldsymbol{0} = \text{E}[\tilde{\boldsymbol{Z}}(Y - \boldsymbol{X}'\boldsymbol{\beta})] = \text{E}(\tilde{\boldsymbol{Z}}Y) - \text{E}(\tilde{\boldsymbol{Z}}\boldsymbol{X}')\boldsymbol{\beta}.$$

Rearranging and solving for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} = [\text{E}(\tilde{\boldsymbol{Z}}\boldsymbol{X}')]^{-1}\,\text{E}(\tilde{\boldsymbol{Z}}Y).$$

Assumption A8.6 ensures that the matrix inverse indeed exists. □

Discussion Question 8.2 (conditional moment restriction)**.** Imagine we find an excluded instrument Z that is exogenous in the sense that $\text{E}(U \mid Z) = 0$. Discuss whether or not each of the following possible excluded instruments is exogenous.

a) Z
b) Z^2
c) Z^3
d) $\sin(Z)$

8.3.3 Estimation, Inference, and Efficiency

Although not particularly important for proper use in practice, the 2SLS estimator can be written as follows. Let $\underline{\boldsymbol{X}}$ be the $n \times k$ regressor matrix, with $\boldsymbol{X}_i'$ as row i. Similarly, let $\underline{\boldsymbol{Z}}$ be the $n \times m + k - 1$ matrix of all instruments (exogenous regressors and excluded instruments), with $\boldsymbol{Z}_i'$ as row i. Let $\underline{\boldsymbol{P}}_{\boldsymbol{Z}}$ be the projection matrix $\underline{\boldsymbol{Z}}(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}})^{-1}\underline{\boldsymbol{Z}}'$, so the $n \times k$ matrix $\tilde{\underline{\boldsymbol{Z}}}$ is

$$\tilde{\underline{\boldsymbol{Z}}} = \underline{\boldsymbol{P}}_{\boldsymbol{Z}}\underline{\boldsymbol{X}}. \tag{8.23}$$

The sample analog of the population $\boldsymbol{\beta}$ formula in Theorem 8.4 is then

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= [\widehat{\mathrm{E}}(\tilde{\boldsymbol{Z}}\boldsymbol{X}')]^{-1}\,\widehat{\mathrm{E}}(\tilde{\boldsymbol{Z}}Y) \\ &= \{(\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}/n)(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}/n)^{-1}(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}/n)\}^{-1}(\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}/n)(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}/n)^{-1}(\underline{\boldsymbol{Z}}'\boldsymbol{Y}/n),\end{aligned} \tag{8.24}$$

where $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)'$. Note all the $1/n$ cancel out and could be omitted to save space (as often done).

Assuming the type of sampling admits some WLLN and CLT, the consistency and asymptotic normality of the IV regression estimator in (8.24) follow readily. All the sample averages in the formula converge in probability to their corresponding population means, and assuming all the matrix inverses exist, then the CMT says their limits combine into the population $\boldsymbol{\beta}$. Asymptotic normality follows the same structure of argument as OLS. That is, we can first plug in $\boldsymbol{Y} = \underline{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{U}$, where $\boldsymbol{U} \equiv (U_1, \ldots, U_n)'$; centering and scaling gives the form

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \{(\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}/n)(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}/n)^{-1}(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}/n)\}^{-1}(\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}/n)(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}/n)^{-1}\sqrt{n}(\underline{\boldsymbol{Z}}'\boldsymbol{U}/n), \tag{8.25}$$

the last term of which is, in summation notation,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{Z}_i U_i, \tag{8.26}$$

to which we can apply a CLT, noting $\mathrm{E}(\boldsymbol{Z}_i U_i) = \boldsymbol{0}$. Again applying a WLLN to the other terms and combining with the CMT yields the final asymptotic normal distribution. That is, defining $\underline{\boldsymbol{\Sigma}} \equiv \mathrm{E}[U^2\boldsymbol{Z}\boldsymbol{Z}']$, $\underline{\boldsymbol{Q}}_{XZ} \equiv \mathrm{E}(\boldsymbol{X}\boldsymbol{Z}')$, and $\underline{\boldsymbol{Q}}_{ZZ} \equiv \mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}')$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}_{XZ}'\}^{-1}\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{\Sigma}}), \tag{8.27}$$

which follows a normal distribution with mean zero and covariance matrix

$$\underline{\boldsymbol{\Omega}} \equiv \{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}_{XZ}'\}^{-1}\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{\Sigma}}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}_{XZ}'\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}_{XZ}'\}^{-1}. \tag{8.28}$$

Confidence intervals (and confidence sets and hypothesis tests) can be constructed based on this asymptotic normal distribution, just as you have seen done for LP coefficients estimated by OLS.

The covariance matrix $\underline{\boldsymbol{\Omega}}$ in (8.28) has a **sandwich form**, with the general structure $A^{-1}BA^{-1}$. This structure usually implies the estimator is not the most efficient possible, at least in theory. However, if the structural error U is homoskedastic in the sense of $\text{Var}(U \mid \boldsymbol{Z}) = \text{Var}(U) \equiv \sigma_U^2$, then using iterated expectations

$$\underline{\boldsymbol{\Sigma}} = \text{E}[U^2 \boldsymbol{Z}\boldsymbol{Z}'] = \text{E}[\text{E}(U^2 \boldsymbol{Z}\boldsymbol{Z}' \mid \boldsymbol{Z})] = \text{E}[\overbrace{\text{E}(U^2 \mid \boldsymbol{Z})}^{=\text{Var}(U|\boldsymbol{Z})=\sigma_U^2} \boldsymbol{Z}\boldsymbol{Z}'] = \sigma_U^2 \overbrace{\text{E}(\boldsymbol{Z}\boldsymbol{Z}')}^{\underline{\boldsymbol{Q}}_{ZZ}}. \tag{8.29}$$

Because σ_U^2 is a scalar, it can move freely throughout $\underline{\boldsymbol{\Omega}}$, so

$$\begin{aligned}
\underline{\boldsymbol{\Omega}} &= \{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\overbrace{\sigma_U^2 \underbrace{\underline{\boldsymbol{Q}}_{ZZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}}_{\text{identity matrix}}}^{\underline{\boldsymbol{\Sigma}}} \underline{\boldsymbol{Q}}'_{XZ}\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1} \\
&= \sigma_U^2\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1} \underbrace{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}}_{\text{identity matrix}} \qquad (8.30) \\
&= \sigma_U^2\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{Q}}_{ZZ}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}. \qquad (8.31)
\end{aligned}$$

That is, under homoskedasticity, the sandwich covariance "collapses," which usually indicates efficiency (within a certain class of estimators). Efficiency is discussed more in Chapter 10.

8.4 General IV Regression

The most general case of linear IV regression allows both multiple instruments and multiple endogenous regressors. Perhaps surprisingly, this does not change much from Section 8.3, at least in terms of identification and estimation. That is, as long as we have enough valid excluded instruments to satisfy Assumptions A8.5 and A8.6, then the identification result holds with the same formula, and we can again use the sample analog as a consistent, asymptotically normal estimator.

As a sanity check, a necessary (but not sufficient) condition for identification is that there are at least as many excluded instruments as endogenous regressors. The intuition for the excluded instruments remains the same as in the simpler settings: they must be unrelated to the structural error term (other determinants of Y besides $\boldsymbol{X}$), and they must be related to the endogenous regressors that they instrument for. For example, if both X_2 and X_3 are endogenous, then we hope to have Z_1 related to X_2 and Z_2 related to X_3 (or in principle both IVs could relate to both regressors, but in practice usually each IV instruments for one particular regressor, at least conceptually).

The biggest difference in this case of multiple endogenous regressors is discuss in Chapter 9.

Chapter 9
IV Diagnostics

Unit learning objectives for this chapter

9.1. Define terms and concepts related to assessment of IV model validity. [TLO 1]

9.2. Describe tests for weak instruments and for model misspecification, both mathematically and intuitively. [TLOs 2 and 3]

9.3. Use diagnostic tests to help judge whether an IV model is valid in real-world examples. [TLO 4]

This chapter discusses common diagnostics of IV identification. Related measures are reported by the user-contributed `ivreg2` Stata command (Baum, Schaffer, and Stillman, 2002), with the help of `ranktest` (Kleibergen, Schaffer, and Windmeijer, 2007), both available through SSC.

9.1 Underidentification

Recall the simple IV regression model of Section 8.1 and its relevance condition (rank condition). In the linear projection of endogenous X onto $(1, Z)$, the coefficient on Z was required to be non-zero. That is, with $\text{LP}(X \mid 1, Z) = \delta_0 + Z\theta$, the relevance condition holds if and only if $\theta \neq 0$. If $\theta = 0$, then we have a problem of **underidentification**: we do not have enough "information" from the IV to help us learn about the coefficient on X. Put differently, there are an infinite number of possible values of the structural slope β_1 that could be consistent with the population distribution of observable variables. From yet another perspective: the denominator in $\text{Cov}(Z, Y)/\text{Cov}(Z, X)$ is zero, thus the expression is undefined; recall this came from $\text{Cov}(Z, Y) = \text{Cov}(Z, X)\beta_1$, so if $\text{Cov}(Z, X) = 0$, the RHS evaluates to zero regardless of β_1, so β_1 can be any value.

We know how to test the null hypothesis that a linear projection slope is zero, i.e., $H_0 \colon \theta = 0$ in the "first stage." Thus, we can interpret this as a test of underidentification,

where "underidentification" (specifically the violation of relevance) is the null hypothesis, so the alternative hypothesis is that relevance is satisfied. (The alternative is not that β_1 is identified, because it is possible that relevance holds but exogeneity does not, in which case identification fails for a different reason.)

Because the null hypothesis (relevance failure) is bad news, we are hoping to have a very small p-value and reject the null. That is, a small p-value suggests the data are not consistent with $\theta = 0$. Rejection requires relatively strong empirical evidence because the (frequentist) hypothesis test must control its **type I error rate**, the probability of incorrectly rejecting when H_0 is actually true. Conversely, if there is just a lot of uncertainty in the data, then the test will default to non-rejection to avoid making too many type I errors. Recall that if the null is true and the test incorrectly fails to reject, then it is a **type II error**; and **power** is 100% minus the type II error rate. Because frequentist tests usually make no claim about type II error rate (power), there may be cases where the type II error rate is very high (i.e., low power). For example, if there is a small sample size n, then we have lots of uncertainty, so tests are prone to have type II errors. Even with large n, if H_0 is false but "close" to true, then there can be a high type II error rate. So, in practice, if the underidentification test rejects, then we have fairly strong evidence that the IV is relevant, but if it fails to reject, then we should not necessarily conclude that relevance fails. But, as seen in Section 9.2, have θ "close" to zero is also problematic.

In the more general IV regression model with a single endogenous regressor but other exogenous regressors and possibly multiple excluded instruments, the relevance or rank condition requires at least one non-zero element of the vector $\boldsymbol{\theta}$ of coefficients on the excluded vectors in the "first stage" LP in (8.20). That is, failure of the rank condition is equivalent to $\boldsymbol{\theta} = \mathbf{0}$. Again, we know how to test such a null hypothesis for LP coefficients estimated by OLS, using a Wald test.

In the most general case with multiple endogenous regressors, there is not such a familiar equivalence of the rank condition on $\mathrm{E}(\boldsymbol{Z}\boldsymbol{X}')$ and LP coefficients. However, there are other tests that can test for full rank of a matrix. For these tests, too, the null hypothesis is the failure of the rank condition, so we hope to get a low p-value and reject the null. The Kleibergen and Paap (2006) test (KP test) has the null hypothesis that the rank is equal to $k-1$, with alternative hypothesis that the rank is k (full column rank). In some cases, the test can perform less well if the true rank is actually $k-2$ (or less), as pointed out by Chen and Fang (2019), who propose a rank test that does not suffer such problems. They also have a Stata command `bootranktest`.[1]

As always, you should also consider your prior beliefs when interpreting statistical results. For example, if you have a set of variables that you don't think relate to X, but you just keep running KP tests until you get a low p-value, this does not mean that you magically found an amazing, valid instrument. Ideally, you should have other real-world reasons you believe the rank condition (relevance) holds, and then use the statistical tests to show others that the data are consistent with your arguments.

[1] As described here: https://arxiv.org/abs/2108.00511

9.2 Weak Identification

Discussion Question 9.1. Consider the IV (Wald) estimator $\hat{\beta}_1 = \hat{\lambda}/\hat{\theta}$ like in Section 8.1, where λ is the slope of $\text{LP}(Y \mid 1, Z)$ and θ is the slope of $\text{LP}(X \mid 1, Z)$. Assume the instrument Z is exogenous.

a) Is β_1 identified if $\theta = 1$? If $\theta = 0$? If $\theta = 0.0001$?
b) For a given sample size n, how might the sampling distribution of $\hat{\beta}_1$ differ across those values (1, 0, 0.0001) of the true population θ?

The problem of **weak identification** occurs when condition(s) for identification are "close" to being violated. In the IV setting, this is sometimes called the problem of **weak instruments** because the weak identification is (roughly speaking) due to the excluded instrument correlation with the endogenous regressor being too weak. More precisely, if $\text{E}(\boldsymbol{Z}\boldsymbol{X}')$ has full column rank k but is "close" to having rank $k-1$, then there can be problems in practice.

Beyond our scope...

There are sometimes multiple asymptotic frameworks that can be used to study an estimator, including with IV regression. Under "conventional" asymptotics, we take the distribution of $(Y, \boldsymbol{X}', \boldsymbol{Z}')$ as fixed as we let sample size $n \to \infty$. In the simple IV regression case, for example, if $\text{Cov}(Z, X) > 0$, then as $n \to \infty$, the estimator $\text{Cov}(Z, Y)/\text{Cov}(Z, X)$ is consistent and asymptotically normal, regardless of how near zero is $\text{Cov}(Z, X)$. But we can see (for example from simulations) that for a given n in practice we have problems when $\text{Cov}(Z, X)$ is near zero. That is, this "conventional asymptotics" fails to capture the real-world performance of the estimator. A more sophisticated asymptotic framework can succeed in representing the weak instrument problem. As initially suggested by Staiger and Stock (1997), the trick is to set $\text{Cov}(Z, X) = c/\sqrt{n}$, where c is a constant, so $c/\sqrt{n}$ is a sequence that goes to zero (at a particular "rate" of $n^{-1/2}$) as $n \to \infty$. When limits are taken using this weak-instrument asymptotic framework, they show the effect of the instrument's strength (c). That is, this framework provides more accurate approximations of real-world estimator properties. Other examples of this phenomenon include many-regressors asymptotics with number of regressors cn proportional to n (or some function of n) and local-to-unit-root asymptotics where the AR(1)'s autoregressive parameter is $1 - c/\sqrt{n}$.

9.2.1 Consequences of Weak Identification

One consequence of weak identification is bias. That is, even with large n, the IV estimator's distribution is not centered at the true parameter value if there are weak instruments. Interestingly, weak IV bias is in the direction of the OLS estimator; when instruments are totally irrelevant (underidentification), the IV estimator is centered at the OLS estimand

(the linear projection coefficient). Moreover, the (true) standard errors can be particularly large with weak IV because the estimator's sampling distribution does not collapse to a single point as $n \to \infty$.

Another consequence of weak identification is incorrect inference (like confidence intervals) if it is based on asymptotic normality, because the IV estimator is not asymptotically normal under the weak IV asymptotics. That is, even with large n, a 95% confidence interval may have actual coverage probability much lower than 95%, a problem called **undercoverage**. Often this problem is phrased in terms of **size distortion**, meaning that a level 5% hypothesis test rejects a true H_0 with probability more than 5% (even with large n).

9.2.2 Assessing Weak Identification

There are methods to gauge the strength of identification for IV regression. Recall from Section 9.1 the test of $H_0 : \boldsymbol{\theta} = \mathbf{0}$ of the first-stage coefficients on the excluded instruments. If we construct the F-statistic for this hypothesis, then comparing to the usual critical value gives us a test of underidentification; but we know that weak identification can still be a problem even if we can reject underidentification. Thus, intuitively, if we use the F-statistic (or something like it) to measure instrument strength, we want it to be even larger than the usual critical value.

There are indeed alternative (higher) critical values that correspond to different levels of bias and size distortion caused by weak identification. The early "$F > 10$" rule-of-thumb was suggested by Staiger and Stock (1997), and it indeed gives a rough sense of instrument strength in most cases (like if you're in a seminar and don't have a detailed critical value table handy). Stock and Yogo (2005) later tabulated critical values that depend on the level of bias or size distortion, as well as depending on the number of endogenous regressors and instruments, and even extending to related "k-class" estimators; see their Tables 5.1–5.4. With multiple endogenous regressors, the F-statistic is replaced by the more general Cragg–Donald statistic. Roughly, the null hypothesis is like "instruments are weak enough that bias may exceed 20%," with the alternative hypothesis that weak IV bias is less than 20%. Some such critical values are reported by Stata commands like `ivreg2`.

Note: "bias" in the weak IV context usually means "relative bias," which means bias as a percentage of the OLS bias. For example, if OLS bias is 8, then 20% relative bias would be $(8)(20\%) = 1.6$. Of course, if we have a case where OLS bias is very small, then it does not actually matter even if we have "80% relative bias," whereas if OLS bias is very large then 10% may still be economically significant.

More recently, for a single endogenous regressor with multiple excluded instruments, some experts recommend[2] using the Stata command `weakivtest` (Pflueger and Wang, 2013), available on SSC, based on the work of Montiel Olea and Pflueger (2013).

[2] https://web.archive.org/web/20230201022532/https://www.nber.org/sites/default/files/2020-12/NBERSI2018_Methods%20Lectures_WeakIV1-2_v4.pdf

9.2.3 Coping with Weak Identification

Failing to reject a weak instrument test does not mean that you should give up on your research, but it does mean that you should be suspicious of your estimated $\hat{\beta}$ and use special confidence intervals that are robust to weak instruments.

The intuition for the possibility of valid weak-IV-robust inference is that you do not need to consistently estimate β in order to test a hypothesis about its value, because the null hypothesis specifies the value for you, like $H_0\colon \beta = 0$. For example, consider the simple IV regression setting of Corollary 8.2, where $\mathrm{LP}(Y \mid 1, Z) = \rho_0 + Z\rho_1$ and $\mathrm{LP}(X \mid 1, Z) = \delta_0 + Z\theta$, and given a valid instrument Z, the structural slope β_1 is identified with $\beta_1 = \rho_1/\theta$. We can have problems estimating β_1 if θ is near zero, but we do not have any problem estimating the linear projection coefficients ρ_1 and θ. That is, OLS estimators $\hat{\rho}_1$ and $\hat{\theta}$ are consistent and jointly asymptotically normal, meaning $\sqrt{n}(\hat{\rho}_1 - \rho_1, \hat{\theta} - \theta)'$ converges in distribution to a bivariate mean-zero normal distribution. If we want to test whether $\beta_1 = 5$, then instead of trying to test whether $\rho_1/\theta = 5$, we can rearrange and equivalently test whether $H_0\colon \rho_1 - 5\theta = 0$. Because of the joint asymptotic normality, $\hat{\rho}_1 - 5\hat{\theta}$ is also approximately normal (because it's a linear combination of jointly normal random variables). After deriving the asymptotic variance, we can use the usual normality-based t-test. This is called the **Anderson–Rubin** (AR) approach to hypothesis testing under weak identification, going back to Anderson and Rubin (1949).

Such a hypothesis test can be "inverted" into a confidence interval, a procedure called **test inversion**. If we have a level α test, then we can derive a confidence level $1 - \alpha$ CI. Specifically, the CI collects all possible values of β_1 that are not rejected by the test. The probability that the CI contains the true β_1 equals the probability that the true β_1 is not rejected. That is,

$$\mathrm{P}(\text{CI contains } \beta_1) = \mathrm{P}(\beta_1 \text{ not rejected}) = 1 - \overbrace{\mathrm{P}(\beta_1 \text{ rejected})}^{\leq \alpha} \geq 1 - \alpha. \tag{9.1}$$

The CI from inverting the AR test is called an Anderson–Rubin CI. Such a CI can equal $(-\infty, \infty)$, specifically when we cannot reject $\theta = 0$. Generally, the AR CI works well in just-identified models (same number of excluded instruments as endogenous regressors), but less well in overidentified models (more excluded instruments than endogenous regressors).

In Stata, as recommended by Isaiah Andrews,[3] weak-IV-robust confidence intervals can be computed by the `weakiv` command (Finlay, Magnusson, and Schaffer, 2013), available on SSC.

[3] https://web.archive.org/web/20230201022416/https://www.nber.org/sites/default/files/2020-12/robustinference_openissues.pdf

9.3 Misspecification

This section provides intuition that is later formalized in the more general GMM context in Section 10.4.2.

There is a type of test sometimes confusingly called an **overidentification test**, or ***J*-test**, or **Sargan–Hansen test** (Hansen, 1982; Sargan, 1958), or less-confusingly called a **test of overidentifying restrictions**. It is a type of **specification test**, where "specification" refers to our structural model and our various identifying assumptions. If we have more restrictions than we need to estimate the parameter of interest, then we can test whether the restrictions are all consistent with each other.

If the system is just-identified (exactly identified), then we need all of our restrictions just to estimate the parameters. For example, in simple IV regression with scalar structural slope β_1, if we have one valid excluded instrument then we have one "restriction" $\text{Cov}(Z, U) = 0$, which is just enough to estimate β_1. But then we have exhausted all the information (all the moment conditions) we have.

If the model is overidentified, then we can use the extra identifying restrictions to test the assumptions we've made, broadly speaking. For example, continuing the simple IV regression example, imagine we now have two excluded instruments, Z_1 and Z_2, and we think/hope both are uncorrelated with the structural error $U \equiv Y - \beta_0 - X\beta_1$. Even without Z_2, we can estimate the parameters and construct residuals

$$\hat{U}_i = Y_i - \hat{\beta}_0 - X_i\hat{\beta}_1 \tag{9.2}$$

for each observation $i = 1, \ldots, n$. If Z_1 is indeed valid, then in large samples, the estimators should (with high probability) be very close to the true values, because they are consistent: $(\hat{\beta}_0, \hat{\beta}_1) \xrightarrow{p} (\beta_0, \beta_1)$. Thus, the residual $\hat{U}_i$ should be very close to the true unobserved structural error term U_i. If Z_2 is exogenous (uncorrelated with the true U), then the sample Z_{2i} should be approximately uncorrelated with the residuals $\hat{U}_i$. Given the asymptotic normality of the estimators, test statistics can be derived with a known asymptotic distribution, so we can compute the corresponding p-values.

Note that if the model is exactly identified, like if we only had Z_1 in the running example, then we cannot learn anything from the sample correlation of Z_{1i} and $\hat{U}_i$ because it is set to be zero (exactly) by the estimator $(\hat{\beta}_0, \hat{\beta}_1)$. So, our test statistic would always be zero. Indeed, in that case `ivreg2` reports the Hansen J-statistic equal to 0.000, and the output also notes "equation exactly identified."

The interpretation of the null hypothesis depends on how confident you are about different parts of your model. For example, if you are confident about Z_1, then you could interpret this as a test of Z_2. Or if you are confident that both Z_1 and Z_2 are independent of variables besides X that affect Y, then you could interpret this as a test of the linear functional form of your structural model $Y = \beta_0 + X\beta_1 + U$. That is, if U also contains terms like X^2, then Z_1 and Z_2 may be correlated with this U even if they are independent of all variables other than X. Most generally, we can interpret the null hypothesis as saying that all of our assumptions are correct, against the alternative hypothesis that at

least one of our assumptions is wrong. For this reason, tests of overidentifying restrictions are often called **omnibus tests**, meaning it's a single test that tests everything mixed together.

Because the null is correct specification, we hope to have a high p-value and not reject the null. If instead the p-value is small, then it suggests our model is not consistent with the observed data, so we should interpret our results cautiously. However, as Box (1979, p. 2) famously wrote, "All models are wrong but some are useful," so we do not necessarily want to throw away our results just because of the specification test's rejection. This is especially true if our sample size n is very large, which gives the test high power: even a very small deviation from our model can be detected by the test, leading to rejection. Conversely, if the sample size is small, then the test has low power (high type II error rate) even against larger violations of the model, so non-rejection does not necessarily tell us much about our model specification.

As another caveat to interpretation, recall the LATE of Section 7.3: different instruments may identify different causal parameters. So, possibly we have two instruments for education that are both valid, but they identify a different causal parameter, like the return to education for the 12th year of education vs. for the 16th year of education. If those population parameters differ, then the J-test may reject even though each instrument is valid, just for a different causal parameter.

In Stata, the `ivreg2` output show's the p-value (and test statistic value, labeled "Hansen J statistic") for a J-test. (Alternatively, after running `ivregress` you could run `estat overid`, but it seems better and easier to just run `ivreg2`.)

Discussion Question 9.2 (J-test)**.** Imagine you are using survey data and have one endogenous regressor and three possible excluded instruments, (Z_1, Z_2, Z_3). The first two (Z_1 and Z_2) are from the main survey and are non-missing for 99% of the sample; the third (Z_3) is from a supplemental survey and only available for 10% of the observations in your sample. You run the J-test three times: with Z_1 only, with Z_1 and Z_2, and with Z_1 and Z_3. The Z_1 test statistic equals zero. The (Z_1, Z_2) test statistic is much larger, with a p-value of 0.04. The (Z_1, Z_3) test statistic is in between, with $p = 0.17$. How do you interpret these results? What do you learn about the models?

Chapter 10
Generalized Method of Moments

Unit learning objectives for this chapter

10.1. Define terms and concepts related to GMM. [TLO 1]

10.2. Describe the GMM estimator mathematically and intuitively. [TLO 3]

This chapter provides a relatively brief overview of the generalized method of moments (GMM). GMM is defined very generally and includes other estimators (like 2SLS) as a special case. Unlike linear IV regression, the asymptotic arguments are qualitatively different than those for OLS.

I recently found some lecture notes by Bent Sørensen that provide many mathematical details while also providing intuition.[1] For even more details, see Chapters 12 and 14 of Wooldridge (2010), or the classic GMM *Handbook of Econometrics* chapter by Newey and McFadden (1994).

10.1 Basic Setting and Notation

Generally, let $\boldsymbol{D}$ (for "data") denote a vector containing all the observable variables (outcome, regressors, instruments, etc.), and let the parameter vector of interest be $\boldsymbol{\theta} \in \Theta$, where Θ is the parameter space (like $\mathbb{R}^k$ or some subset of $\mathbb{R}^k$). Similar to Section 8.1.4, our identifying assumptions lead to moment conditions of the form

$$\mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{\theta})] = \mathbf{0}, \tag{10.1}$$

where the **moment function** $\boldsymbol{g}(\cdot)$ is a vector-valued function. (Sometimes the mean of $\boldsymbol{g}$ is called the moment function, but see top page 2116 of Newey and McFadden (1994).)

[1] http://web.archive.org/web/20230201194603/https://uh.edu/~bsorense/EconometricsII_GMM_2016.pdf

Assume that $\boldsymbol{\theta}$ is point identified by these moment conditions, meaning the true value is the only value in Θ that satisfies (10.1) by setting all moment conditions equal to zero.

Discussion Question 10.1 (GMM notation for IV)**.** Consider the linear IV regression model from Theorem 8.4, in particular the moment conditions. Re-write it in the notation of (10.1). That is, in (10.1), what are $\boldsymbol{g}(\cdot)$, $\boldsymbol{D}$, and $\boldsymbol{\theta}$?

10.2 Simple Examples

Consider the following contrived but insightful example. Continuing the notation from Section 10.1, let $\boldsymbol{D} = (X, Y)'$, two independent random variables (without the usual connotation of "Y" as a dependent variable or "X" as an independent variable). We assume that X and Y have the same mean, which is also our parameter of interest, scalar θ. The moment function is $\boldsymbol{g}(\boldsymbol{D}, t) = (X - t, Y - t)'$, where t is a generic possible value of the parameter (a "dummy variable" in the math sense but not econ sense). Thus, our moment conditions are

$$\mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \theta)] = \mathrm{E}\left[\begin{pmatrix} X - \theta \\ Y - \theta \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{10.2}$$

Why bother with Y when we could easily just estimate $\theta = \mathrm{E}(X)$? Generally: why bother with extra overidentifying restrictions when we could simply estimate the just-identified model? Indeed, it does not help with identification, nor with consistency; the goal is to improve estimation "efficiency." This is equivalent to improving the estimator's "precision," or decreasing the standard error, or decreasing the (asymptotic) variance.

Consider trying to set $\hat{\theta}$ to solve the sample analog of (10.2),

$$\mathbf{0} = \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \hat{\theta})] = \widehat{\mathrm{E}}\left[\begin{pmatrix} X - \hat{\theta} \\ Y - \hat{\theta} \end{pmatrix}\right] = \begin{bmatrix} \widehat{\mathrm{E}}(X) - \hat{\theta} \\ \widehat{\mathrm{E}}(Y) - \hat{\theta} \end{bmatrix}, \tag{10.3}$$

where $\widehat{\mathrm{E}}(X)$ and $\widehat{\mathrm{E}}(Y)$ are the respective sample means, also written $\widehat{\mathrm{E}}(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\widehat{\mathrm{E}}(Y) = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. In (10.3) we have two equations with one unknown, θ. Solving the top equation yields $\hat{\theta} = \widehat{\mathrm{E}}(X)$. But plugging this into the bottom equation yields $\widehat{\mathrm{E}}(Y) = \hat{\theta} = \widehat{\mathrm{E}}(X)$. Even if in the population $\mathrm{E}(Y) = \mathrm{E}(X)$, it is unlikely (zero probability if X and Y are continuous) that in the sample $\widehat{\mathrm{E}}(X) = \widehat{\mathrm{E}}(Y)$.

Instead of setting $\widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \hat{\theta})]$ equal to zero exactly, we can try to make it "close" to zero. A common measure of "close" is Euclidean distance; how close a vector $\boldsymbol{c} = (c_1, c_2, \ldots)'$ is to zero can be measured by the Euclidean norm, also called the L^2 norm, $\|\boldsymbol{c}\|_2 \equiv \sqrt{c_1^2 + c_2^2 + \cdots} = \sqrt{\boldsymbol{c}'\boldsymbol{c}}$. Because it is equivalent to minimize the square (because all values are non-negative), this proposal is to set

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{t}} \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})]' \, \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})]. \tag{10.4}$$

Discussion Question 10.2 (GMM for overidentified mean)**.** Continue the example with $\theta = \mathrm{E}(X) = \mathrm{E}(Y)$ and $X \perp\!\!\!\perp Y$, and assume iid sampling. Solve (10.4) for $\hat{\theta}$:

$$\hat{\theta} = \arg\min_t \widehat{\mathrm{E}}[(X - t, Y - t)]\, \widehat{\mathrm{E}}[(X - t, Y - t)'] = \arg\min_t (\bar{X} - t, \bar{Y} - t)(\bar{X} - t, \bar{Y} - t)'$$
$$= \ldots.$$

(Hint: the SOC holds, so just solve the FOC.)

a) What is $\hat{\theta}$?
b) Sanity check: does this seem reasonable?
c) Show that $\bar{X}$, $\bar{Y}$, and $\hat{\theta}$ are all unbiased.
d) What is the variance of $\hat{\theta}$ compared to the variances of $\bar{X}$ and $\bar{Y}$?
e) Among $\bar{X}$, $\bar{Y}$, and $\hat{\theta}$, which estimator has the best MSE? (Recall Section 3.10.2.)

More generally, we can add a weight matrix in the middle of (10.4), which gives the **GMM criterion function** (also called the GMM objective function)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{t}} \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})]' \underline{\hat{\boldsymbol{W}}}\, \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})], \tag{10.5}$$

where the "hat" on $\underline{\hat{\boldsymbol{W}}}$ indicates it can (optionally) be computed using the data. This $\underline{\hat{\boldsymbol{W}}}$ is assumed symmetric and positive definite; it can be relaxed to positive semidefinite under certain conditions (and extra complication), but the intuition is the same.

Discussion Question 10.3 (weights with exact identification)**.** Consider an exactly-identified model. To be concrete (and more familiar), consider the linear IV model with one endogenous regressor and one excluded IV, so $\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t}) = \boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{t})$.

a) In principle, what's the smallest possible numerical value of the quadratic form on the RHS of (10.5)?
b) Is there any $\hat{\boldsymbol{\theta}}$ that can achieve that value? (If so, how?)
c) How does the weight matrix change your answers?

Discussion Question 10.4 (weights with overidentified mean)**.** Continue from DQ 10.2. Further assume $\mathrm{Var}(X) = 1$ and $\mathrm{Var}(Y) = 4$, and assume both X and Y are normal, still with mean $\mathrm{E}(X) = \mathrm{E}(Y) = \theta$. Let $n = 1$.

a) Explain why the sampling distributions of the separate mean estimators are $\bar{X} \sim \mathrm{N}(\theta, 1)$ and $\bar{Z} \sim \mathrm{N}(\theta, 4)$.
b) Explain intuitively whether you would prefer $\bar{X}$ or $\bar{Y}$ (if you had to pick only one or the other), and whether this agrees with the mean squared error (MSE) criterion in this case.
c) Intuitively, if we take a weighted average $\hat{\theta} = (1 - w)\bar{X} + w\bar{Y}$, should we want $w > 0.5$, $w = 0.5$, or $w < 0.5$? Why?
d) Mathematically, given $\hat{\theta} = (1 - w)\bar{X} + w\bar{Y}$, show $\mathrm{Var}(\hat{\theta}) = 5w^2 - 2w + 1$ and solve for the w that minimizes the variance.

Discussion Question 10.5 (GMM weights with overidentified mean)**.** Continue the setup of DQ 10.4. Consider the GMM estimator defined in (10.5) with

$$\hat{\underline{\boldsymbol{W}}} = \begin{pmatrix} 1-w & 0 \\ 0 & w \end{pmatrix}.$$

a) Show that the GMM estimator simplifies to $\hat{\theta} = \arg\min_t (1-w)(\bar{X}-t)^2 + w(\bar{Y}-t)^2$, and that solving the FOC yields $\hat{\theta} = (1-w)\bar{X} + w\bar{Y}$.
b) What's the "optimal" weighting matrix $\hat{\underline{\boldsymbol{W}}}$ (with this diagonal form) that minimizes the MSE of $\hat{\theta}$? (Use DQ 10.4.)

The weight matrix allows the GMM estimator to improve efficiency, at least asymptotically (but usually in practice, too). Note that the optimal weighting matrix in Discussion Question 10.5 depends on unknown population values, specifically $\text{Var}(X)$ and $\text{Var}(Y)$, but those values can be estimated consistently. That is, letting $\underline{\boldsymbol{W}}$ denote the optimal weight matrix (with the true population values),

$$\hat{\underline{\boldsymbol{W}}} \xrightarrow{p} \underline{\boldsymbol{W}}. \tag{10.6}$$

It turns out that the estimation error in $\hat{\underline{\boldsymbol{W}}}$ does not appear in the (first-order) asymptotic normal distribution of the GMM estimator, at least in the most common cases, so it is generally better to try to use the optimal weight matrix.

Often the estimated weights require an estimate of $\boldsymbol{\theta}$ itself (the parameter of interest). This sounds circular. However, recall that all this weighting is only to improve efficiency, not to achieve consistency. That is, we could simply use the identity matrix as $\hat{\underline{\boldsymbol{W}}}$ to get an initial consistent estimator $\check{\boldsymbol{\theta}}$, then use $\check{\boldsymbol{\theta}}$ to compute an efficient $\hat{\underline{\boldsymbol{W}}}$, and use that $\hat{\underline{\boldsymbol{W}}}$ to compute our "real" estimator $\hat{\boldsymbol{\theta}}$. This is known as the **two-step GMM estimator**. The Stata command `ivreg2` has a `gmm2s` option to automatically compute the two-step GMM estimator.

This begs the question: why not three-step? Or four-step? Indeed, you could keep iterating to compute an **iterative GMM** estimator, but it does not affect the first-order asymptotic distribution and does not seem to make much improvement in practice, either. There is also a **continuously updated estimator** (CUE) that solves for $\hat{\boldsymbol{\theta}}$ accounting for the dependence of $\hat{\underline{\boldsymbol{W}}}$ on $\hat{\boldsymbol{\theta}}$. This is more difficult to solve, and it does not improve the asymptotic distribution, but there is some evidence that it improves finite-sample properties in some settings. That said, two-step GMM is a practical default choice.

10.3 2SLS as GMM

This section shows how 2SLS is a GMM estimator with a particular weight matrix that is efficient under homoskedastic structural errors. Because the 2SLS moment function is linear in the parameter vector, we can explicitly solve for the parameter, which simplifies the asymptotic theory (basically like OLS). Nonlinear models are included in the general treatment in Section 10.4.

From Assumption A8.5, the full instrument vector $\boldsymbol{Z}$ is assumed to satisfy the moment conditions $\mathrm{E}(\boldsymbol{Z}U) = \boldsymbol{0}$, where $U = Y - \boldsymbol{X}'\boldsymbol{\beta}$ is the structural error (where $\boldsymbol{\beta}$ is the true value). To put this into GMM notation, let

$$\boldsymbol{D} \equiv (Y, \boldsymbol{X}', \boldsymbol{Z}')', \quad \boldsymbol{g}(\boldsymbol{D}, \boldsymbol{b}) = \boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{b}), \tag{10.7}$$

where $\boldsymbol{b}$ is a generic possible value of the parameter vector whose true population value is $\boldsymbol{\beta}$. Let $\underline{\boldsymbol{Z}}$ be the matrix with n rows whose row i equals $\boldsymbol{Z}_i'$, and similarly let $\underline{\boldsymbol{X}}$ be the matrix with n rows whose row i equals $\boldsymbol{X}_i'$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. As in (10.5),

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{b}} \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{b})]' \hat{\underline{\boldsymbol{W}}}\, \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{b})] \\
&= \arg\min_{\boldsymbol{b}} \widehat{\mathrm{E}}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{b})]' \hat{\underline{\boldsymbol{W}}}\, \widehat{\mathrm{E}}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{b})] \\
&= \arg\min_{\boldsymbol{b}} \widehat{\mathrm{E}}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{b})]' \hat{\underline{\boldsymbol{W}}}\, \widehat{\mathrm{E}}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{b})] \\
&= \arg\min_{\boldsymbol{b}} [\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\boldsymbol{b})/n]' \hat{\underline{\boldsymbol{W}}} [\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\boldsymbol{b})/n].
\end{aligned} \tag{10.8}$$

The $1/n$ can be removed without changing the minimizer. The second-order condition is satisfied, so the minimizer $\hat{\boldsymbol{\beta}}$ solves the first-order condition. Using

$$\frac{\partial}{\partial \boldsymbol{b}'} [\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\boldsymbol{b})]' = -\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}} \tag{10.9}$$

and applying the generic vector calculus derivative $\frac{\partial \boldsymbol{x}'\underline{\boldsymbol{a}}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\underline{\boldsymbol{a}} + \underline{\boldsymbol{a}}')\boldsymbol{x}$, along with the assumed symmetry of $\hat{\underline{\boldsymbol{W}}}$ such that $\hat{\underline{\boldsymbol{W}}} + \hat{\underline{\boldsymbol{W}}}' = 2\hat{\underline{\boldsymbol{W}}}$, and applying the chain rule, the derivative of the GMM criterion function is

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{b}'} \{[\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\boldsymbol{b})]' \hat{\underline{\boldsymbol{W}}} [\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\boldsymbol{b})]\} \\
&= [\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\boldsymbol{b})]' (\hat{\underline{\boldsymbol{W}}} + \hat{\underline{\boldsymbol{W}}}')(-\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}) \\
&= -2[\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\boldsymbol{b})]' \hat{\underline{\boldsymbol{W}}}\, \underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}.
\end{aligned} \tag{10.10}$$

Setting the transpose of (10.10) to zero and solving,

$$\begin{aligned}
\boldsymbol{0} = \{-2[\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})]' \hat{\underline{\boldsymbol{W}}}\, \underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}\}' &= -2\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}[\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})], \\
\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}\underline{\boldsymbol{Z}}'\boldsymbol{Y} &= \underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}, \\
\hat{\boldsymbol{\beta}} &= (\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}\underline{\boldsymbol{Z}}'\boldsymbol{Y}.
\end{aligned} \tag{10.11}$$

A special case of (10.11) with $\hat{\underline{\boldsymbol{W}}} = (\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}/n)^{-1}$ is… 2SLS! That is, the formula reduces to (8.24).

Because we have a closed-form expression for $\hat{\boldsymbol{\beta}}$ in terms of sample moments, the asymptotic theory follows the same type of arguments as for OLS, nearly identical to the derivations in Section 8.3.3 but with general weight matrix $\hat{\underline{\boldsymbol{W}}}$ instead of $(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}/n)^{-1}$.

That is, we can insert $1/n$ in the right places and plug in $\boldsymbol{Y} = \underline{\boldsymbol{X}}\boldsymbol{\beta}+\boldsymbol{U}$ from the structural model to get

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}\underline{\boldsymbol{Z}}'\boldsymbol{U} \tag{10.12}$$

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= (n^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}n^{-1}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}})^{-1}n^{-1}\underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{W}}}n^{-1/2}\underline{\boldsymbol{Z}}'\boldsymbol{U} \\ &\xrightarrow{d} \{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{\Sigma}}),\end{aligned} \tag{10.13}$$

using the notation from (8.27). That is, altogether $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a mean-zero normal distribution with covariance matrix

$$\underline{\boldsymbol{\Omega}} \equiv \{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{\Sigma}}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}. \tag{10.14}$$

The special case of 2SLS in (8.28) is the same but with $\underline{\boldsymbol{Q}}_{ZZ}^{-1}$ instead of $\underline{\boldsymbol{W}}$.

If $\underline{\boldsymbol{W}} = \underline{\boldsymbol{\Sigma}}^{-1}$, then the covariance "sandwich" collapses:

$$\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{\Sigma}}^{-1}\overbrace{\underline{\boldsymbol{\Sigma}}\underline{\boldsymbol{\Sigma}}^{-1}}^{\text{cancels}}\underline{\boldsymbol{Q}}'_{XZ}\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1} = \{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}. \tag{10.15}$$

It can be shown that (in general) such "collapsed" covariance matrices are "smaller" than the corresponding sandwich form, in the sense that the sandwich matrix minus the collapsed matrix is positive semidefinite; for example, see the claim on the top of page 218 of Wooldridge (2010). That is, the collapsed version corresponds to a more efficient (better) estimator.

To achieve $\underline{\boldsymbol{W}} = \underline{\boldsymbol{\Sigma}}^{-1}$, in practice we use $\hat{\underline{\boldsymbol{W}}} = \hat{\underline{\boldsymbol{\Sigma}}}^{-1}$. Recall $\underline{\boldsymbol{\Sigma}} = \mathrm{E}[U^2\boldsymbol{Z}\boldsymbol{Z}']$. If we have any consistent estimator (even if not efficient) $\check{\boldsymbol{\beta}}$, then we can compute residuals $\hat{U}_i = Y_i - \boldsymbol{X}_i'\check{\boldsymbol{\beta}}$, and with iid sampling use

$$\hat{\underline{\boldsymbol{\Sigma}}} = \frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2\boldsymbol{Z}_i\boldsymbol{Z}_i' \xrightarrow{p} \underline{\boldsymbol{\Sigma}}. \tag{10.16}$$

With other types of sampling, we would need variations on this estimator that appropriately account for dependence in order to achieve consistency. There are usually such options in Stata (or R, etc.), where your job is to choose the most appropriate type of sampling given your empirical setting, and then Stata will use the appropriate formula.

Finally, note that with "homoskedasticity" in the sense of

$$\mathrm{Var}(U \mid \boldsymbol{Z}) = \mathrm{Var}(U), \tag{10.17}$$

the original 2SLS estimator is efficient. Recalling also that $\mathrm{E}(U) = 0$, $\mathrm{Var}(U) = \mathrm{E}(U^2)$, so homoskedasticity can also be written as $\mathrm{E}(U^2 \mid \boldsymbol{Z}) = \mathrm{E}(U^2)$. Using this,

$$\mathrm{E}[U^2\boldsymbol{Z}\boldsymbol{Z}'] = \mathrm{E}[\mathrm{E}(U^2\boldsymbol{Z}\boldsymbol{Z}' \mid \boldsymbol{Z})] = \mathrm{E}[\overbrace{\mathrm{E}(U^2 \mid \boldsymbol{Z})}^{=\mathrm{E}(U^2)}\boldsymbol{Z}\boldsymbol{Z}'] = \mathrm{E}(U^2)\,\mathrm{E}[\boldsymbol{Z}\boldsymbol{Z}'] \equiv \sigma_U^2\underline{\boldsymbol{Q}}_{ZZ}. \tag{10.18}$$

Thus, noting σ_U^2 is a scalar that can move freely around,

$$\begin{aligned}\underline{\boldsymbol{\Omega}} &= \{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\overbrace{\sigma_U^2\underline{\boldsymbol{Q}}_{ZZ}}^{\underline{\boldsymbol{\Sigma}}}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1} \\ &= \sigma_U^2\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}_{ZZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\{\underline{\boldsymbol{Q}}_{XZ}\underline{\boldsymbol{W}}\underline{\boldsymbol{Q}}'_{XZ}\}^{-1}, \qquad (10.19)\end{aligned}$$

so simply setting $\underline{\boldsymbol{W}} = \underline{\boldsymbol{Q}}_{ZZ}^{-1}$ collapses the sandwich, implying $\hat{\underline{\boldsymbol{W}}} = (\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}/n)^{-1}$, which (again) makes (10.11) simplify to the 2SLS estimator in (8.24).

Conversely, with heteroskedasticity, 2SLS is not efficient, so we may improve asymptotic efficiency by using two-step GMM. In Stata, the `gmm2s` option does just that. (The `center` option also seems useful to use in that case.)

10.4 General Estimator

Unlike in Section 10.3, where the parameter vector enters the moment function linearly and allows a closed-form expression of the estimator, if the parameter vector does not enter linearly, then there is generally not a closed-form expression for the GMM estimator. That is, we cannot write the estimator as a function of various sample moments, but only as the solution to a minimization problem. This makes the asymptotic theory much different, so it is worth describing, although this is not a central focus of this class.

Additionally, in Section 10.4.2 more details are given about testing overidentifying restrictions.

10.4.1 Asymptotic Theory

The general GMM estimator was defined in (10.5) as the minimizer of a quadratic form of the sample moments with weight matrix $\hat{\underline{\boldsymbol{W}}}$. We can think of the quadratic form as a function of generic vector $\boldsymbol{t}$:

$$\hat{Q}(\boldsymbol{t}) \equiv \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D},\boldsymbol{t})]'\hat{\underline{\boldsymbol{W}}}\,\widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D},\boldsymbol{t})], \qquad (10.20)$$

so the GMM estimator is $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{t}} \hat{Q}(\boldsymbol{t})$.

Without a closed-form expression for $\hat{\boldsymbol{\theta}}$, in order to learn about the asymptotic properties of $\hat{\boldsymbol{\theta}}$ we must learn about the asymptotic properties of the function $\hat{Q}(\cdot)$. This is more challenging because $\hat{Q}(\cdot)$ is a function, rather than a vector like the OLS or 2SLS estimators.

The general idea is to show that the GMM criterion function converges to the corresponding population criterion whose unique solution is the true parameter value, in order to show consistency; and then given consistency, an expansion around the true value provides a linear approximation that facilitates the asymptotic normal distribution. Such results are more readily obtained if the moment function is "smooth," for example if $\boldsymbol{g}(\boldsymbol{D}_i,\boldsymbol{t})$ is continuous in $\boldsymbol{t}$ for any $\boldsymbol{D}_i$ and further is continuously differentiable in a (small) neighborhood around the true value $\boldsymbol{\theta}$. However, such smoothness assumptions

can be relaxed (with extra work in the proofs). Similarly, iid sampling is a sufficient condition that simplifies proofs, but it is not necessary for either consistency or asymptotic normality.

The other general point to notice is that the GMM estimator's asymptotic covariance matrix depends on the (plim of the) weighting matrix used and in general has a "sandwich" form. Similar to the introduction of the two-step GMM estimator in Section 10.2, in this general case we can follow a two-step approach in which we first get any consistent estimator of $\boldsymbol{\theta}$ and use it to estimate the efficient weighting matrix that causes the sandwich to "collapse." As before, this "efficiency" is only within the scope of choosing different weighting matrices; it assumes we are stuck with using whichever moment conditions we have. More generally, it may be possible to improve efficiency further by using different moment conditions themselves, but that is beyond our scope. For example, see Section 14.4.3 of Wooldridge (2010) regarding optimal instruments (moment conditions) when we assume the stronger conditional-mean form of exogeneity, $\mathrm{E}(U \mid \boldsymbol{Z}) = 0$ instead of the weaker $\mathrm{E}(\boldsymbol{Z}U) = \mathbf{0}$.

Some technical details are in the appendix, if you are curious (not required).

10.4.2 Testing Overidentifying Restrictions

This section generalizes the idea introduced in Section 9.3, of using the "extra" information from additional moment conditions (if there is overidentification) to test the overall model specification. The intuition and interpretation is the same here, so only some mathematical details are provided.

Consider iid sampling, to allow the following simplifications, but remembering that iid sampling is not necessary to test overidentifying restrictions. Given iid sampling, and recalling $\mathrm{E}[\boldsymbol{g}(\boldsymbol{D}_i, \boldsymbol{\theta})] = \mathbf{0}$, the usual CLT applies to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_i, \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Omega}}), \quad \underline{\boldsymbol{\Omega}} \equiv \mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{\theta}) \boldsymbol{g}(\boldsymbol{D}, \boldsymbol{\theta})'], \tag{10.21}$$

like in (10.31). As in (10.32), a consistent estimator of $\underline{\boldsymbol{\Omega}}$ is

$$\hat{\underline{\boldsymbol{\Omega}}} = \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \hat{\boldsymbol{\theta}}) \boldsymbol{g}(\boldsymbol{D}, \hat{\boldsymbol{\theta}})'] = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{D}_i, \hat{\boldsymbol{\theta}}) \boldsymbol{g}(\boldsymbol{D}_i, \hat{\boldsymbol{\theta}})'.$$

Let $\hat{\boldsymbol{\theta}}$ be the two-step GMM estimator, and as above $\hat{\underline{\boldsymbol{\Omega}}} \xrightarrow{p} \underline{\boldsymbol{\Omega}}$. Then, under the null hypothesis $\mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{\theta})] = \mathbf{0}$, the test statistic

$$\hat{J} \equiv n \, \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \hat{\boldsymbol{\theta}})]' \hat{\underline{\boldsymbol{\Omega}}}^{-1} \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \hat{\boldsymbol{\theta}})] \xrightarrow{d} \chi^2_{m-k}, \tag{10.22}$$

a chi-squared distribution with degrees of freedom $m - k$, where $m - k$ is the degree of overidentification (m moment conditions, k parameters). Note when there is exact identification with $m = k$, then $m = k = 0$, and the χ^2_0 is a degenerate distribution with

all probability at value zero, i.e., $\mathrm{P}(\hat{J} = 0) = 1$. That is, the estimator $\hat{\boldsymbol{\theta}}$ will perfectly set all sample moments equal to zero, so $\hat{J} = 0$ and we cannot learn anything about possible misspecification. Only if $m > k$ (overidentification) can we learn something about misspecification here.

See also Section 9.5 ("Tests for overidentifying restrictions") of Newey and McFadden (1994).

Appendix to Chapter 10

10.A Technical Details: GMM Consistency

This section shows technical details for deriving consistency of the GMM estimator, continuing from Section 10.4.1.

The population criterion function corresponding to (10.20) is

$$Q(\boldsymbol{t}) \equiv \mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})]' \underline{\boldsymbol{W}} \, \mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})], \tag{10.23}$$

simply replacing sample expectations with population expectations, and replacing the sample weight matrix with its probability limit. The identification assumption is that only the true value $\boldsymbol{\theta}$ sets the moment conditions all equal to zero; thus if $\underline{\boldsymbol{W}}$ is positive definite, $Q(\boldsymbol{t}) = 0$ iff $\boldsymbol{t} = \boldsymbol{\theta}$. (Again, this can be weakened to positive semidefinite if $\underline{\boldsymbol{W}} \, \mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})] \neq \boldsymbol{0}$ for all $\boldsymbol{t} \neq \boldsymbol{\theta}$, but this point is usually not helpful in practice.) In nonlinear models, it can be difficult to provide conditions for such (global) identification. That is, it's possible to show that the true $\boldsymbol{\theta}$ satisfies all the moment conditions, but it's difficult to show that no other possible values also solve the moment conditions. GMM identification is further discussed in Section 2.2.3 of Newey and McFadden (1994), also a practical summary would be their statement, "A practical 'solution' to the problem of global GMM identification. . . is to simply assume identification" (p. 2127).

The function $\hat{Q}(\cdot)$ must converge uniformly in probability to the population $Q(\cdot)$, meaning

$$\sup_{\boldsymbol{t} \in \Theta} |\hat{Q}(\boldsymbol{t}) - Q(\boldsymbol{t})| = o_p(1). \tag{10.24}$$

This type of result is called a **uniform (weak) law of large numbers** (ULLN). While iid sampling is a sufficient condition that makes it easier to establish a ULLN, as in Lemma 2.4 of Newey and McFadden (1994, p. 2129), ULLNs can also hold with dependent data (under certain restrictions, of course). Results like Theorem 5.7 of van der Vaart (1998) prove that this uniform convergence in probability of the sample criterion function to the population criterion function is sufficient for the sample minimizer to converge to the population minimizer, also assuming the population function cannot get arbitrarily close to zero (except in a neighborhood of the true $\boldsymbol{\theta}$). Theorem 2.1 of Newey and McFadden

(1994) is also a (slightly less) general consistency result based on uniform convergence, which is condition (iv) of their theorem.

Newey and McFadden (1994) provide lower-level conditions that essentially imply (10.24), which in turn implies consistency. The following is a slightly simplified version of Theorem 2.6 of Newey and McFadden (1994). Another slight variant is Theorem 14.1 of Wooldridge (2010).

Theorem 10.1 (GMM consistency)**.** *If i) data $\boldsymbol{D}_i$ are sampled iid; ii) $\hat{\underline{\boldsymbol{W}}} \xrightarrow{p} \underline{\boldsymbol{W}}$ as in (10.6), where both matrices are symmetric and positive definite; iii) the moment conditions are uniquely solved by the true $\boldsymbol{\theta}$ that satisfies $\mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{\theta})] = \mathbf{0}$; iv) the parameter space Θ is a compact set; v) the moment function $\boldsymbol{g}(\boldsymbol{D}, \cdot)$ is continuous given any $\boldsymbol{D}$; vi) the elements of $\mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})]$ are all finite for every $\boldsymbol{t} \in \Theta$; then the GMM estimator in (10.5) is consistent: $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$.*

Proof. See page 2132 of Newey and McFadden (1994). □

Example 10.1 (2SLS consistency)**.** Consider the conditions of Theorem 10.1 for the 2SLS estimator. Condition (i) is iid sampling, unrelated to the particular estimator; again, iid is sufficient but not necessary here. For 2SLS, $\hat{\underline{\boldsymbol{W}}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i' \xrightarrow{p} \mathrm{E}[\boldsymbol{Z}\boldsymbol{Z}'] = \underline{\boldsymbol{W}}$ given the assumed iid sampling, and also assuming $\boldsymbol{Z}$ has a finite second moment. Condition (iii) is the identification assumption that only $\boldsymbol{\theta}$ solves all the moment conditions, which is true given instrument exogeneity and relevance (rank condition). Condition (iv) requires us to limit the possible parameter values to a compact (finite) set, not allowing any value in $\mathbb{R}^k$ (again a sufficient but not necessary condition); this is often reasonable because it does not require Θ to be small, just bounded. For example, if θ_2 is the return to education, then we should feel comfortable with $-999 \leq \theta_2 \leq 999$. Condition (v) requires that $\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{\theta})$ is continuous, which clearly it is. Condition (vi) requires finite $\mathrm{E}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{\theta})]$, which holds if $\boldsymbol{Z}$, Y, and $\boldsymbol{X}$ all have finite second moments.

Discussion Question 10.6 (IVQR GMM)**.** Consider the IV quantile regression model based on moment conditions $\mathrm{E}[\boldsymbol{Z}(\mathbb{1}\{Y \leq \boldsymbol{X}'\boldsymbol{\theta}\} - \tau)] = \mathbf{0}$.

a) Write the moment function $\boldsymbol{g}(\cdot, \cdot)$.

b) Explain why the moment function is not continuous in $\boldsymbol{\theta}$, for any values of $\boldsymbol{Z}$, Y, and $\boldsymbol{X}$.

c) Does this violation imply that the corresponding GMM estimator is not consistent? Why/not? (Hint: recall Chapter 2.)

Beyond our scope...

There are ways to prove consistency and asymptotic normality of quantile estimators, including IV quantile regression, whose $\boldsymbol{g}(\cdot)$ includes an indicator function and thus violates the usual smoothness assumptions. One approach is to replace the indicator function with a sequence of smooth functions that approaches the indicator

function asymptotically. However, this takes some "manual" labor (not just invoking an existing theorem); see de Castro, Galvao, Kaplan, and Liu (2019).

10.B Technical Details: GMM Asymptotic Normality

This section shows technical details for deriving asymptotic normality of the GMM estimator, given that consistency has already been established.

The derivation of the asymptotic distribution uses the fact that (given consistency) the estimator is asymptotically within a small neighborhood of the true value with probability approaching one. Because of this, the asymptotic behavior of $\hat{\boldsymbol{\theta}}$ (the asymptotic sampling distribution) depends only on the behavior of the criterion function "locally" (near the true value). This makes the theory easier: we do not need the asymptotic distribution of the entire sample criterion function, only the function evaluated at the true $\boldsymbol{\theta}$, for which often a standard CLT applies.

To develop intuition, first consider the exactly identified model for which the (G)MM estimator solves $\widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \hat{\boldsymbol{\theta}})] = \mathbf{0}$. Define

$$\hat{\boldsymbol{M}}(\boldsymbol{t}) \equiv \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})], \quad \boldsymbol{M}(\boldsymbol{t}) \equiv \mathrm{E}[\boldsymbol{g}(\boldsymbol{D}, \boldsymbol{t})], \quad \nabla\hat{\boldsymbol{M}}(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \boldsymbol{t}'}\hat{\boldsymbol{M}}(\boldsymbol{t})\bigg|_{\boldsymbol{t}=\boldsymbol{\theta}}, \tag{10.25}$$

where each row of $\nabla\hat{\boldsymbol{M}}(\boldsymbol{\theta})$ refers to a different element of the column vector $\hat{\boldsymbol{M}}(\boldsymbol{\theta})$ and each column of $\nabla\hat{\boldsymbol{M}}(\boldsymbol{\theta})$ refers to a different element of $\boldsymbol{\theta}$. Consider the mean value expansion

$$\mathbf{0} = \hat{\boldsymbol{M}}(\boldsymbol{\theta}) + (\dot{\underline{\boldsymbol{M}}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \tag{10.26}$$

where matrix $\dot{\underline{\boldsymbol{M}}}$ contains the derivatives evaluated at the "mean values" $\tilde{\boldsymbol{\theta}}_{(1)}, \tilde{\boldsymbol{\theta}}_{(2)}, \ldots$ that are all on the line segment between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$,

$$\dot{\underline{\boldsymbol{M}}} \equiv \begin{bmatrix} \nabla\hat{M}_1(\tilde{\boldsymbol{\theta}}_{(1)}) \\ \nabla\hat{M}_2(\tilde{\boldsymbol{\theta}}_{(2)}) \\ \vdots \end{bmatrix}, \tag{10.27}$$

where $\hat{M}_j$ refers to element j in the vector. Because $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, the mean values are also converging in probability to the true $\boldsymbol{\theta}$, so (given enough "smoothness")

$$\dot{\underline{\boldsymbol{M}}} \xrightarrow{p} \nabla\boldsymbol{M}(\boldsymbol{\theta}). \tag{10.28}$$

Rearranging (10.26) and solving for the centered and scaled estimator,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -[\nabla\boldsymbol{M}(\boldsymbol{\theta})]^{-1}\sqrt{n}\hat{\boldsymbol{M}}(\boldsymbol{\theta}). \tag{10.29}$$

Recall the last term (including the $\sqrt{n}$) can be written

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{D}_i,\boldsymbol{\theta}), \tag{10.30}$$

to which a CLT applies under the usual (relatively weak) sampling dependence and finite-moment conditions, because $\mathrm{E}[\boldsymbol{g}(\boldsymbol{D}_i,\boldsymbol{\theta})]=\mathbf{0}$. With iid sampling, more specifically

$$\sqrt{n}\hat{\boldsymbol{M}}(\boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0},\underline{\boldsymbol{\Omega}}), \quad \underline{\boldsymbol{\Omega}} \equiv \mathrm{E}[\boldsymbol{g}(\boldsymbol{D},\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{D},\boldsymbol{\theta})']. \tag{10.31}$$

If we have any consistent estimator $\hat{\boldsymbol{\theta}}$, then we can estimate $\underline{\boldsymbol{\Omega}}$ by

$$\hat{\underline{\boldsymbol{\Omega}}} = \widehat{\mathrm{E}}[\boldsymbol{g}(\boldsymbol{D},\hat{\boldsymbol{\theta}})\boldsymbol{g}(\boldsymbol{D},\hat{\boldsymbol{\theta}})'] = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{D}_i,\hat{\boldsymbol{\theta}})\boldsymbol{g}(\boldsymbol{D}_i,\hat{\boldsymbol{\theta}})'. \tag{10.32}$$

This is useful for computing both standard errors and the two-step GMM estimator (see below). If sampling is not iid, then this particular formula is not correct, but in Stata you can simply tell it the appropriate type of sampling and it has the proper formulas implemented.

Combining (10.26) and (10.28), recalling that notation $X_n \xrightarrow{p} c$ is equivalent to $X_n = c + o_p(1)$, we can also write

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) &= -[\dot{\underline{\boldsymbol{M}}}]^{-1}\sqrt{n}\hat{\boldsymbol{M}}(\boldsymbol{\theta}) = -[\nabla\boldsymbol{M}(\boldsymbol{\theta})+o_p(1)]^{-1}\overbrace{\sqrt{n}\hat{\boldsymbol{M}}(\boldsymbol{\theta})}^{=O_p(1)} \\ &= -[\nabla\boldsymbol{M}(\boldsymbol{\theta})]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{D}_i,\boldsymbol{\theta})+o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{D}_i)+o_p(1)\end{aligned} \tag{10.33}$$

given $\boldsymbol{\psi}(\boldsymbol{D}_i) \equiv -[\nabla\boldsymbol{M}(\boldsymbol{\theta})]^{-1}\boldsymbol{g}(\boldsymbol{D}_i,\boldsymbol{\theta})$. An estimator that (when centered and scaled) can be written with this structure is called **asymptotically linear**, and this form is also called the **influence function representation**, where here $\boldsymbol{\psi}(\cdot)$ is the influence function. For more, see for example page 2142 and (3.3) of Newey and McFadden (1994).

Newey and McFadden (1994) provide a general GMM asymptotic normality result in their Theorem 3.2 (p. 2145). The idea is similar to above but with a mean value expansion of the GMM first-order condition, which is complicated by more terms but retains the same intuition. The following is a slightly simplified version. Another slight variation is Theorem 14.2 on page 527 of Wooldridge (2010).

Theorem 10.2. *If i) the GMM estimator in (10.5) is consistent, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$; ii) $\hat{\underline{\boldsymbol{W}}} \xrightarrow{p} \underline{\boldsymbol{W}}$ as in (10.6), where both matrices are symmetric and positive definite; iii) the true parameter value $\boldsymbol{\theta}$ is in the interior of parameter space Θ; iv) the sample function $\hat{\boldsymbol{M}}(\cdot)$ in (10.25) is continuously differentiable in a (small) neighborhood of the true θ; v) a CLT holds: $\sqrt{n}\hat{\boldsymbol{M}}(\boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0},\underline{\boldsymbol{\Omega}})$; vi) the sample Jacobian matrix converges in probability to the*

population Jacobian, in that $\nabla \hat{\boldsymbol{M}}(\boldsymbol{t}) \xrightarrow{p} \nabla \boldsymbol{M}(\boldsymbol{t})$ *uniformly over a neighborhood of* $\boldsymbol{\theta}$*, where the limiting function is continuous in* $\boldsymbol{t}$*; vii) defining* $\underline{\boldsymbol{G}} \equiv \nabla \boldsymbol{M}(\boldsymbol{\theta})$*, the matrix* $\underline{\boldsymbol{G}}'\underline{\boldsymbol{W}}\underline{\boldsymbol{G}}$ *is invertible; then the GMM estimator in* (10.5) *is asymptotically normal*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Sigma}}), \quad \underline{\boldsymbol{\Sigma}} \equiv (\underline{\boldsymbol{G}}'\underline{\boldsymbol{W}}\underline{\boldsymbol{G}})^{-1}\underline{\boldsymbol{G}}'\underline{\boldsymbol{W}}\underline{\boldsymbol{\Omega}}\underline{\boldsymbol{W}}\underline{\boldsymbol{G}}(\underline{\boldsymbol{G}}'\underline{\boldsymbol{W}}\underline{\boldsymbol{G}})^{-1}.$$

Proof. See page 2145 of Newey and McFadden (1994). □

The covariance matrix in Theorem 10.2 has the familiar sandwich form, and it "collapses" if $\underline{\boldsymbol{W}} = \underline{\boldsymbol{\Omega}}^{-1}$. That is, if we can consistently estimate the asymptotic covariance matrix from condition (v) by $\hat{\underline{\boldsymbol{\Omega}}} \xrightarrow{p} \underline{\boldsymbol{\Omega}}$, then we can use its inverse as our weighting matrix, $\hat{\underline{\boldsymbol{W}}} = [\hat{\underline{\boldsymbol{\Omega}}}]^{-1}$. This achieves **efficiency** among all possible weighting matrices, in the sense of minimizing the GMM estimator's asymptotic variance (in the matrix sense of $\underline{\boldsymbol{a}} \leq \underline{\boldsymbol{b}}$ meaning $\underline{\boldsymbol{a}} - \underline{\boldsymbol{b}}$ is negative semidefinite). This is exactly what the **two-step GMM estimator** does, as introduced in Section 10.2. That is, we can use any positive definite matrix (like the identity matrix) as $\hat{\underline{\boldsymbol{W}}}$ to get an initial consistent estimator $\check{\boldsymbol{\theta}}$, then use $\check{\boldsymbol{\theta}}$ to compute the efficient $\hat{\underline{\boldsymbol{W}}} = [\hat{\underline{\boldsymbol{\Omega}}}]^{-1}$, and use that $\hat{\underline{\boldsymbol{W}}}$ to compute our two-step GMM estimator $\hat{\boldsymbol{\theta}}$.

Exercises

In Stata, the `ivreg2` command (Baum, Schaffer, and Stillman, 2002), available on SSC, runs the IV regression estimator and helps automatically run weak identification and over-identification tests, with the help of `ranktest` (Kleibergen, Schaffer, and Windmeijer, 2007), as well as GMM estimators (and yet other estimators). If you have a single endogenous regressor, then you can also use `weakivtest` (Pflueger and Wang, 2013) for the weak identification testing, with the help of `avar` (Baum and Schaffer, 2013); `weakivtest` provides a somewhat different test statistic than `ivreg2` as well as different critical values (in terms of maximal relative bias instead of size distortion), so it is helpful to look at both sets of results. Finally, you can use `weakiv` (Finlay, Magnusson, and Schaffer, 2013) to compute weak-IV-robust AR confidence intervals. You can install all of these from SSC with:

```
ssc install ranktest
ssc install ivreg2
ssc install avar
ssc install weakivtest
ssc install weakiv
```

Exercise II.1. This exercise looks at the impact of participation in a 401(k) retirement plan (dummy variable `p401k`) on an individual's net total financial assets (`nettfa`), using 401(k) eligibility (dummy variable `e401k`) as an instrument.

a. For this example, describe an individual's potential outcomes.

b. For this example, describe who is a "complier" and who is a "never-taker."

c. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

d. Load the data (and look at variable labels to see descriptions and units of measure):
`bcuse 401ksubs , clear`

e. Regress net total financial assets on 401(k) participation
`reg nettfa p401k , vce(robust)`
and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction.

f. Regress net total financial assets on 401(k) eligibility:
`reg nettfa e401k , vce(robust)` and interpret the estimated coefficient on `e401k`

g. Explain what it would mean for 401(k) eligibility to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

h. Compute the IV estimator, CI, and corresponding tests:
`ivreg2 nettfa (p401k = e401k) , robust`

   i. Describe the IV estimand in this example.
   ii. Discuss the economic significance of the estimate.
   iii. Explain why you are or are not worried about weak instruments in this case; in addition to the `ivreg2` output, run `weakivtest` and refer to specific output (and how to interpret it).
   iv. Run `weakiv` and explain which confidence interval you think is more appropriate as well as what that CI tells us about our uncertainty about the true population value; be precise and explicit.
   v. Explain what the J-test results suggest about the model.

i. Run

```
ivreg2 nettfa (p401k=e401k) , robust gmm2s center
gmm (nettfa - p401k*{p401k} - {_cons}) , instruments(e401k) nolog
    vce(robust) twostep
```

and briefly compare with the estimate/CI from part (h).

j. Run `ivreg2 nettfa (p401k = e401k) inc age marr fsize , robust` and say briefly if the change (compared to above without control variables) in the estimated effect is economically significant, as well as if/how this changes our uncertainty about the true population value.

Exercise II.2. The following analyzes data originally from Graddy (1995). The goal is to estimate the demand curve for a particular type of fish (whiting) in a particular (large) fish market in New York City. Prices and quantities are in logs (so the slope is approximately an elasticity); specifically, the (average) daily price was measured in dollars per pound of fish, and the daily quantity in pounds sold, and the natural log was taken of each. The weather is used as the (hopefully) exogenous supply shifter: bad weather (specifically wind and waves) makes it more difficult to fish, which moves the supply curve inward. (Not needed for this exercise, but if you're curious, see Graddy's fish papers on her website,[2] like page 210 "How the Market Worked at Fulton Street" of her 2006 *JEP* paper.)

a. Load the data with (remove line break)

[2] https://www.kathryngraddy.org/research#pubfish

```
use https://raw.githubusercontent.com/kaplandm/stata/main/data/
    fishdata.dta , clear
```

b. Rename variables to be more intuitive[3]:
```
rename price lnp
rename qty lnq
```

c. Run `reg lnq lnp` and explain why this estimator of the demand curve is not consistent.

d. Explain what it would mean for the weather to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

e. Compute the IV estimator and corresponding tests:
```
ivreg2 lnq (lnp=stormy mixed) , robust
```

   i. Describe the IV estimand in this example.

   ii. Discuss the economic significance of the estimate.

   iii. Explain why you are or are not worried about weak instruments in this case; in addition to the `ivreg2` output, run `weakivtest` and refer to specific output (and how to interpret it).

   iv. Run `weakiv` and explain which confidence interval you think is more appropriate as well as what that CI tells us about our uncertainty about the true population value; be precise and explicit.

   v. Explain what the J-test results suggest about the model.

f. Run `ivreg2 lnq (lnp=windspd) , robust` and briefly compare with the previous IV results (slope estimate, weak IV test, J-test).

g. Run
```
ivreg2 lnq (lnp=stormy mixed) , robust gmm2s center
gmm (lnq - lnp*{lnp} - {_cons}) , instruments(stormy mixed) nolog
    vce(robust) twostep
```
and briefly compare with the slope estimate/CI from part (e).

h. What do you think about a model with a constant slope in this case? That is, a model where the shock/error shifts the demand curve up and down but does not change its slope?

Exercise II.3. The data are originally from Card (1995), with individual-level observations of (log) wages, years of education, and other variables. Note the dataset lacks variable labels, but they can be found online.[4] This is the same dataset as previously in Exercise I.4.

[3] Unfortunately, there are no variable labels, so there is no way to know these are in logs unless you look back at the original paper.

[4] http://fmwww.bc.edu/ec-p/data/wooldridge/card.des

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse card , clear`

c. Create a dummy to compare high-school (12 years education) and college (16 years education):
```
gen d_coll = .
replace d_coll=0 if educ==12
replace d_coll=1 if educ==16
```

d. Regress log wage on years of education `reg lwage educ , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction.

e. Explain what it would mean for `nearc4` to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

f. Run `ivreg2 lwage (educ = nearc4 nearc2 ) , robust`

   i. Describe the IV estimand in this example.
   ii. Discuss economic significance of the estimated slope ("returns to education").
   iii. Explain why you are or are not worried about weak instruments in this case; in addition to the `ivreg2` output, run `weakivtest` and refer to specific output (and how to interpret it).
   iv. Run `weakiv` and explain which confidence interval you think is more appropriate as well as what that CI tells us about our uncertainty about the true population value; be precise and explicit.
   v. Explain what the J-test results suggest about the model.

g. Run `ivreg2 lwage (educ = nearc4 ) , robust` and briefly compare with the previous IV results (slope estimate, weak IV test, J-test).

h. Run
```
ivreg2 lwage (educ = nearc4) , gmm2s center robust
ivreg2 lwage (educ = nearc4 nearc2) , gmm2s center robust
```
and comment on differences among these and previous estimates of the return to education.

i. Run `gmm (lwage - educ*{educ} - {_cons}) , instruments(nearc4 nearc2) nolog vce(robust) twostep` and compare with the corresponding estimate/CI from part (h).

j. Run `ivreg2 lwage (d_coll = nearc4 nearc2 ) , robust`

   i. Describe the IV estimand in this example.
   ii. Discuss economic significance of the estimated coefficient on `d_coll`.
   iii. Explain why you are or are not worried about weak instruments in this case; in addition to the `ivreg2` output, run `weakivtest` and refer to specific output (and how to interpret it).

iv. Run `weakiv` and explain which confidence interval you think is more appropriate as well as what that CI tells us about our uncertainty about the true population value; be precise and explicit.

v. Explain what the J-test results suggest about the model.

Exercise II.4. This is another "returns to education" example but with parents' education as the instrument. Note the dataset lacks variable labels, but they can be found online.[5]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse mroz , clear`

c. Regress log wage on years of education `reg lwage educ , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous in your argument for the direction.

d. Explain what it would mean for `motheduc` to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

e. Run `ivreg2 lwage (educ = motheduc fatheduc) , robust`

i. Describe the IV estimand in this example.

ii. Discuss economic significance of the estimated slope ("returns to education").

iii. Explain why you are or are not worried about weak instruments in this case; in addition to the `ivreg2` output, run `weakivtest` and refer to specific output (and how to interpret it).

iv. Run `weakiv` and explain which confidence interval you think is more appropriate as well as what that CI tells us about our uncertainty about the true population value; be precise and explicit.

v. Explain what the J-test results suggest about the model.

f. Run `ivreg2 lwage (educ = motheduc fatheduc) exper expersq , gmm2s center robust` and briefly compare the estimate and CI for the coefficient on education with the previous estimates/CIs above.

g. Run `gmm (lwage - educ*{educ} - exper*{exper} - expersq*{expersq} - {_cons}) , instruments(motheduc fatheduc exper expersq) nolog vce(robust) twostep` and compare with the estimate/CI from part (f).

Exercise II.5. The following IV analysis uses cigarette prices to instrument for how much a mother smoked while pregnant, in hopes of estimating the causal effect of cigarette smoking on birthweight (which when too low is associated with other negative health outcomes for infants). Note the dataset lacks variable labels, but they can be found online.[6]

[5] http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.des

[6] http://fmwww.bc.edu/ec-p/data/wooldridge/bwght.des

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse bwght , clear`

c. Run `reg lbwght cigs male parity lfaminc , vce(robust)` and explain one potential source of omitted variable bias along with the direction of bias; be precise and rigorous.

d. Explain what it would mean for `cigprice` to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

e. Run `ivreg2 lbwght (cigs=cigprice) male parity lfaminc , robust`

   i. Describe the IV estimand in this example.

   ii. Discuss the economic significance of the estimated slope on `cigs`.

   iii. Explain why you are or are not worried about weak instruments in this case; in addition to the `ivreg2` output, run `weakivtest` and refer to specific output (and how to interpret it).

   iv. Run `weakiv` and explain which confidence interval you think is more appropriate as well as what that CI tells us about our uncertainty about the true population value; be precise and explicit.

   v. Explain what the J-test results suggest about the model.

f. Run `ivreg2 lbwght (cigs=cigprice) male parity lfaminc , robust gmm2s center` and briefly compare with your previous estimate and CI.

g. Run `gmm (lbwght - cigs*{cigs} - male*{male} - parity*{parity} - lfaminc*{lfaminc} - {_cons}) , instruments(cigprice male parity lfaminc) nolog vce(robust) twostep` and briefly compare with your previous estimates/CIs.

Exercise II.6. The following example uses data from Blackburn and Neumark (1992), specifically a cross-section of men in the year 1980, originally from the National Longitudinal Survey (NLS). The analysis uses birth order (1 means first-born in family / oldest child in family; 2 means second-born / second-oldest child in family; etc.) to instrument for how much education someone gets, in hopes of estimating the causal effect of education on (log) wage. Note the dataset lacks variable labels, but they can be found online.[7]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse wage2 , clear`

c. Run `reg lwage educ exper exp2 married , vce(robust)` and explain one potential source of omitted variable bias (for the coefficient on education) along with the direction of bias; be precise and rigorous in your argument for the direction.

d. Explain what it would mean for `brthord` to be a helpful "proxy" variable, and a potential (real-world) reason it may not be.

[7] http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.des

e. Explain what it would mean for `brthord` to be an "exogenous" instrument, and a potential (real-world) reason it may not be exogenous.

f. Run `ivreg2 lwage (educ=brthord) c.exper##c.exper married , robust`

   i. Describe the IV estimand in this example.

   ii. Discuss the economic significance of the estimated slope on `educ`.

   iii. Explain why you are or are not worried about weak instruments in this case; in addition to the `ivreg2` output, run `weakivtest` and refer to specific output (and how to interpret it).

   iv. Run `weakiv` and explain which confidence interval you think is more appropriate as well as what that CI tells us about our uncertainty about the true population value; be precise and explicit.

   v. Explain what the J-test results suggest about the model.

g. How many of your previous answers would change if we used two-step GMM estimation (instead of IV/2SLS regression)? Explain. (Feel free to re-run the `ivreg2` command with additional options `gmm2s center` to check.)

h. Run `gen expersq = exper^2` and then `gmm (lwage - educ*{educ} - exper*{exper} - expersq*{expersq} - married*{married} - {_cons}) , instruments(brthord exper expersq married) nolog vce(robust) twostep` and briefly compare with your previous estimates/CIs.

Part III

Panel Data

Introduction

One common way economists try to avoid omitted variable bias is with panel data, in which the same "individuals" (firms, counties, etc.) are observed in different time periods. Essentially, if there are omitted variables that are constant over time, then we can control for them even without observing them.

Again, concepts and intuition are developed in both a potential outcomes framework as well as a structural model framework. Identification is the focus, although some topics in estimation and inference (like cluster-robust standard errors) are also mentioned.

Chapter 11
Difference-in-Differences

Unit learning objectives for this chapter

11.1. Define and discuss the difference-in-differences approach to identification in the potential outcomes framework, both mathematically and intuitively, along with related concepts. [TLOs 1–3]

11.2. Judge whether or not the key identifying assumptions hold in specific real-world examples. [TLO 4]

This chapter describes the **difference-in-differences** (DiD) approach to causal identification with **panel data**, where the same individuals are observed across multiple time periods. The "canonical" DiD model consists of many individuals split into two groups and two time periods: nobody is treated in the first period, and (only) everyone in the treated group is treated in the second period.

Optional resources for this chapter

- Stata: commands `csdid` (Callaway, Rios-Avila, and Sant'Anna, 2021) and `drdid` (Naqvi, Rios-Avila, and Sant'Anna, 2021) available through SSC
- R: package `did` (Callaway and Sant'Anna, 2021a)
- Difference-in-differences (Masten video)
- Parallel trends (Masten video)
- Parallel trends example: immigration and unemployment (Masten videos)
- Diff-in-diff example: immigration and unemployment (Masten videos)
- Diff-in-diff example: minimum wage (Masten video)
- Diff-in-diff example: posting calorie counts (Masten video)

11.1 Panel Data Basics

This section contains some terms and notation related to panel data generally.

11.1.1 Basic Terms and Notation

The phrase **panel data** (same as **longitudinal data**) refers to observations of the same individuals in multiple time periods. As usual, these "individuals" could be firms, counties, hospitals, etc., and are labeled as $i = 1, \ldots, n$. The time periods $t = 1, \ldots, T$ could be years, months, weeks, or other periods of time; for example, $t = 1$ could mean the year 2019, $t = 2$ mean 2020, etc. Sometimes n is called the **cross-sectional dimension** of the data, and T the **time-series dimension** of the data. A **balanced panel** has exactly T observations for all n individuals, whereas an **unbalanced panel** has T observations for some individuals but fewer than T for others. For simplicity, I only consider balanced panel data, but in practice you should be aware if you have an unbalanced panel and how that might influence your results. With a balanced panel, there are nT total observations because each of the n individuals has T observations. A representative value is written like Y_{it} for the value of outcome variable Y for individual i at time t, or similarly $\boldsymbol{X}_{it}$ or U_{it}.

Terms from time series apply to the time dimension. For the outcome variable Y_{it}, the **first lag** or **lagged** outcome is Y_{it-1}, i.e., the value for the same individual in the previous time period. The **first difference** is $\Delta Y_{it} \equiv Y_{it} - Y_{it-1}$. The time-series average for individual i is $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$, sometimes also written $\bar{Y}_{i\cdot}$.

For identification, at least in the fixed-T microeconometric setting, instead of simply writing Y as a representative individual in the population, with panel data we usually write Y_t to denote a particular period (but the i remains implicit). That is, we imagine sampling individuals from the population like before, but now instead of each individual being represented by $(Y, \boldsymbol{X}')$, each individual is represented by $(Y_1, \ldots, Y_T, \boldsymbol{X}_1', \ldots, \boldsymbol{X}_T')$.

11.1.2 Asymptotic Frameworks

There are multiple possible asymptotic frameworks for panel data. Recall from Section 9.2 the weak-instrument asymptotic framework that can provide more accurate approximations of estimators' finite-sample properties when IVs are weak. Similarly, different asymptotic frameworks can provide more accurate approximations of estimators' finite-sample properties depending on the features of the panel dataset. For example, in macroeconomics, we may have quarterly data from 1960Q1 through 2019Q4 for Canada, Mexico, and the US; with relatively large $T = 240$ but only $n = 3$, the most accurate asymptotic approximations take $T \to \infty$ while holding $n = 3$ fixed. That is, we treat this type of panel data essentially as multiple time series. Alternatively, we may have a couple years of daily stock return data for a few hundred stocks; with n and T similarly in the hundreds, it may be best to use **joint asymptotics** with $n \to \infty$ and $T \to \infty$

together, written $n, T \to \infty$, or (usually less accurate but mathematically simpler) **sequential asymptotics** where first $n \to \infty$ with T fixed and then $T \to \infty$ (or vice-versa). In applied microeconomics, often n is large but T is as small as $T = 2$, in which case the most accurate asymptotic approximation takes $n \to \infty$ while holding T fixed. This framework is the focus of this textbook. Asymptotics with $n \to \infty$ are often called "large-n" asymptotics, and similarly $T \to \infty$ is called "large-T," whereas keeping n fixed is called either "small-n" or "fixed-n" asymptotics, or similarly "small-T" or "fixed-T" if T does not change asymptotically. Because n and/or T goes to infinity asymptotically, it suffices to say **fixed-T asymptotics** without explicitly saying "and large-n," and similarly it suffices to say **fixed-n asymptotics** without explicitly saying "and large-T."

With large-T asymptotics, we must explicitly model and make assumptions about the time-series properties, whereas with small-T asymptotics, we do not need to. That is, although potentially it could improve efficiency to explcitly model the time-series properties, with small-T asymptotics we can still achieve identification, consistency, and valid confidence intervals without such assumptions, so often they are avoided.

Still, the idea of **serial correlation** (or **autocorrelation**) is often discussed, meaning statistical correlation of values across time within the same individual. This is a common property of economic variables like monthly employment status, annual earnings, quarterly inflation rate, etc.

11.2 Some Intuition

⟹ Kaplan video: Diff-in-Diff Intuition

This section is largely from Section 9.7 of Kaplan (2022a), although with different notation.

Imagine some individuals were exposed to some "treatment," like a training program or law or other policy, whose causal effect we want to learn. The treatment wasn't randomized, but there's a group of untreated individuals whose outcomes can be used to form a **counterfactual**: what's the mean outcome of treated individuals in the parallel universe where they weren't treated?

If there is a plausible way to create such a counterfactual, to get an "as good as randomized" comparison, then such setups are sometimes called **natural experiments** or **quasi-experiments**. Generally, without randomization, it's invalid to simply compare treated and untreated outcomes, as seen in Section 11.2.1. However, we assume there is enough randomness that a valid comparison can be found, with some additional work.

Example 11.1 introduces a running example used throughout this section.

Example 11.1 (Kaplan video)**.** Let Y be annual labor income, and we are interested in the effect of minimum wage. Imagine our city recently implemented a large minimum wage increase. The goal is to learn the effect of this particular minimum wage increase on Y (income), for individuals in our city. Notationally, $X = 1$ if the individual lives in

our city (and $X = 0$ otherwise), and $t = 2$ if the observation is from the year after the minimum wage increase (and $t = 1$ if before the increase).

Notationally, $X = 1$ is the "treated group" and $X = 0$ the "untreated group"; $t = 1$ is the time period "before" treatment and $t = 2$ is "after."

11.2.1 Bad Approaches

One bad approach is to use only data from the treated group, comparing before and after. That is, we could try to estimate $\mathrm{E}(Y_2 \mid X = 1) - \mathrm{E}(Y_1 \mid X = 1)$, where Y_t is the observed outcome in time period t. Part of this mean difference is indeed due to the effect of the treatment. However, there are almost always other important determinants of Y that change over time. In that case, there is omitted variable bias: the mean difference is a combination of the treatment effect plus many other effects, so it is wrong to interpret the mean difference as only the effect of the treatment.

Example 11.2 (Kaplan video)**.** Continuing the minimum wage example (Example 11.1), one bad approach is to use only data from our city, before and after the minimum wage increase. Coincidentally, there may have been a national (or global) recession right after the minimum wage law was passed. This may make everybody's income lower in the year after. It would look like the minimum wage hurt incomes, but really it was the recession. Alternatively, there may have been great national (macroeconomic) conditions that made incomes go up, which would make us incorrectly conclude that the law increased incomes greatly.

Another bad approach is to use only data from the "after" period, comparing the treated group to an untreated group. That is, we could try to estimate $\mathrm{E}(Y_2 \mid X = 1) - \mathrm{E}(Y_2 \mid X = 0)$. Part of this mean difference is indeed due to the effect of the treatment. However, there are almost always other important determinants of Y that differ between the treated and untreated groups. In that case, there is again omitted variable bias.

Example 11.3 (Kaplan video)**.** Again continuing the minimum wage example (Example 11.1), this bad approach compares incomes in our city and another city, in the year after our law passed. By using the other city as a sort of control group, we avoid the problem of misinterpreting macroeconomic changes as treatment effects. However, it's hard to know which other city to pick. We could pick one that has the same population, for example, but our city may still have much higher (or lower) income for reasons other than our minimum wage. For example, San Francisco and Columbus, OH have very similar populations, but they have (and have for a while had) very different incomes. If San Francisco happens to have a higher minimum wage, it is wrong to attribute the entire mean difference in income as a causal effect of their higher minimum wage. There may indeed be a minimum wage effect, but it's mixed with the effects of education, industry types, geography, etc.

Discussion Question 11.1 (bad panel approaches: Mariel boatlift)**.** Consider the basic setup from Card (1990). Due to a seemingly random/exogenous political decision, Cubans were temporarily permitted to immigrate to the U.S. for a few months in 1980. About half settled in Miami, FL, while the other half went to live in other cities around the U.S.

a) We could compare wages of native-born workers in Miami in 1979 (before boatlift) vs. 1981 (after). Explain why this change in average wage would not be a good estimate of the causal effect of the Mariel boatlift on native worker wage. (Hint: are 1979 Miami and 1981 Miami the same except for how many Cubans live there, or might something else have changed?)

b) We could compare 1981 wages of native workers in Miami vs. Houston, TX, a city that did not receive a large influx of Cuban immigrants in 1980. Explain why this difference (Miami minus Houston) in average wage would not be a good estimate of the causal effect of the Mariel boatlift on native worker wage. (Hint: are 1981 Miami and Houston the same except for how many Cubans live there, or might there be other differences between the cities that might cause omitted variable bias?)

Discussion Question 11.2 (bad panel approaches: fracking)**.** This is based loosely on the setting of Street (2022), who of course uses much better approaches. For counties in the U.S. state of North Dakota, let Y denote crime rate. "Fracking" was a new technology that allowed extraction of certain underground oil and natural gas reserves that were previously infeasible or unprofitable to extract.

a) We could compare the crime rate among counties that eventually started fracking activity, before vs. after the fracking started. Explain why this before/after change in crime would not be a good estimate of the causal effect of the fracking activity on crime rate.

b) We could compare the "after" crime rates in North Dakota counties with fracking vs. those without fracking. Explain why this difference (fracking minus non-fracking) in crime would not be a good estimate of the causal effect of fracking on crime rate.

11.2.2 Counterfactuals and Parallel Trends

The difference-in-differences idea is to combine the before vs. after comparison with the treated vs. untreated comparison.

Conceptually, the goal is to construct a **counterfactual** (link to pronunciation), like what our city's mean income would have been if there were not a minimum wage increase. Thinking of the potential outcomes framework, the counterfactual is the parallel universe where the treatment never happened.

The key identifying assumption for DiD is called **parallel trends**. Conceptually, in the running example, parallel trends says that without the minimum wage law, our city's mean income would have increased by exactly the same amount as the other city's mean income. Mathematically, with

$$m_t(x) \equiv \mathrm{E}(Y_t \mid X = x), \tag{11.1}$$

the other city's mean income increase (i.e., "after" minus "before") is

$$m_2(0) - m_1(0) = \mathrm{E}(Y_2 \mid X = 0) - \mathrm{E}(Y_1 \mid X = 0). \tag{11.2}$$

Parallel trends assumes that adding this increase to the "before" mean income in our city, $m_1(1) = \mathrm{E}(Y_1 \mid X = 1)$, gives us the counterfactual income for our city in the "after" time period.

Given parallel trends, we can learn about causality by comparing

$$\overbrace{\mathrm{E}(Y_2 \mid X = 1)}^{\text{actual (our city, after)}} \quad \text{vs.}$$

$$\overbrace{\underbrace{\mathrm{E}(Y_1 \mid X = 1)}_{\text{our city, before}} + \underbrace{\mathrm{E}(Y_2 \mid X = 0) - \mathrm{E}(Y_1 \mid X = 0)}_{\text{increase in other city over time}}}^{\text{counterfactual}}, \tag{11.3}$$

$$\overbrace{m_2(1)}^{\text{actual}} - \overbrace{\{m_1(1) + [m_2(0) - m_1(0)]\}}^{\text{counterfactual}} = \overbrace{[m_2(1) - m_1(1)] - [m_2(0) - m_1(0)]}^{\text{difference-in-differences}}. \tag{11.4}$$

Figure 11.1 visualizes this effect. Note the notation is from Kaplan (2022a, §9.7), where $m(a, b)$ translates to $m_{b+1}(a)$ in our notation. We can think of constructing the counterfactual outcome, and then subtracting it from the actual outcome; or equivalently we can think of taking the before/after difference for our city and subtracting off the before/after difference in the other city.

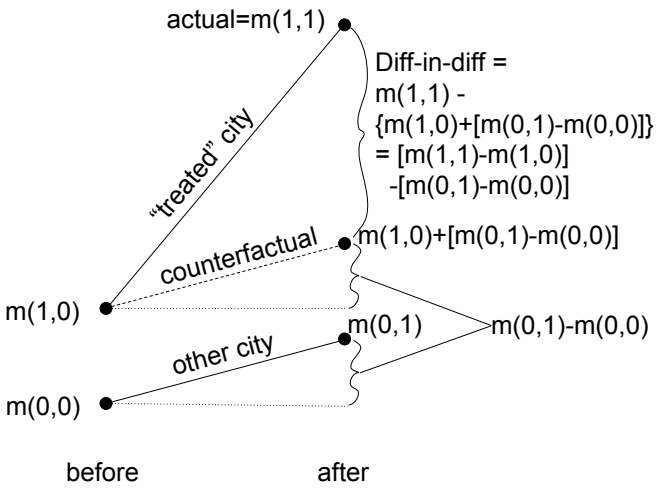


Figure 11.1: Difference-in-differences.

11.3 ATT Identification

This section formalizes the above intuition, within the potential outcomes framework.

We need to define treated and untreated potential outcomes in both the before and after periods. To avoid confusion with time period t, now superscript 1 denotes the treated potential outcomes, and superscript 0 for untreated; and the t subscript denotes the time period. That is, period t treated and untreated potential outcomes are respectively Y_t^1 and Y_t^0. Thus, the period t ATT is

$$\text{ATT}_t \equiv \text{E}(Y_t^1 - Y_t^0 \mid X = 1). \tag{11.5}$$

Assumption A11.1 (parallel trends)**.** The **parallel trends** assumption is that

$$\begin{aligned} &\text{E}(Y_2^0 \mid X = 1) - \text{E}(Y_1^0 \mid X = 1) \\ &= \text{E}(Y_2^0 \mid X = 0) - \text{E}(Y_1^0 \mid X = 0). \end{aligned} \tag{11.6}$$

Discussion Question 11.3 (before-before trends)**.** Imagine you also observe the period before the "before" period, $t = 0$. In your data, when you compare the change in average Y from the before-before period to the before period, you notice that the eventually-treated group changes by almost exactly the same amount as the untreated group (and confidence intervals are very short/precise). That is, $\widehat{\text{E}}(Y_1 \mid X = 1) - \widehat{\text{E}}(Y_0 \mid X = 1) \approx \widehat{\text{E}}(Y_1 \mid X = 0) - \widehat{\text{E}}(Y_0 \mid X = 0)$.

a) Why is this related to parallel trends (A11.1)?
b) Why does this not prove that the parallel trends assumption holds true (approximately)?

Theorem 11.1 (DiD identification)**.** *Under Assumptions A4.1 and A11.1 and notation from (11.1), the ATT at $t = 2$ defined in (11.5) is identified by*

$$\text{ATT}_2 = [m_2(1) - m_1(1)] - [m_2(0) - m_1(0)].$$

Proof. Starting from (11.5) and using linearity,

$$\begin{aligned} \text{ATT}_2 \equiv \text{E}(Y_2^1 - Y_2^0 \mid X = 1) &= \text{E}(Y_2^1 \mid X = 1) - \text{E}(Y_2^0 \mid X = 1) \\ &= \overbrace{\text{E}(Y_2 \mid X = 1)}^{=m_2(1)} - \text{E}(Y_2^0 \mid X = 1), \end{aligned} \tag{11.7}$$

using the fact that we observe $Y_2 = Y_2^1$ given $X = 1$. Rearranging (11.6),

$$\begin{aligned} \text{E}(Y_2^0 \mid X = 1) &= \text{E}(Y_1^0 \mid X = 1) + \text{E}(Y_2^0 \mid X = 0) - \text{E}(Y_1^0 \mid X = 0) \\ &= \text{E}(Y_1 \mid X = 1) + \text{E}(Y_2 \mid X = 0) - \text{E}(Y_1 \mid X = 0) \\ &= m_1(1) + m_2(0) - m_1(0), \end{aligned}$$

using the fact that the untreated potential outcome is the observed outcome for everyone in $t = 1$ and for $X = 0$ in $t = 2$. Plugging back into (11.7),

$$\begin{aligned} \text{ATT}_2 &= m_2(1) - \text{E}(Y_2^0 \mid X = 1) = m_2(1) - [m_1(1) + m_2(0) - m_1(0)] \\ &= [m_2(1) - m_1(1)] - [m_2(0) - m_1(0)]. \end{aligned}$$

□

Discussion Question 11.4 (DiD with targeted treatment)**.** Let $Y = 1$ if an individual is employed sometime during the month, otherwise $Y = 0$. Imagine there is a job training program, but it has zero effect on anybody, so untreated and treated potential outcomes are equal to each other and thus equal to the observed outcome: $Y_t^0 = Y_t^1 = Y_t$ for $t = 1, 2$. Assume that "after" outcomes are independent of "before" outcomes, $Y_2 \perp\!\!\!\perp Y_1$. Finally, imagine the program targets unemployed individuals, so that everyone with $Y_1 = 0$ then receives treatment while everyone with $Y_1 = 1$ does not.

a) Explain why Assumption A11.1 fails.
b) Will the DiD estimator have positive or negative bias? Why?
c) How would such bias affect our policy decision if we falsely assume our DiD estimator is accurate? Specifically, will we be more likely to continue this ineffective program, or will it actually help us realize the program doesn't work?

11.4 Estimation by Regression

⟹ Kaplan video: Fully Saturated Model Interpretation

The statistical object from Theorem 11.1 can be estimated by OLS. This section is largely from Section 9.3 of Kaplan (2022a).

Notationally, let X_{it1} be the treatment group dummy with $X_{it1} = 1$ if individual i is in the treated group and $X_{it1} = 0$ otherwise, and let $X_{it2} = \mathbb{1}\{t = 2\}$ be a time dummy with $X_{it2} = 1$ in period $t = 2$ and $X_{it2} = 0$ otherwise (in period $t = 1$).

Consider the following CMF model in error form,

$$Y_{it} = \overbrace{\beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + \beta_3 X_{it1} X_{it2}}^{\equiv m(X_{it1}, X_{it2})} + V_{it}, \quad \mathrm{E}(V_{it} \mid X_{it1}, X_{it2}) = 0. \tag{11.8}$$

This is an example of a **fully saturated** model because it is flexible enough to allow a different CMF value for each value of (X_1, X_2), so the linear-in-parameters functional form is not restrictive. Logically, having the same number (four) of possible values of (X_1, X_2) as β_j parameters is necessary but not sufficient for the model to be fully saturated.

Interpretation of the coefficients requires writing them in terms of different CMF values. First, each CMF value can be written in terms of the β_j:

$$\begin{aligned}
m(x_1, x_2) &= \beta_0 + (\beta_1)(x_1) + (\beta_2)(x_2) + (\beta_3)(x_1)(x_2), \\
m(0,0) &= \beta_0 + (\beta_1)(0) + (\beta_2)(0) + (\beta_3)(0)(0) = \beta_0, && (11.9)\\
m(0,1) &= \beta_0 + (\beta_1)(0) + (\beta_2)(1) + (\beta_3)(0)(1) = \beta_0 + \beta_2, && (11.10)\\
m(1,0) &= \beta_0 + (\beta_1)(1) + (\beta_2)(0) + (\beta_3)(1)(0) = \beta_0 + \beta_1, && (11.11)\\
m(1,1) &= \beta_0 + (\beta_1)(1) + (\beta_2)(1) + (\beta_3)(1)(1) = \beta_0 + \beta_1 + \beta_2 + \beta_3. && (11.12)
\end{aligned}$$

From (11.9)–(11.12) and their differences,

$$\overbrace{\beta_0 = m(0,0)}^{(11.9)}, \tag{11.13}$$

$$\beta_1 = \overbrace{(\beta_0 + \beta_1) - \beta_0 = m(1,0) - m(0,0)}^{(11.11)\text{ minus }(11.9)}, \tag{11.14}$$

$$\beta_2 = \overbrace{(\beta_0 + \beta_2) - \beta_0 = m(0,1) - m(0,0)}^{(11.10)\text{ minus }(11.9)}, \tag{11.15}$$

$$\begin{aligned}\beta_3 = [\beta_2 + \beta_3] - [\beta_2] &= \overbrace{[(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1)]}^{(11.12)\text{ minus }(11.11)} - \overbrace{[(\beta_0 + \beta_2) - (\beta_0)]}^{(11.10)\text{ minus }(11.9)} \\ &= \overbrace{\overbrace{[m(1,1) - m(1,0)]}^{\text{difference}} - \overbrace{[m(0,1) - m(0,0)]}^{\text{difference}}}^{\text{difference-in-differences}} \qquad (11.16) \\ &= [m(1,1) - m(0,1)] - [m(1,0) - m(0,0)] \qquad (11.17) \\ &= \overbrace{[(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2)]}^{(11.12)\text{ minus }(11.10)} - \overbrace{[(\beta_0 + \beta_1) - (\beta_0)]}^{(11.11)\text{ minus }(11.9)}.\end{aligned}$$

This shows the same difference-in-differences structure seen in (11.4) and Figure 11.1.

Using (11.13)–(11.17), the four β_j in (11.8) have the following interpretations.

- $\beta_0 = m(0,0)$ is the mean Y in the subpopulation with $X_1 = 0$ and $X_2 = 0$ (untreated group, before period). Caution: generally $\beta_0 \neq \mathrm{E}(Y)$.
- $\beta_1 = m(1,0) - m(0,0)$ is the mean Y difference between $X_1 = 1$ and $X_1 = 0$ individuals (treated vs. untreated group) within the $X_2 = 0$ subpopulation (before period). Caution: generally $\beta_1 \neq \mathrm{E}(Y \mid X_1 = 1) - \mathrm{E}(Y \mid X_1 = 0)$; it additionally conditions on $X_2 = 0$.
- $\beta_2 = m(0,1) - m(0,0)$ is the mean Y difference between $X_2 = 1$ and $X_2 = 0$ individuals (after vs. before) within the $X_1 = 0$ subpopulation (untreated group). Caution: generally $\beta_2 \neq \mathrm{E}(Y \mid X_2 = 1) - \mathrm{E}(Y \mid X_2 = 0)$; it additionally conditions on $X_1 = 0$.
- $\beta_3 = [m(1,1) - m(1,0)] - [m(0,1) - m(0,0)]$ is the mean Y difference associated with X_2 in the $X_1 = 1$ subpopulation *minus* the mean Y difference associated with X_2 in the $X_1 = 0$ subpopulation, i.e., the mean before/after difference in the treated group *minus* the mean before/after difference in the untreated group.
- $\beta_3 = [m(1,1) - m(0,1)] - [m(1,0) - m(0,0)]$ is also the mean Y difference associated with X_1 in the $X_2 = 1$ subpopulation *minus* the mean Y difference associated with X_1 in the $X_2 = 0$ subpopulation, i.e., the mean treated/untreated difference in the after period *minus* the mean treated/untreated difference in the before period.

Discussion Question 11.5 (diff-in-diff CMF)**.** Consider the running minimum wage example, where Y_{it} is annual labor income of individual i in year t (in dollars), $X_{it1} = 1$ if the individual lives in our city with the minimum wage increase (and $X_{it1} = 0$ if they live in the other city without the minimum wage change), and $X_{it2} = \mathbb{1}\{t = 2\}$ is the after-period time dummy. Consider the estimated CMF $\hat{m}(x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$.

a) Interpret $\hat{\beta}_0$.
b) Interpret $\hat{\beta}_1$.
c) Interpret $\hat{\beta}_2$.
d) Interpret $\hat{\beta}_3$.

Chapter 12

Fixed Effects Regression

Unit learning objectives for this chapter

12.1. Describe the fixed effects approach to solving OVB with panel data, both mathematically and intuitively. [TLOs 1–3]

12.2. Judge whether or not the key identifying assumptions hold in specific real-world examples. [TLO 4]

This chapter describes the **fixed effects** (FE) regression approach with panel data, to identify structural coefficients in the presence of a particular type of omitted variable bias.

Optional resources for this chapter

- Stata: built-in commands `xtset` and `xtreg`
- R: package `plm` (Croissant and Millo, 2008) with function `plm()`, where argument `index=c('id','year')` specifies the individual and year identifier variables (i and t), `model='pooling'` runs pooled OLS, and `model='within'` runs FE, with time effects most easily added by argument `effect='twoways'`; and cluster-robust SE can be computed through the `vcovHC()` function in the `sandwich` package (Zeileis, 2004) with for example `cluster='group'` and `method='arellano'` (and `type='HC0'`, although they differ from Stata's by a degree-of-freedom adjustment (Stata's are slightly larger); or you can just use Stata and it's just one line of code with zero packages.

12.1 Structural Model

Consider the following structural model, with notation as in Section 11.1.1:

$$Y_t = \beta_0 + \boldsymbol{X}_t'\boldsymbol{\beta} + \overbrace{C + V_t}^{\text{unobserved}}, \tag{12.1}$$

where (unlike in Parts I and II) regressor vector $\boldsymbol{X}_t$ does not include an intercept term. Assume the **idiosyncratic error term** V_t is exogenous (where "idiosyncratic" refers to it varying across both individuals and time), so the only threat to identification is C. This model allows omitted variable bias through the unobserved C, but it imposes other restrictions: the model is linear-in-parameters, C is additive (not interacting with $\boldsymbol{X}_t$), and C is constant over time (no t subscript). More precisely: the structural error term V_t includes any deviations from linearity, additive separability of C, or time-invariance of C, which may make exogeneity less plausible. More optimistically, even if we don't fully believe these conditions, small violations may not introduce much bias; for example, maybe C is not literally time-invariant, but it varies very slowly over decades, and we have a three-year sample. As before, we could also add an i subscript to each random variable in (12.1), with the same meaning. The goal is to identify and estimate structural parameter vector $\boldsymbol{\beta}$, despite the OVB caused by C.

Discussion Question 12.1 (FE: wage model)**.** Consider the structural model (12.1) where Y is log wage, $\boldsymbol{X}$ includes years of education and experience (as well as their squares and interaction/product), and C aggregates unobserved abilities (beyond those gained through education) that determine wage, like social skills, perseverance, etc.

a) Explain one real-world reason we might think C interacts with $\boldsymbol{X}$, rather than being additively separable.
b) What do you think about the assumption that C is time-invariant in this example? Explain.

Model (12.1) has a variety of names. It could be called an **unobserved effects model**, as in Wooldridge (2010, Ch. 10). Besides "unobserved effect," C can also be called **unobserved heterogeneity** (a more general term of which this form is a special case) or individual-specific heterogeneity or individual effects or a fixed effect, or it can be seen as a random intercept, using the terminology of random coefficient models (Section 4.2.2). Sometimes (12.1) is called a fixed effects regression model, but "fixed effects" refers more specifically to assumptions about C rather than the structural model itself. To any of these names may be added any subset of the additional terms "panel," "panel data," and "regression."

Beyond our scope...

What of the constant $\boldsymbol{\beta}$ is replaced by a vector of random coefficients (Section 4.2.2)? Under certain conditions, the FE approach can still identify the mean

(across individuals) of the individual-specific coefficient vector $\boldsymbol{\beta}_i$. For example, see Section 11.7 of Wooldridge (2010) and references therein.

12.2 Pooled OLS and Random Effects

What happens if we simply run OLS regression to estimate (12.1)? We can use our OVB results from Theorem 6.1 and Corollary 6.2, where here $\gamma = 1$ and $Q = C$. Those results say we have asymptotic bias if and only if C is correlated with $\boldsymbol{X}_t$, $\text{Cov}(C, \boldsymbol{X}_t) \neq 0$.

To think about such correlation, recall each individual in the population is represented by $(C, \boldsymbol{X}_1', \ldots, \boldsymbol{X}_T', Y_1, \ldots, Y_T)$. For example, with scalar X for simplicity, if individuals with high C tend to have high X_1 and generally high X_t, and individuals with low C tend to have low X_t, then $\text{Cov}(C, X_t) > 0$. That is, we are thinking about correlation in the cross-sectional dimension, because C is time-invariant.

Conversely, if C affects Y but is not correlated with $\boldsymbol{X}$, then the structural coefficients equal LP coefficients that OLS can estimate consistently.

The **pooled OLS** (POLS) estimator simply uses all the Y_{it} and $\boldsymbol{X}_{it}$ observations together, without regard for the i or t value, and can be written in the following notation. First, rewrite the structural model (12.1) as

$$Y_{it} = \boldsymbol{X}_{it}'\boldsymbol{\beta} + \overbrace{C_i + V_{it}}^{\equiv U_{it}}, \tag{12.2}$$

where now each $\boldsymbol{X}_{it}$ includes 1 as the first element (so the first element of $\boldsymbol{\beta}$ is the intercept). Let

$$\boldsymbol{Y} \equiv (Y_{11}, Y_{12}, \ldots, Y_{1T}, Y_{21}, \ldots, Y_{2T}, \ldots, Y_{nT})',$$

an $nT \times 1$ column vector, and define $\boldsymbol{U}$ in the same order. Let

$$\underline{\boldsymbol{X}}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT})'$$

be the $n \times \dim(\boldsymbol{\beta})$ regressor matrix for individual i that stacks the row vectors $\boldsymbol{X}_{it}'$ with $t = 1, \ldots, T$, and stacking these $\underline{\boldsymbol{X}}_i$ gives

$$\underline{\boldsymbol{X}} \equiv \begin{pmatrix} \underline{\boldsymbol{X}}_1 \\ \vdots \\ \underline{\boldsymbol{X}}_n \end{pmatrix}.$$

Given these definitions, the pooled OLS estimator has the familiar form

$$\hat{\boldsymbol{\beta}}_{\text{POLS}} = (\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'\boldsymbol{Y}. \tag{12.3}$$

Assumption A12.1 (contemporaneous exogeneity)**.** The error term U_t is **contemporaneously exogenous** in the sense that $\text{E}(\boldsymbol{X}_t U_t) = \boldsymbol{0}$ for each $t = 1, \ldots, T$, where "contemporaneous" refers to having the same time period t subscript on both the $\boldsymbol{X}$ and U inside the expectation.

Proposition 12.1 (pooled OLS)**.** *Given structural model* (12.2)*, if Assumption A12.1 holds, then the structural* $\boldsymbol{\beta}$ *is also the vector of linear projection coefficients in* $\text{LP}(Y_t \mid \boldsymbol{X}_t)$ *(assuming they are well-defined), which can be estimated under relatively general conditions (like iid sampling of individuals from the population) by OLS regression of* Y_{it} *on* $\boldsymbol{X}_{it}$*.*

Proof. Plugging (12.2) into (12.3) and using notation above,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{POLS}} &= (\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'\boldsymbol{Y} \\ &= (\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'(\underline{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{U}) \\ &= \boldsymbol{\beta} + (\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}'\boldsymbol{U} \\ &= \boldsymbol{\beta} + (\underline{\boldsymbol{X}}'\underline{\boldsymbol{X}}/n)^{-1}\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{X}_{it}U_{it}.\end{aligned}$$

As in the rest of this chapter, use the fixed-T asymptotic framework that lets $n \to \infty$ while keeping T constant. To emphasize this, while averaging over n (to apply WLLN/CLT), we simply sum over t because no WLLN/CLT is used in the time dimension. Centering,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{POLS}} - \boldsymbol{\beta} &= \Bigg[\sum_{t=1}^{T}\overbrace{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{it}\boldsymbol{X}_{it}'}^{\xrightarrow{p}\text{E}(\boldsymbol{X}_t\boldsymbol{X}_t')}\Bigg]^{-1}\sum_{t=1}^{T}\overbrace{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{it}U_{it}}^{\xrightarrow{p}\text{E}(\boldsymbol{X}_tU_t)=\boldsymbol{0}} \\ &\xrightarrow{p} \Bigg[\sum_{t=1}^{T}\text{E}(\boldsymbol{X}_t\boldsymbol{X}_t')\Bigg]^{-1}\sum_{t=1}^{T}\boldsymbol{0} \\ &= \boldsymbol{0},\end{aligned}$$

so the POLS estimator is consistent.

For asymptotic normality, scaling by $\sqrt{n}$ like usual,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{POLS}} - \boldsymbol{\beta}) = \Bigg[\sum_{t=1}^{T}\overbrace{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{it}\boldsymbol{X}_{it}'}^{\xrightarrow{p}\text{E}(\boldsymbol{X}_t\boldsymbol{X}_t')}\Bigg]^{-1}\sum_{t=1}^{T}\overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_{it}U_{it}}^{\xrightarrow{d}\text{N}(\boldsymbol{0},\underline{\boldsymbol{\Sigma}}_t)} \xrightarrow{d} \text{N}(\boldsymbol{0}, \underline{\boldsymbol{\Omega}}), \tag{12.4}$$

where the asymptotic covariance matrix depends on the $\text{E}(\boldsymbol{X}_t\boldsymbol{X}_t')$ as well as both $\underline{\boldsymbol{\Sigma}}_t \equiv \text{E}(U_t^2\boldsymbol{X}_t\boldsymbol{X}_t')$ (assuming iid across i, or else a different formula depending on the dependence across i) and the serial correlation across t of the $\boldsymbol{X}_tU_t$ vectors, which is generally non-zero. Section 12.8 further discusses how to account for such correlation when computing standard errors. □

The **random effects** (RE) estimator makes even stronger assumptions than POLS, so it is not discussed here. For example, see Assumption RE.1 (p. 292) of Wooldridge

(2010), which assumes the stronger "strict exogeneity" condition (Assumption A12.2 below) instead of the weaker contemporaneous exogeneity condition of Assumption A12.1. If the assumptions hold, then the RE estimator is more efficient than POLS (or other estimators), but if not, then the RE estimator is not even consistent. In line with this textbook's focus on identification, the focus below is instead on how to identify the structural coefficients under weaker assumptions than required by POLS.

12.3 Two-Period Case

Now assume the unobserved heterogeneity C causes OVB. Let $T = 2$ for intuition and to connect with Chapter 11.

Because C is time-invariant and enters the structural model (12.1) additively, we can first-difference the model to remove it, i.e., subtract the model for $t = 1$ from the model for $t = 2$. It is also very important that $\boldsymbol{\beta}$ is time-invariant. Using the notation from Section 11.1.1 like $\Delta Y_t = Y_t - Y_{t-1}$, and noting we only observe the first-differences at $t = 2$ (because we do not observe $t = 0$), the model becomes

$$\Delta Y_2 = (\beta_0 - \beta_0) + \Delta \boldsymbol{X}_2'\boldsymbol{\beta} + (C - C) + \Delta V_2 = \Delta \boldsymbol{X}_2'\boldsymbol{\beta} + \Delta V_2. \tag{12.5}$$

This looks promising: the troublesome C is gone, and the vector $\boldsymbol{\beta}$ is the same structural coefficient from our original model in (12.1). The natural question is: what if we simply regress ΔY_2 on $\Delta \boldsymbol{X}_2$ by OLS? As usual, the OLS estimand is fundamentally the linear projection coefficient vector, so our identification question reduces to whether or not ΔV_2 satisfies the LP error property in (12.5).

Beyond our scope...

The elegant removal of C in (12.5) does not work for "nonlinear" panel models such as quantile regression (and other non-quantile models, too). Essentially, unlike for the mean, the difference of medians does not equal the median difference, and similarly for other quantiles. If we try to control for C by including a dummy variable for each individual i, then we have n (or $n - 1$) such dummies and $n - 1$ parameters, which is problematic if we only have $T = 2$ and thus $2n$ total observations, although letting $T \to \infty$ allows us to estimate the individual FE values themselves (the individual dummy coefficients). And, for quantile models, there are yet further complications; for example, see Section 7.3 of Kaplan (2021) and references therein.

The desired LP error property is

$$\begin{aligned} \mathbf{0} &= \mathrm{E}[\Delta \boldsymbol{X}_2 \Delta V_2] = \mathrm{E}[(\boldsymbol{X}_2 - \boldsymbol{X}_1)(V_2 - V_1)] \\ &= \mathrm{E}(\boldsymbol{X}_2 V_2) - \mathrm{E}(\boldsymbol{X}_2 V_1) - \mathrm{E}(\boldsymbol{X}_1 V_2) + \mathrm{E}(\boldsymbol{X}_1 V_1). \end{aligned} \tag{12.6}$$

The contemporaneous exogeneity in Assumption A12.1 makes the first and last terms zero, but it does not restrict the middle two terms. That is, even if the idiosyncratic

error V_t is contemporaneously exogenous, the first-differenced model may not be a linear projection. Thus, we need a stronger type of exogeneity called **strict exogeneity** that requires the regressors and idiosyncratic errors uncorrelated across any two time periods. (Sometimes it is defined more strongly in terms of a conditional mean restriction rather than correlation.)

Assumption A12.2 (strict exogeneity)**.** The error term V_t is strictly exogenous in the sense that $\mathrm{E}(\boldsymbol{X}_s V_t) = \mathbf{0}$ for any combination of $s = 1, \ldots, T$ and $t = 1, \ldots, T$.

Discussion Question 12.2 (exogeneity example)**.** Let Y be a grocery store's weekly sales revenue in dollars, and let X be the store's weekly average discount percentage (like $X = 10$ if every product is on sale all week with a 10% discount). The structural model from (12.1) is $Y_t = \beta_0 + \beta_1 X_t + C + V_t$.

a) Explain why a grocery store's V_1 value (i.e., week $t = 1$ sales shock) might affect their X_2 (week $t = 2$ discounts), specifically why a negative V_1 might cause the store to choose larger X_2. (Hint: for example, consider perishable items that spoil after 2 weeks in the store.)
b) How does your argument relate to strict exogeneity (A12.2)?
c) Assume the V_t are iid, the X_t are only determined by V_{t-1}, and $\beta_1 = 0$ and $C = 0$. Explain why if V_1 is very negative, then we tend to see high (positive) $Y_2 - Y_1$ values and high $X_2 - X_1$. (A similar argument should suggest that if V_1 is very positive, then we'll tend to see negative $Y_2 - Y_1$ and negative $X_2 - X_1$.)
d) Given that, if we regress ΔY_2 on ΔX_2, what type of bias do we expect for our estimator of the true structural $\beta_1 = 0$?

The other OLS/LP assumption (rank condition) that is usually innocuous requires more thought here. In the basic cross-sectional linear projection model with Y and $\boldsymbol{X}$, the rank condition is that matrix $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')$ is invertible, so that the LPC $[\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')]^{-1}\boldsymbol{X}Y$ is well-defined. This essentially requires that no regressor is "redundant" in the formal sense of being a linear combination of other regressors; for example, we cannot have X_1 in miles and $X_2 = 0.6X_1$ the same variable in kilometers. In the first-differenced model, this is violated by any regressor that is time-invariant for all individuals, $X_{i1} = X_{i2}$ for all $i = 1, \ldots, n$. For example, if our sample includes older adults whose years of education is no longer changing, then we cannot include education, and we certainly cannot learn about the original intercept β_0 from (12.1). Intuitively, first-differencing not only removes the time-invariant C effect, but also any other time-invariant effects, so we cannot learn about the causal effect of any time-invariant regressor. Additionally, if our model includes a time dummy for period $t = 2$, then its first-differenced value is 1 for all individuals, so we cannot include any other regressor that automatically increases by 1 every period for every individual. For example, if t is in years, then we cannot include an individual's age in years as a regressor, because age increases by 1 for all individuals between $t = 1$ and $t = 2$. Intuitively, we cannot distinguish between the effect of the world changing from $t = 1$ to $t = 2$ and the the effect of everyone being one year older.

One "exception" is that we can learn about how the coefficient on a time-invariant regressor changes across time periods. This is not actually an "exception" because we can't learn about the actual level/value of the coefficient, only it's relative value across different t. To do this, we can interact the time-invariant regressor with time dummies. For example, in our $T = 2$ case, we can include $W_{it} \equiv X_i \mathbb{1}\{t = 2\}$ as a regressor because it is time-varying: its value is $W_{i1} = 0$ in the first period (for all i), whereas generally $W_{i2} \neq 0$.

Assumption A12.3 (FD rank condition)**.** The matrix $\mathrm{E}(\Delta X_2 \Delta X_2')$ is invertible.

The estimator running OLS on (12.5) is called the **first-difference estimator** (FD estimator), and the related **fixed effects estimator** (FE estimator) is equivalent with $T = 2$ but differs with $T > 2$. Before running OLS, instead of applying the FD transformation to the structural model, the FE estimator applies the **within transformation** (or **fixed effects transformation**), demeaning each observation by its within-individual average (across $t = 1, \ldots, T$). Recall such averages are denoted $\bar{Y}_i$, $\bar{\boldsymbol{X}}_i$, etc.; with $T = 2$, they are simply $\bar{Y}_i = (Y_1 + Y_2)/2$, etc. The demeaned Y values are

$$\begin{aligned} Y_{i2} - \bar{Y}_i &= Y_{i2} - \frac{Y_{i1} + Y_{i2}}{2} = \frac{Y_{i2} - Y_{i1}}{2} \\ Y_{i1} - \bar{Y}_i &= Y_{i1} - \frac{Y_{i1} + Y_{i2}}{2} = \frac{Y_{i1} - Y_{i2}}{2} = -(Y_{i2} - \bar{Y}_i). \end{aligned}$$

Similarly,

$$\boldsymbol{X}_{i2} - \bar{\boldsymbol{X}}_i = \frac{\boldsymbol{X}_{i2} - \boldsymbol{X}_{i1}}{2}, \quad \boldsymbol{X}_{i1} - \bar{\boldsymbol{X}}_i = -(\boldsymbol{X}_{i2} - \bar{\boldsymbol{X}}_i).$$

That is, for both Y and $\boldsymbol{X}$, the $t = 1$ demeaned values are exactly the negative of the $t = 2$ demeaned values, so they do not provide any additional data. Note also that the $t = 2$ values are simply the first-differenced values divided by 2, so in this $T = 2$ case, regressing the demeaned $Y_{it} - \bar{Y}_i$ on the demeaned $\boldsymbol{X}_{it} - \boldsymbol{X}_i$ is identical to regressing the first-differenced versions, so Theorem 12.2 applies equally to each. Later sections will distinguish the two in the general T setting.

Theorem 12.2 (two-period panel identification)**.** *Under Assumptions A12.2 and A12.3, the structural slope coefficient vector $\boldsymbol{\beta}$ in the structural model (12.1) is identified and equal to the linear projection coefficient vector of* $\mathrm{LP}(\Delta Y_2 \mid \Delta \boldsymbol{X}_2)$*:*

$$\boldsymbol{\beta} = [\mathrm{E}(\Delta \boldsymbol{X}_2 \Delta \boldsymbol{X}_2')]^{-1} \mathrm{E}(\Delta \boldsymbol{X}_2 \Delta Y_2).$$

Proof. From (12.6), the linear projection error property holds in the first-differenced model if

$$\mathbf{0} = \mathrm{E}(\boldsymbol{X}_2 V_2) - \mathrm{E}(\boldsymbol{X}_2 V_1) - \mathrm{E}(\boldsymbol{X}_1 V_2) + \mathrm{E}(\boldsymbol{X}_1 V_1),$$

all four of which equal zero by Assumption A12.2. The linear projection coefficients are well-defined because of Assumption A12.3. □

This model can be more explicitly connected to the DiD model in Chapter 11. Continue with $T = 2$. Let $D_{it} = 1$ if individual i is treated in period t, otherwise $D_{it} = 0$. Let $W_{it} = \mathbb{1}\{t = 2\}$ be a time dummy for the "after" period. Let $G_{it} = D_{i2}$, so $G_{it} = 1$ in both periods if individual i is ever treated, otherwise $G_{it} = 0$ in both periods; that is, G_{it} is the group dummy. The DiD model had $Y_{it} = \beta_0 + \beta_1 G_{it} + \beta_2 W_{it} + \beta_3 D_{it} + V_{it}$ with $\mathrm{E}(V_{it} \mid G_{it}, W_{it}) = 0$, noting that $G_{it} W_{it} = D_{it}$. Noting G_{it} does not depend on t, nor does W_{it} depend on i, we can write this equivalently as $Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 W_t + \beta_3 D_{it} + V_{it}$. This is almost a special case of (12.1), except groups of individuals have the same G_i rather than everyone having (potentially) a different fixed effect C_i. First-differencing and using $\Delta W_2 = W_2 - W_1 = 1 - 0 = 1$,

$$\Delta Y_{i2} = \beta_2 + \beta_3 \Delta D_{i2} + \Delta V_{i2},$$

where $\Delta D_{i2} = 1$ for treated individuals and $\Delta D_{it} = 0$ for untreated individuals. So the FD/FE estimator can also estimate the ATT parameter β_3. Recalling results for simple regression with a single binary regressor like in Section 6.3.2 of Kaplan (2022a), $\beta_2 = \mathrm{E}(\Delta Y_2 \mid \Delta D_2 = 0)$ and $\beta_2 + \beta_3 = \mathrm{E}(\Delta Y_2 \mid \Delta D_2 = 1)$. Noting $\Delta D_2 = 1$ is equivalent to being in the treated group $G = 1$, and using linearity, $\mathrm{E}(\Delta Y_2 \mid \Delta D_2 = 0) = \mathrm{E}(Y_2 - Y_1 \mid G = 0) = \mathrm{E}(Y_2 \mid G = 0) - \mathrm{E}(Y_1 \mid G = 0)$, which is the same as β_2 from Chapter 11, and

$$\mathrm{E}(\Delta Y_2 \mid \Delta D_2 = 1) = \mathrm{E}(Y_2 - Y_1 \mid G = 1) = \mathrm{E}(Y_2 \mid G = 1) - \mathrm{E}(Y_1 \mid G = 1),$$

the before/after mean difference for the treated group. Subtracting β_2 yields β_3 and the same "difference-in-differences" interpretation: the before/after mean difference for the treated group *minus* the before/after mean difference for the untreated group.

12.4 Efficiency

The identification advantage of FE/FD over POLS (allowing C to be correlated with regressors) generally comes at the price of efficiency. FE and FD use only the **within variation**, i.e., the within-individual variation of Y_{it} and $\boldsymbol{X}_{it}$ over time. POLS uses within variation as well as **between variation**, i.e., the variation of Y_{it} and $\boldsymbol{X}_{it}$ across individuals. In the extreme, as discussed, for regressors that have only between variation and not within variation (i.e., time-invariant regressors), FE and FD cannot identify their effect, whereas there is at least some hope of learning about their effect from cross-sectional variation. Less extreme, maybe the variables are not constant over time but have very little variation over time (within individual) compared with the amount of between variation. In that case, the FE and FD estimators have a lot of uncertainty, similar to how the simple regression estimator of the linear projection slope $\mathrm{Cov}(Y, X)/\mathrm{Var}(X)$ has a lot of uncertainty (all else equal) if $\mathrm{Var}(X)$ is near zero.

In fact, recalling Section 3.10.2, it is possible to have real-world applications where POLS is biased but its variance is so much smaller than FE/FD's that the POLS mean squared error is smaller than FE/FD's. However, in that case the POLS confidence

intervals would not be valid (even approximately), whereas the FE confidence intervals would be. That is, FE/FD's lower efficiency may mean wider confidence intervals, but they are still valid (asymptotically). Generally, economists tend to use FE or FD because usually the threat of OVB (due to unobserved C) is perceived as the most important consideration.

With $T > 2$, either FE or FD may be more efficient depending on the properties of the idiosyncratic error. Essentially, if the idiosyncratic error is highly serially correlated, like a random walk in the extreme case, then FD is more efficient, whereas if the idiosyncratic error has low serial correlation, like iid in the extreme case, then FE is more efficient. For example, see Assumptions FE.3 (p. 304) and FD.3 (p. 318) of Wooldridge (2010), although note that even his exogeneity assumption (FE.1) is stronger (conditional mean restrictions) than Assumption A12.2 here.

However, beyond efficiency, FE and FD can also differ in bias, which again is usually what economists are more concerned about. Specifically, if only contemporaneous exogeneity holds (not strict exogeneity), then generally the FE bias is smaller than FD bias, at least for larger T. Formal results under some additional assumptions are given on pages 322–323 of Wooldridge (2010), showing that the FE bias is $O(1/T)$ and thus decreasing to zero as T increases, whereas the FD bias is $O(1)$ and thus not decreasing with T. However, if the idiosyncratic error is a unit-root process like a random walk, then actually FD tends to be better than FE, even in terms of bias.

12.5 Two-Way Fixed Effects

With any $T \geq 2$, the **two-way fixed effects** (TWFE) estimator is essentially the FE estimator when **time effects** (dummies for each time period) are included. That is, the "two ways" are 1) individual effects and 2) time effects. More precisely, there must be $T - 1$ time dummies (not T) to avoid violating the rank condition (i.e., to avoid perfect multicollinearity), but as usual Stata will drop the extra one if you accidentally include it, so you don't need to worry too much about this. Usually the $t = 1$ dummy is excluded, so $t = 1$ is interpreted as the base period or reference period, unless some other t is more salient (like the period before a major policy change). In the large-n, fixed-T framework, including time effects does not require any special consideration like the individual effects do (FE or FD transformation); you simply include the corresponding time dummies.

The FE and FD estimators are the same as in the $T = 2$ case of Section 12.3. That is, FD runs OLS on the first-differenced model, regressing ΔY_{it} on $\Delta \boldsymbol{X}_{it}$ for $t = 2, \ldots, T$ (and $i = 1, \ldots, n$); and FE runs OLS on the within-transformed model, regressing $Y_{it} - \bar{Y}_i$ on $\boldsymbol{X}_{it} - \bar{\boldsymbol{X}}_i$, using all combinations of $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

Note that strict exogeneity is sufficient for FD identification, though somewhat stronger than necessary. With general T, we need the same condition as in (12.6), but replacing the subscript values 1 and 2 with $t - 1$ and t:

$$\mathbf{0} = \mathrm{E}(\Delta \boldsymbol{X}_t \Delta V_t) = \mathrm{E}(\boldsymbol{X}_t V_t) - \mathrm{E}(\boldsymbol{X}_t V_{t-1}) - \mathrm{E}(\boldsymbol{X}_{t-1} V_t) + \mathrm{E}(\boldsymbol{X}_{t-1} V_{t-1}).$$

These terms all equal zero if

$$\mathrm{E}(\boldsymbol{X}_s V_t) = \mathbf{0}, \quad |s-t| \leq 1, \quad (s,t) \in \{1,\ldots,T\}^2, \tag{12.7}$$

whereas strict exogeneity says this must hold for any (s,t), not only $|s-t| \leq 1$.

Assumption A12.4 (FD rank condition)**.** Using notation $\Delta \boldsymbol{X}_t \equiv \boldsymbol{X}_t - \boldsymbol{X}_{t-1}$, the matrix $\sum_{t=2}^{T} \mathrm{E}(\Delta \boldsymbol{X}_t \Delta \boldsymbol{X}_t')$ is invertible (full rank).

Assumption A12.5 (FE rank condition)**.** Using notation $\bar{\boldsymbol{X}} \equiv \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{X}_t$, the matrix $\sum_{t=1}^{T} \mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})']$ is invertible (full rank).

Theorem 12.3 (FE/FD identification)**.** *Given the structural model in* (12.1) *with coefficient vector* $\boldsymbol{\beta}$*, let strict exogeneity Assumption A12.2 hold. i) If additionally Assumption A12.4 holds, then* $\boldsymbol{\beta}$ *is identified and equal to*

$$\boldsymbol{\beta} = \left[\sum_{t=2}^{T} \mathrm{E}(\Delta \boldsymbol{X}_t \Delta \boldsymbol{X}_t')\right]^{-1} \sum_{t=2}^{T} \mathrm{E}(\Delta \boldsymbol{X}_t \Delta Y_t).$$

ii) If Assumption A12.5 holds instead of A12.4, then $\boldsymbol{\beta}$ *is identified and equal to*

$$\boldsymbol{\beta} = \left\{\sum_{t=1}^{T} \mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})']\right\}^{-1} \sum_{t=1}^{T} \mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(Y_t - \bar{Y})],$$

using notation $\bar{Y} \equiv \frac{1}{T}\sum_{t=1}^{T} Y_t$ *and* $\bar{\boldsymbol{X}} \equiv \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{X}_t$*.*

Proof. i) Applying the first-difference transformation to the structural model in (12.1) yields $\Delta Y_t = \Delta \boldsymbol{X}_t' \boldsymbol{\beta} + \Delta V_t$, and plugging that into the expression provided in Theorem 12.3,

$$\begin{aligned}
&\left[\sum_{t=2}^{T} \mathrm{E}(\Delta \boldsymbol{X}_t \Delta \boldsymbol{X}_t')\right]^{-1} \sum_{t=2}^{T} \mathrm{E}(\Delta \boldsymbol{X}_t \Delta Y_t) \\
&= \left[\sum_{t=2}^{T} \mathrm{E}(\Delta \boldsymbol{X}_t \Delta \boldsymbol{X}_t')\right]^{-1} \sum_{t=2}^{T} \mathrm{E}[\Delta \boldsymbol{X}_t (\Delta \boldsymbol{X} t' \boldsymbol{\beta} + \Delta V_t)] \\
&= \boldsymbol{\beta} + \left[\sum_{t=2}^{T} \mathrm{E}(\Delta \boldsymbol{X}_t \Delta \boldsymbol{X}_t')\right]^{-1} \sum_{t=2}^{T} \mathrm{E}[\Delta \boldsymbol{X}_t \Delta V_t].
\end{aligned} \tag{12.8}$$

For any $t = 2,\ldots,T$,

$$\begin{aligned}
\mathrm{E}[\Delta \boldsymbol{X}_t \Delta V_t] &= \mathrm{E}[(\boldsymbol{X}_t - \boldsymbol{X}_{t-1})(V_t - V_{t-1})] \\
&= \mathrm{E}[\boldsymbol{X}_t V_t - \boldsymbol{X}_t V_{t-1} - \boldsymbol{X}_{t-1} V_t + \boldsymbol{X}_{t-1} V_{t-1}] \\
&= \mathrm{E}[\boldsymbol{X}_t V_t] - \mathrm{E}[\boldsymbol{X}_t V_{t-1}] - \mathrm{E}[\boldsymbol{X}_{t-1} V_t] + \mathrm{E}[\boldsymbol{X}_{t-1} V_{t-1}] \\
&= \mathbf{0} - \mathbf{0} - \mathbf{0} + \mathbf{0}
\end{aligned}$$

by Assumption A12.2. Plugging this back into (12.8), the far-right term is zero, which zeroes out that entire term assuming the matrix inverse indeed exists (as in Assumption A12.4), leaving only the structural $\boldsymbol{\beta}$ as desired.

ii) Applying the "within transformation" to the structural model in (12.1) yields

$$Y_t - \bar{Y} = (\boldsymbol{X}_t - \bar{\boldsymbol{X}})'\boldsymbol{\beta} + (V_t - \bar{V}),$$

and plugging that into the expression provided in Theorem 12.3,

$$\begin{aligned}
&\left\{\sum_{t=1}^{T} \mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})']\right\}^{-1} \sum_{t=1}^{T} \mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(Y_t - \bar{Y})] \\
=&\left\{\sum_{t=1}^{T} \mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})']\right\}^{-1} \sum_{t=1}^{T} \mathrm{E}\{(\boldsymbol{X}_t - \bar{\boldsymbol{X}})[\overbrace{(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'\boldsymbol{\beta} + (V_t - \bar{V})}^{=Y_t - \bar{Y}}]\} \\
&= \boldsymbol{\beta} + \left\{\sum_{t=1}^{T} \mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})']\right\}^{-1} \sum_{t=1}^{T} \mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(V_t - \bar{V})]. \qquad (12.9)
\end{aligned}$$

Expanding the final summand (for any $t = 1, \ldots, T$) and using the linearity property of the expectation operator,

$$\begin{aligned}
\mathrm{E}[(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(V_t - \bar{V})] &= \mathrm{E}[(\boldsymbol{X}_t - \frac{1}{T}\sum_{s=1}^{T} \boldsymbol{X}_s)(V_t - \frac{1}{T}\sum_{s=1}^{T} V_s)] \\
&= \overbrace{\mathrm{E}(\boldsymbol{X}_t V_t)}^{=\mathbf{0}} - \frac{1}{T}\sum_{s=1}^{T} \overbrace{\mathrm{E}(\boldsymbol{X}_t V_s)}^{=\mathbf{0}} - \frac{1}{T}\sum_{s=1}^{T} \overbrace{\mathrm{E}(\boldsymbol{X}_s V_t)}^{=\mathbf{0}} \\
&\quad + \frac{1}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T} \overbrace{\mathrm{E}(\boldsymbol{X}_r V_s)}^{=\mathbf{0}} \\
&= \mathbf{0},
\end{aligned}$$

where all the zeros are due to Assumption A12.2. Plugging this back into (12.8), the far-right term is zero, which zeroes out that entire term assuming the matrix inverse indeed exists (as in Assumption A12.5), leaving only the structural $\boldsymbol{\beta}$ as desired. □

The derivation of asymptotic properties (consistency and asymptotic normality) is complicated somewhat by the multiple time periods, but the general strategy is the same as for OLS: write the centered (or centered and scaled) estimator, then use exogeneity to apply a WLLN or CLT to the term involving the idiosyncratic error V_{it}. Because T is finite and does not change with n, we just need to sum over t but otherwise use the standard WLLN/CLT as $n \to \infty$.

12.6 Other Approaches

The main alternative to the FE approach is the **correlated random effects** (CRE) approach, going back (at least) to Mundlak (1978). As with FE, CRE allows C to be correlated with $\boldsymbol{X}$. However, unlike FE, the dependence is restricted in some way, often by assuming a particular model of the dependence that can be estimated. This is particularly helpful with "nonlinear" models for which the FE or FD transformation cannot be used to remove the unobserved C from the model. It can also be used to allow time-invariant regressors, unlike FE/FD.

The Hausman and Taylor (1981) model can also allow for some time-invariant regressors as long as they are uncorrelated with the fixed effect; for example, see Section 11.3 of Wooldridge (2010).

12.7 Other Considerations

There have been many methodological papers recently about two-way fixed effects and difference-in-differences. One consideration is what to do when there is a binary treatment but different individuals begin treatment in different periods, as addressed by Callaway and Sant'Anna (2021b). If you end up using TWFE/DiD for research, you would want to learn more about the latest developments. Thankfully, many authors now provide Stata and/or R packages implementing their methodology, such as the Stata package `csdid` by Callaway, Rios-Avila, and Sant'Anna (2021) or the R package `did` by Callaway and Sant'Anna (2021a).

12.8 Cluster-Robust Standard Errors

With panel data, you almost always want to use **cluster-robust standard errors**. Similar to heteroskedasticity-robust standardard errors, this means you want standard errors that are still accurate even when you have clusters of related observations, like for all the time periods for a given individual. Other related terms include "clustering your standard errors" or "clustered standard errors," or "clustered sampling." Unfortunately, "clustering" has a completely different meaning in the world of statistics and machine learning (Wikipedia link here).

12.8.1 Types of Sampling

⟹ Kaplan video: Types of Sampling

For intuition, imagine we want to estimate the population mean using a sample of four observations. It helps me to think of four empty buckets; our sampling procedure with fill each bucket with a single number (realization) from the population. If we have iid

sampling, then each bucket takes a realization from the full population, without regard for the other buckets or their values.

If we have **stratified sampling**, then before sampling we label each bucket with the name of a particular group (like undergraduate or graduate student), and each bucket takes a realization from the corresponding subpopulation (**stratum**; plural **strata**), not the full population. Compared with iid, we could still sample independently, but these bucket labels violate the "identically distributed" property, assuming our variable has a different distribution in the two subpopulations (like GPA for undergraduate and graduate students). The benefit is that we can better enforce that the sample is representative of the population, at least in terms of group proportions, or we can "over-sample" certain groups who may be difficult to study if we simply took an iid sample; for example, if we randomly sample 50 people in Columbia then we may have zero unhoused individuals, which is not helpful if our research question pertains to that subpopulation.

If we have **clustered sampling**, then before sampling we might label buckets 1 and 2 as coming from the same individual (in different time periods) and label buckets 3 and 4 as coming from another individual (in different periods). That is, before sampling, we do not know which individual will be sampled for buckets 1 and 2, but we know that it will be the same individual, so the buckets are linked together.

The following examples are modified from Section 3.2 of Kaplan (2022a), which includes additional discussion of iid, stratified, and clustered sampling.

Example 12.1 (Kaplan video)**.** Imagine randomly picking a Mizzou student ID number, then randomly picking a 2nd, then 3rd, then 4th. The corresponding Y_i are both independent and identically distributed (iid). They are independent because each ID number is randomly drawn without any consideration of how the other numbers are drawn, and without any consideration of the other observed Y_i values. They are identically distributed because each ID number is drawn from the same population (anyone who has a Mizzou student ID).

Example 12.2 (Kaplan video)**.** The following procedure illustrates stratified sampling. Each Mizzou student is classified as either a resident of Missouri ("in-state") or not ("non-resident"). Imagine buckets 1 and 2 say "in-state," while buckets 3 and 4 say "non-resident": observations Y_1 and Y_2 are from in-state students, while Y_3 and Y_4 are from non-resident students. For most variables, the in-state distribution differs from the non-resident distribution, so the distribution of Y_1 and Y_2 (in-state) differs from the distribution of Y_3 and Y_4 (non-resident). Thus, sampling is not identically distributed. The observations could still be independently sampled, although they may not be.

Example 12.3 (Kaplan video)**.** The following procedure illustrates clustered sampling with panel data. Imagine randomly picking two Mizzou students (like with random ID numbers), then observing them this semester and next semester. For example, imagine bucket 1 contains the first student's GPA this semester, bucket 2 contains the same student's GPA next semester, and buckets 3 and 4 contain the other student's GPAs from this semester and next semester. Buckets 1 and 2 (Y_1 and Y_2) are probably both

high or both low, rather than one high and one low, and similarly for buckets 3 and 4 (Y_3 and Y_4). That is, buckets 1 and 2 are correlated, and 3 and 4 are correlated.

Example 12.4 (Kaplan video)**.** The following procedure illustrates clustered sampling without panel data. Imagine randomly picking a class (like my intro econometrics class) at Mizzou, and filling the first two buckets (Y_1 and Y_2) with two random students from that class; then randomly picking another class (like intro philosophy), and another two students for the other buckets (Y_3 and Y_4). Observations are identically distributed (because each Y_i has the same probability of getting any particular student in the population) but probably not independent. For example, dependence may come from students in the same class being similarly affected by their shared experience. Here, buckets 1 and 2 are correlated, and 3 and 4 are correlated, but not 1 and 3, nor 2 and 4, etc.

Discussion Question 12.3 (sampling types)**.** For each of the following, explain why you think sampling is clustered, stratified, iid, or something else or not sure.

a) To decrease the per-respondent survey cost, a survey team in a rural area randomly selects five villages and then surveys each individual within that village.
b) In a survey of 1000 respondents, to help the sample be representative of the full US population, a survey team decides to have the first 19 individuals randomly sampled from Missouri, and the other 981 randomly sampled from other states.
c) When you load a survey dataset with $n = 20$ individuals, you notice exactly half are female.
d) You randomly select 40 firms from the Fortune 500 and track their monthly sales revenue over one calendar year.

12.8.2 SE for Panel Regression

Below, first intuition is developed in the simplest intercept-only model; then POLS and FE/FD results are given.

Notation reminder: because we have iid sampling across individuals, the i subscript is dropped when writing population random variables. For example, in the population, we have Y_t and $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$. Because sampling is iid over i, for any i, $\mathrm{E}(Y_{it}) = \mathrm{E}(Y_t)$, and similarly $\mathrm{E}(\boldsymbol{Y}_i) = \mathrm{E}(\boldsymbol{Y})$. However, we do not have iid across t, so dropping the t subscript has a different meaning. Specifically, it refers to aggregating observations of $t = 1, \ldots, T$ into a single vector or matrix. For example, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$ stacks individual i's Y_{it} values, or $\underline{\boldsymbol{X}}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT})'$ stacks individual i's regressor vectors X_{it}'.

Intercept-Only Model

To develop intuition, first consider POLS with $X_{it} = 1$, meaning we are simply trying to estimate the population mean $\mu \equiv \mathrm{E}(Y)$, assuming it is the same for each $t = 1, \ldots, T$,

meaning $\mathrm{E}(Y_{it}) = \mu$ for each t (and i). The estimator is

$$\hat{\mu} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T} Y_{it}. \tag{12.10}$$

Centering and scaling,

$$\sqrt{n}(\hat{\mu} - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \mu). \tag{12.11}$$

As we continue to use the fixed-T asymptotics (with $n \to \infty$), for simplicity let $T = 2$, so

$$\sqrt{n}(\hat{\mu} - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{2}[(Y_{i1} - \mu) + (Y_{i2} - \mu)]. \tag{12.12}$$

To connect more explicitly with familiar results, we can define a mean-zero random variable indexed only by i (not t). Specifically, let $Z_i \equiv (Y_{i1} - \mu) + (Y_{i2} - \mu)$, and recall that we (assume we) sample individuals i iid from the population. Thus, regardless of correlation of Y_{it} across t, the Z_i are iid with mean zero. Dropping the i subscript to write Z as the representative population individual,

$$\mathrm{E}(Z) = \mathrm{E}[(Y_1 - \mu) + (Y_2 - \mu)] = \overbrace{\mathrm{E}(Y_1)}^{=\mu} - \mu + \overbrace{\mathrm{E}(Y_2)}^{=\mu} - \mu = 0. \tag{12.13}$$

Thus, we can apply the Lindeberg–Lévy CLT. Keeping the $1/T = 1/2$ outside to better parallel regression results later:

$$\sqrt{n}(\hat{\mu} - \mu) = \frac{1}{2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \xrightarrow{d} \frac{1}{2}\mathrm{N}\big(0, \mathrm{Var}(Z)\big), \tag{12.14}$$

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{E}(Z^2) = \mathrm{E}\big\{[(Y_1 - \mu) + (Y_2 - \mu)]^2\big\} \\ &= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + 2\,\mathrm{Cov}(Y_1, Y_2).\end{aligned} \tag{12.15}$$

If we had iid sampling across both i and t, then the covariance would be zero, and this would simplify to $\mathrm{Var}(Y_t)/2$, where the $1/2$ is because earlier we only scaled by $\sqrt{n}$ even though total we have $2n$ observations. But with panel data, usually $\mathrm{Cov}(Y_1, Y_2) > 0$, in which case iid-based standard errors would be too small (and thus confidence intervals too short, making it seem like less uncertainty than there really is).

To estimate $\mathrm{Var}(Z) = \mathrm{E}(Z^2)$, we can plug in our estimator $\hat{\mu}$ to get

$$\widehat{\mathrm{Var}}(Z) = \widehat{\mathrm{E}}(Z^2) = \frac{1}{n}\sum_{i=1}^{n} Z_i^2 = \frac{1}{n}\sum_{i=1}^{n}[(Y_{i1} - \hat{\mu}) + (Y_{i2} - \hat{\mu})]^2. \tag{12.16}$$

We could also think of this in terms of residuals in the model $Y_{it} = \mu + V_{it}$. The error term is $V_{it} = Y_{it} - \mu$, so the residuals are $\hat{V}_{it} = Y_{it} - \hat{\mu}$, and the variance estimator is equivalently

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{V}_{i1} + \hat{V}_{i2})^2. \tag{12.17}$$

For general T, we can write results more compactly in vector notation if we also explicitly write the "regressor" $X_{it} = 1$. Let $\boldsymbol{V}_i \equiv (V_{i1}, \ldots, V_{iT})'$ and $\boldsymbol{X}_i \equiv (X_{i1}, \ldots, X_{iT})' = (1, \ldots, 1)'$. Note

$$Z_i = \sum_{t=1}^{T} V_{it} = \boldsymbol{X}_i'\boldsymbol{V}_i = \boldsymbol{V}_i'\boldsymbol{X}_i.$$

Thus, the variance (second moment) of Z and its estimator are

$$\mathrm{Var}(Z) = \mathrm{E}\{ \overbrace{\boldsymbol{X}'\boldsymbol{V}}^{=\sum_{t=1}^{T} V_{it}} \overbrace{\boldsymbol{V}'\boldsymbol{X}}^{=\sum_{t=1}^{T} V_{it}} \}, \quad \widehat{\mathrm{Var}}(Z) = \widehat{\mathrm{E}}(Z^2) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{\boldsymbol{X}_i'\hat{\boldsymbol{V}}_i}^{=\sum_{t=1}^{T} \hat{V}_{it}} \overbrace{\hat{\boldsymbol{V}}_i'\boldsymbol{X}_i}^{=\sum_{t=1}^{T} \hat{V}_{it}}. \tag{12.18}$$

Because Z_i is a scalar, we could write the expressions in multiple equivalent ways, but $\boldsymbol{X}_i'\boldsymbol{V}_i\boldsymbol{V}_i'\boldsymbol{X}_i$ parallels the more general regression case below.

POLS Regression

Consider again the POLS estimator and notation from Section 12.2. Here we assume U_{it} is a linear projection error and $\boldsymbol{\beta}$ is the linear projection coefficient vector; whether or not $\boldsymbol{\beta}$ also has a causal interpretation is irrelevant for the following. The centered and scaled estimator was given in (12.4). Swapping the summations over i and t,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{POLS}} - \boldsymbol{\beta}) = \left[\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\boldsymbol{X}_{it}\boldsymbol{X}_{it}'\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T}\boldsymbol{X}_{it}U_{it}. \tag{12.19}$$

As in the intercept-only case, given that T is fixed as $n \to \infty$, we can think of $\sum_{t=1}^{T}\boldsymbol{X}_{it}\boldsymbol{X}_{it}'$ as a single matrix-valued random variable associated with i, to which a WLLN applies when we average over $i = 1, \ldots, n$. That is, with $\boldsymbol{X}_t$ the representative population individual's period-t regressor vector,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\boldsymbol{X}_{it}\boldsymbol{X}_{it}' \xrightarrow{p} \mathrm{E}\left(\sum_{t=1}^{T}\boldsymbol{X}_t\boldsymbol{X}_t'\right) = \sum_{t=1}^{T}\mathrm{E}(\boldsymbol{X}_t\boldsymbol{X}_t'). \tag{12.20}$$

Similarly, we can think of $\boldsymbol{Z}_i \equiv \sum_{t=1}^{T}\boldsymbol{X}_{it}U_{it}$ as a mean-zero random vector that depends only on i (not t) and is thus iid and admits a CLT. This is the generalization of

$Z_i = \sum_{t=1}^{T} V_{it}$ from the intercept-only model (i.e., where $X_{it} = 1$). Thus,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{t=1}^{T} \boldsymbol{X}_{it} U_{it} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{Z}_i \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{\Omega}}), \tag{12.21}$$

$$\underline{\boldsymbol{\Omega}} = \mathrm{Var}(\boldsymbol{Z}) = \mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}') = \mathrm{E}\left[\left(\sum_{t=1}^{T} \boldsymbol{X}_t U_t \right) \left(\sum_{t=1}^{T} \boldsymbol{X}_t U_t \right)' \right]. \tag{12.22}$$

For example, with $T = 2$, this would be

$$\mathrm{E}(\boldsymbol{X}_1 \boldsymbol{X}_1' U_1^2) + \mathrm{E}(\boldsymbol{X}_1 \boldsymbol{X}_2' U_1 U_2) + \mathrm{E}(\boldsymbol{X}_2 \boldsymbol{X}_1' U_2 U_1) + \mathrm{E}(\boldsymbol{X}_2 \boldsymbol{X}_2' U_2^2). \tag{12.23}$$

If we erroneously assume independence across t, then the middle two terms zero out. Thus, if we only use heteroskedasticity-robust standard errors (that allow heteroskedasticity but still assume independence across each observation), generally our confidence intervals will not have the desired coverage probability, even with large samples (asymptotically).

As in the intercept-only case, we can estimate $\underline{\boldsymbol{\Omega}}$ by first computing residuals $\hat{U}_{it} = Y_{it} - \boldsymbol{X}_{it}' \hat{\boldsymbol{\beta}}$ and then plugging into the sample version of the formula for $\underline{\boldsymbol{\Omega}}$ (replacing E with $\widehat{\mathrm{E}}$).

The expressions for $\underline{\boldsymbol{\Omega}}$ and $\hat{\underline{\boldsymbol{\Omega}}}$ can be written more compactly with matrix notation. Let

$$\boldsymbol{U}_i \equiv (U_{i1}, \ldots, U_{iT})', \quad \underline{\boldsymbol{X}}_i \equiv (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT})'. \tag{12.24}$$

Thus,

$$\boldsymbol{Z}_i = \underline{\boldsymbol{X}}_i' \boldsymbol{U}_i, \quad \boldsymbol{Z}_i \boldsymbol{Z}_i' = \underline{\boldsymbol{X}}_i' \boldsymbol{U}_i \boldsymbol{U}_i' \underline{\boldsymbol{X}}_i, \tag{12.25}$$

and

$$\underline{\boldsymbol{\Omega}} = \mathrm{E}(\underline{\boldsymbol{X}}' \boldsymbol{U} \boldsymbol{U}' \underline{\boldsymbol{X}}), \quad \hat{\underline{\boldsymbol{\Omega}}} = \frac{1}{n} \sum_{i=1}^{n} \underline{\boldsymbol{X}}_i' \hat{\boldsymbol{U}}_i \hat{\boldsymbol{U}}_i' \underline{\boldsymbol{X}}_i. \tag{12.26}$$

For completeness, although it is not particularly important, the following result combines the above components into the full asymptotic distribution.

Theorem 12.4 (panel POLS asymptotic normality)**.** *Let* $\mathrm{LP}(Y_{it} \mid \boldsymbol{X}_{it}) = \boldsymbol{X}_{it}' \boldsymbol{\beta}$ *and* $U_{it} \equiv Y_{it} - \mathrm{LP}(Y_{it} \mid \boldsymbol{X}_{it})$*. Assume individuals* i *are sampled iid from the population, but dependence across* t *is not restricted. Assuming the needed population moments are finite and that the matrix inverse exists,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{POLS} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{\Sigma}}),$$

$$\underline{\boldsymbol{\Sigma}} \equiv \left[\sum_{t=1}^{T} \mathrm{E}(\boldsymbol{X}_t \boldsymbol{X}_t') \right]^{-1} \mathrm{E}\left[\left(\sum_{t=1}^{T} \boldsymbol{X}_t U_t \right) \left(\sum_{t=1}^{T} \boldsymbol{X}_t U_t \right)' \right] \left[\sum_{t=1}^{T} \mathrm{E}(\boldsymbol{X}_t \boldsymbol{X}_t') \right]^{-1}.$$

Proof. From (12.19),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{POLS}} - \boldsymbol{\beta}) = \left[\overbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\boldsymbol{X}_{it}\boldsymbol{X}_{it}'}^{\xrightarrow{p}\sum_{t=1}^{T}\mathrm{E}(\boldsymbol{X}_t\boldsymbol{X}_t') \text{ by (12.20)}} \right]^{-1} \overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T}\boldsymbol{X}_{it}U_{it}}^{\xrightarrow{d}\mathrm{N}(\mathbf{0},\underline{\boldsymbol{\Omega}}) \text{ by (12.21)}} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Sigma}}),$$

with $\underline{\boldsymbol{\Sigma}}$ as defined in Theorem 12.4. □

FE/FD Regression

FE/FD regression is essentially the same as POLS, except after the within transformation or first-difference transformation. That is, we get the same result as Theorem 12.4, but either replacing $\boldsymbol{X}_t$ with $\boldsymbol{X}_t - \bar{\boldsymbol{X}}$ and replacing U_t with $V_t - \bar{V}$ (for FE), or replacing $\boldsymbol{X}_t$ with $\Delta\boldsymbol{X}_t$ and replacing U_t with ΔV_t and replacing $\sum_{t=1}^{T}$ with $\sum_{t=2}^{T}$ (for FD). Practically, the main point is that correlation across t adds terms to the asymptotic variance that would be zero if we have independence across t (iid over it), so it is important to use cluster-robust standard errors with FE/FD.

Stata

If you have already run `xtset` to tell Stata your panel variables (i and t variables), then adding `vce(robust)` to any regression command computes analytic cluster-robust standard errors, using the i variable as the cluster variable. You can also specify a different cluster variable (or if `xtset` has not been run) with `vce(cluster` *clustervar*`)`.

Beyond our scope...

The analytic cluster-robust SE in Stata may not be accurate if there are few and/or very different (heterogeneous) clusters, and sometimes you may need "two-way" (or multi-way) clustering in multiple dimensions. In that case, there are other methods that can perform better, like randomization inference (permutation tests) or various bootstraps. On Colin Cameron's website,[1] there are some very detailed slides from November 2022 about many different aspects of clustered standard errors (including some Stata commands), and there is a corresponding survey paper with Doug Miller coming in 2023(?).

Chapter 13
Dynamic Panel Models

Unit learning objectives for this chapter

13.1. Define concepts and terms related to dynamic panel models, both mathematically and intuitively. [TLOs 1–3]

13.2. Judge whether or not identifying assumptions hold in specific real-world examples. [TLO 4]

This chapter introduces the idea of a dynamic panel data regression model, the failure of the FE approach in Chapter 12, and a related approach that can work.

A **dynamic panel model** includes the lagged outcome Y_{it-1} as a regressor. (It could include additional lags, too.) This contrasts with the **static panel model** from Chapter 12.

The inclusion of Y_{it-1} violates strict exogeneity, so we need another type of exogeneity. Although weaker, this exogeneity still gives us moment conditions that we can use to estimate the structural parameters.

13.1 Types of Exogeneity

Besides contemporaneous exogeneity and strict exogeneity (A12.1 and A12.2), there is also **sequential exogeneity**. Recall that "contemporaneous" only requires that V_t is uncorrelated with the same time period's $\boldsymbol{X}_t$, whereas "strict" requires that V_t is uncorrelated with all past, present, and even future $\boldsymbol{X}_s$, $s = 1, \ldots, T$. Sequential is in between, requiring that V_t is uncorrelated with past and present but not future $\boldsymbol{X}_s$, $s = 1, \ldots, t$. This is useful in cases where we may suspect that next periods $\boldsymbol{X}_{t+1}$ is chosen partly in consideration of this period's V_t, possibly through the effect of V_t on Y_t. This is directly the case with a dynamic panel, where $\boldsymbol{X}_{t+1}$ includes Y_t, which is clearly affected by V_t.

Assumption A13.1 (sequential exogeneity)**.** Error term V_t is sequentially exogenous: $\mathrm{E}(\boldsymbol{X}_s V_t) = \mathbf{0}$ for any $t = 1, \ldots, T$ and any $s = 1, \ldots, t$.

Discussion Question 13.1 (panel exogeneity)**.** Consider the three types of exogeneity: contemporaneous (A12.1), strict (A12.2), and sequential (A13.1).

a) Using the notation/terms of Section 2.1, write the relationships among the three assumptions in terms of $\implies$ and (separately) in terms of "weaker than."
b) Practically, what do these relationships mean, in terms of when we can apply result in practice? For example, if I have two estimators, one of which requires strict exogeneity and one of which requires sequential exogeneity, then what's the relationship between empirical settings in which the first estimator is valid vs. settings in which the second is valid?
c) Can POLS be used in more settings than FE? Explain.

Discussion Question 13.2 (AR(1) exog)**.** Consider the AR(1) model $Y_t = \rho Y_{t-1} + V_t$, $V_t \overset{iid}{\sim} \mathrm{N}(0, 1)$, where V_t is independent of all past values $Y_{t-1}, Y_{t-2}, \ldots$, and $|\rho| < 1$. Define $X_t \equiv Y_{t-1}$. Show the following mathematically (verbal explanation optional).

a) Does contemporaneous exogeneity (Assumption A12.1) hold? Why/not?
b) Does sequential exogeneity (Assumption A13.1) hold? Why/not?
c) Does strict exogeneity (Assumption A12.2) hold? Why/not?

Even with a static panel model, sequential exogeneity may seem more plausible than strict exogeneity, like in Example 13.1.

Example 13.1 (exogeneity: R&D)**.** Imagine Y_t is the number of patents granted to a firm in year t, X_t is their R&D spending in year $t-1$ (assuming some lag between the expenditure and the tangible product), and V_t is idiosyncratic shocks to patents. If we think that having a positive V_t makes a firm want to invest more in R&D, then even if V_t is unrelated to past R&D spending (sequential exogeneity holds), it could be correlated with future spending, which violates strict exogeneity.

13.2 FE/FD Failure in Simple Model

Consider the simple model from DQ 13.2,

$$Y_t = \rho X_t + V_t, \quad X_t = Y_{t-1}, \quad t = 1, \ldots, T, \tag{13.1}$$

where V_t is independent of all past values $Y_{t-1}, Y_{t-2}, \ldots$ and thus all current and past $X_t, X_{t-1}, \ldots$. Note we observe Y_0 so that we can observe X_t for all $t = 1, \ldots, T$. Recall $\bar{Y} \equiv \frac{1}{T}\sum_{t=1}^{T} Y_t$, $\bar{X} \equiv \frac{1}{T}\sum_{t=1}^{T} X_t$, and $\bar{V} \equiv \frac{1}{T}\sum_{t=1}^{T} V_t$. The FE estimator regresses $Y_t - \bar{Y}$ on $X_t - \bar{X}$, based on the within-transformed model

$$Y_t - \bar{Y} = \rho(X_t - \bar{X}) + (V_t - \bar{V}). \tag{13.2}$$

For ρ to be the linear projection coefficient, $V_t - \bar{V}$ must satisfy the linear projection error property

$$0 = \mathrm{E}[(X_t - \bar{X})(V_t - \bar{V})] = \mathrm{E}\left[\left(Y_{t-1} - \frac{1}{T}\sum_{t=0}^{T-1} Y_t\right)\left(V_t - \frac{1}{T}\sum_{t=1}^{T} V_t\right)\right]. \tag{13.3}$$

However, this includes terms like

$$\mathrm{E}(Y_1 V_1) = \mathrm{E}[(\rho Y_0 + V_1)V_1] = \rho \overbrace{\mathrm{E}(Y_0 V_1)}^{=0} + \mathrm{E}(V_1^2) = \mathrm{Var}(V_1) > 0. \tag{13.4}$$

In general, for any $s \geq t$, $\mathrm{E}(Y_s V_t) \neq 0$.

Intuitively, the problem is that the within transformation includes all the V_t ($t = 1, \ldots, T$) and most of the Y_t ($t = 0, \ldots, T-1$) on the RHS, so it requires strict exogeneity to zero out all the cross-terms.

The FD transformation has a less egregious version of the same problem. Again with $X_t = Y_{t-1}$,

$$\begin{aligned}
\mathrm{E}[\Delta X_t \Delta V_t] &= \mathrm{E}[(X_t - X_{t-1})(V_t - V_{t-1})] \\
&= \mathrm{E}(X_t V_t) - \mathrm{E}(X_t V_{t-1}) - \mathrm{E}(X_{t-1} V_t) + \mathrm{E}(X_{t-1} V_{t-1}) \\
&= \overbrace{\mathrm{E}(Y_{t-1} V_t)}^{=0} - \overbrace{\mathrm{E}(Y_{t-1} V_{t-1})}^{\neq 0} - \overbrace{\mathrm{E}(Y_{t-2} V_t)}^{=0} + \overbrace{\mathrm{E}(Y_{t-2} V_{t-1})}^{=0} \\
&= -\mathrm{E}(Y_{t-1} V_{t-1}) = -\mathrm{Var}(V_{t-1}) < 0
\end{aligned} \tag{13.5}$$

using (13.4). Thus, both the standard FE and FD approaches fail to identify the structural $\boldsymbol{\beta}$ for dynamic models that include Y_{t-1} as a regressor.

13.3 Moment Conditions

Although we cannot use the standard FD "moment conditions"

$$\mathbf{0} = \mathrm{E}(\Delta \boldsymbol{X}_t \Delta V_t) = \mathrm{E}[\Delta \boldsymbol{X}_t (\Delta Y_t - \Delta \boldsymbol{X}_t' \boldsymbol{\beta})] \tag{13.6}$$

due to the failure of strict exogeneity in dynamic panel models, we can use sequential exogeneity to find alternative moment conditions. Then, as in Chapter 10, we can use GMM for estimation and inference.

Initial work on this approach came from Anderson and Hsiao (1982), Holtz-Eakin, Newey, and Rosen (1988), and Arellano and Bond (1991).

We first use the FD transformation to remove the fixed effects, then see which moment conditions are implied by sequential exogeneity. These are written in terms of ΔV_t, which in practice is replaced by $\Delta Y_t - \Delta \boldsymbol{X}_t' \boldsymbol{\beta}$ to get moment conditions in terms of only model parameters and observable variables.

Consider which regressors are uncorrelated with the period $t = 2$ first-differenced idiosyncratic error, $\Delta V_2 = V_2 - V_1$. Recall that sequential exogeneity (A13.1) says $\mathrm{E}(\boldsymbol{X}_s V_t) = \mathbf{0}$ for any $s \leq t$, so $\mathrm{E}(\boldsymbol{X}_1 V_2) = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{X}_1 V_1) = \mathbf{0}$, but $\mathrm{E}(\boldsymbol{X}_s V_1)$ for $s > 1$ are not assumed to be zero. Thus, the only moment condition for ΔV_2 is

$$\mathrm{E}(\boldsymbol{X}_1 \Delta V_2) = \mathrm{E}[\boldsymbol{X}_1 (V_2 - V_1)] = \overbrace{\mathrm{E}(\boldsymbol{X}_1 V_2)}^{=\mathbf{0}} - \overbrace{\mathrm{E}(\boldsymbol{X}_1 V_1)}^{=\mathbf{0}}. \tag{13.7}$$

At $t = 3$, there are more moment conditions:

$$\mathrm{E}[\boldsymbol{X}_1 \Delta V_3] = \overbrace{\mathrm{E}(\boldsymbol{X}_1 V_3)}^{=\mathbf{0}} - \overbrace{\mathrm{E}(\boldsymbol{X}_1 V_2)}^{=\mathbf{0}} = \mathbf{0}, \tag{13.8}$$

$$\mathrm{E}[\boldsymbol{X}_2 \Delta V_3] = \overbrace{\mathrm{E}(\boldsymbol{X}_2 V_3)}^{=\mathbf{0}} - \overbrace{\mathrm{E}(\boldsymbol{X}_2 V_2)}^{=\mathbf{0}} = \mathbf{0}, \tag{13.9}$$

$$\mathrm{E}[\Delta \boldsymbol{X}_2 \Delta V_3] = \overbrace{\mathrm{E}(\boldsymbol{X}_2 V_3)}^{=\mathbf{0}} - \overbrace{\mathrm{E}(\boldsymbol{X}_2 V_2)}^{=\mathbf{0}} - \overbrace{\mathrm{E}(\boldsymbol{X}_1 V_3)}^{=\mathbf{0}} + \overbrace{\mathrm{E}(\boldsymbol{X}_1 V_2)}^{=\mathbf{0}} = \mathbf{0}. \tag{13.10}$$

The larger t is, the more $\boldsymbol{X}_s$ and $\Delta \boldsymbol{X}_s$ there are uncorrelated with ΔV_t, due to the nature of sequential exogeneity. Generally, for any $t \geq 3$,

$$\mathbf{0} = \mathrm{E}(\boldsymbol{X}_s \Delta V_t) = \mathrm{E}(\boldsymbol{X}_{s-1} \Delta V_t) = \mathrm{E}(\Delta \boldsymbol{X}_s \Delta V_t), \quad s = 2, \ldots, t-1. \tag{13.11}$$

To think about interpreting this structure of moment condition, recall the linear projection and IV regression moment conditions. For LP,

$$\mathbf{0} = \mathrm{E}[\boldsymbol{X} V] = \mathrm{E}[\boldsymbol{X}(Y - \boldsymbol{X}'\boldsymbol{\beta})],$$

which yields the usual LPC formula when solved for $\boldsymbol{\beta}$. If $\boldsymbol{X}$ includes endogenous regressors that are replaced by excluded instruments in the full instrument vector $\boldsymbol{Z}$, then the moment conditions become

$$\mathbf{0} = \mathrm{E}[\boldsymbol{Z} V] = \mathrm{E}[\boldsymbol{Z}(Y - \boldsymbol{X}'\boldsymbol{\beta})].$$

Compare this structure with one of the moment conditions in (13.11) like

$$\mathbf{0} = \mathrm{E}[\boldsymbol{X}_s \Delta V_t] = \mathrm{E}[\boldsymbol{X}_s (\Delta Y_t - \Delta \boldsymbol{X}_t' \boldsymbol{\beta})].$$

Compared to the IV moment conditions, $\boldsymbol{X}_s$ plays the role of the IV $\boldsymbol{Z}$ that intruments for the regressor vector $\Delta \boldsymbol{X}_t$ in the regression of ΔY_t on $\Delta \boldsymbol{X}_t$. Indeed, this approach to estimating dynamic panel models is sometimes called FD–IV.

Considering the above can also help us think about which of the many available moment conditions are most helpful for estimation. Recall from Section 9.2 that problems with both estimation and inference can arise with weak instruments or generally weak identification. Given the above interpretation, we want our "instruments" like $\boldsymbol{X}_s$ or $\Delta \boldsymbol{X}_s$ to be strongly correlated with "regressors" $\Delta \boldsymbol{X}_t$. Usually as $t - s$ grows, this correlation

weakens, so even if technically $\mathrm{E}(\boldsymbol{X}_1 \Delta V_9) = \mathbf{0}$, we may not want to include that moment condition in practice. Further, if $\boldsymbol{X}_t$ is strongly persistent, then $\Delta \boldsymbol{X}_s$ may be a relatively weak "instrument." For example, in the extreme case where scalar X_t is a random walk, the first difference ΔX_t is an iid (or white noise) sequence of random variables, in which case ΔX_{t-1} is independent of ΔX_t and is thus not a "relevant" instrument. Often it is reasonable to use (only) $\boldsymbol{X}_{t-1}$ and $\boldsymbol{X}_{t-2}$ in the moment conditions involving ΔV_t, although other strategies may be better given knowledge about a particular empirical setting.

In Stata, for many typical dynamic panel models, `xtabond` will suffice. It also supports two-step estimation with option `twostep` and (bias-corrected) cluster-robust standard errors with option `vce(robust)`.

Exercises

Exercise III.1. You will analyze data on driving laws and fatal accident rates, originally from Freeman (2007). In particular, you'll compare weekend driving fatality (death) rates for states that adopted a 0.08 blood alcohol content (BAC) law and states that didn't, comparing rates before and after the law adoption.

a. Load the data with (remove the line break)

```
use https://raw.githubusercontent.com/kaplandm/stata/main/data/
    driving.dta , clear
```

and read the variable labels (including units of measure): `describe`

b. Keep only years 1980 and 1990: `keep if year==1980 | year==1990`

c. Create an "after" period dummy variable: `gen after = (year==1990)`

d. Create variable `bac` equal to `1` if there's any BAC law that year:
`gen bac = (bac08 + bac10 >= 1)`

e. Drop states that already had a BAC law in the "before" period (1980), leaving only states that never had the law or adopted it between 1980 and 1990:

```
generate dropflag = ((!after) & bac)
bysort state : egen dropst = max(dropflag)
drop if dropst
```

f. Create a treatment dummy equal to 1 for states that adopted a BAC law by 1990:
`bysort state : egen treat = max(bac)`

g. Run a difference-in-differences regression with the intercept, "after" dummy, treatment dummy, and interaction term. Below, the `##` automatically generates the desired interaction term: `reg wkndfatrte treat##after , vce(robust)`

h. To see how the OLS coefficient estimates relate to the conditional means (CMF estimates), compute the sample mean weekend driving fatality rate within each of the four groups defined by the time period and "treatment" status:
`tabulate treat after , summarize(wkndfatrte) means missing`

i. Display the CMF-based replication of the OLS estimates:

```
collapse (mean) wkndfatrte , by(treat after)
display wkndfatrte[1]
display wkndfatrte[3]-wkndfatrte[1]
display wkndfatrte[2]-wkndfatrte[1]
display (wkndfatrte[4]-wkndfatrte[3])-(wkndfatrte[2]-wkndfatrte[1])
```

j. Repeat part (g) but with a different outcome variable to replace `wkndfatrte`, like the weekend fatalities per 100 million miles driven (instead of population), or the total fatality rate (not just weekends), etc.

k. Repeat parts (d)–(g) but replacing your `bac` treatment variable created in part (d) with a treatment dummy equal to 1 if `perse` (a different driving law) equals 1 (and equal to 0 otherwise).

Exercise III.2. The dataset here has an observation for each state (plus DC) in the U.S. ($i = 1, \ldots, 51$) in years 1987, 1990, and 1993 ($t = 1, 2, 3$). The dependent variable `mrdrte` is the number of murders per 10,000 people (in state i during year t). The `d90` and `d93` are time dummies to include year effects. The two regressors are the unemployment rate (in state i, year t) and the number of executions in state i in years $t-2$, $t-1$, and t combined. (Note: this is not intended to be a sophisticated, definitive analysis upon which you should base your beliefs.)

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse murder , clear`

c. Run `reg mrdrte d90 d93 exec unem , vce(cluster state)`

d. Run `xtset id year`

e. Run `xtreg mrdrte d90 d93 exec unem, fe cluster(id)`

f. Report the pooled OLS and FE estimated coefficients on `exec`, and explain (both mathematically and in real-world terms) what this suggests about the relationship between `exec` and the unobserved state effects.

g. Discuss the economic significance of the FE estimated coefficient.

h. Explain what the corresponding confidence interval tells us about our uncertainty about the true population value; be precise and explicit.

i. Think of one additional (unobserved) time-varying variable that might also be correlated with `exec`. Explain which sign (positive or negative) you think the correlation might have, and in which direction this would bias the FE estimator.

Exercise III.3. The following dataset is not a panel but a repeated cross-section that includes years 1978 and 1981 ($t = 1, 2$), between which a new garbage incinerator was built in a particular neighborhood. Interest is in the causal effect on house prices; variable `lrprice` has log real house prices. Note `y81` is a dummy for year 1981, and `nearinc` is a dummy for being "near" the incinerator's location (even if it's 1978 and the incinerator itself does not yet exist).

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse kielmc , clear`

c. Run `reg lrprice nearinc if y81 , vce(robust)` and say what that code estimates as well as a specific real-world reason you think this is a biased estimator of the causal effect of being near the incinerator on housing price.

d. Run `reg lrprice y81 if nearinc , vce(robust)` and say what that code estimates as well as a specific real-world reason you think the estimator is biased.

e. Run `reg lrprice nearinc##y81 , vce(robust)`

   i. Report the number that is the difference-in-differences estimator of the effect of interest, as well as the units of measure.

   ii. What is the population estimand of this diff-in-diff estimator? Provide both math and real-world description (including definitions of the potential outcomes).

   iii. Discuss the economic significance of the estimate.

   iv. Explain what the confidence interval tells us about our uncertainty about the true population value: be precise and explicit.

   v. Recall that here we only have a repeated cross-section (not panel), and house prices are only observed when a house is sold. Assume conditions are relatively normal, so houses not near the incinerator (`nearinc=0`) are essentially sold at random (somebody gets a job in another state, somebody moves into a retirement home, etc.), so our dataset has a random sample of such house prices, and similarly in 1978 for all houses. Why might the 1981 near-incinerator prices not be a random sample, i.e., why might those houses not just be sold randomly? In which direction might this bias the diff-in-diff estimator? (There are many possible aspects to consider, but if you're having trouble getting started: imagine usually 5% of houses in a neighborhood sell in a typical year; 20% of homeowners are extremely opposed to living near a garbage incinerator while 80% don't care at all; recall basic supply and demand, how price responds to an increase in supply that shifts the supply curve; etc.)

Exercise III.4. The following analyzes county-year level crime data from North Carolina. The variable descriptions can be found online.[1] (Some of the descriptions are still vague; for research you would want to understand the variables much better, but we'll focus on other issues for now.)

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse crime4 , clear`

c. Run `xtset county year`

[1] http://fmwww.bc.edu/ec-p/data/wooldridge/crime4.des

d. Run `reg lcrmrte lpolpc if year==87 , vce(robust)` and explain one specific reason you don't think the slope coefficient can be interpreted as a causal effect; say in which direction you think it is biased, and why.

e. Run `reg lcrmrte lpolpc d8* , vce(robust)` and explain why this does not address your above concern (or if it does, come up with a different reason you don't think this estimates a causal effect).

f. Run `xtreg lcrmrte lpolpc d8* , fe cluster(county)`

   i. Explain what type of omitted variable (bias) the county-level fixed effects capture.

   ii. Discuss the economic significance of the FE estimated coefficient on `lpolpc`.

   iii. Explain what the corresponding confidence interval tells us about our uncertainty about the true population value; be precise and explicit.

   iv. Explain why this FE model still does not identify a causal effect in this example, including the direction of bias. (Feel free to try `reg lpolpc lcrmrte d8*` while you're thinking.)

Exercise III.5. The following analyzes data on manufacturing scrap rates for firms that did or did not receive grant money to improve. The variable descriptions can be found online.[2]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse jtrain , clear`

c. Run `xtset fcode year`

d. Run `reg grant L.lscrap` and briefly say what this suggests about which firms receive a grant. (Note: for real research, you would want to read about the grant program itself, not just run a simple regression.)

e. Run `reg lscrap L.lscrap` and briefly say what this suggests about firms' scrap rates over time.

f. Run `reg lscrap grant grant_1 if year==1989 , vce(robust)` and explain one specific reason you don't think the slope coefficient can be interpreted as a causal effect; say in which direction you think it is biased, and why. (Hint: think about your previous two results.)

g. Run `xtreg lscrap grant grant_1 d88 d89 , fe cluster(fcode)`

   i. Explain what type of omitted variable (bias) the firm-level fixed effects capture.

   ii. Discuss the economic significance of the FE estimated coefficients on `grant` and `grant_1`.

[2] http://fmwww.bc.edu/ec-p/data/wooldridge/jtrain.des

iii. Explain what the corresponding confidence intervals tell us about our uncertainty about the true population values; be precise and explicit.

iv. Explain what would need to be true for strict exogeneity to be satisfied here.

h. Run `lincom grant + grant_1` to get the estimate and confidence interval for the sum of these coefficients; how do you interpret this sum economically?

i. Run `reg D(lscrap grant grant_1 d89) , vce(cluster fcode)` to compute the FD estimator and briefly compare with the FE results.

Exercise III.6. The following analyzes crime data from Norway. The "clear-up percentage" is how many reported crimes were resolved by charging an individual with the crime (most commonly), which may be a deterrent to future crime. The variable descriptions can be seen in the variable labels.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse norway , clear`

c. Run `xtset district year , delta(6)` noting that the `delta(6)` tells it to treat year 1972 as $t = 1$ and 1978 as $t = 2$.

d. Run `reg lcrime clrprc1 clrprc2 if year==78 , vce(robust)` and explain one specific reason you don't think the slope coefficient can be interpreted as a causal effect; say in which direction you think it is biased, and why.

e. Run `xtreg lcrime clrprc1 clrprc2 d78 , fe cluster(district)`

i. Explain what type of omitted variable (bias) the district-level fixed effects capture.

ii. Discuss the economic significance of the FE estimated coefficients on `clrprc1` and `clrprc2`.

iii. Explain what the corresponding confidence intervals tell us about our uncertainty about the true population values; be precise and explicit.

iv. Explain one possible reason that strict exogeneity might be violated here.

f. Run `reg D(lcrime clrprc1 clrprc2) , vce(cluster district)` to compute the FD estimator and briefly compare with the FE results.

Exercise III.7. The following examines the relationship between low infant birthweight (a bad health outcome) and participation in a welfare program (that hopes to help pregnant women through nutrition programs and prenatal care). The specific program is the Aid to Families with Dependent Children (AFDC). The panel data is aggregated at the state-year level. Other control variables try to proxy for general quality of health care and income level in the state. The variable descriptions can be found online.[3]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

[3] http://fmwww.bc.edu/ec-p/data/wooldridge/lowbrth.des

b. Load the data: `bcuse jtrain , clear`

c. Run `encode state , gen(state_id)` to get a numeric identifier for the states (because `xtset` does not allow strings).

d. Run `xtset state_id year , delta(3)` noting that the `delta(3)` tells it to treat year 1987 as $t = 1$ and 1990 as $t = 2$.

e. Run `reg lowbrth afdcprc if year==1990 , vce(robust)` and explain one specific reason you don't think the slope coefficient can be interpreted as a causal effect; say in which direction you think it is biased, and why.

f. Run `xtreg lowbrth afdcprc d90 , fe cluster(state_id)`

   i. Explain what type of omitted variable (bias) the state-level fixed effects capture.

   ii. Discuss the economic significance of the FE estimated coefficient on `afdcprc`.

   iii. Explain what the corresponding confidence interval tells us about our uncertainty about the true population value; be precise and explicit.

   iv. Explain one possible reason that strict exogeneity might be violated here.

g. Run `reg D(lowbrth afdcprc) , vce(cluster state_id)` to compute the FD estimator and briefly compare with the FE results.

h. Run `xtreg lowbrth afdcprc d90 lphypc lbedspc lpcinc lpopul , fe cluster(state_id)` and briefly compare with the earlier FE results that did not include control variables.

Part IV

Probit

Introduction

This part looks at the probit model and related topics. Specifically, there is some discussion of binary response models, of (quasi) maximum likelihood, and of optimal prediction. Unlike previous parts, there is little emphasis on causal identification.

Chapter 14
Binary Response Models

Unit learning objectives for this chapter

14.1. Define ideas surrounding binary response models, including for both description and prediction. [TLOs 1–3]

14.2. Describe different methods for estimating binary response models, including their assumptions and the interpretation of model parameters. [TLOs 3 and 4]

Optional resources for this chapter

- Optimal prediction: Hastie, Tibshirani, and Friedman (2009, §2.4)

14.1 Binary Basics

Some properties of binary variable apply to all **binary response models**, meaning models of a binary outcome Y. The word "binary" most generally refers to anything with two possible values, but in this context, such values are coded as 0 and 1. For example, the underlying binary variable may have values "has bachelor's degree" and "does not have bachelor's degree," but this would be coded as $Y = 1$ (degree) and $Y = 0$ (if no degree). More precisely, such a Y is a **Bernoulli random variable** with parameter $p \equiv \mathrm{P}(Y = 1)$: $Y \sim \text{Bernoulli}(p)$.

A convenient property is

$$\begin{aligned} \mathrm{E}(Y) &= (0)\,\mathrm{P}(Y = 0) + (1)\,\mathrm{P}(Y = 1) = \mathrm{P}(Y = 1), \\ \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathrm{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \equiv p(\boldsymbol{x}), \end{aligned} \tag{14.1}$$

so the conditional mean function (CMF) can be interpreted as the **response probability** (function). This means that anything we know about identification, estimation, and

inference for CMF models can in principle apply to binary response models (including IV, FE, etc.). There are two main caveats. First, the simple linear-in-variables model is often especially implausible, although we could just use more flexible nonlinear or even nonparametric estimators. Second, we are often most interested in an underlying economic model that explains choices rather than just predicting choices themselves.

14.2 Linear Probability Model

The **linear probability model** (LPM) is essentially a linear regression model in which Y happens to be binary. Often "linear" refers to **linear-in-variables**, meaning that only the raw covariates appear in vector $\boldsymbol{X}$ (including an intercept) and the model is $p(\boldsymbol{x}) = \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$. However, sometimes "linear" refers to **linear-in-parameters**, meaning that $\boldsymbol{X}$ can also contain interaction terms and quadratic (and other polynomial) and log and other transformations, as long as we can still write $p(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$. This is much more flexible than the linear-in-variables model, and just as easy to estimate, but also more complicated to interpret.

14.2.1 Model Interpretation and Partial Effects

With the stricter linear-in-variables LPM, each individual coefficient has a simple interpretation. The intercept β_0 is the probability of $Y = 1$ if all regressors are equal to zero. (As usual, if the regressors have been demeaned first, then this intercept is a "centercept" equal to the probability of $Y = 1$ when the raw regressors are all at their mean values; otherwise, it may be meaningless if we are talking about the employment probability of an individual with zero years of education and zero income, who is zero years old.) The coefficient β_j on regressor X_j provides the change in the predicted value of $Y = 1$ associated with a one-unit increase in X_j. For example, if $\beta_j = 0.02$, then a one-unit increase in X_j is associated with a 2 **percentage point**pp increase in the probability of $Y = 1$, recalling that $\mathrm{P}(Y = 1) = 1$ means 100%. Note this differs from a 2% increase; for example, going from $\mathrm{P}(Y = 1) = 0.10$ to 0.12 is a 2pp increase but a $(0.12 - 0.10)/0.10 = 0.2 = 20\%$ increase. The phrase **associated with** describes a statistical relationship without claiming any causal interpretation.

As usual, the linear-in-variables model heavily restricts the partial derivative with respect to any regressor X_j: it is constant, not allowed to vary based on the initial level of X_j, nor based on the values of any other regressors. In contrast, the linear-in-parameters LPM allows both types of heterogeneity. In that case, you can take the partial derivative with respect to X_j (which may appear in multiple terms) to see how it depends on both the parameters and variable values. More precisely, define the **partial effect** (on the reponse probability, but not in the causal sense of "effect") at $\boldsymbol{X} = \boldsymbol{x}$ of continuous and

binary X_j respectively:

$$\mathrm{PE}_j(\boldsymbol{x}) \equiv \frac{\partial p(\boldsymbol{x})}{\partial x_j}, \quad \mathrm{PE}_j(\boldsymbol{x}) \equiv p(x_1, \ldots, x_{j-1}, 1, x_{j+1}, \ldots) - p(x_1, \ldots, x_{j-1}, 0, x_{j+1}, \ldots). \tag{14.2}$$

However, as is often true, this flexibility of letting the PE depend on $\boldsymbol{x}$ makes results more difficult to summarize and communicate.

As a compromise, often the more flexible model is used but then heterogeneous partial derivatives are summarized by a single number. Perhaps the two most common summaries are the **partial effect at the average** (PEA) and the **average partial effect** (APE). The PEA simply plugs in the sample average $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ when computing the partial effect of X_j:

$$\mathrm{PEA}_j \equiv \mathrm{PE}_j(\mathrm{E}(\boldsymbol{X})) = \left.\frac{\partial p(\boldsymbol{x})}{\partial x_j}\right|_{\boldsymbol{x}=\mathrm{E}(\boldsymbol{X})}. \tag{14.3}$$

The APE instead takes the mean PE by averaging over the population distribution of $\boldsymbol{X}$:

$$\mathrm{APE}_j \equiv \mathrm{E}[\mathrm{PE}_j(\boldsymbol{X})]. \tag{14.4}$$

Writing $\widehat{\mathrm{PE}}_j(\boldsymbol{x})$ for the estimated PE for X_j as a function of $\boldsymbol{x}$, which is some function of the estimated parameter vector $\hat{\boldsymbol{\beta}}$, the typical PEA and APE estimators are

$$\widehat{\mathrm{PEA}}_j = \widehat{\mathrm{PE}}_j(\bar{\boldsymbol{X}}), \quad \widehat{\mathrm{APE}}_j = \widehat{\mathrm{E}}[\mathrm{PE}_j(\boldsymbol{X})] = \frac{1}{n}\sum_{i=1}^n \mathrm{PE}_j(\boldsymbol{X}_i). \tag{14.5}$$

In Stata, whether the underlying model is the LPM or something else, you can use the `margins` command after your initial estimation command to compute either PEA (option `atmeans`) or APE.

The APE tends to be more commonly used and is more similar to other objects of interest like the ATE. One downside of the PEA is that the mean is difficult to interpret for binary X_j (like $X_j = \mathbb{1}\{\text{urban}\}$, and maybe $\bar{X}_j = 0.74$).

Example 14.1. Let $p(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. The PE of x is the partial derivative, $\beta_1 + 2x\beta_2$, which depends on the initial value of x. The PEA is $\beta_1 + 2\,\mathrm{E}(X)\beta_2$, which is estimated by $\hat{\beta}_1 + 2\,\widehat{\mathrm{E}}(X)\hat{\beta}_2$. The APE is $\mathrm{E}(\beta_1 + 2X\beta_2) = \beta_1 + 2\,\mathrm{E}(X)\beta_2$, which happens to be identical in this special case. More fundamentally, the APE is estimated by plugging in X_i to the PE and averaging over $i = 1, \ldots, n$: $\frac{1}{n}\sum_{i=1}^n \mathrm{PE}(X_i) = \frac{1}{n}\sum_{i=1}^n \beta_1 + 2X_i\beta_2 = \beta_1 + 2\beta_2\frac{1}{n}\sum_{i=1}^n X_i$, identical to the PEA estimator in this special case.

If either type of LPM is misspecified, then we can use the general misspecified CMF interpretations reviewed in Section 3.7.1. That is, we have a linear projection model, which provides the "best" (wrt MSE) linear predictor of Y given $\boldsymbol{X}$, and the "best" linear approximation of the true CMF, which here is also the response probability function.

14.2.2 Limitations

The main complaint against the LPM is that it can generate impossible predicted probabilities, $\hat{p}(\boldsymbol{x}) > 1$ or $\hat{p}(\boldsymbol{x}) < 0$. However, this is often not a major concern, as the next discussion questions help illustrate.

Discussion Question 14.1. Let $Y = \mathbb{1}\{\text{employed}_i\}$, and X is years of education. Imagine you estimate $\hat{\beta}_0 = 0.56$ and $\hat{\beta}_1 = 0.02$ in $\hat{p}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$.

a) Would these estimates ever predict employment probability $> 100\%$? Or $< 0\%$? If so, when?
b) Does this imply that our model is bad? Why/not?

Discussion Question 14.2. Let $Y = \mathbb{1}\{\text{employed}\}$, $X = \mathbb{1}\{\text{female}\}$, $\mathrm{P}(Y = 1 \mid X = x) = \beta_0 + \beta_1 x$. (Hint: can treat as CMF; or, just plug in x values.)

a) What's the interpretation of β_0 and β_1?
b) What are the sample analogs $\hat{\beta}_0$ and $\hat{\beta}_1$?
c) Is $\hat{p}(x) < 0$ or > 1 possible? Why/not?

Nonetheless, if we can simply use a model that enforces $0 \leq \hat{p}(\boldsymbol{x}) \leq 1$, then we don't need to worry about the issue. There are also other models that may better represent underlying economic decisions.

14.3 Binary Prediction

⟹ Kaplan video: Intuition for Prediction

This subsection provides a brief introduction to concepts about optimal prediction, with a focus on binary Y. The underlying model (LPM, probit, etc.) is irrelevant. See Sections 2.4–2.5 of Kaplan (2022a) for additional details.

Although we usually think of data as fundamental to prediction, prediction is more fundamentally formulated mathematically in the population. Then, after deriving the optimal population predictor, we can use data to estimate it. (There are more sophisticated ways to try to incorporate empirical uncertainty into optimal prediction, but those are beyond our scope.)

Consider binary $Y = 1$ if it rains on Tuesday, otherwise $Y = 0$, and our job is to guess whether or not it rains on Tuesday. Our guess g is non-random because we have to make a single guess before Y is realized (and it is not optimal to randomize guessing). Because Y is binary, it only makes sense for us to guess either $g = 1$ (rain) or $g = 0$ (not). There are two important components: the probability $\mathrm{P}(Y = 1)$, and the consequences of being wrong.

14.3.1 Loss Function

The consequences of being wrong are quantified through a **loss function** that says how bad it is to have guessed g when the true realization is y, $L(y, g)$. Higher values (closer to

$+\infty$) of loss indicate worse consequences. Usually the loss function is normalized to have $L(y, y) = 0$ (zero loss for correct prediction), so that "how bad" is relative to a correct guess, and all loss values are non-negative.

Example 14.2. Imagine you make a bet with your friend: if you correctly predict the rain variable Y, then you win \$1, but if you're wrong then you must pay your friend \$1. Without normalizing, this means you lose \$1 when $y \neq g$, so $L(0, 1) = L(1, 0) = 1$, and winning \$1 is -1 loss so $L(0, 0) = L(1, 1) = -1$. Alternatively, for reasons that will be clearer later, we could normalize by adding 1 to each value: $L(0, 0) = L(1, 1) = 0$ and $L(0, 1) = L(1, 0) = 2$. If you could cheat and see already that it's raining ($Y = 1$), then intuitively you'd guess $g = 1$ to make sure you win the bet; indeed, $L(1, 1) = 0 < 2 = L(1, 0)$, so your optimal choice of g minimizes your loss.

Example 14.3. Imagine you again make a bet with your friend, but now your friend will pay you \$10 if you correctly predict rain. Without normalizing, as before $L(0, 1) = L(1, 0) = 1$ and $L(0, 0) = -1$, but now $L(1, 1) = -10$. Compared to correctly guessing $g = 1$ to get $L(1, 1) = -10$, incorrectly guessing $g = 0$ to get $L(1, 0) = 1$ is $+11$ higher loss, so we could normalize $L(1, 1) = 0$ and $L(1, 0) = 11$. As before, we can also normalize $L(0, 0) = 0$ and $L(0, 1) = 2$.

The **0–1 loss function** is

$$L_0(Y, g) = \mathbb{1}\{Y \neq g\} = \begin{cases} 0 & \text{if } Y = g, \\ 1 & \text{if } Y \neq g. \end{cases} \tag{14.6}$$

It only cares whether we are exactly correct or not. A more general version is the **weighted 0–1 loss function**,

$$L_w(Y, g) = \mathbb{1}\{Y = 0, g = 1\} + w\,\mathbb{1}\{Y = 1, g = 0\}, \tag{14.7}$$

which allows $L_w(1, 0) > L_w(0, 1)$ if $w > 0$, or $L_w(1, 0) < L_w(0, 1)$ if $w < 0$. This is similar to type I errors being regarded as worse than type II errors in the hypothesis testing context. This weighted 0–1 loss is fully general given binary Y and g.

14.3.2 Unconditional Optimal Prediction in the Population

With random Y, we cannot choose g to minimize loss for every realization $Y = y$, so often instead the "optimal" prediction is defined as minimizing our mean loss. Given a non-random guess g, **mean loss** (or **expected loss**) is $\mathrm{E}[L(Y, g)]$, where the mean is with respect to the distribution of Y. In this context, mean loss is also called **risk**. With binary Y and g, this is simply

$$\mathrm{E}[L(Y, g)] = \mathrm{P}(Y = 0)L(0, g) + \mathrm{P}(Y = 1)L(1, g). \tag{14.8}$$

With the normalization $L(0,0) = L(1,1) = 0$ and weighted 0–1 loss, and writing $p \equiv \mathrm{P}(Y=1) = 1 - \mathrm{P}(Y=0)$,

$$\begin{aligned} \mathrm{E}[L_w(Y,0)] &= \overbrace{\mathrm{P}(Y=0)}^{=1-p}\overbrace{L_w(0,0)}^{=0} + \overbrace{\mathrm{P}(Y=1)}^{=p}\overbrace{L_w(1,0)}^{=w} = pw, \\ \mathrm{E}[L_w(Y,1)] &= \overbrace{\mathrm{P}(Y=0)}^{=1-p}\overbrace{L_w(0,1)}^{=1} + \overbrace{\mathrm{P}(Y=1)}^{=p}\overbrace{L_w(1,1)}^{=0} = 1-p. \end{aligned} \tag{14.9}$$

Given (14.9), we prefer to guess $g=1$ when $1-p < pw$, or equivalently $p(w+1) > 1$, and conversely we prefer $g=0$ when $p(w+1) < 1$, with indifference at $p(w+1) = 1$. Holding w fixed, we are more likely to prefer $g=1$ as $p \uparrow 1$. This is intuitive: all else equal, as $Y=1$ becomes more likely, we should be more likely to guess $g=1$. Holding p fixed, we are more likely to prefer $g=1$ as w increases. This is also intuitive: all else equal, as the consequences of a wrong $g=0$ guess become worse, we have more incentive to avoid $g=0$ and instead guess $g=1$.

Given w, the optimal prediction follows a probability threshold. Let g_w^* be the optimal prediction given $L_w(\cdot)$,

$$g_w^* \equiv \underset{g\in\{0,1\}}{\arg\min}\, \mathrm{E}[L_w(Y,g)] = \mathbb{1}\{p > 1/(1+w)\}. \tag{14.10}$$

Discussion Question 14.3 (optimal prediction)**.** Wait: isn't the mean the best predictor of Y? (For example, (3.6) in Section 3.7.1 replacing $\boldsymbol{X} = 1$.)

a) What is the mean of Y, and how does it compare to g_w^*?
b) Given $L_w(\cdot)$, how does the risk of the mean of Y compare to the risk of g_w^*?

Discussion Question 14.4 (threshold)**.** Consider the optimal prediction in (14.10).

a) When would a lower threshold $1/(1+w)$ help better predict $Y=1$? (Meaning, improve prediction accuracy for cases where the true value is $Y=1$.)
b) When would a lower threshold $1/(1+w)$ help better predict $Y=0$?

14.3.3 Conditional Optimal Prediction in the Population

All the unconditional intuition continues to hold conditional on $\boldsymbol{X} = \boldsymbol{x}$. For example, generalizing (14.8),

$$\mathrm{E}[L(Y,g) \mid \boldsymbol{X}=\boldsymbol{x}] = \mathrm{P}(Y=0 \mid \boldsymbol{X}=\boldsymbol{x})L(0,g) + \mathrm{P}(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})L(1,g), \tag{14.11}$$

and following the same derivations leads to the optimal prediction

$$g_w^*(\boldsymbol{x}) \equiv \underset{g\in\{0,1\}}{\arg\min}\, \mathrm{E}[L_w(Y,g) \mid \boldsymbol{X}=\boldsymbol{x}] = \mathbb{1}\{p(\boldsymbol{x}) > 1/(1+w)\}, \tag{14.12}$$

continuing the notation $p(\boldsymbol{x}) \equiv \mathrm{P}(Y=1 \mid \boldsymbol{X}=\boldsymbol{x})$. Again, this has the simple form of a probability threshold that depends on the relative loss w.

Discussion Question 14.5 (prediction and policy)**.** Imagine that you estimated crime probabilities for different neighborhoods for each 8-hour police shift of every day. There's only enough budget to patrol 10 of the 20 neighborhoods each shift. Assume you can perfectly prevent all crime in the patrolled neighborhoods. How should you choose which neighborhoods to patrol? Explain your procedure, any additional assumptions you make, and anything that might go wrong if you actually implemented this procedure in the real world.

14.3.4 Prediction in Practice

If we knew the true response probability function $p(\cdot)$, then given loss function $L_w(\cdot)$, we could compute $g^*_w(\boldsymbol{x}) = \mathbb{1}\{p(\boldsymbol{x}) > 1/(1+w)\}$. In practice, we can plug in our estimated $\hat{p}(\cdot)$ to get

$$\hat{g}_w(\boldsymbol{x}) = \mathbb{1}\{\hat{p}(\boldsymbol{x}) > 1/(1+w)\}. \tag{14.13}$$

Note having $\hat{p} > 1$ or $\hat{p} < 0$ is not a problem for prediction. A formula like (14.13) is sometimes called a **decision rule** in the context of statistical decision theory. In large samples, our $\hat{p}(\cdot)$ is close to the true $p(\cdot)$, so this should generate predictions that are close to minimizing risk. Again, the threshold $1/(1+w)$ depends on the relative loss of incorrectly guessing $g = 1$ versus incorrectly guessing $g = 0$, so the decision rule depends on both the predicted probability as well as the loss function.

Beyond our scope...

In small samples, the estimation uncertainty may play an important role. For example, consider unconditional prediction in the extreme case $n = 1$. Consider very large $w = 999$, so it is much worse to incorrectly guess $g = 0$ when $Y = 1$ than to incorrectly guess $g = 1$ when $Y = 0$. Imagine you observe $Y_1 = 0$; which g would you choose for out-of-sample prediction? According to (14.13), $\hat{g}_w = 0$ because $0 = \hat{p} < 1/(1+999) = 0.001$. However, with $n = 1$, we have lots of uncertainty about the estimated $\hat{p} = 0$, and more specifically we are not sure that $p < 0.001$. For example, even if the true $p = 0.1 > 0.001$, we are still very likely to observe $Y_1 = 0$. There are more sophisticated approaches (both frequentist and Bayesian) to try to incorporate such uncertainty into our decision-making.

If prediction is the ultimate goal (rather than description or causal inference), then it is helpful to separately report the percent correctly predicted for individuals in the sample with $Y_i = 0$ and $Y_i = 1$. That is, given (14.13), the percent correctly predicted for $Y_i = 0$ is

$$\frac{\sum_{i=1}^n \mathbb{1}\{Y_i = \hat{g}_w(\boldsymbol{X}_i) = 0\}}{\sum_{i=1}^n \mathbb{1}\{Y_i = 0\}} = \frac{\sum_{i=1}^n \mathbb{1}\{Y_i = 0\}\,\mathbb{1}\{\hat{p}(\boldsymbol{X}_i) < 1/(1+w)\}}{\sum_{i=1}^n \mathbb{1}\{Y_i = 0\}}, \tag{14.14}$$

and similarly for $Y_i = 1$ replacing all the $= 0$ with $= 1$ and replacing $<$ with $>$.

14.4 Probit Model

Consider the general single index model

$$\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = G(\boldsymbol{x}'\boldsymbol{\beta}), \tag{14.15}$$

where $G(\cdot)$ is a known, increasing, nonlinear function. (This is related to the "link function" in a "generalized linear model.") In our case the LHS equals $p(\boldsymbol{x})$, so it makes sense to restrict $0 \leq G(r) \leq 1$ for all $r \in \mathbb{R}$. One class of such functions is CDFs. The use of a CDF for $G(\cdot)$ can thus be thought of as a convenient way to enforce $0 \leq p(\boldsymbol{x}) \leq 1$, but it is also derived from a more economic model below.

The **probit model** uses $G(\cdot) = \Phi(\cdot)$, the standard normal distribution's CDF, while the **logit model** uses $G(\cdot) = \Lambda(\cdot)$, the standard logistic distribution's CDF.

14.4.1 Interpretation and Partial Effects

For the probit, the partial effect defined in (14.2) for continuous x_j is

$$\mathrm{PE}_j(\boldsymbol{x}) \equiv \frac{\partial p(\boldsymbol{x})}{\partial x_j} = \phi(\boldsymbol{x}'\boldsymbol{\beta})\beta_j, \tag{14.16}$$

assuming none of the other elements in $\boldsymbol{x}$ involve x_j, and similar to before the PE for binary x_j is the difference in response probability when plugging in $x_j = 1$ versus $x_j = 0$, holding all other elements of $\boldsymbol{x}$ constant. Although PEs can be more convenient mathematically for providing intuition, even for a continuous x_j, the most accurate way to compute the change in response probability associated with a change from $\boldsymbol{x}$ to $\boldsymbol{x} + \boldsymbol{d}$ is

$$p(\boldsymbol{x} + \boldsymbol{d}) - p(\boldsymbol{x}) = \Phi((\boldsymbol{x} + \boldsymbol{d})'\boldsymbol{\beta}) - \Phi(\boldsymbol{x}'\boldsymbol{\beta}). \tag{14.17}$$

Discussion Question 14.6 (probit features)**.** Consider a simple probit model $p(x) = \Phi(\beta_0 + \beta_1 x)$.

a) Why is the derivative with respect to x (partial effect) equal to $\phi(\beta_0+\beta_1 x)\beta_1$, where $\phi(\cdot)$ is the standard normal PDF?
b) Can the derivative depend on x (unlike LPM)?
c) Can the derivative be constant (wrt x)?
d) Can the derivative have a different sign ($\pm$) at different x?

Although the probit model is nonlinear (in both variables and parameters), "nonlinear" does not necessarily imply "flexible" or "robust to misspecification." Given $\boldsymbol{X}$, the probit has the same number of parameters in $\boldsymbol{\beta}$ as the LPM, so it is not really "more" flexible, just differently flexible, as DQ 14.6 illustrates. (But, arguably it is "differently flexible" in a way that's more appropriate for binary Y.)

The APE and PEA are defined the same way as in (14.3) and (14.4). For continuous X_j, assuming again no other element of $\boldsymbol{X}$ involves X_j,

$$\begin{aligned} \mathrm{APE}_j &\equiv \mathrm{E}[\mathrm{PE}_j(\boldsymbol{X})] = \mathrm{E}[\phi(\boldsymbol{X}'\boldsymbol{\beta})\beta_j] = \mathrm{E}[\phi(\boldsymbol{X}'\boldsymbol{\beta})]\beta_j, \\ \mathrm{PEA}_j &\equiv \mathrm{PE}_j(\mathrm{E}(\boldsymbol{X})) = \phi(\mathrm{E}(\boldsymbol{X})'\boldsymbol{\beta})\beta_j. \end{aligned} \tag{14.18}$$

Although these expressions are complex, the ratio of APEs or PEAs across two regressors X_j and X_k are simple:

$$\frac{\text{APE}_j}{\text{APE}_k} = \frac{\text{E}[\phi(\boldsymbol{X}'\boldsymbol{\beta})]\beta_j}{\text{E}[\phi(\boldsymbol{X}'\boldsymbol{\beta})]\beta_k} = \frac{\beta_j}{\beta_k} = \frac{\text{PEA}_j}{\text{PEA}_k}. \tag{14.19}$$

Example 14.4. You estimate a probit of employment (Y) given education (X_1) and experience (X_2). Although the estimated $\hat{\beta}_1$ and $\hat{\beta}_2$ do not have a direct interpretation, $\hat{\beta}_1/\hat{\beta}_2 = 3.1$, which you can interpret as the APE due to education being 3.1 times larger than the APE due to experience. (This approximately means that one extra year of education is associated with the same increase in employment probability as an extra 3.1 years of experience.)

If a covariate enters nonlinearly, then the PE/APE/PEA have additional terms. For example, let

$$p(x_1, x_2) = \Phi(\beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3 + x_1^2\beta_4 + x_2^2\beta_5).$$

Then the PE for x_1 is

$$\frac{\partial p(x_1, x_2)}{\partial x_1} = \phi(\beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3 + x_1^2\beta_4 + x_2^2\beta_5)(\beta_1 + x_2\beta_3 + 2x_1\beta_4), \tag{14.20}$$

which feeds into the PEA/APE accordingly.

Thus in Stata, although you do not need to manually compute derivatives, it is very important to make sure Stata knows you have nonlinear functions of a particular regressor, otherwise the APE/PEA from `margins` will be wrong. For example, if you have a raw variable `age`, then you would want to include `c.age\#\#c.age` in your probit model. If instead you generate `gen agesq = age^2` and include `age` and `agesq`, then Stata will incorrectly interpret them as separate variables and fail to compute the APE/PEA properly.

14.4.2 Prediction

Prediction is the same as with the LPM: nothing in Section 14.3 depended on the underlying model. Given the probit $\hat{\boldsymbol{\beta}}$, the predicted Y given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\hat{p}(\boldsymbol{x}) = \Phi(\boldsymbol{x}'\hat{\boldsymbol{\beta}}), \tag{14.21}$$

which can then be plugged into the decision rule in (14.13).

14.4.3 Structural Model and Interpretation

Consider a **latent** (unobserved) continuous outcome Y^*. This Y^* is often interpreted as an individual's utility, or specifically the difference in their utility from choosing $Y = 1$ instead of $Y = 0$. (So $Y^* > 0$ means more utility from $Y = 1$, while $Y^* < 0$ means

more utility from $Y = 0$.) If people maximize utility, then $Y = \mathbb{1}\{Y^* > 0\}$. For example, maybe Y^* is the utility "gain" (possibly negative) from getting married and Y is the observed choice about getting married; or Y^* is the utility gain from buying organic milk (vs. non-organic) and Y is the observed milk type choice; or going to college, or serving in the military, etc. As usual, the "individual" could also be a firm or country or other unit, so we could model the firm's benefit from going public (vs. staying privately held), or a city's benefit from having a free (vs. paid) bus system, etc. For the probit, the full structural model is

$$Y^* = \boldsymbol{X}'\boldsymbol{\beta} + U, \quad U \sim \mathrm{N}(0,1), \quad Y = \mathbb{1}\{Y^* > 0\}. \tag{14.22}$$

Alternatively, we could change the distribution of U to any other parametric distribution, like a standard logistic (for the logit model), or any other continuous, symmetric distribution with known CDF $G(\cdot)$.

Given "exogeneity" in the sense of $U \perp\!\!\!\perp \boldsymbol{X}$, the structural $\boldsymbol{\beta}$ are identified and equal to the parameters in the response probability model:

$$\begin{aligned}
\mathrm{P}(\overbrace{Y=1}^{\text{use (14.22)}} \mid \boldsymbol{X}=\boldsymbol{x}) &= \mathrm{P}(\overbrace{Y^*>0}^{Y=1} \mid \boldsymbol{X}=\boldsymbol{x}) \\
&= \mathrm{P}(\overbrace{\boldsymbol{X}'\boldsymbol{\beta}+U}^{Y^* \text{ in (14.22)}} > 0 \mid \boldsymbol{X}=\boldsymbol{x}) \\
&= \mathrm{P}(U > \overbrace{-\boldsymbol{x}'\boldsymbol{\beta}}^{\text{non-random}} \mid \overbrace{\boldsymbol{X}=\boldsymbol{x}}^{\text{no effect on } U \text{ b/c } U \perp\!\!\!\perp \boldsymbol{X}}) \\
&= \mathrm{P}(U > -\boldsymbol{x}'\boldsymbol{\beta}) \\
&= 1 - \overbrace{\mathrm{P}(U \le -\boldsymbol{x}'\boldsymbol{\beta})}^{\text{use } U\sim\mathrm{N}(0,1) \text{ from (14.22)}} \\
&= 1 - \Phi(-\boldsymbol{x}'\boldsymbol{\beta}) \\
&= \Phi(\boldsymbol{x}'\boldsymbol{\beta}),
\end{aligned} \tag{14.23}$$

where the last equality uses the symmetry of the standard normal distribution.

The reason $\mathrm{Var}(U) = 1$ in (14.22) is because even if we let $\mathrm{Var}(U) = \sigma^2$, we could not learn about σ^2 given that we only observe binary Y. This also means we cannot learn about the absolute magnitude of $\boldsymbol{\beta}$, only the relative magnitudes and signs. That is, $\boldsymbol{\beta}$ is only **identified up to scale**, as DQ 14.7 helps illustrate.

Discussion Question 14.7 (identification up to scale)**.** Consider the intercept-only model with

$$\mathrm{P}(Y=1) = \mathrm{P}(Y^*>0) = \mathrm{P}(U > -\beta) = \mathrm{P}(U \le \beta)$$

with $U \sim \mathrm{N}(0, \sigma^2)$. Imagine $\mathrm{P}(Y = 0.95)$. Consider β as a function of σ, $\beta(\sigma)$. (Hint: recall if $Z \sim \mathrm{N}(0,1)$, then $\sigma Z \sim \mathrm{N}(0, \sigma^2)$.)

a) With $\sigma = 1$, what is $\beta(1)$? (If you haven't memorized z-tables, it's fine to just write a formula instead of a number.)

b) With $\sigma = 2$, what is $\beta(2)$?
c) What's the relationship between $\beta(2)$ and $\beta(1)$?
d) What's the general formula for the ratio $\beta(\sigma)/\beta(1)$?
e) Does the sign ($\pm$) of $\beta(\sigma)$ depend on σ? Is this true for other $\mathrm{P}(Y = 0.95)$?

If we look back at the identification argument but write $U = \sigma Z$ with $Z \sim \mathrm{N}(0, \sigma^2)$, then altogether we get

$$\mathrm{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \mathrm{P}(U \leq -\boldsymbol{x}'\boldsymbol{\beta}) = \mathrm{P}(\sigma Z \leq -\boldsymbol{x}'\boldsymbol{\beta}) = \mathrm{P}(Z \leq -\boldsymbol{x}'\boldsymbol{\beta}/\sigma) = \Phi(-\boldsymbol{x}'\boldsymbol{\beta}/\sigma), \tag{14.24}$$

meaning we can only identify $\boldsymbol{\beta}/\sigma$.

Thankfully, everything meaningful is not scale-dependent. The relative magnitudes do not depend on σ because it's the same scalar σ dividing each element of $\boldsymbol{\beta}$, so $(\beta_j/\sigma)/(\beta_k/\sigma) = \beta_j/\beta_k$, regardless of σ. Similarly, estimated probabilities $\hat{p}(\boldsymbol{x})$ are not scale-dependent, nor are the PE/APE/PEA. This is another reason we focus on partial effects rather than individual probit coefficient estimates.

Discussion Question 14.8 (probit economic significance)**.** Let $p(x) = \Phi(\beta_0 + \beta_1 x)$, so the partial effect is $\phi(\beta_0 + \beta_1 x)\beta_1$. Let $Y = \mathbb{1}\{\text{employed}\}$, X is income of the individual's spouse measured in \$1000s (like $x = 50$ means \$50,000). Let $\hat{\beta}_0 = 4$. Discuss the economic significance of $\hat{\beta}_1 = -0.1$. (Hint #1: is the partial effect the same at all x?) (Hint #2: $\phi(0) \approx 0.4$, $\phi(2) \approx 0.05$, $\phi(4) \approx 0.0001$.)

14.5 Logit Model

The logistic distribution is relatively similar to the standard normal distribution (unimodal, symmetric, not fat-tailed, etc.), so the logit model that uses the single index model from (14.15) with $G(\cdot) = \Lambda(\cdot)$ (logistic CDF) is relatively similar to the probit model that uses $G(\cdot) = \Phi(\cdot)$. For example, thinking of the latent model, specifying normal instead of logistic errors is not a big difference economically. Estimates of the APE or PEA tend to be very similar between probit and logit. However, there are some qualitative differences.

Due to the logistic CDF formula, logit can be interpreted as modeling the **log odds ratio**. (I think this interpretation is the primary reason some statisticians deride economists for using probit instead of logit.) The "odds ratio" refers to $\mathrm{P}(Y = 1)/\mathrm{P}(Y = 0)$; with $p \equiv \mathrm{P}(Y = 1)$, this is $p/(1 - p)$. The log odds ratio is the (natural) log of this. The transformation from p to $\log(p/(1 - p))$ is called the **logit function**, which is also the inverse CDF (i.e., quantile function) of the standard logistic distribution. This can be

derived from the standard logistic CDF $\Lambda(r) = 1/(1 + e^{-r})$ and the single index model:

$$\begin{aligned} p(\boldsymbol{x}) &= \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{1}{1 + e^{-\boldsymbol{x}'\boldsymbol{\beta}}}, \\ 1 - p(\boldsymbol{x}) &= \frac{e^{-\boldsymbol{x}'\boldsymbol{\beta}}}{1 + e^{-\boldsymbol{x}'\boldsymbol{\beta}}}, \\ \frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} &= \frac{1/(1 + e^{-\boldsymbol{x}'\boldsymbol{\beta}})}{e^{-\boldsymbol{x}'\boldsymbol{\beta}}/(1 + e^{-\boldsymbol{x}'\boldsymbol{\beta}})} = e^{\boldsymbol{x}'\boldsymbol{\beta}}, \\ \log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) &= \boldsymbol{x}'\boldsymbol{\beta}. \end{aligned}$$

That is, our linear-in-parameters "index" $\boldsymbol{x}'\boldsymbol{\beta}$ models the log odds ratio.

The logit model also provides some convenient cancellation in a panel model with fixed effects, and generalizes nicely to a multinomial choice models (more than two choices), although such topics are beyond our scope.

Beyond our scope...

If you want to determine the "better" model for a given dataset, you can apply one of the many "model selection" procedures to judge between probit and logit; for example, see Chapter 18 ("Model Selection") of Kaplan (2021).

Chapter 15
(Quasi) Maximum Likelihood

Unit learning objectives for this chapter

15.1. Describe terms and concepts related to quasi-maximum likelihood, both mathematically and intuitively. [TLOs 1–3]

If you've seen anything about maximum likelihood before, it most likely started with a joint PDF for all n observations, which under iid sampling factors into a product of individual PDFs, and we want to find the parameter values that maximize the product. Despite seeming intuitive, this heuristic argument gives us no formal justification for the estimator being consistent. Section 15.2 provides a more insightful view, which also helps us understand performance under misspecification. Before that, some basic terms and formulas are given.

15.1 Basics

First consider Y_i whose marginal PDF we assume has the form $f(\cdot \mid \theta)$ for some true $\theta \in \Theta$. That is, we know the "family" $f(\cdot \mid \cdot)$ and consider only possible PDFs $f(\cdot \mid t)$ for some $t \in \Theta$. If our assumption is correct, then our model is **properly specified**; if there is no $t \in \Theta$ such that $f(\cdot \mid t)$ is the PDF of Y_i, then our model is **misspecified**. Sometimes such a PDF (evaluated at point $Y = y$) is also written as $f_t(y)$, or $f(y; t)$. The PDF is the **likelihood function**. The likelihood function is sometimes written as $\mathcal{L}(t \mid y)$ to emphasize that we are trying to pick among possible $t \in \Theta$ values, but that form can also generate confusion. The **log-likelihood function** is simply the (natural) log, $\log[f(y \mid t)]$. Evaluated at $y = Y_i$, this is sometimes written $\ell_i(t)$.

The general idea of **maximum likelihood** (ML) is to characterize the estimator $\hat{\theta}$ as the value that maximizes the likelihood. More specifically, the **maximum likelihood**

estimator (MLE) is

$$\hat{\theta} \equiv \arg\max_{t \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log[f(Y_i \mid t)]. \tag{15.1}$$

If the second-order condition holds, then the maximizer can be written as the solution to the first-order condition

$$0 = \frac{1}{n} \sum_{i=1}^{n} s_i(\hat{\theta}), \quad s_i(t) \equiv \frac{\partial \log[f(Y_i \mid t)]}{\partial t}, \tag{15.2}$$

where $s_i(\cdot)$ is the **score function** for observation i.

Example 15.1. Let $Y_i \sim \mathrm{N}(\theta, 1)$, so $f(y \mid t) = (2\pi)^{-1/2} e^{-(y-t)^2/2}$, with $t \in \Theta = \mathbb{R}$. The log-likelihood and score are

$$\log[f(y \mid t)] = -\frac{1}{2}\log(2\pi) - \frac{(y-t)^2}{2}, \quad \frac{\partial \log[f(y \mid t)]}{\partial t} = y - t. \tag{15.3}$$

As the log-likelihood is quadratic in t and thus satisfies the SOC, the MLE solves the FOC:

$$0 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\theta}) \implies \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{15.4}$$

Interestingly, we know from the WLLN that this MLE is consistent for the true population mean $\mathrm{E}(Y)$ even if the Y_i are not normally distributed, but this is not a general property of MLE.

Example 15.2. Let $Y_i \sim \text{Bernoulli}(\theta)$. (Assuming the Y_i are iid or at least stationary, this must be properly specified because there are no other possible distributions for binary random variables.) Instead of a PDF, we have a PMF, but the idea is the same: $f(1 \mid t) = \mathrm{P}(Y = 1 \mid t) = t$ and $f(0 \mid t) = 1 - t$, so for $y \in \{0, 1\}$, $f(y \mid t) = t^y + (1-t)^{1-y}$. The log-likelihood and score are

$$\begin{aligned} \log[f(y \mid t)] &= y \log(t) + (1-y)\log(1-t), \\ \frac{\partial \log[f(y \mid t)]}{\partial t} &= y t^{-1} + (1-y)(1-t)^{-1}(-1). \end{aligned} \tag{15.5}$$

Solving the FOC, remembering either $Y_i = 0$ or $Y_i = 1$,

$$0 = \frac{1}{n} \sum_{i=1}^{n} [Y_i \hat{\theta}^{-1} - (1 - Y_i)(1 - \hat{\theta})^{-1}] \hat{\theta}^{-1} \bar{Y} - (1 - \hat{\theta})^{-1}(1 - \bar{Y}),$$

so

$$\frac{1 - \bar{Y}}{1 - \hat{\theta}} = \frac{\bar{Y}}{\hat{\theta}} \implies \hat{\theta} - \hat{\theta}\bar{Y} = \bar{Y} - \hat{\theta}\bar{Y} \implies \hat{\theta} = \bar{Y}.$$

We know from the WLLN that indeed $\hat{\theta} = \bar{Y} \xrightarrow{p} \theta$.

Beyond our scope...

There are also nonparametric approaches to maximum likelihood, where you do not need to specify the distribution up to a finite number of unknown parameters. These include the Kiefer and Wolfowitz (1956) nonparametric maximum likelihood estimator (NPMLE) and sieve maximum likelihood, which allows the parametric model to be more complex when more data is available (with "more complex" being implemented in a precise way, and trying to choose the optimal level of complexity given a particular dataset); for example, see Sections 2.2.2 and 4.2.4 and other references throughout Chen (2007).

The ideas above extend readily to a vector of parameters. Because Y is still a scalar, the likelihood $f(y \mid \boldsymbol{t})$ is still a scalar, as is the log-likelihood, but the score is a vector with the same dimension as $\boldsymbol{t}$.

Also, the **Hessian** $\underline{\boldsymbol{H}}_i(\boldsymbol{t})$ is the matrix of second derivatives of the log-likelihood for observation i, which is the derivative of the score $\boldsymbol{s}_i(\boldsymbol{t})$ with respect to $\boldsymbol{t}$. Another important matrix is the expected outer product of the score, $\mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{\theta})']$.

15.2 Identification and Quasi-Maximum Likelihood

The terms **quasi-maximum likelihood** (QML) and **pseudo maximum likelihood** (PML) both refer to ML when the parametric family of distributions is misspecified, meaning the true distribution is not a member of the family. (The only difference is "quasi" is Latin, whereas "pseudo" is Greek.)

The situation is similar in spirit to when we specify a linear regression model when the true CMF is not linear. Recall from Section 3.7.1 that instead of the CMF, we end up estimating the "best linear approximation" of the CMF. That is, we estimate the function that's "closest" to the true CMF within our erroneously restricted set of possible functions (like linear functions), but we must also remember that "closest" does not mean "close." (For example, the city in Missouri closest to New Zealand is still not close to New Zealand.) Similarly with QML, we estimate the distribution in our specified family that's "closest" to the true one, but it may not be "close."

First, we must define "closest" quantitatively. For QML, it is defined in terms of the **Kullback–Leibler** (KL) divergence, also known as the KL information or **Kullback–Leibler information criterion** (KLIC). Note the word "distance" is not used because it is not symmetric, meaning the KL divergence from $f(\cdot)$ to $g(\cdot)$ generally does not equal the KL divergence from $g(\cdot)$ to $f(\cdot)$.

For notational simplicity, consider the unconditional case with scalar Y with support $\mathbb{R}$. Let $g(\cdot)$ be the true PDF of Y. Thus, the expectation operator refers to integrating against $g(\cdot)$, like $\mathrm{E}(Y) = \int_{\mathbb{R}} y g(y)\,\mathrm{d}y$. Let $f(\cdot \mid \boldsymbol{t})$ be the specified PDF family. If there is some $\boldsymbol{\theta}$ such that $g(\cdot) = f(\cdot \mid \boldsymbol{\theta})$, then it is properly specified; if not, then it is misspecified.

With proper specification, the population $\boldsymbol{\theta}$ maximizes the population expected log-likelihood, as the (long) argument below shows. First write the KL divergence,

$$\mathrm{E}[f(Y \mid \boldsymbol{t})/f(Y \mid \boldsymbol{\theta})] = \int_{\mathbb{R}} \frac{f(y \mid \boldsymbol{t})}{f(y \mid \boldsymbol{\theta})} f(y \mid \boldsymbol{\theta}) \, \mathrm{d}y = \int_{\mathbb{R}} f(y \mid \boldsymbol{t}) \, \mathrm{d}y = 1$$

because $f(\cdot \mid \boldsymbol{t})$ is a PDF and thus integrates to 1. Because $\log(1) = 0$, then

$$\log\big\{\mathrm{E}[f(Y \mid \boldsymbol{t})/f(Y \mid \boldsymbol{\theta})]\big\} = 0. \tag{15.6}$$

By Jensen's inequality, because $\log(\cdot)$ is concave,

$$\mathrm{E}\big[\log\{f(Y \mid \boldsymbol{t})/f(Y \mid \boldsymbol{\theta})\}\big] \leq 0, \tag{15.7}$$

and with $\boldsymbol{t} = \boldsymbol{\theta}$ the upper bound of zero is attained:

$$\mathrm{E}\big[\log\{f(Y \mid \boldsymbol{\theta})/f(Y \mid \boldsymbol{\theta})\}\big] = \mathrm{E}[\log(1)] = \mathrm{E}[0] = 0. \tag{15.8}$$

Further, using $\log(a/b) = \log(a) - \log(b)$,

$$0 \geq \mathrm{E}\big[\log\{f(Y \mid \boldsymbol{t})/f(Y \mid \boldsymbol{\theta})\}\big] = \mathrm{E}\{\log[f(Y \mid \boldsymbol{t})]\} - \mathrm{E}\{\log[f(Y \mid \boldsymbol{\theta})]\},$$
$$\mathrm{E}\{\log[f(Y \mid \boldsymbol{\theta})]\} \geq \mathrm{E}\{\log[f(Y \mid \boldsymbol{t})]\}.$$

If for all $\boldsymbol{t} \neq \boldsymbol{\theta}$ (and $\boldsymbol{t} \in \Theta$) a) $\mathrm{P}\{f(Y \mid \boldsymbol{\theta}) \neq f(Y \mid \boldsymbol{t})\} > 0$, and b) $\mathrm{E}\{|\log[f(Y \mid \boldsymbol{t})]|\} < \infty$, then the true $\boldsymbol{\theta}$ is the unique maximizer of the expected log-likelihood; see Lemma 2.2 of Newey and McFadden (1994).

More generally, even under misspecification, we can consider the population estimand of the MLE as the maximizer of the expected log-likelihood,

$$\boldsymbol{\theta} \equiv \operatorname*{arg\,max}_{\boldsymbol{t} \in \Theta} \mathrm{E}\{\log[f(Y \mid \boldsymbol{t})]\}. \tag{15.9}$$

Even if $f(\cdot \mid \boldsymbol{\theta})$ is not the true PDF of Y, this is still a well-defined population object, although again it may not have much meaning if there is a lot of misspecfication. Also note that the measure of "close" is not related to any particular distributional feature like the mean or median, but rather the KL divergence of the PDF as a whole.

15.2.1 Asymptotic Theory

The definition of the (Q)ML population parameter in (15.9) suggests estimation by

$$\hat{\boldsymbol{\theta}} = \operatorname*{arg\,max}_{\boldsymbol{t} \in \Theta} \widehat{\mathrm{E}}\{\log[f(Y \mid \boldsymbol{t})]\} = \operatorname*{arg\,max}_{\boldsymbol{t} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \{\log[f(Y_i \mid \boldsymbol{t})]\}, \tag{15.10}$$

which is indeed the MLE. The leading $1/n$ can be removed without changing the maximizer, and the increasing transformation $\exp\{\cdot\}$ can be applied to yield $\prod_{i=1}^{n} f(Y_i \mid \boldsymbol{t})$,

which looks like the joint PDF of n iid random variables Y_i, but such an interpretation is not fundamental or helpful for the asymptotic theory. Also, the key to consistency is (similar to with GMM) uniform (over $\boldsymbol{t} \in \Theta$) convergence in probability of $\hat{Q}(\boldsymbol{t}) \equiv \widehat{\mathrm{E}}\{\log[f(Y \mid \boldsymbol{t})]\}$ to the population $Q(\boldsymbol{t}) \equiv \mathrm{E}\{\log[f(Y \mid \boldsymbol{t})]\}$. For this, iid sampling is not necessary. Nor are we required to model the dependence across observations; using the marginal PDF is sufficient. For example, this is helpful with panel data: we do not need to model the joint likelihood over $t = 1, \ldots, T$, only the marginal distribution in each period t. It may be possible to increase efficiency by modeling the dependence, but as usual, this generally makes the estimator less robust: we are adding information that could reduce our uncertainty, but if the information is wrong then we may introduce bias.

The asymptotic theory for MLE is actually similar to that for GMM: both estimators maximize (or minimize) a criterion function that depends on the data. We need the criterion function (which for finite n is a random function, in the sense that it depends on the data) to converge uniformly in probability to a fixed population criterion function whose unique solution is the true population parameter value, and then we need to look at the function's behavior local to the true parameter to derive the asymptotic normal distribution. Actually, parallel to the GMM results in Newey and McFadden (1994) are MLE results, because the underlying theory is closely related. For example, Theorem 2.5 of Newey and McFadden (1994) provides a consistency result for MLE, and Theorem 3.3 has an asymptotic normality result, given proper specification. As they sketch at the beginning of Section 3 (page 2141), the strategy is similar to that we saw for GMM: take a second-order mean value theorem expansion of the FOC around the true $\boldsymbol{\theta}$, then isolate $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, which yields the negative inverse Hessian (which converges in probability to a fixed matrix) times the normalized sum of scores, to which a CLT applies. That is, with some abuse of notation and $\tilde{\boldsymbol{\theta}}$ values between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ from the mean value theorem expansion, we have an expansion like

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\theta}) + \frac{1}{n}\sum_{i=1}^{n} \underline{\boldsymbol{H}}_i(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{15.11}$$

as in (3.1) of Newey and McFadden (1994), which yields something like their (3.2):

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\overbrace{\left[\frac{1}{n}\sum_{i=1}^{n} \underline{\boldsymbol{H}}_i(\tilde{\boldsymbol{\theta}})\right]}^{\xrightarrow{p} \underline{\boldsymbol{H}}}{}^{-1} \overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\theta})}^{\xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{J}})} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{H}}^{-1}\underline{\boldsymbol{J}}\underline{\boldsymbol{H}}^{-1}). \tag{15.12}$$

However, the asymptotic details are not particularly useful in practice and thus are omitted.

15.2.2 Conditional MLE

We can follow the same MLE approach after conditioning on $\boldsymbol{X}$. That is, we specify a family of conditional PDFs $f(\cdot \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{t})$, taking the log to get the log-likelihood and

then taking the derivative with respect to $\boldsymbol{t}$ to get the score. Specifying the conditional distribution is weaker than specifying the full joint distribution of $(Y, \boldsymbol{X}')$ because a joint distribution uniquely determines the conditional distribution, but there are many joint distributions possible for a given conditional distribution.

Example 15.3. Consider the probit model from (14.22), $Y^* = \boldsymbol{X}'\boldsymbol{\beta} + U$ with $U \sim \mathrm{N}(0,1)$ and $U \perp\!\!\!\perp \boldsymbol{X}$, and observed $Y = \mathbb{1}\{Y^* > 0\}$. From (14.23), $\mathrm{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta})$. Thus, conditional on $\boldsymbol{X} = \boldsymbol{x}$, Y follows a Bernoulli distribution with parameter $\Phi(\boldsymbol{x}'\boldsymbol{\beta})$. Thus, we can adapt the unconditional Bernoulli structure from Example 15.2: for $y \in \{0, 1\}$,

$$f(y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{t}) = [\Phi(\boldsymbol{x}'\boldsymbol{t})]^y [1 - \Phi(\boldsymbol{x}'\boldsymbol{t})]^{1-y}, \tag{15.13}$$

$$\log[f(y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{t})] = y \log[\Phi(\boldsymbol{x}'\boldsymbol{t})] + (1 - y) \log[1 - \Phi(\boldsymbol{x}'\boldsymbol{t})], \tag{15.14}$$

$$\frac{\partial \log[f(y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{t})]}{\partial \boldsymbol{t}} = y \frac{\phi(\boldsymbol{x}'\boldsymbol{t})\boldsymbol{x}}{\Phi(\boldsymbol{x}'\boldsymbol{t})} + (1 - y) \frac{-\phi(\boldsymbol{x}'\boldsymbol{t})\boldsymbol{x}}{1 - \Phi(\boldsymbol{x}'\boldsymbol{t})}. \tag{15.15}$$

Often MLE is used for structural models that hope to have enough detail that the remaining error term is plausibly independent of $\boldsymbol{X}$, and assuming a particular parametric distribution for the error term leads to a tractable MLE. Counterfactual policy simulations can then be run by changing the model parameters and/or $\boldsymbol{X}$ distribution and drawing error terms from the assumed distribution, to generate simulated outcomes.

15.2.3 Efficiency and Standard Errors

Assuming you know the conditional distribution of Y up to just a finite number of unknown parameters is a much stronger assumption than we have seen elsewhere in this book, which has two implications. First, recall that our results are essentially a combination of the data with information we bring to the data in the form of "assumptions." If we bring lots of information, then our results have less uncertainty. One way to phrase this is that our estimator will be more efficient. Indeed, there are results about MLE being asymptotically efficient (attaining the Cramér–Rao lower bound).

Second, however, such strong assumptions are almost surely false. Wooldridge (2010) writes, "efficiency usually comes at the price of nonrobustness" (p. 469). He notes, "there are cases in which MLE turns out to be robust to failure of certain assumptions, but these must be examined on a case-by-case basis" (p. 470). My favorite quip on the topic is from Andres Santos (teaching ECON 220C in Spring 2009 at UCSD), who said something to the effect of, "if you're smart enough to correctly specify a model up to only a finite number of unknown parameters, then you should be doing something much more important than econometrics." The point being: if any mere mortal claims to have properly specified a ML model, then we should not believe them.

The efficiency can be seen as getting the "sandwich form" asymptotic variance in (15.12) to collapse. In GMM, this "collapse" was achieved by a particular choice of the weighting matrix. In ML, this collapse comes if the model is properly specified, in which

case (in the notation of (15.12)) $-\underline{\boldsymbol{H}} = \underline{\boldsymbol{J}}$, or in terms of notation elsewhere in this chapter, $-\mathrm{E}[\underline{\boldsymbol{H}}_i(\boldsymbol{\theta})] = \mathrm{E}[\boldsymbol{s}_i(\boldsymbol{\theta})\boldsymbol{s}_i(\boldsymbol{\theta})']$, known as the **information matrix equality**.

When you compute standard errors, you should use standard errors that are robust to misspecification. That is, you should ask Stata to estimate the full $\underline{\boldsymbol{H}}^{-1}\underline{\boldsymbol{J}}\underline{\boldsymbol{H}}^{-1}$ asymptotic covariance matrix without first assuming the information matrix equality holds, i.e., without assuming proper specification. Otherwise, your standard errors will be too small, even in large samples (asymptotically), because you have assumed something to be true (proper specification) that is not.

Exercises

Exercise IV.1. Consider the binary variable (`inlf` below) of whether or not a married woman is in the labor force, and its relationship with other socioeconomic variables. Note the dataset lacks variable labels, but they can be found online.[1]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse mroz , clear`

c. Run `reg inlf educ exper kidslt6 kidsge6 nwifeinc , vce(robust)`

   i. Describe how to interpret the population model you are estimating.

   ii. Interpret the estimated coefficient on `educ`, and comment on its economic significance.

   iii. Explain what the confidence interval tells us about our uncertainty about the true population value; be precise and explicit.

   iv. Explain one reason (specific to this economic example) that you doubt the true conditional mean function is linear-in-variables.

d. Run `probit inlf educ exper kidslt6 kidsge6 nwifeinc , vce(robust)`

e. Run `margins, dydx(educ exper) atmeans` and `margins, dydx(educ exper)` and explain the difference between the two commands; then compare the results with the OLS estimated slopes.

f. Run `logit inlf educ exper kidslt6 kidsge6 nwifeinc` followed by `margins, dydx(educ exper)` and briefly comment on the economic significance of the difference with the probit-estimated average partial "effects."

g. Consider the following very stylized hypothetical predication application. Imagine you work for a company that offers services for married women in the labor force, and your job is to write code to decide whether or not to buy an online ad for each user that visits another website (that allows you to buy ads for a fixed price). The other website collects all the regressors (predictors) used above, but cannot observe

[1] http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.des

`inlf`, so you need to guess (predict). Given your estimated model from above, you can compute the predicted (conditional) probability of being in the labor force for user i, denoted $\hat{p}_i$. Each ad costs \$0.001; if the user is indeed in the labor force, then expected revenue is \$0.003 (because most people don't click through, etc.), otherwise expected revenue is zero. Assuming your goal is to maximize expected profit, which is a better prediction of being in the labor force (y_i), $\hat{y}_i = \mathbb{1}\{\hat{p}_i > 0.25\}$ or $\hat{y}_i = \mathbb{1}\{\hat{p}_i > 0.75\}$? (That is, you generate binary $\hat{y}_i$, then run the ad if $\hat{y}_i = 1$ but not if $\hat{y}_i = 0$.) Try to find an even better prediction rule for $\hat{y}_i$ as a function of $\hat{p}_i$, and explain why your prediction generates higher expected profit than the two above.

Exercise IV.2. The following models whether an individual is arrested or not in a particular year, given their past criminal justice involvement, demographics, and current employment and income. Variable descriptions are provided in the variable labels in the dataset, originally studied by Grogger (1991). Section II of the original paper provides more details about the data, like covering men in California who were arrested at least once since 1972 and who were born in either 1960 or 1962.

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse grogger , clear`

c. Create the dependent variable: `gen d_arr86 = (narr86>0)`

d. Run `reg d_arr86 pcnv avgsen tottime black hispan , vce(robust)`

   i. Describe how to interpret the population model you are estimating.

   ii. Interpret the estimated coefficients on `pcnv` and `avgsen`, and comment on their economic significance.

   iii. Explain what the confidence intervals tell us about our uncertainty about the true population values; be precise and explicit.

   iv. Explain one reason (specific to this economic example) that you doubt the true conditional mean function is linear-in-variables.

e. Run `probit d_arr86 pcnv avgsen tottime black hispan , vce(robust)`

f. Run `margins, dydx(pcnv avgsen) atmeans` and `margins, dydx(pcnv avgsen)` and explain the difference between the two commands; then compare the results with the OLS estimated slopes.

g. Run `logit d_arr86 pcnv avgsen tottime black hispan` followed by `margins , dydx(pcnv avgsen)` and briefly comment on the economic significance of the difference with the probit-estimated average partial "effects" of `pcnv` and `avgsen`.

h. Now consider trying to predict whether or not an individual will be arrested over the next 12 months for the purpose of targeting an intervention that includes 1-on-1 mentoring, job training, and subsidized housing, and imagine you only care about reducing arrests (not any other outcome). There is no budget constraint, but

the opportunity cost of spending $1 on this program is not spending that $1 on a different program to help reduce arrests. After running your `probit` command from above, pretend we then loaded a new dataset that includes only the predictor variables but not `d_arr86`, and then generate the predicted arrest probabilities with `predict phat` along with two possible binary predictions of being arrested using two different probability thresholds:

```
gen target25 = (phat>0.25)
gen target48 = (phat>0.48)
```

Finally, because actually we do know the true `d_arr86` values, compare the true and predicted values:

```
tab d_arr86 target25
tab d_arr86 target48
```

i. For the 25% threshold: how many "false negatives" (`target25=0` but they are arrested) and "false positives" (`target25=1` but they are not arrested) are there? How many are there for the 48% threshold?

ii. Qualitatively, what is the cost of a false negative? What's the cost of a false positive?

iii. Adding whatever additional details you need (about costs, benefits, etc.) for the following to be true: why might the higher threshold be preferred here?

iv. Would a 50% threshold be even better? 60%? Explain why/not.

Exercise IV.3. Consider the relationship between whether or not somebody reports being in good health (`gdhlth`) and other variables. This dataset is from 1975. Note the dataset lacks variable labels, but they can be found online.[2]

a. As usual, make sure the command `bcuse` is installed: `ssc install bcuse`

b. Load the data: `bcuse sleep75 , clear`

c. Run `reg gdhlth c.age##c.age male##yngkid sleep totwrk educ , vce(robust)`

i. Describe how to interpret the population model you are estimating.

ii. Interpret the estimated coefficients on `age` and its square, and comment on their economic significance.

iii. Explain what the confidence intervals tell us about our uncertainty about the true population values; be precise and explicit.

iv. Explain one reason (specific to this economic example) that you doubt the true conditional mean function has this exact functional form.

d. Run `margins , dydx(age) at(age=(30(15)60)) vsquish`

[2] http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.des

e. Run `probit gdhlth c.age##c.age male##yngkid sleep totwrk educ , vce(robust)` and then `margins , dydx(age) at(age=(30(15)60)) vsquish` and compare with the OLS results.

f. Repeat part (e) but with `logit` instead of `probit` and briefly compare to the `probit` results.

g. Now imagine you work for a health insurance company and want to predict if an individual is in good health; if not, the insurance company will call them with a reminder to visit the doctor. After running your `probit` command from above, pretend we then loaded a new dataset that includes only the predictor variables but not `gdhlth`, and then generate the predicted arrest probabilities with `predict phat` along with two possible binary predictions of being arrested using two different probability thresholds:

```
gen target50 = (phat>0.50)
gen target80 = (phat>0.80)
```

Finally, because actually we do know the true `gdhlth` values, compare the true and predicted values:

```
tab gdhlth target50
tab gdhlth target80
```

   i. For the 50% threshold: how many extraneous phone calls would be made (`target50=0` but `gdhlth=1`)? How many individuals not in good health fail to get called (`target50=1` but `gdhlth=0`)? How many of each for the 80% threshold?

   ii. Qualitatively, what is the cost of calling somebody who's actually in good health? What's the cost of failing to call somebody in bad health?

   iii. Adding whatever additional details you need (about costs, benefits, etc.) for the following to be true: why might the higher 80% threshold be preferred here?

Bibliography

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1):265–296. URL https://doi.org/10.3982/ECTA12675. [25]

Anderson, T. W. and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1):46–63. URL https://doi.org/10.1214/aoms/1177730090. [123]

Anderson, T.W. and Cheng Hsiao. 1982. "Formulation and estimation of dynamic models using panel data." *Journal of Econometrics* 18 (1):47–82. URL https://doi.org/10.1016/0304-4076(82)90095-1. [185]

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80 (3):313–336. URL https://www.jstor.org/stable/2006669. [104]

Arellano, Manuel and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58 (2):277–297. URL https://doi.org/10.2307/2297968. [185]

Baum, Christopher F. 2012. "BCUSE: Stata module to access instructional datasets on Boston College server." Statistical Software Components S457508, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s457508.html. Revised 12 Nov 2021. [9]

Baum, Christopher F. and Mark E. Schaffer. 2013. "BCUSE: Stata module to perform asymptotic covariance estimation for iid and non-iid data robust to heteroskedasticity, autocorrelation, 1- and 2-way clustering, and common cross-panel autocorrelated disturbances." Statistical Software Components S457689, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s457689.html. Revised 30 July 2015. [143]

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2002. "IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation." Statistical Software Components S425401, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s425401.html. Revised 10 May 2022. [9, 119, 143]

Blackburn, McKinley and David Neumark. 1992. "Unobserved Ability, Efficiency Wages,

and Interindustry Wage Differentials." *Quarterly Journal of Economics* 107 (4):1421–1436. URL https://www.jstor.org/stable/2118394. [148]

Box, G. E. P. 1979. "Robustness in the Strategy of Scientific Model Building." Tech. Rep. 1954, Mathematics Research Center, University of Wisconsin–Madison. URL http://www.dtic.mil/docs/citations/ADA070213. [125]

Callaway, Brantly, Fernando Rios-Avila, and Pedro H. C. Sant'Anna. 2021. "CSDID: Stata module for the estimation of Difference-in-Difference models with multiple time periods." Statistical Software Components S458976, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s458976.html. Revised 04 Nov 2022. [155, 176]

Callaway, Brantly and Pedro H.C. Sant'Anna. 2021a. "did: Difference in Differences." URL https://bcallaway11.github.io/did/. R package version 2.1.2. [155, 176]

———. 2021b. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225 (2):200–230. URL https://doi.org/10.1016/j.jeconom.2020.12.001. [176]

Card, David. 1990. "The impact of the Mariel boatlift on the Miami labor market." *Industrial & Labor Relations Review* 43 (2):245–257. [159]

———. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, edited by Louis N. Christophides, E. Kenneth Grant, and Robert Swidinsky. University of Toronto Press, 201–222. [92, 145]

Chen, Qihui and Zheng Fang. 2019. "Improved inference on the rank of a matrix." *Quantitative Economics* 10 (4):1787–1824. URL https://doi.org/10.3982/QE1139. [120]

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." In *Handbook of Econometrics*, vol. 6B, edited by James J. Heckman and Edward E. Leamer, chap. 76. Elsevier, 5549–5632. URL https://doi.org/10.1016/S1573-4412(07)06076-X. [213]

Chernozhukov, Victor and Christian Hansen. 2005. "An IV Model of Quantile Treatment Effects." *Econometrica* 73 (1):245–261. URL https://www.jstor.org/stable/3598944. [108]

Croissant, Yves and Giovanni Millo. 2008. "Panel Data Econometrics in R: The plm Package." *Journal of Statistical Software* 27 (2):1–43. URL https://doi.org/10.18637/jss.v027.i02. [165]

de Castro, Luciano, Antonio F. Galvao, David M. Kaplan, and Xin Liu. 2019. "Smoothed GMM for quantile models." *Journal of Econometrics* 213 (1):121–144. URL https://doi.org/10.1016/j.jeconom.2019.04.008. [139]

Deming, W. Edwards and Frederick F. Stephan. 1941. "On the Interpretation of Censuses as Samples." *Journal of the American Statistical Association* 36 (213):45–49. URL https://www.jstor.org/stable/2278811. [26]

Finlay, Keith, Leandro Maschietto Magnusson, and Mark E. Schaffer. 2013. "WEAKIV: Stata module to perform weak-instrument-robust tests and confidence intervals for instrumental-variable (IV) estimation of linear, probit and tobit models." Statistical

Software Components S457684, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s457684.html. Revised 18 Oct 2016. [123, 143]

Freeman, Donald G. 2007. "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws." *Contemporary Economic Policy* 25 (3):293–308. URL https://doi.org/10.1111/j.1465-7287.2007.00039.x. [189]

Graddy, Kathryn. 1995. "Testing for imperfect competition at the Fulton fish market." *RAND Journal of Economics* 26 (1):75–92. URL https://www.jstor.org/stable/2556036. [144]

Grogger, Jeffrey. 1991. "Certainty vs. Severity of Punishment." *Economic Inquiry* 29 (2):297–309. URL https://doi.org/10.1111/j.1465-7295.1991.tb01272.x. [220]

Hansen, Bruce E. 2020a. "Econometrics." URL https://www.ssc.wisc.edu/~bhansen/econometrics. Textbook draft. [xv, 3]

———. 2020b. "Introduction to Econometrics." URL https://www.ssc.wisc.edu/~bhansen/probability. Textbook draft. [xv, 3]

Hansen, Lars Peter. 1982. "Large sample properties of generalized method of moments estimators." *Econometrica* 50 (4):1029–1054. URL https://www.jstor.org/stable/1912775. [124]

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd ed. URL https://web.stanford.edu/~hastie/ElemStatLearn. Corrected 12th printing, January 13, 2017. [199]

Hausman, Jerry A. and William E. Taylor. 1981. "Panel Data and Unobservable Individual Effects." *Econometrica* 49 (6):1377–1398. URL https://doi.org/10.2307/1911406. [176]

Holtz-Eakin, Douglas, Whitney Newey, and Harvey S. Rosen. 1988. "Estimating vector autoregressions with panel data." *Econometrica* 56 (6):1371–1395. URL https://www.jstor.org/stable/1913103. [185]

Imbens, Guido and Jeffrey M. Wooldridge. 2007. "What's New in Econometrics: Estimation of Average Treatment Effects Under Unconfoundedness." NBER summer lecture notes, available at https://www.nber.org/WNE/lect_1_match_fig.pdf. [55]

Kaplan, David M. 2019. "distcomp: Comparing distributions." *Stata Journal* 19 (4):832–848. URL https://doi.org/10.1177/1536867x19893626. [11]

———. 2021. "Distributional and Nonparametric Econometrics." URL https://kaplandm.github.io/teach.html. Textbook draft. [xv, 28, 35, 39, 41, 50, 56, 67, 101, 108, 111, 169, 210]

———. 2022a. *Introductory Econometrics: Description, Prediction, and Causality*. Columbia, MO: Mizzou Publishing, 3rd ed. URL https://www.themizzoustore.com/p-236916-introductory-econometrics-description-prediction-and-causality.aspx. [xv, 17, 24, 25, 27, 28, 31, 35, 36, 38, 43, 44, 47, 51, 52, 53, 54, 76, 79, 81, 85, 157, 160, 162, 172, 177, 202]

———. 2022b. "Smoothed instrumental variables quantile regression." *Stata Journal*

22 (2):379–403. URL https://doi.org/10.1177/1536867X221106404. [11, 108]

Kaplan, David M. and Yixiao Sun. 2017. "Smoothed Estimating Equations for Instrumental Variables Quantile Regression." *Econometric Theory* 33 (1):105–157. URL https://doi.org/10.1017/S0266466615000407. [108]

Kaplan, David M. and Longhao Zhuo. 2021. "Frequentist properties of Bayesian inequality tests." *Journal of Econometrics* 221 (1):312–336. URL https://doi.org/10.1016/j.jeconom.2020.05.015. [31]

Kiefer, J. and J. Wolfowitz. 1956. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters." *Annals of Mathematical Statistics* 27 (4):887–906. URL https://projecteuclid.org/euclid.aoms/1177728066. [213]

Kleibergen, Frank and Richard Paap. 2006. "Generalized reduced rank tests using the singular value decomposition." *Journal of Econometrics* 133 (1):97–126. URL https://doi.org/10.1016/j.jeconom.2005.02.011. [120]

Kleibergen, Frank, Mark E. Schaffer, and Frank Windmeijer. 2007. "RANKTEST: Stata module to test the rank of a matrix." Statistical Software Components S456865, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s456865.html. Revised 29 Sep 2020. [9, 119, 143]

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4):604–620. URL https://www.jstor.org/stable/1806062. [91]

Lewbel, Arthur. 2019. "The Identification Zoo: Meanings of Identification in Econometrics." *Journal of Economic Literature* 57 (4):835–903. URL https://doi.org/10.1257/jel.20181361. [37]

Lucas, Robert E., Jr. 1976. "Econometric policy evaluation: A critique." In *Carnegie–Rochester Conference Series on Public Policy*, vol. 1. North-Holland, 19–46. [36]

Montiel Olea, José Luis and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics* 31 (3):358–369. URL https://doi.org/10.1080/00401706.2013.806694. [122]

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46 (1):69–85. URL https://doi.org/10.2307/1913646. [176]

Naqvi, Asjad, Fernando Rios-Avila, and Pedro H. C. Sant'Anna. 2021. "DRDID: Stata module for the estimation of Doubly Robust Difference-in-Difference models." Statistical Software Components S458977, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s458977.html. Revised 18 Oct 2022. [155]

Newey, Whitney K. and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." In *Handbook of Econometrics*, vol. 4, edited by Robert F. Engle and Daniel L. McFadden, chap. 36. Elsevier, 2111–2245. [127, 135, 137, 138, 140, 141, 214, 215]

Pflueger, Carolin E. and Su Wang. 2013. "WEAKIVTEST: Stata module to perform weak instrument test for a single endogenous regressor in TSLS and LIML." Statistical Software Components S457732, Boston College Department of Economics. URL https://ideas.repec.org/c/boc/bocode/s457732.html. Revised 13 Nov 2020. [122, 143]

Sargan, J. D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica* 26 (3):393–415. URL https://doi.org/10.2307/1907619. [124]

Staiger, Douglas and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3):557–586. URL https://doi.org/10.2307/2171753. [121, 122]

StataCorp. 2017. "Stata Statistical Software: Release 15." College Station, TX: StataCorp LP. [9]

Stock, James H. and Mark W. Watson. 2015. *Introduction to Econometrics*. Pearson, 3rd updated ed. URL https://www.pearson.com/us/higher-education/product/Stock-Introduction-to-Econometrics-Update-3rd-Edition/9780133486872.html. [69, 70]

Stock, James H. and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by Donald W. K. Andrews and James H. Stock, chap. 5. Cambridge University Press, 80–108. URL https://www.cambridge.org/9780521844413. [122]

Street, Brittany. 2022. "The Impact of Economic Opportunity on Criminal Behavior: Evidence from the Fracking Boom." Working Paper, available at https://sites.google.com/site/brittanyrstreet/research. [159]

van der Vaart, Aad W. 1998. *Asymptotic Statistics*. Cambridge: Cambridge University Press. URL https://books.google.com/books?id=UEuQEM5RjWgC. [137]

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd ed. URL https://www.worldcat.org/oclc/831625495. [38, 51, 79, 81, 84, 85, 99, 109, 127, 132, 134, 138, 140, 166, 167, 168, 173, 176, 216]

———. 2020. *Introductory Econometrics: A Modern Approach*. Cengage, 7th ed. [91]

Zeileis, Achim. 2004. "Econometric Computing with HC and HAC Covariance Matrix Estimators." *Journal of Statistical Software* 11 (10):1–17. URL https://doi.org/10.18637/jss.v011.i10. [165]

Index