Notes to Self for Teaching ECON 9473 with Wooldridge's *Econometric Analysis of Cross Section and Panel Data*

David M. Kaplan

February 28, 2019

Chapter 0: First Days of Class

0.1 Basic Information

1. As students walk in, say hi and ask their names. (Look at myZou roster with photos the night before to try to get head start.)

2. Say something about me: background, time at Mizzou, research, etc.

3. Mention course website, should read syllabus there, hopefully have access already (if I didn't forget to hit Publish again).

4. Some highlights of syllabus: textbook, exams (in-class), ES grading (mostly effort-based; can work together, but learn more if try on your own first), etc.

5. In-class structure: lecture on blackboard punctuated by discussions. I'll write "Q:" on the board, then check if it's clear, then give you a few minutes (1–4, depending on the question) to discuss w/ your neighbor(s), then share your (or your neighbor's) ideas with full class, and/or you may ask questions that came up during your discussions.

6. Do example discussion question: what's one thing you remember from the last econometrics or statistics class you took? (Some non-econ-PhD students, so may not have been ECON 9472.) Should be loud and noisy in here!

0.2 Notation

In-class (blackboard) notation: uppercase random, lowercase non-random; underbar for vector, under-tilde for matrix, nothing for scalar. Vectors usually column vectors, except in Wooldridge the regressor vectors are row vectors so I'll do that to avoid confusion (and avoids needing the transpose symbol as much).

In these typed notes: Y is scalar rv, y is scalar value, $\mathbf{X}$ is random row vector, $\mathbf{x}$ is non-random row vector, $\boldsymbol{\beta}$ is non-random column vector, $\underset{\sim}{\mathbf{M}}$ is random matrix, $\underset{\sim}{\mathbf{m}}$ is non-random matrix. Also e.g. $\mathbf{X} = (X_1, X_2, X_3, \ldots)$, and row-$i$ column-j element of $\underset{\sim}{\mathbf{M}}$ is M_{ij}, etc.

Figure 1 shows one perspective (similar to 9472?) of what econometrics is about.

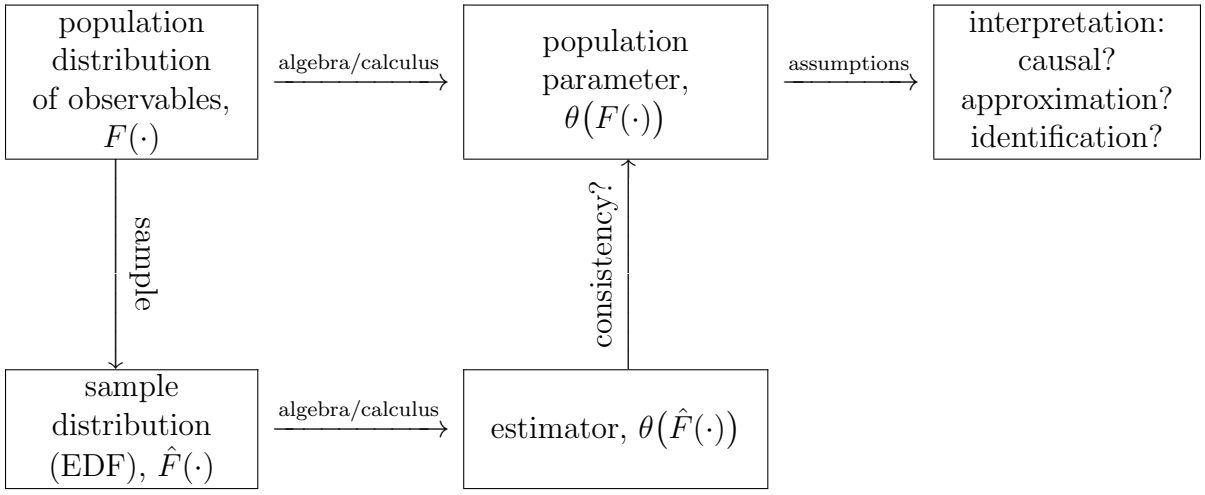

Figure 1: Map of the world of econometrics (Mercator projection).

0.3 Some Dichotomies

0.3.1 Statistical vs. causal/structural

Before saying the name of this subsection, ask the following.

Discussion Question 0.1. Interpret $Y = X\beta + U$; in particular, what does β mean? U? Concrete example: Y wage, $X = (1, \mathbb{1}\{\text{college}\})$.

Statistical models: CEF, LP/BLA/BLP. Just summarize features of joint distribution of observables (assuming no missing data etc.). The U has precise mathematical properties, as does β, but maybe no economic meaning.

Causal/structural model: U and β have economic meaning, but maybe not $\mathrm{E}(XU) = 0$.

0.3.2 Prediction/forecasting vs. policy analysis

Some work in economics focuses on prediction. Forecasting is a special case (predicting the future). For example, what will the price of oil be next quarter, or unemployment, or GDP, etc. We are not changing anything in the world; we just predict some value we don't know.

For prediction, often learning the conditional expectation function (CEF) is sufficient, or even the best linear predictor (BLP). The CEF of Y given $\mathbf{X}$ is the "best" predictor of Y given $\mathbf{X}$ if the loss function is quadratic. (With other loss functions, the optimal predictor may be the conditional median, or other conditional quantiles, or the conditional mode, or something else altogether.)

However, this does not suffice for policy analysis. "Policy" refers to changing the world, specifically changing X values. This could be national monetary policy, or an individual decision to attend college, or anything in between. In each case, the policy decision is about how to change some X. So, it is important to know (or guess) how Y would respond to

different possible changes in X (different policies). For values of X other than the current value, this is called counterfactual analysis.

Prediction is not sufficient for policy analysis because we want to hold U fixed while changing X. E.g., I want to hold fixed my ability U when considering whether to attend college ($X = 1$) or not ($X = 0$). In contrast, optimal prediction of Y would account for the fact that perhaps U is on average higher for individuals with $X = 1$.

0.3.3 Structural vs. causal

I still need to understand this better myself (`https://www.aeaweb.org/articles?id=10.1257/jel.20181361`), but I think the simplest way to distinguish structural and causal models is the following.

The causal approach (aka "reduced form" approach, but don't get confused w/ 1st-stage of 2SLS) focuses on the causal effect of a certain change in X on Y. It does not worry about modeling the underlying mechanism. For example, the causal approach might ask, for a certain population, what's the average effect on salary of going to college? It does not try to model the individual's choice in terms of future salary, consumption value of college, tuition and other costs, etc.

In a structural model, the parameters have economic meaning. For example, there could be a risk aversion parameter, elasticity parameters, a demand curve, etc.

Often, there is a tradeoff between internal and external validity. If we could really learn how economic agents make decisions and estimate all the parameters involved, then we could consider all sorts of counterfactuals. However, modeling economic behavior in such detail usually requires stronger (less realistic) assumptions. The causal approach often has more plausible assumptions (greater internal validity), but it is hard to generalize results to other settings or populations or policies.

Ideally, both approaches can be used. In this class we focus on the reduced form approach, but some models can be used for structural estimation, too.

0.3.4 General equilibrium vs. partial equilibrium

Both general equilibrium (GE) and partial equilibrium (PE) approaches are useful. GE tries to model entire markets (or multiple related markets), whereas PE takes the current market equilibria as given (e.g., price of childcare, wages, etc.).

Often, there is a tradeoff between internal validity and external validity. To evaluate a policy that will indeed affect markets significantly, then GE is clearly necessary; hence, GE models are used by the Fed. GE models explicitly try to achieve great external validity. In contrast, PE models may lack external validity if the results are sensitive to the current market conditions. However, PE models often have much weaker (more realistic) assumptions, since they don't try to model as much.

For this class, we primarily consider PE models.

0.3.5 Identification vs. estimation (and inference)

Much of ECON 9472 focused on estimation, like OLS, and deriving properties of estimators like consistency and asymptotic normality. It also discussed how to derive valid confidence intervals and hypothesis tests from these results. Here, we won't worry about proving such results; we'll think critically about the assumptions needed for consistency and asymptotic normality, but trust somebody else has proved the theorems.

Instead, we'll focus more on identification. One way to think of identification is: if we had infinite data, could we learn the true value of a parameter? With infinite data, it seems like we should know everything, but in economics, often we still can't learn what we want to know. For example, even if you knew the true population salary distributions for individuals with and without college degrees, you still wouldn't be able to learn the causal effect of going to college.

Identification tends to be emphasized more in econometrics than other branches of statistics (that emphasize prediction).

0.4 Some Details

If $Y = X\beta + U$ is structural or causal model, and if $U \perp\!\!\!\perp X$, then can estimate β using OLS if iid sample. Also sufficient to have $\mathrm{E}(U \mid X) = 0$.

Random coefficients model: could also have $Y = U_0 + U_1 X$, where U_0 is random intercept and U_1 is random slope. "Random" allows different individuals to have different coefficients. So, the effect on Y of increasing X by one unit is U_1, which can be different across individuals. The CEF is

$$\begin{aligned}&\mathrm{E}(Y \mid X)\\ &= \mathrm{E}(U_0 + U_1 X \mid X)\\ &= \mathrm{E}(U_0 \mid X) + \mathrm{E}(U_1 \mid X)X.\end{aligned}$$

If $(U_0, U_1) \perp\!\!\!\perp X$ (or again mean independence is sufficient), then

$$\begin{aligned}&= \mathrm{E}(U_0 \mid X) + \mathrm{E}(U_1 \mid X)X\\ &= \overbrace{\mathrm{E}(U_0)}^{\equiv\beta_0} + \overbrace{\mathrm{E}(U_1)}^{\equiv\beta_1} X.\end{aligned}$$

Thus the CEF is linear, so (β_0, β_1) can be estimated by OLS. The identifying assumption $(U_0, U_1) \perp\!\!\!\perp X$ allows us to *interpret* the CEF slope β_1 as the average effect of X on Y, i.e., $\mathrm{E}(U_1)$.

ACE and CIA: end of Chapter 2 in Hansen (2018). Assume Y is determined as $Y = h(X_1, X_2, U)$, where U is vector of unobservables. We're interested in the effect of X_1 on Y; X_2 is just a vector of control variables. Imagine X_1 is binary, 0 or 1. Then define causal effect of X_1 on Y as

$$\mathrm{C}(x_2, u) = h(1, x_2, u) - h(0, x_2, u), \tag{0.1}$$

i.e., how Y changes if we change X_1 from 0 to 1 while holding constant $X_2 = x_2$ and $U = u$. Define the average causal effect as

$$\mathrm{ACE}(x_2) = \mathrm{E}[\mathrm{C}(X_2, U) \mid X_2 = x_2]. \tag{0.2}$$

It averages the causal effect over different values of U for "individuals" with the same X_2. The conditional independence assumption (CIA) in this case is

$$U \perp\!\!\!\perp X_1 \mid X_2, \tag{0.3}$$

i.e., conditional on X_2 (like, among everyone w/ same X_2), U and X_1 are independent. This identifying assumption can be shown to link the CEF (a statistical object) to the ACE (a causal object). Specifically,

$$\text{ACE}(x_2) = \text{E}[Y \mid X_1 = 1, X_2 = x_2] - \text{E}[Y \mid X_1 = 0, X_2 = x_2]. \tag{0.4}$$

Many more identification results; just trying to get a sense of what they look like. Usually there is some "identifying assumption" (usually not testable in data, just have to argue) that allows us to equate our causal object of interest with some statistical object that can be consistently estimated.

0.5 Stata/R

Reserve computer lab (like Middlebush 7 or 8 [basement], ideally) for a class (or two) to do Stata and/or R: `https://25live.collegenet.com/missouri/`

Chapter 1

In §1.1, clarify: the "control variables" $\mathbf{c}$ is like Hansen's $\mathbf{X}_2$; want enough that CIA holds, so can interpret "partial effect" (CEF derivative) as ACE.

Skip §1.2

In §1.3, look at Example 1.1. Say that the model is specifying a particular $h(\cdot)$ (in Hansen notation) that is 1) parameteric, 2) linear-in-parameters, 3) linear-in-variables, 4) additively separable in U.

Clarify that structural error U in Wooldridge is *not* the CEF or LP error e from Hansen.

Discussion Question 1.1 (Example 1.1)**.** Why might you doubt the functional form of the structural model in (1.1)? Use your intuition/knowledge, from both economics and the real world. Focus on the linearity in regressors and the lack of interaction between U and regressors.

Emphasize the importance of population of interest vs. population studied. Different policy questions have different populations of interest.

Discussion Question 1.2 (college subsidies)**.** Imagine you're considering a national policy to increase college graduation rates by increasing subsidies to students attending college. You want to know how wages would be affected by the policy change. What do you think is the appropriate population of interest? E.g., everyone in the country, or only people who currently have college degrees, or only people who currently *don't* have college degrees, etc.

Chapter 2

§2.1: again "structural CEF" can think of like Hansen, "effect of $w\ldots$" like ACE.

"the notion of conditional expectation is fundamental": not sure—I think quantiles can also be quite helpful =)

Example 2.1: describe why (2.2) is linear-in-variables (b/c linear combination of $1, x_1, x_2$), linear-in-parameters (b/c lincom of $\beta_0, \beta_1, \beta_2$), and parametric (b/c only 3 unknown parameter values, $3 < \infty$; no unknown functions). Nonparametric: $h(x_1)$, unknown function $h(\cdot)$; more in ECON 9476. Confusing: parametric CEF can correspond to a "semiparametric" model $Y = \mathbf{X}\boldsymbol{\beta} + U$ if we only specify $\mathrm{E}(U \mid \mathbf{X}) = 0$ (but not whether U is normal, etc.).

Discussion Question 2.1 (linearity)**.** For each of (2.3), (2.4), and (2.5), say why it is or isn't a) linear-in-variables, b) linear-in-parameters, c) parametric.

§2.2.2: "partial effect" is not necessarily "causal" (poor choice of terminology); and, for nonlinear functions, derivatives aren't as accurate as just taking differences of the function at different $\mathbf{X}$ values.

Discussion Question 2.2 (misspecification)**.** Let X_1 be years of education, X_2 years of experience, Y annual income/salary. For (2.2), the PE wrt x_2 is just β_2, a constant, which does not seem realistic: each additional year of experience is associated with a β_2 higher average salary, regardless of initial experience, regardless of education, etc. 1) compute the partial effect wrt x_2 in models (2.2), (2.3), (2.4). 2) interpret the PE "economically." 3) argue why it might be more realistic than (2.2), but still unrealistic.

pp. 16–17: be careful with log models and elasticities.

Chapter 4

p. 53, "structural model": I think I'd define it somewhat differently (see intro material)

Show basic identification result: if structural model $Y = \mathbf{X}\boldsymbol{\beta} + U$ and $\text{Cov}(X_j, U) = 0$ for all j, and $\text{E}(U) = 0$, then $\boldsymbol{\beta}$ is LPC vector; and LPC is statistical object that OLS consistently estimates. "Identification" meaning the joint dist of $(Y, \mathbf{X})$ uniquely determines the value of the structural parameter vector $\boldsymbol{\beta}$, since $\boldsymbol{\beta}$ equals the LPC (which is uniquely determined, up to finite moments/invertibility at least).

In linear model, endogenous means $\text{Cov}(X_j, U) \neq 0$, exogenous means $\text{Cov}(X_j, U) = 0$.

Discussion Question 4.1 (OVB)**.** Let $X = 1$ if college degree ($X = 0$ o/w), Y wage. Consider structural model $Y = \beta_0 + \beta_1 X + U$. Describe one individual characteristic that both affects an individual's wage and influences her decision to get a college degree. If this were the only variable in U, then how would it bias the OLS estimate of β_1? That is, would it make it look like college increases wage more than it really does, or less?

Discussion Question 4.2 (simultaneity)**.** Let Y be a city's homicide rate (per capita per year) and X its number of police officers per capita. How do you think a city decides what X to have? Do you think large Y would cause cities to choose larger or small X? In which direction might this bias the OLS estimate of β_1 in structural model $Y = \beta_0 + \beta_1 X + U$?

Go over three sources of endogeneity part: OVB, measurement error, simultaneity. Maybe add simultaneous/reverse causality, since it's a little different than OVB or simultaneity? OVB: mention self-selection and ability/edu example. Simultaneity: also supply and demand, price and quantity simultaneously determined.

Discussion Question 4.3 (frequentist vs. Bayesian inference)**.** Discuss `https://xkcd.com/1132`. Null hypothesis H_0: sun not exploded. Alternative hypothesis H_1: sun exploded. Do you think the sun exploded? Why/not? Was the p-value computed incorrectly?

Discussion Question 4.4 (jelly beans)**.** Discuss `https://xkcd.com/882`. Why does the newspaper claim "only 5% chance of coincidence"? Is that claim correct? Why/not? Hint(?): $0.95^{20} = 0.36$.

Section 4.3.1

(4.19,4.20) structural CEF; (4.21) structural (same parameters from structural model 4.19) but maybe not CEF. Plug LP in (4.23) into (4.19), argue that $(v + \gamma r)$ satisfies LP error property, so $(\beta_j + \gamma\delta_j)$ are the LP coefficients that OLS consistently estimates, i.e., plim $\hat{\beta}_j = \beta_j + \gamma\delta_j$.

"Asymptotic bias" defined here as $\operatorname{plim}\hat{\beta} - \beta$, vs. "bias" $\mathrm{E}(\hat{\beta}) - \beta$. Asymptotic bias is worse problem: still big problem even w/ infinite data, whereas bias may go to zero (i.e., may be very small w/ large enough dataset).

Asy bias is $\gamma\delta_j$. This is zero if $\gamma = 0$, i.e., if omitted variable Q isn't actually a causal determinant of Y. Also zero if $\delta_j = 0$, i.e., Q and X_j are not statistically related after "partialing out" other regressors (i.e., not partially correlated). So, OVB is problem only if *both* Q determines Y causally *and* statistical relationship b/w Q and some X_j.

If $\gamma > 0$ and $\delta_j > 0$, then positive (upward) bias: systematically overestimate. (Note: could have $\beta < \hat{\beta} < 0$, so smaller absolute value/magnitude even though "upward" bias; can be confusing.) Idea: since $\delta_j > 0$, Q tends to be large when X_j is large. Then since $\gamma > 0$, the larger Q tend to make Y larger. Altogether, larger X_j are associated with larger Y values, but because of Q effect, not X_j effect.

Alternatively, if $\delta_j < 0$, then large X_j assoc'd with small Q, which makes Y smaller; negative (downward) bias.

Discussion Question 4.5 (Ex 4.2)**.** Imagine structural wage equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \gamma Q + U$, where Y is log wage, X_1 years of experience, X_2 years edu, Q "ability," U structural error. Assume $\mathrm{E}(V \mid X_1, X_2, Q) = 0$ and LP of Q given regressors is $Q = \delta_0 + 0 + 0 + \delta_3 X_2 + R$, so R is uncorrelated with all regressors. What's asymptotic bias of $\hat{\beta}_3$? Why? Hint: for each of δ_3, β_3, γ, is it > 0, < 0, or $= 0$? Then use formula to compare $\operatorname{plim}\hat{\beta}_3$ to true β_3.

Section 4.3.2 (through p. 69)

Proxy variable: hope to find observable variable Z that can "proxy" for unobservable Q in regression. Hope Z is very closely related to Q, so it can control for Q and reduce (or remove) OVB.

First property of proxy var: redundant/ignorable in structural model: $\mathrm{E}(Y \mid \mathbf{X}, Q, Z) = \mathrm{E}(Y \mid \mathbf{X}, Q)$. Once we know $\mathbf{X}$ and Q, essentially no additional information about (mean of) Y coming from Z. Or, imagine Z is in the structural CEF, but its coefficient is zero. Note that this property implies $\mathrm{E}(V \mid \mathbf{X}, Q, Z) = 0$.

Second: more complicated. Idea: same computations for asymptotic bias as before, but now we have Z in the LP. So LP of Q is

$$Q = \mathbf{X}\boldsymbol{\rho} + \theta Z + R, \tag{4.1}$$

where by definition $\mathrm{Cov}(Z, R) = 0$ and $\mathrm{Cov}(\mathbf{X}, R) = \mathbf{0}$. Plug into structural CEF, and rearrange:

$$Y = \mathbf{X}(\boldsymbol{\beta} + \gamma\boldsymbol{\rho}) + \gamma\theta Z + (\gamma R + V). \tag{4.2}$$

Since V is a CEF error for $\mathrm{E}(Y \mid \mathbf{X}, Q, Z)$, then $\mathrm{Cov}(\mathbf{X}, V) = \mathbf{0}$ and $\mathrm{Cov}(Z, V) = 0$. Similarly for R, per above. Thus, this shows the linear projection of Y onto $(\mathbf{X}, Z)$ in error form, and the LPCs are $\boldsymbol{\beta} + \gamma\boldsymbol{\rho}$ and $\gamma\theta$. Thus, the plim of the OLS estimators are like $\operatorname{plim}\hat{\beta}_j = \beta_j + \gamma\rho_j$.

If Z is a perfect proxy, then $\boldsymbol{\rho} = \mathbf{0}$, so OLS consistently estimates the entire structural parameter vector $\boldsymbol{\beta}$. The interpretation of $\boldsymbol{\rho} = \mathbf{0}$ is that Z is so closely related to Q, once we

"partial out" Z from Q, what's left is uncorrelated with any regressor. Seen differently: if we take the linear projection of Q onto $(1, Z)$, that LP error is uncorrelated with $\mathbf{X}$ regressors. So e.g. Z could be a noisy measurement of Q as long as the noise is uncorrelated with the regressors.

More realistically, we can hope Z is an imperfect proxy, where $\boldsymbol{\rho} \neq \mathbf{0}$, but $|\rho_j| < |\delta_j|$ (the δ_j from the previous section), so the magnitude of asymptotic bias is reduced.

Discussion Question 4.6 (Ex 4.3)**.** Consider (4.29), structural model $Y = \mathbf{X}\boldsymbol{\beta} + \gamma Q + V$, where Y is log wage, $\mathbf{X}$ includes a constant, experience, tenure, dummies for marriage, South, urban, and black, and finally education (regressor of interest), and Q is "ability." Consider Z as IQ score, a possible proxy for ability. 1) argue Z is redundant; 2) argue $\rho_7 \neq 0$ (imperfect proxy; X_7 is education); 3) should we include Z, or just reg Y $\mathbf{X}$?

Section 4.4.1

True but unobserved ("latent") Y^*, measured as Y (observable). Measurement error $M \equiv Y - Y^*$, so $Y = Y^* + M$.

Simple case: want to learn $\mathrm{E}(Y^*)$. If we observed Y^*, then just take sample average: consistent for population mean; with infinite data, we'd learn $\mathrm{E}(Y^*)$ perfectly. With $M \neq 0$, can we learn $\mathrm{E}(Y^*)$ with infinite data? That is: is $\mathrm{E}(Y^*)$ "identified"?

With infinite data, we can learn $\mathrm{E}(Y)$, the population mean of the observable measure Y. So, e.g., $\mathrm{E}(Y^*)$ is identified if $\mathrm{E}(Y^*) = \mathrm{E}(Y)$.

Plugging in: $\mathrm{E}(Y) = \mathrm{E}(Y^* + M) = \mathrm{E}(Y^*) + \mathrm{E}(M)$. So if $\mathrm{E}(M) = 0$, then identified; $\mathrm{E}(M) = 0$ is "identifying assumption." But if $\mathrm{E}(M) > 0$, then $\mathrm{E}(Y) > \mathrm{E}(Y^*)$, so the sample average of Y_i converges in probability to a value that's too high: upward (positive) asymptotic bias. If $\mathrm{E}(M) < 0$, then $\mathrm{E}(Y) < \mathrm{E}(Y^*)$, so downward (negative) asymptotic bias.

Discussion Question 4.7 (exercise)**.** Imagine you ask people how many minutes they exercised last week, to try to learn how much exercise people do each week. What's Y? What's Y^*? What's M? Explain a reason we might see $\mathrm{E}(M) > 0$. Explain a reason we might see $\mathrm{E}(M) < 0$.

This extends to regression. Imagine the true LP model (with Y^*) is $Y^* = \mathbf{X}\boldsymbol{\beta} + R$, so $\boldsymbol{\beta}$ is the LPC vector we want to learn, and by definition $\mathrm{E}(\mathbf{X}R) = \mathbf{0}$. Plugging in $Y^* = Y - M$,

$$\begin{aligned} Y - M &= \mathbf{X}\boldsymbol{\beta} + R, \\ Y &= \mathbf{X}\boldsymbol{\beta} + (R + M). \end{aligned}$$

If (and only if) $R + M$ satisfies the LP error property $\mathrm{E}(\mathbf{X}(R + M)) = \mathbf{0}$, then $\boldsymbol{\beta}$ is the LPC for the LP of Y onto $(1, \mathbf{X})$ so we can just reg Y X in Stata.

Since $\mathrm{E}(\mathbf{X}R) = \mathbf{0}$ by definition of R, the key for identification is whether or not $\mathrm{E}(\mathbf{X}M) = 0$. As usual, if either we don't care about the intercept or $\mathrm{E}(M) = 0$, then we instead need $\mathrm{Cov}(X_j, M) = 0$ for all non-constant X_j. This is the identifying assumption required to identify $\boldsymbol{\beta}$.

Discussion Question 4.8 (exercise and gym membership)**.** Imagine again Y^* is minutes of exercise per week, and Y is the self-reported value. Let $X = 1$ if somebody is a gym member, and $X = 0$ otherwise. First, explain a reason we might see $\text{Cov}(X, M) \neq 0$, and whether this would make it $>$ or $<$. Then, compare the LP slope of Y on $(1, X)$, $\text{Cov}(X, Y)/\text{Var}(X)$, with the LP slope of Y^* on $(1, X)$, which is $\beta_1 = \text{Cov}(X, Y^*)/\text{Var}(X)$. What's the direction of asymptotic bias?

So far we have assumed additive measurement error, but we could imagine multiplicative error, too. That is, let $Y = MY^*$. Then, $\ln(Y) = \ln(M)+\ln(Y^*)$. Thus, for a regression with a logged outcome and multiplicative measurement error, there's just a $\ln(M)$ term floating around, so we check if $\text{Cov}(X_j, \ln(M)) = 0$ or not.

Discussion Question 4.9 (log with additive measurement error)**.** What if the true LP is $\ln(Y^*) = \beta_0 + \beta_1 X + V$, but there's additive measurement error, $Y = Y^* + M$? Generally, do you think β_1 is also the LP slope for Y on $(1, X)$? Why/not? Try to justify your answer mathematically.

Discussion Question 4.10 (Ex 4.7: self-reported scrap rate)**.** Imagine the government wants to help increase the efficiency of chalk manufacturing firms. Specifically, Y^* is a firm's "scrap rate": what proportion of their output has to be "scrapped" (trashed/not sold) due to manufacturing defects? For example, $Y^* = 0.04$ means 4% scrap rate. The government randomly assigns firms to a control group and treatment group, to run an experiment. On January 1, the treated firms receive grant money, which they are supposed to use to improve efficiency. All firms self-report their scrape rates on December 31; this is $Y = Y^* + M$. (Note: Y^* and Y could also be log scrap rates, and everything else would remain identical.) 1) Describe a reason why treated firms might systematically overreport ($M > 0$) or underreport ($M < 0$) their scrap rates. 2) In that case, and assuming untreated firms report accurately ($M = 0$), would we overestimate or underestimate the treatment effect of a grant? Why? 3) If the government uses these incorrect estimates to decide whether or not to continue the program, what incorrect decision might they make? Why?

Section 4.4.2

"Errors-in-variables" refers to measurement error in $\mathbf{X}$ (regressors) instead of Y. Again the star (asterisk) indicates true value: X^* true (but unobserved), observe $X = X^* + M$.

Whereas before there were reasonable cases where M did not affect the plim of OLS est, here there is almost certainly some type of bias.

For example, consider the simple LP of Y (now error-free) onto $(1, X^*)$: $Y = \beta_0+\beta_1 X^*+ R$, $\text{E}(R) = \text{Cov}(X^*, R) = 0$. Assume X is "redundant" so that $\text{Cov}(X, R) = 0$, too. Note $\beta_1 = \text{Cov}(Y, X^*)/\text{Var}(X^*)$. The LP slope from reg Y X is $\text{Cov}(Y, X)/\text{Var}(X)$. Note

$$\begin{aligned} Y &= \beta_0 + \beta_1 X^* + R \\ &= \beta_0 + \beta_1 (X - M) + R \\ &= \beta_0 + \beta_1 X + (R - \beta_1 M). \end{aligned}$$

Since $\text{Cov}(X, R) = 0$ by redundancy and $\text{Cov}(X, -\beta_1 M) = -\beta_1 \text{Cov}(X, M)$, then

$$\text{Cov}(X, R - \beta_1 M) = 0 \iff \text{Cov}(X, M) = 0. \tag{4.3}$$

So, is it realistic to assume $\text{Cov}(X, M) = 0$? This is an identifying assumption that would allow us to consistently estimate β_1 by reg Y X.

We could also see this from

$$\begin{aligned}
\text{Cov}(X, Y) &= \text{Cov}(X, \beta_0 + \beta_1 X^* + R) \\
&= \overbrace{\text{Cov}(X, \beta_0)}^{=0} + \text{Cov}(X, \beta_1 X^*) + \overbrace{\text{Cov}(X, R)}^{=0} \\
&= \beta_1 \text{Cov}(X, X^*) \\
&= \beta_1 \text{Cov}(X, X - M) \\
&= \beta_1 [\overbrace{\text{Cov}(X, X)}^{=\text{Var}(X)} - \text{Cov}(X, M)] \\
&= \beta_1 \text{Var}(X) - \beta_1 \text{Cov}(X, M).
\end{aligned}$$

Thus, the LP slope from regressing Y on the mismeasured X and a constant is

$$\begin{aligned}
\frac{\text{Cov}(X, Y)}{\text{Var}(X)} &= \frac{\beta_1 \text{Var}(X) - \beta_1 \text{Cov}(X, M)}{\text{Var}(X)} \\
&= \beta_1 - \beta_1 \frac{\text{Cov}(X, M)}{\text{Var}(X)} \\
&= \beta_1 \left[1 - \frac{\text{Cov}(X, M)}{\text{Var}(X)} \right].
\end{aligned} \tag{4.4}$$

Discussion Question 4.11 (EIV example)**.** Imagine $X^* \in \{1, 2, 3\}$. 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. Is $\text{Corr}(X, M) = 0$, > 0, or < 0? Why? What type of bias does this cause?

More generally, consider the classical errors-in-variables (CEV) assumption:

$$\text{Cov}(X^*, M) = 0. \tag{4.5}$$

If this is true, then

$$\text{Cov}(X, M) = \text{Cov}(X^* + M, M) = \overbrace{\text{Cov}(X^*, M)}^{=0 \text{ by CEV}} + \text{Cov}(M, M) = \text{Var}(M). \tag{4.6}$$

Unlesss there is never any measurement error ($M = 0$ always), then $\text{Var}(M) > 0$, so $\text{Cov}(X, M) > 0$. Equation (4.5) also implies

$$\text{Var}(X) = \text{Var}(X^* + M) = \text{Var}(X^*) + \text{Var}(M) + 2\overbrace{\text{Cov}(X^*, M)}^{=0 \text{ by CEV}} = \text{Var}(X^*) + \text{Var}(M). \tag{4.7}$$

Since variances are always positive, this implies

$$\text{Var}(X) > \text{Var}(M). \tag{4.8}$$

Thus, going back to (4.4),

$$\begin{aligned}\frac{\text{Cov}(X,Y)}{\text{Var}(X)} &= \beta_1\left[1 - \frac{\text{Cov}(X,M)}{\text{Var}(X)}\right] \\ &= \beta_1\left[1 - \frac{\text{Var}(M)}{\text{Var}(X)}\right].\end{aligned}$$

Since variances are positive, $\text{Var}(M) \geq 0$ and $\text{Var}(X) \geq 0$, so $1 - \text{Var}(M)/\text{Var}(X) < 1$. Further, by (4.8), $\text{Var}(M) < \text{Var}(X)$ so $\text{Var}(M)/\text{Var}(X) < 1$. Thus,

$$0 \leq 1 - \frac{\text{Var}(M)}{\text{Var}(X)} \leq 1. \tag{4.9}$$

Thus, the plim of the OLS estimate lies somewhere between β_1 and zero. This is called **attenuation bias**, meaning $|\hat{\beta}_1| < |\beta_1|$. This makes our estimates systematically "too small" in terms of magnitude, whereas positive bias makes $|\hat{\beta}_1| > |\beta_1|$ if $\beta_1 > 0$, and negative bias makes $|\hat{\beta}_1| > |\beta_1|$ if $\beta_1 < 0$.

We could also write, using (4.7),

$$\begin{aligned}1 - \frac{\text{Var}(M)}{\text{Var}(X)} &= \frac{\text{Var}(X) - \text{Var}(M)}{\text{Var}(X)} \\ &= \frac{\text{Var}(X^*)}{\text{Var}(X^*) + \text{Var}(M)}.\end{aligned} \tag{4.10}$$

Is the bias "big"? It depends on how "much" measurement error there is *relative to* the variation in X. If X does not vary a lot (small variance), then even small amounts of measurement error can cause lots of bias. If X varies a lot, then it can withstand larger measurement errors. (Of course, if the bias only depended on $\text{Var}(M)$ and not the variance ratio, then we could just change the units of X and M to get smaller bias.)

Equation (4.47) shows the attenuation bias formula (plim) when there are multiple regressors, using the form from (4.10). Here it's not just $\text{Var}(X)$, but the residual variance in X^* after partialing out the other regressors.

This attenuation result is very pleasing, both elegant and practical. For example, if we estimate the returns to schooling are 20% (ignoring the endogeneity issue), but we think people mis-report their years of education, we could interpret our estimate as an estimated lower bound. That is, we know attenuation bias means (asymptotically) $|\hat{\beta}| < |\beta|$, so if $\hat{\beta} = 20\%$, the true value is at least as big. Of course, if $\hat{\beta} = 1\%$, this may not be interesting.

However, the attenuation result is based on some strong assumptions. First, CEV was assumed. Second, the result from (4.47) assumes only the regressor of interest is mismeasured. Third, the result is only for a linear-in-variables regression model. If any of these is not true, there may be a different type of bias.

Chapter 5

Section 5.1.1

Consider structural model $Y = \beta_0 + \beta_1 X + U$. We want to estimate β_1, but $\text{Cov}(X, U) \neq 0$, so β_1 is not LPC, can't use OLS to reg Y X.

More precisely, reg Y X consistently estimates the LP slope $\text{Cov}(X, Y)/\text{Var}(X)$. But

$$\begin{aligned}
\text{Cov}(X, Y) &= \text{Cov}(X, \beta_0 + \beta_1 X + U) \\
&= \overbrace{\text{Cov}(X, \beta_0)}^{=0} + \overbrace{\text{Cov}(X, \beta_1 X)}^{\text{use linearity}} + \text{Cov}(X, U) \\
&= \beta_1 \text{Cov}(X, X) + \text{Cov}(X, U) \\
&= \beta_1 \text{Var}(X) + \text{Cov}(X, U),
\end{aligned}$$

so

$$\begin{aligned}
\frac{\text{Cov}(X, Y)}{\text{Var}(X)} &= \frac{\beta_1 \text{Var}(X) + \text{Cov}(X, U)}{\text{Var}(X)} \\
&= \beta_1 + \frac{\text{Cov}(X, U)}{\text{Var}(X)}.
\end{aligned}$$

So if $\text{Cov}(X, U) = 0$ (exogenous X), then OLS consistently estimates β_1, but otherwise there is asymptotic bias.

One way to think of the IV strategy is to separate the "endogenous part" of X from the "exogenous part." To accomplish this, the instrumental variable Z should vary with X, but not with U. Conceptually, we can see how Y varies with Z, then see how X varies with Z, and then infer how much Y varies with X by dividing.

Discussion Question 5.1 (simple IV)**.** Consider the statistical object

$$\frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}. \tag{5.1}$$

It's "statistical" since it depends only on the joint probability distribution of (Y, X, Z). It is consistently estimated by simply replacing population covariances with sample covariances. But, does it have any economic meaning? Show that it in fact can equal β_1 from structural model $Y = \beta_0 + \beta_1 X + U$ if you make certain assumptions, and explain what those are.

The two important properties are **relevance** and **exogeneity**. Relevance generally means the IV has to vary with the endogenous regressor; here, $\text{Cov}(Z, X) \neq 0$. Exogeneity generally means the IV is unrelated to the structural error; here, $\text{Cov}(Z, U) = 0$.

Discussion Question 5.2 (JTPA IV)**.** Consider a simplified version of the Job Training Partnership Act (JTPA), a federal program from decades past. The program finds a pool of eligible workers and randomizes invitations to a special job training. Let $Z = 1$ if invited, $Z = 0$ if not. However, only 60% of invited workers attend. Let $X = 1$ if attend, $X = 0$ if not. Let Y be labor income over the year after the program ends. 1) Why is OLS reg Y X biased? Hint: consider the structural model $Y = \beta_0 + \beta_1 X + U$; what's in U, and is it correlated with X? 2) Is Z relevant? exogenous? Why/not?

Consider another derivation of the IV estimator in this simple model. The LP of X onto $(1, Z)$ is (5.4),

$$X = \delta_0 + \theta Z + R, \tag{5.2}$$

where by definition $\text{Cov}(Z, R) = 0$. Plugging this into the structural model,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + U \\ &= \beta_0 + \beta_1(\delta_0 + \theta Z + R) + U \\ &= (\beta_0 + \beta_1\delta_0) + \beta_1\theta Z + (\beta_1 R + U). \end{aligned} \tag{5.3}$$

The error term satisfies

$$\text{Cov}(Z, \beta_1 R + U) = \beta_1 \overbrace{\text{Cov}(Z, R)}^{=0} + \overbrace{\text{Cov}(Z, U)}^{=0 \text{ b/c exog}} = 0,$$

so the coefficient on Z ($\beta_1\theta$) is the linear projection slope in the LP of Y onto $(1, Z)$. That is, if like (5.6) we define the LP of Y onto $(1, Z)$ as

$$Y = \alpha_0 + \lambda Z + V, \tag{5.4}$$

then $\lambda = \beta_1\theta$. Thus, the structural β_1 is identified by the ratio of two linear projection slopes: $\beta_1 = \lambda/\theta$, where θ was the LP slope of X onto $(1, Z)$. The LP slopes are sometimes called **reduced form parameters**, meaning they are just "statistical" parameters that can (usually) be estimated consistently by standard methods (here, OLS). Sometimes (5.2) is called the **reduced form equation** (p. 90). Sometimes this ratio estimator is called the **Wald estimator**.

Discussion Question 5.3 (pretty-much identified?)**.** What might happen to the estimator $\hat{\beta}_1 = \hat{\lambda}/\hat{\theta}$ in practice if $\theta = 0.1$? What if $\theta = 0.001$? Think about identification, estimation, and confidence intervals (for $\hat{\beta}_1$), and explain why each is/not adversely affected.

If an instrument is technically relevant, but just barely, it is a **weak instrument**. In this case β_1 is still identified, but it is **weakly identified**; the **weak IV** problem is a problem of so-called **weak identification**. This can cause two problems: biased estimation, and over-rejection of hypothesis tests. Unlike exogeneity, which cannot directly be tested since it is about the unobservable U (and Z), we *can* examine the relevance assumption directly because it's only about observables (Z and X). In the simplest form, you just look at the "first stage" LP of the endogenous regressor (in this case X) onto all the exogenous variables (in this case 1 and Z), and compute the F-statistic for testing whether all coefficients on excluded instruments (here, just Z) are zero. If $F < 10$, then you have a weak IV problem

and should Google what to do. If $F > 10$, you might still have a problem but it's not severe enough for people to bother you to do something about it. See also p. 108 in Wooldridge.

Instead of the ratio, we could also regress Y on the "exogenous part" of X. Considering (5.2), Z is exogenous, so $\delta_0 + \theta Z$ is exogenous, which must mean that R contains all the endogeneity. Let $X^* \equiv \delta_0 + \theta Z$. Let the LP of Y onto $(1, \hat{X})$ be

$$Y = \gamma_0 + \gamma_1 X^* + V = \gamma_0 + \gamma_1(\delta_0 + \theta Z) + V = (\gamma_0 + \gamma_1\delta_0) + (\gamma_1\theta)Z + V. \tag{5.5}$$

Since

$$\overbrace{0 = \operatorname{Cov}(V, X^*)}^{\text{b/c } V \text{ is LP error}} = \operatorname{Cov}(V, \delta_0 + \theta Z) = 0 + \theta \operatorname{Cov}(V, Z), \tag{5.6}$$

it must be that either $\theta = 0$ or $\operatorname{Cov}(V, Z) = 0$ (or both). Assuming Z is "relevant," then $\theta \neq 0$, so it must be that $\operatorname{Cov}(V, Z) = 0$. Thus, the slope $\gamma_1\theta$ is the slope of the LP of Y onto $(1, Z)$. By comparison with (5.3), it must be that $\gamma_1\theta = \beta_1\theta$. This implies $\gamma_1 = \beta_1$, which is the structural coefficient of interest. Recall that γ_1 is the slope of the LP of Y onto $(1, X^*)$. So, regressing Y onto $(1, X^*)$ estimates β_1 consistently. That is, once we have constructed the exogenous part of X, we can just run OLS like usual. As before, we need Z to be exogenous so that X^* is exogenous, and we need Z to be relevant ($\theta \neq 0$) otherwise $X^* = \delta_0$ is just a constant.

Even with this simplest model, there's yet another way to think of the IV estimator. Let $\mathbf{Z} = (1, Z)$ be the **full instrument vector**. In this simple model, Z is the only **excluded instrument** (meaning it doesn't appear in the structural model $Y = \beta_0 + \beta_1 X + U$), and the constant 1 is the only **included instrument** (meaning it does appear in the structural model, i.e., it's an exogenous regressor). The exogeneity property is

$$\mathrm{E}(\mathbf{Z}'U) = \mathbf{0}. \tag{5.7}$$

That is, $\mathrm{E}(U) = 0$ [this doesn't really matter; not restrictive to just assume] and

$$\operatorname{Cov}(Z, U) = \mathrm{E}[(Z - \mathrm{E}(Z))(U - \overbrace{\mathrm{E}(U)}^{=0})] = \mathrm{E}[ZU] - \mathrm{E}[\mathrm{E}(Z)U] = \mathrm{E}(ZU) - \mathrm{E}(Z)\overbrace{\mathrm{E}(U)}^{=0} = 0.$$

Given the exogeneity condition for the full vector of instruments, we can plug in for U and solve for the vector $\boldsymbol{\beta} = (\beta_0, \beta_1)'$:

$$\begin{aligned} \mathbf{0} = \mathrm{E}[\mathbf{Z}'(Y - \mathbf{X}\boldsymbol{\beta})] = \mathrm{E}[\mathbf{Z}'Y] - \mathrm{E}[\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}] = \mathrm{E}[\mathbf{Z}'Y] - \mathrm{E}[\mathbf{Z}'\mathbf{X}]\boldsymbol{\beta}, \\ \mathrm{E}[\mathbf{Z}'\mathbf{X}]\boldsymbol{\beta} = \mathrm{E}[\mathbf{Z}'Y], \\ \boldsymbol{\beta} = \{\mathrm{E}[\mathbf{Z}'\mathbf{X}]\}^{-1}\mathrm{E}[\mathbf{Z}'Y]. \end{aligned} \tag{5.8}$$

This has the same look of $\operatorname{Cov}(Z, Y)/\operatorname{Cov}(Z, X)$, but more general. But, similar to before, it expresses the vector of structural parameters in terms of moments (expected values) of the population distribution of observables $(Y, \mathbf{X}, \mathbf{Z})$. To estimate $\boldsymbol{\beta}$, simply replace the population expectation $\mathrm{E}(\cdot)$ by sample expectation (sample average) $\hat{\mathrm{E}}(\cdot)$.

Note that the formula for $\boldsymbol{\beta}$ includes an inverted matrix. The instruments are only valid if this matrix is indeed invertible. The matrix $\mathrm{E}[\mathbf{Z}'\mathbf{X}]$ is invertible if it is of full rank, so such an assumption is often called the **rank condition**. It is the same idea as "relevance."

The expression in (5.7), or more precisely the formula

$$\mathbf{0} = \mathrm{E}[\mathbf{Z}'(Y - \mathbf{X}\boldsymbol{\beta})],$$

is called a **moment condition**. The LHS is zero and the RHS is some moment involving observable variables and the parameters of interest. The estimated $\hat{\boldsymbol{\beta}}$ try to satisfy the sample moment condition, where the population expectation is replaced by sample expectation. More on this in the later GMM (generalized method of moments) chapter.

The following is from Example 5.1, page 93. Consider the structural model (5.12),

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + U, \tag{5.9}$$

where Y is log wage, X_1 is years of experience, X_2 is years of education, and U is the structural error term. We imagine $\mathrm{Corr}(X_2, U) > 0$, so we can't run OLS. For simplicity assume X_1 is exogenous.

Discussion Question 5.4 (Ex 5.1, mother's education)**.** Consider the instrument Z of mother's education (in years). Do you think it's relevant? Exogenous? Why/not?

Discussion Question 5.5 (Ex 5.1, SSN last digit)**.** Consider the instrument Z of the last digit of an individual's social security number (SSN). Do you think it's relevant? Exogenous? Why/not?

Discussion Question 5.6 (Ex 5.1, birth quarter)**.** Consider the instrument Z of an individual's quarter of birth; i.e., $Z = 1$ if born in January through March, $Z = 2$ if born in April through June, $Z = 3$ if July–Sept, $Z = 4$ otherwise. Do you think it's relevant? Exogenous? Why/not? Hint: many U.S. states require you to attend school until a certain age, but the grade level corresponding to your age depends on which month you were born in.

With the birth quarter example, even if it were valid, it may have a different interpretation than we hope. LATE: "local" for the subpopulation who only attend more school if forced by law. Maybe very different effect than you grad students. See more in Ch. 21.

Pages 94–95: **natural experiment** (often from political policy, or geography) can produce "exogenous variation." Examples of IV like Vietnam war lottery, election timing, rivers, college proximity (Ex 5.2).

Section 5.1.2

Readily generalizes to multiple instruments, multiple regressors, but still single endogenous regressor for now. Structural model is

$$Y = \mathbf{X}\boldsymbol{\beta} + U, \tag{5.10}$$

where $\mathbf{X} = (1, X_1, X_2, \ldots, X_K)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$, U is structural error. Regressor X_K is endogenous, others are exogenous, $\mathrm{Cov}(X_j, U) = 0$ for $j < K$. Excluded IVs are $Z_1, \ldots, Z_M$, where $\mathrm{Cov}(Z_j, U) = 0$.

The population "first-stage regression" is the linear projection of the endogenous X_K onto all exogenous variables:

$$X_K = \delta_0 + \delta_1 X_1 + \cdots + \delta_{K-1} X_{K-1} + \theta_1 Z_1 + \cdots + \theta_M Z_M + R, \tag{5.11}$$

where by definition the LP error R is uncorrelated w/ all RHS variables. Again, the rule-of-thumb for when weak IV is a problem is determined by the F-statistic for testing the joint hypothesis

$$H_0 \colon \theta_1 = \theta_2 = \cdots = \theta_M = 0. \tag{5.12}$$

If $F < 10$, then you need to do something to account for weak identification, or else people won't believe your results. If $F > 10$, then people probably won't bother you, although if you have $F = 11$ it might still be good to try methods robust to weak identification, since in reality the problems (bias, non-normality) increase in magnitude as F decreases, but they don't magically disappear suddenly when $F > 10$. Although people talk about "testing" for weak IV, it's really not a "test," but an assessment or measure of the strength of identification. When F is big, identification is strong; when F is small, identification is weak. The fact that F is itself a test statistic (although here we don't care about its p-value) probably confuses things further.

The exogenous part of X_K is then

$$X_K^* = \delta_0 + \delta_1 X_1 + \cdots + \delta_{K-1} X_{K-1} + \mathbf{Z}\boldsymbol{\theta}. \tag{5.13}$$

Since all RHS are exogenous (uncorr'd with structural U), then $\text{Cov}(X_K^*, U) = 0$, too (applying linearity of Cov).

Like in the simple model, there are different (equivalent) ways we could think about the 2SLS (two-stage least squares, also TSLS) estimator.

First, we could imagine a **second-stage regression** where we regress (i.e., LP) Y onto $(1, X_1, \ldots, X_{K-1}, X_K^*)$, and the coefficient on X_K^* is β_K. Actually, it is not obvious that this will work. More immediately clear is that X_K^* can be used as an instrument for X_K; at least, it's clear that it's exogenous (as already argued) and intuitive that the rank condition/relevance should also hold. Let

$$\mathbf{Z} = (1, X_1, \ldots, X_{K-1}, Z_1, \ldots, Z_M) \tag{5.14}$$

be the full instrument vector. Define vector

$$\mathbf{X}^* \equiv (1, X_1, \ldots, X_{K-1}, X_K^*). \tag{5.15}$$

Since every element except X_K^* is also in $\mathbf{Z}$, they are also their own linear projections onto $\mathbf{Z}$. Thus the whole vector $\mathbf{X}^*$ is the linear projection of $\mathbf{X}$ onto $\mathbf{Z}$:

$$\mathbf{X}^* = \mathbf{Z}[\text{E}(\mathbf{Z}'\mathbf{Z})]^{-1} \text{E}(\mathbf{Z}'\mathbf{X}). \tag{5.16}$$

Using (5.8), since $\mathbf{X}^*$ is the same length as $\mathbf{X}$,

$$\boldsymbol{\beta} = \{\text{E}[\mathbf{X}^{*\prime}\mathbf{X}]\}^{-1} \text{E}[\mathbf{X}^{*\prime}Y].$$

Coincidentally, as seen below, $\mathrm{E}[\mathbf{X}^{*\prime}\mathbf{X}] = \mathrm{E}[\mathbf{X}^{*\prime}\mathbf{X}^*]$, so the formula reduces to the formula for the LPC of Y onto $\mathbf{X}^*$. This is fun for me to see but not worth showing students:

$$\boldsymbol{\beta} = \{\mathrm{E}[\mathbf{X}^{*\prime}\mathbf{X}^*]\}^{-1}\,\mathrm{E}[\mathbf{X}^{*\prime}Y]. \tag{5.17}$$

Using

$$\mathbf{X}^{*\prime} = \{\mathbf{Z}[\mathrm{E}(\mathbf{Z}'\mathbf{Z})]^{-1}\,\mathrm{E}(\mathbf{Z}'\mathbf{X})\}' = \mathrm{E}(\mathbf{X}'\mathbf{Z})[\mathrm{E}(\mathbf{Z}'\mathbf{Z})]^{-1}\mathbf{Z}', \tag{5.18}$$

then

$$\begin{aligned}
\mathrm{E}(\mathbf{X}^{*\prime}\mathbf{X}^*) &= \mathrm{E}\{\mathrm{E}(\mathbf{X}'\mathbf{Z})[\mathrm{E}(\mathbf{Z}'\mathbf{Z})]^{-1}\mathbf{Z}'\mathbf{Z}[\mathrm{E}(\mathbf{Z}'\mathbf{Z})]^{-1}\,\mathrm{E}(\mathbf{Z}'\mathbf{X})\} \\
&= \mathrm{E}(\mathbf{X}'\mathbf{Z})[\mathrm{E}(\mathbf{Z}'\mathbf{Z})]^{-1}\,\mathrm{E}[\mathbf{Z}'\mathbf{Z}][\mathrm{E}(\mathbf{Z}'\mathbf{Z})]^{-1}\,\mathrm{E}(\mathbf{Z}'\mathbf{X}) \\
&= \overbrace{\mathrm{E}(\mathbf{X}'\mathbf{Z})[\mathrm{E}(\mathbf{Z}'\mathbf{Z})]^{-1}}^{\text{move inside last expectation}}\ \mathrm{E}(\mathbf{Z}'\mathbf{X}) \\
&= \mathrm{E}\{\overbrace{\mathrm{E}(\mathbf{X}'\mathbf{Z})[\mathrm{E}(\mathbf{Z}'\mathbf{Z})]^{-1}\mathbf{Z}'}^{=\mathbf{X}^{*\prime}}\mathbf{X}\} \\
&= \mathrm{E}(\mathbf{X}^{*\prime}\mathbf{X}).
\end{aligned}$$

So, although the "second stage" happens to be equivalent to linear projection, it is more fundamentally a second stage IV estimation.

If I haven't said yet: this first/second stage stuff is just to help our intuition. When Stata/R computes the estimator, it just uses the big, single formula. And if you actually ran OLS twice, your SE would be wrong.

Note that since $\mathrm{E}[\mathbf{X}^{*\prime}\mathbf{X}]$ needs to be invertible, $\mathrm{E}(\mathbf{Z}'\mathbf{X})$ must have full rank. This rank condition happens to be equivalent to at least one of the θ_j being non-zero (p. 98). Hence the "weak IV" measure (the F-statistic) can also be seen as measuring how far from singular (not invertible) that matrix is.

Terminology: when $M > 1$ (and only 1 endog regressor) **overidentified**. "Over" sounds like a bad thing ("too much," "more than optimal") but actually it's good: we have even more identifying restrictions than we need to estimate $\boldsymbol{\beta}$. When $M = 1$, $\boldsymbol{\beta}$ is **exactly identified**. If $M = 0$, then it's **underidentified** (or, not identified). In the first case, there are $M - 1$ **overidentifying restrictions**. Generally these are helpful: they can help us estimate $\boldsymbol{\beta}$ more precisely, and they can be used to test whether our model seems properly specified, including whether our instruments seem valid.

Discussion Question 5.7. Assume $\mathrm{E}(U \mid Z) = 0$ and Z is a valid instrument. Is Z^2 a valid instrument? Is Z^3? $\sin(Z)$?

Section 5.2.1

Just formally write Assumptions 2SLS.1–2 and say they imply consistency. Basically consistency follows since we can write $\boldsymbol{\beta}$ in terms of moments of the data, and the moments can be consistently estimated (WLLN) and combined by continuous mapping theorem, assuming the matrices are invertible.

Chapter 8

Section 8.3: 2SLS as GMM

Ignore the "system" part; imagine same model and dimensions as Chapter 5. (In Chapter 8 notation: imagine $G = 1$.) Dimensions: Y is scalar, $\mathbf{X}$ is $1 \times K$ row vector, $\boldsymbol{\beta}$ is $K \times 1$ column vector of structural parameters, $\mathbf{Z}$ is $1 \times L$ row vector of all instruments (included and excluded). Structural model is $Y = \mathbf{X}\boldsymbol{\beta} + U$.

Recall that one of the many ways to think about IV was that the structural parameter vector $\boldsymbol{\beta}$ satisfies

$$\mathbf{0} = \mathrm{E}[\mathbf{Z}'(Y - \mathbf{X}\boldsymbol{\beta}], \tag{8.1}$$

as in (8.23). This holds if all variables in $\mathbf{Z}$ are indeed exogenous. With only exogeneity (not relevance), there may be multiple $\boldsymbol{\beta}$ values that satisfy the moment condition, but if the rank condition holds, then there is a unique solution to (8.1).

Page 214: by analogy principle, could try to solve sample analog of (8.1), i.e., try to find $\hat{\boldsymbol{\beta}}$ that solves

$$\mathbf{0} = \hat{\mathrm{E}}[\mathbf{Z}'(Y - \mathbf{X}\hat{\boldsymbol{\beta}}] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i'(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}]. \tag{8.2}$$

Recall that $\mathbf{Z}'$ is $L \times 1$, so the LHS $\mathbf{0}$ is also $L \times 1$, so (8.2) is really a set of L equations. Meanwhile, $\hat{\boldsymbol{\beta}}$ is $K \times 1$, so we have a system of L equations with K variables (K unknowns).

As we saw in Chapter 5, if $L = K$ (exact identification), and if the rank condition holds, then there is a unique solution to (8.2). There is even a closed-form expression for $\hat{\boldsymbol{\beta}}$ in that case. This is called **method of moments** estimation.

If $L < K$ (underidentification), then there are an infinite number of solutions. At first this might sound good: in school, we always want to find the solution, so if there are infinite solutions, it should be even easier for us to find one. However, there's only one true value of $\boldsymbol{\beta}$, and we don't know which "solution" is the right one. For example, imagine $Y = \beta_0 + \beta_1 + U$, $\mathrm{E}(U) = 0$. There is only a single moment condition: $\mathrm{E}(Y - \beta_0 - \beta_1) = 0$. In the sample, any estimate such that $\hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}$ is a "solution." If the sample average $\bar{Y} = 0$, then we have no way to know if $\beta_0 = \beta_1 = 0$, or if $\beta_0 = 4$ and $\beta_1 = -4$, etc.

If $L > K$ (overidentification), then the sample moment conditions define L equations with only K unknowns, which generally has no solution. This may sound bad, and it begs some questions about computation, but it's actually good. It means we have even more identifying power than we need to estimate $\boldsymbol{\beta}$.

Discussion Question 8.1. Imagine $L > K$. What are some possible ways to get a vector of K variables that are all exogenous?

Consider overidentification in the following contrived example. Consider an iid sample of Y_i, $i = 1, \ldots, 2n$ (yes, $2n$, not n). For $i = 1, \ldots, n$, define $X_i = Y_i$ and $Z_i = Y_{i+n}$. Let $\theta = \mathrm{E}(Y_i) = \mathrm{E}(X_i) = \mathrm{E}(Z_i)$. If we only use Y_i, then our moment condition is $\mathrm{E}(Y - \theta) = 0$, and the sample moment $\hat{\mathrm{E}}(Y - \hat{\theta}) = 0$ can be solved uniquely by $\hat{\theta} = \hat{\mathrm{E}}(Y)$, the sample average of the $2n$ values of Y_i. If instead we used the X and Z data, then we have two moment conditions: $\mathrm{E}(X - \theta) = 0$ and $\mathrm{E}(Z - \theta) = 0$. If desired, we could write this as a single equation with vectors:

$$\mathrm{E}\left[\begin{pmatrix} X \\ Z \end{pmatrix} - \begin{pmatrix} \theta \\ \theta \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{8.3}$$

There are two equations ($L = 2$) but only one unknown parameter ($K = 1$), so the system is overidentified. Imagine we try to solve both sample moments simultaneously:

$$0 = \hat{\mathrm{E}}(X - \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} X_i - \hat{\theta},$$
$$0 = \hat{\mathrm{E}}(Z - \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} Z_i - \hat{\theta}.$$

The first equation implies $\hat{\theta} = \bar{X}$, but the second implies $\hat{\theta} = \bar{Z}$. If $\bar{X} = \bar{Z}$, then this is fine. But even if in the population $\mathrm{E}(X) = \mathrm{E}(Z)$, the probability of the sample averages being equal is very small and goes to zero as n increases. We could throw away Z and only use X; we'd get a consistent estimator, but throwing away data means larger SE.

Since we can't make (8.2) exactly zero, maybe we can just make it as close to zero as possible. One natural way to measure "close" is Euclidean distance. The $\hat{\boldsymbol{\beta}}$ that minimizes the Euclidean distance also minimizes the squared Euclidean distance:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}} \|\hat{\mathrm{E}}[\mathbf{Z}'(Y - \mathbf{X}\mathbf{b})]\|^2 = \arg\min_{\mathbf{b}} \hat{\mathrm{E}}[\mathbf{Z}'(Y - \mathbf{X}\mathbf{b})]'\hat{\mathrm{E}}[\mathbf{Z}'(Y - \mathbf{X}\mathbf{b})]. \tag{8.4}$$

More generally, we could use the weighted (squared) Euclidean distance:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}} \hat{\mathrm{E}}[\mathbf{Z}'(Y - \mathbf{X}\mathbf{b})]'\underline{\hat{\mathbf{W}}}\hat{\mathrm{E}}[\mathbf{Z}'(Y - \mathbf{X}\mathbf{b})], \tag{8.5}$$

where $\underline{\hat{\mathbf{W}}}$ is an $L \times L$ weighting matrix, usually chosen to be symmetric and positive definite (Wooldridge says PSD but I think that just confuses students?), and usually dependent on the observed data, hence the "hat" (bottom p. 214). The estimator in (8.5) is called the **generalized method of moments** (GMM) estimator.

Discussion Question 8.2. Let $L = K$. What's the smallest possible value of the function on the RHS of (8.5) that's being minimized? Is there any $\hat{\boldsymbol{\beta}}$ that can achieve such a small value? How does the weighting matrix affect your answers?

Discussion Question 8.3. Reconsider the contrived example with $\theta = \mathrm{E}(X) = \mathrm{E}(Z)$, and $X_i \perp\!\!\!\perp Z_i$ with both X_i and Z_i iid sampled from the sample distribution for $i = 1, \ldots, n$. Let

$\hat{\underline{\mathbf{W}}}$ be the identity matrix. Compute the GMM estimator

$$\begin{aligned}\hat{\theta} &= \arg\min_b \hat{\mathrm{E}}[(X,Z)-(b,b)]\hat{\underline{\mathbf{W}}}\hat{\mathrm{E}}[(X,Z)-(b,b)]' \\ &= \arg\min_b (b-\bar{X}, b-\bar{Z})(b-\bar{X}, b-\bar{Z})' \\ &= \cdots\end{aligned}$$

(Hint: the SOC holds, so just solve the FOC.) Does this $\hat{\theta}$ make sense?

Discussion Question 8.4. Let $X_i \perp\!\!\!\perp Z_i$, both iid normal with means $\mathrm{E}(X) = \mathrm{E}(Z) = \theta$, but $\mathrm{Var}(X) = 1$ and $\mathrm{Var}(Z) = 4$. Let $n = 1$. 1) What's the sampling distribution of $\bar{X}$? 2) What's the sampling distribution of $\bar{Z}$? 3) Would you prefer (1) or (2)? 4) If you took a weighted average of $\bar{X}$ and $\bar{Z}$, like $(1-w)\bar{X} + w\bar{Z}$, which one would you give more weight to? Think intuitively first; if you have time, try some math (bias, variance, MSE).

Discussion Question 8.5. Continuing the same example. Imagine the GMM estimator with

$$\hat{\underline{\mathbf{W}}} = \begin{pmatrix} 1-w & 0 \\ 0 & w \end{pmatrix},$$

so

$$\hat{\theta} = \arg\min_b (1-w)(\bar{X}-b)^2 + w(\bar{Z}-b)^2.$$

Solve the FOC to show that $\hat{\theta} = (1-w)\bar{X} + w\bar{Z}$. What's the optimal weighting matrix that minimizes MSE of $\hat{\theta}$?

With the linear-in-parameters model, the GMM first-order condition can be solved for $\hat{\boldsymbol{\beta}}$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, let $\underline{\mathbf{Z}}$ be the $n \times L$ matrix of stacked $\mathbf{Z}_i$, similarly let $\underline{\mathbf{X}}$ be $n \times K$. [Should just skip these details and get to the punchline.] First,

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{b}} \underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b}) = -\underline{\mathbf{Z}}'\underline{\mathbf{X}}. \tag{8.6}$$

Using the product rule,

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\mathbf{b}}[\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b})]'\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b}) \\ &= \{\frac{\mathrm{d}}{\mathrm{d}\mathbf{b}}[\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b})]'\}\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b}) + [\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b})]'\hat{\underline{\mathbf{W}}}\{\frac{\mathrm{d}}{\mathrm{d}\mathbf{b}}\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b})\} \\ &= [-\underline{\mathbf{Z}}'\underline{\mathbf{X}}]'\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b}) + [\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b})]'\hat{\underline{\mathbf{W}}}\{-\underline{\mathbf{Z}}'\underline{\mathbf{X}}\} \\ &= 2[-\underline{\mathbf{X}}'\underline{\mathbf{Z}}]\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\mathbf{b}).\end{aligned}$$

Setting this equal to zero and solving,

$$\begin{aligned}0 &= -\underline{\mathbf{X}}'\underline{\mathbf{Z}}\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'(\mathbf{Y} - \underline{\mathbf{X}}\hat{\boldsymbol{\beta}}) \\ &= -\underline{\mathbf{X}}'\underline{\mathbf{Z}}\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'\mathbf{Y} + \underline{\mathbf{X}}'\underline{\mathbf{Z}}\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'\underline{\mathbf{X}}\hat{\boldsymbol{\beta}}, \\ \underline{\mathbf{X}}'\underline{\mathbf{Z}}\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'\underline{\mathbf{X}}\hat{\boldsymbol{\beta}} &= \underline{\mathbf{X}}'\underline{\mathbf{Z}}\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'\mathbf{Y}.\end{aligned}$$

[Stop skipping and show the next result.] So,

$$\hat{\boldsymbol{\beta}} = \{\underline{\mathbf{X}}'\underline{\mathbf{Z}}\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'\underline{\mathbf{X}}\}^{-1}\underline{\mathbf{X}}'\underline{\mathbf{Z}}\hat{\underline{\mathbf{W}}}\underline{\mathbf{Z}}'\mathbf{Y}. \tag{8.7}$$

Pages 215–216: to show consistency and asymptotic normality, the same tools of WLLN, CLT, and CMT can be applied. For example,

$$\underline{\mathbf{X}}'\underline{\mathbf{Z}}/n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{Z}_i \xrightarrow{p} \mathrm{E}[\mathbf{X}'\mathbf{Z}],$$

etc.; just need to add $1/n$ in the right spots. Here, also need $\hat{\underline{\mathbf{W}}} \xrightarrow{p} \underline{\mathbf{W}}$ for some PD matrix $\underline{\mathbf{W}}$ (in practice always symmetric, too).

Discussion Question 8.6. Show 2SLS is a special case of GMM with $\hat{\underline{\mathbf{W}}} = (\underline{\mathbf{Z}}'\underline{\mathbf{Z}}/n)^{-1}$.

The asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is in (8.29),

$$(\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\mathbf{C}})^{-1}\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\boldsymbol{\Lambda}}\underline{\mathbf{W}}\underline{\mathbf{C}}(\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\mathbf{C}})^{-1}, \quad \underline{\boldsymbol{\Lambda}} \equiv \mathrm{E}(U^2\mathbf{Z}'\mathbf{Z}) = \mathrm{Var}(\mathbf{Z}'U). \tag{8.8}$$

The $\underline{\boldsymbol{\Lambda}}$ comes from the CLT. This type of covariance formula is called a **sandwich form**, where $(\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\mathbf{C}})^{-1}$ is the bread.

Page 218: picking a good weighting matrix gets the sandwich to **collapse**, which happens to correspond to lower variance of $\hat{\boldsymbol{\beta}}$. In particular, if $\underline{\mathbf{W}} = \underline{\boldsymbol{\Lambda}}^{-1}$, then $\underline{\mathbf{W}}\underline{\boldsymbol{\Lambda}} = \underline{\mathbf{I}}$ (identity matrix), so

$$\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\boldsymbol{\Lambda}}\underline{\mathbf{W}}\underline{\mathbf{C}} = \underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\mathbf{C}},$$

which is the inverse of the bread matrix. So the asymptotic variance collapses to

$$(\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\mathbf{C}})^{-1} = (\underline{\mathbf{C}}'\underline{\boldsymbol{\Lambda}}^{-1}\underline{\mathbf{C}})^{-1}. \tag{8.9}$$

This is "smaller" than before in the sense that

$$(\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\mathbf{C}})^{-1}\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\boldsymbol{\Lambda}}\underline{\mathbf{W}}\underline{\mathbf{C}}(\underline{\mathbf{C}}'\underline{\mathbf{W}}\underline{\mathbf{C}})^{-1} - (\underline{\mathbf{C}}'\underline{\boldsymbol{\Lambda}}^{-1}\underline{\mathbf{C}})^{-1} \tag{8.10}$$

is positive semi-definite (top p. 218).

In practice, we can estimate $\underline{\boldsymbol{\Lambda}}$ by Procedure 8.1 (p. 218). Such an estimator has $\hat{\underline{\mathbf{W}}} \xrightarrow{p} \underline{\boldsymbol{\Lambda}}^{-1}$.

Discussion Question 8.7. When/why might the estimated $\hat{\underline{\boldsymbol{\Lambda}}}$ not actually minimize the finite-sample variance of $\hat{\boldsymbol{\beta}}$?

Could go through pages 226–228 or wait till Chapter 14.

Chapter 14

Section 14.1

Can generalize GMM to more general structural models. Often economic theory implies some "moment condition" based on (expected) utility maximization or rational expectations or something. (Although, often have to assume some parametric utility function to get the moment condition.) GMM also includes as special cases things like OLS, 2SLS.

Go over notation page 525.

Discussion Question 14.1. Write the 2SLS model and moments in the Chapter 14 notation.

Discussion Question 14.2. Observe $Y_i \overset{iid}{\sim} F$, want to est $\theta_o = \mathrm{E}(Y)$. What are: $\mathbf{W}_i$, $\mathbf{g}(\mathbf{W}_i, \boldsymbol{\theta})$, L, P, and the moment conditions? What's $\hat{\theta}$?

Page 526: theory much more complicated. Need not just "pointwise" limit, but entire $Q_N(\cdot)$ function to "uniformly" converge to the population GMM criterion function. Consequently, Theorem 14.1 has more technical assumptions. We won't worry about these for now. Assuming they hold, the familiar sort of consistency and asymptotic normality (with consistently estimable variance) results hold. Mostly we'll focus on efficiency and overidentification testing.

Just like before, the theoretically optimal weight matrix is any consistent estimator of a certain asymptotic covariance matrix. In the new notation, and assuming iid sampling, this asy cov matrix is in (14.12):

$$\underset{\sim}{\boldsymbol{\Lambda}}_o \equiv \mathrm{E}[\mathbf{g}_i(\boldsymbol{\theta}_o)\mathbf{g}_i(\boldsymbol{\theta}_o)'] = \mathrm{Var}[\mathbf{g}_i(\boldsymbol{\theta}_o)]. \tag{14.1}$$

One estimator is in (14.13):

$$\hat{\underset{\sim}{\boldsymbol{\Lambda}}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\hat{\boldsymbol{\theta}})\mathbf{g}_i(\hat{\boldsymbol{\theta}})'. \tag{14.2}$$

After such estimation, GMM is run again with the newly estimated weighting matrix. This is called **two-step GMM** since there are two steps. This process can also be iterated more than two times, to get iterated GMM. You can also try to solve for a fixed-point with continuous updating GMM. It seems like iterating can indeed improve results, especially with iid data; with time series it seems less clear.

Discussion Question 14.3. Consider a previous example in new notation: iid Y_i and Z_i, $i = 1, \ldots, n$, $Y_i \perp\!\!\!\perp Z_i$, $\mathrm{E}(Y) = \mathrm{E}(Z) = \theta_o$. What are the moment conditions and $\mathbf{g}_i(\cdot)$ function? What is $\underset{\sim}{\boldsymbol{\Lambda}}$? What's the optimal weighting matrix, and does it make sense intuitively?

Discussion Question 14.4. Look at (14.19). Can it be negative, positive, zero? What values do you expect to see, typically? What if $L = P$ (in Ch8 notation, $L = K$)? What if L is much greater than P?

Can use (14.19) for over-ID test. See also page 228 and (8.58) for simpler 2SLS-type case. Basically tests if there exists $\boldsymbol{\theta}_o$ that can satisfy all the specified moment conditions, H_0: $\mathrm{E}[\mathbf{g}(\mathbf{W}, \boldsymbol{\theta}_o)] = \mathbf{0}$. Can imagine it uses the first P moments to estimate $\boldsymbol{\theta}_o$, then checks whether the remaining $L - P$ sample moments are close to zero. Note that in the 2SLS case, it's possible that the instruments are ok but the structural model is misspecified: e.g., if you leave out an X^2 term in the structural model, then it ends up in the structural error U, but then Z is correlated with U, so $\mathrm{E}(ZU) \neq 0$ and the overidentification test should reject. It's also possible that your instruments are all invalid but happen to identify the same population "pseudo-true" parameter. It's also possible that tests make errors in practice, both type I and type II.

Discussion Question 14.5. You have two possible instruments for X_K. You doubt the exogeneity of the 2nd IV, but your overidentification test doesn't reject at a 5% level. How much/does this change your belief about the 2nd IV's exog? Are there reasons the test might not reject H_0 even if really it's false?

Briefly discuss (14.20).

Chapter 10

Section 10.1

Goal: solve a specific type of OVB.

Motivation: Example 10.1 on page 289: if self-selection based on unobserved individual characteristics, then OLS biased.

Back to page 281: basically $C = \gamma Q$ from Chapter 4. (10.2) is "structural CEF." Note that linearity is strong assumption. So is C entering additively: no interaction b/w C and regressors. Notation: note $\mathbf{X}$ excludes intercept here.

Discussion Question 10.1. Is (10.16) a reasonable model for Example 10.1 on page 289? Hint: what's in C, and does it interact with observed regressors?

Page 282: "Also, suppose that the omitted variable C is *time constant*." This is a big assumption! More realistically perhaps, we hope C changes very slowly over time, compared to the sampling frequency.

Discussion Question 10.2. In Ex 10.1, do you think C is time constant? Why/not?

Bottom 282, (10.4), model in error form:

$$Y_t = \beta_0 + \mathbf{X}_t\boldsymbol{\beta} + C + U_t, \quad \mathrm{E}(U_t \mid \mathbf{X}_t, C) = 0,\ t = 1, 2. \tag{10.1}$$

If additionally assume $\mathrm{E}(\mathbf{X}_t'C) = \mathbf{0}$, then $\mathrm{E}[\mathbf{X}_t'(C + U_t)] = 0$, so LP model, so just run OLS. But if C correlated with regressors, then can't.

(10.7): FD notation $\Delta Y = Y_2 - Y_1$, etc.,

$$\Delta Y = \Delta\mathbf{X}\boldsymbol{\beta} + \Delta U. \tag{10.2}$$

Hope (10.2) is just LP model; then OLS consistent for $\boldsymbol{\beta}$. Check: is $\mathrm{E}[\Delta\mathbf{X}'\Delta U] = \mathbf{0}$? Well, (10.10),

$$\begin{aligned}
\mathrm{E}[\Delta\mathbf{X}'\Delta U] &= \mathrm{E}[(\mathbf{X}_2 - \mathbf{X}_1)'(U_2 - U_1)] \\
&= \mathrm{E}[\mathbf{X}_2'U_2 - \mathbf{X}_1'U_2 - \mathbf{X}_2'U_1 + \mathbf{X}_1'U_1] \\
&= \overbrace{\mathrm{E}[\mathbf{X}_2'U_2] + \mathrm{E}[\mathbf{X}_1'U_1]}^{=0+0 \text{ by } (10.1)} - \mathrm{E}[\mathbf{X}_1'U_2] - \mathrm{E}[\mathbf{X}_2'U_1] \\
&= -\mathrm{E}[\mathbf{X}_1'U_2] - \mathrm{E}[\mathbf{X}_2'U_1].
\end{aligned}$$

Not necessarily zero! E.g., if U_1 influences $\mathbf{X}_2$, then second term may be non-zero. Need strict exog for these to be zero: all U_t uncorrelated with all $\mathbf{X}_t$.

Discussion Question 10.3. 1) Explain why low (or high) U_1 might make $prog_2 = 1$ more likely. 2) Explain what could convince you of strict exogeneity.

Page 284: if time-constant X_j, then $\Delta X_j = 0$ for everyone, no variation. Problem: can't distinguish effect of C from effect of time-constant X. Can only distinguish effects of time-varying X, using **within variation** (within i, over t).

Page 284: assume **balanced panel** for simplicity.

Pages 284–285: different possible asymptotic approximations. Recall that asymptotic results use limits like $n \to \infty$ as approximating tools, not statements about gathering more data; in practice, we just have a single dataset. Since now N and T, three possibilities: large N and small T meaning $N \to \infty$ but fixed T; small N and large T meaning fixed N and $T \to \infty$; or large N and large T, meaning $N, T \to \infty$. (Actually, also difference b/w sequential and joint $N, T \to \infty$, but don't need to mention.)

Section 10.2.1

Sampling: if only observe Y, randomly sample individuals independently from population, then Y_i iid. If observe (Y, X), then again iid if sample individuals independently from pop; but, allows $\text{Corr}(Y, X) \neq 0$. Now, again sample individuals independently from pop, but for each individual, we observe $(Y_1, Y_2, \ldots, Y_T, \mathbf{X}_1, \ldots, \mathbf{X}_T)$. Similar to allowing $\text{Corr}(Y, X) \neq 0$, now we allow $\text{Corr}(Y_1, Y_2) \neq 0$, etc.; can be corr'd across time (**autocorrelated**, **serially correlated**).

(10.11) model and terminology in paragraph after, plus **firm effect** etc.

Page 286: beware that statistics literature uses "random effects" and "fixed effects" differently. Econometrics: RE assumes $\text{Cov}(\mathbf{X}_{it}, C_i) = \mathbf{0}$ for all $t = 1, \ldots, T$, but usually we worry this is non-zero (or else we'd just run OLS). FE does not make this assumption.

Discussion Question 10.4. Example 10.1, do you think RE or FE seems more reasonable? Why?

In short: FE is more robust to OVB, but potentially less efficient (since only use "within variation") and can't est coeffs for time-invariant regressors (sex, race, etc.).

Section 10.2.2

See pages 164–165 for definitions of contemporaneous exog; sequential exog; strict exog. Contemp: $\text{E}(U_t \mid X_t) = 0$. Seq: $\text{E}(U_t \mid X_t, X_{t-1}, \ldots, X_1) = 0$. Strict: $\text{E}(U_t \mid X_1, \ldots, X_T) = 0$.

Discussion Question 10.5. Explain why strict exog is "stronger than" sequential exog (i.e., strict implies seq, but not vice-versa). Also explain why sequential is stronger than contemp.

Discussion Question 10.6. Imagine an AR(1) model with $Y_t = \rho Y_{t-1} + U_t$, $U_t \overset{iid}{\sim} \text{N}(0, 1)$, where U_t is independent of all past values $Y_{t-1}, Y_{t-2}, \ldots$, and $|\rho| < 1$. Define $X_t \equiv Y_{t-1}$. Does contemporaneous exog hold, $\text{E}(U_t \mid X_t) = 0$? Seq? Strict? Why/not?

Section 10.2.3

Already did Ex 10.1.

Ex 10.2: show (10.17), explain. "If shocks to patents today (changes to U_{it}) influence R& D spending at future dates, then strict exog can fail.

Ex 10.3: show (10.18), (10.19), let $X_{it} = Y_{it-1}$.

Discussion Question 10.7. Does Ex 10.3 have strict exog? Why/not?

There are "dynamic" panel FE models that allow lags Y_{it-s} on RHS (using GMM), but beyond our scope; we're just considering "static" models.

Section 10.3

As said before, if $\mathrm{E}[\mathbf{X}_{it}'(C_i + U_{it})] = \mathbf{0}$, then OLS consistent since structural coeffs are also LPCs. But, usual HC SE not correct. The errors $C_i + U_{it}$ may be independent across i, but are not independent across t. Need **cluster-robust** SE to allow for intra-i statistical dependence. Stata and R notes on cluster-robust SE below (end of FE section).

Section 10.4

I just skip RE.

Section 10.5.1

(10.41): model. (I sometimes call this Assumption FE.0, but I guess it's not really an "assumption" since you can always write a model like this—if it's misspecified, that just means FE.1 will be violated.)

Assumption FE.1: strict exog

Page 302, (10.43)–(10.44): point out how $\mathbf{Z}_i$ is not time-varying so its coefficient can't be identified w/ FE, but $d2_t$ is time-varying and so is $d2_t\mathbf{Z}_t$.

Discussion Question 10.8. Ignore C_i, but say $Y_{it} = \theta_1 \mathbb{1}\{i = 1\} + \cdots + \Theta_N \mathbb{1}\{i = N\} + U_{it}$. Can we consistently estimate $\hat{\theta}_1 \xrightarrow{p} \theta_1$? Why/not? State any additional assumptions you make.

Fixed effects transformation, within transformation: (10.45), (10.46). Then if (10.46) is LP (and rank condition, FE.2 on page 304), just run OLS on transformed model. Show how strict exog implies LP error property, and how seq exog is not sufficient.

Note FE.2 rules out time-invariant regressors.

Stata: assume i identifier variable named `id`, and `year` is t. Tell Stata this: `xtset id year`. Then use `xtreg` command with `fe` option (for FE; or don't, for OLS) and `vce(cluster id)` for SE. Time effects: add `i.year` as regressor. Note: can also use `areg` but R^2 is computed differently (I forget details); but I think estimates are same.

R: `plm` package. The `plm()` function is similar to `lm()` but with additional arguments: `index` tells it the names of the "individual" and "time" identifier variables (the i and t), and `model` says which type of estimator to use (pooled OLS, FE, etc.). For example, if "individuals" identified by variable `id` and time period `year`, then `index=c('id','year')`. Pooled OLS: `model='pooling'`. FE: `model='within'`. To get cluster-robust SE that match Stata's, you could do something like the following, assuming your data.frame is named my.df:

```
plm.ret <- plm(...)
G <- length(unique(my.df$id))
N <- length(plm.ret$residuals)
dfa <- G/(G-1) * (N-1) / (N - length(plm.ret$coefficients))
vc <- dfa * vcovHC(x=plm.ret, type='HC0',
                   cluster='group', method='arellano')
summary(object=plm.ret, vcov=vc)
```

This would match Stata's standard errors for the pooled OLS; for fixed effects, in the expression for `dfa`, replace the second `N` with `N-1`. Easiest way to add time effects: argument `effect='twoways'`.

Section 10.5.3

Mathematically equivalent to running regression with dummy variable for $N-1$ individuals, but computationally better to do transformed OLS.

Section 10.6.1

Instead of FE transformation, can do FD transformation, $\Delta Y_t = Y_t - Y_{t-1}$, etc. Then run OLS. Can show LP error property using strict exog, like in 10.1.

Assumption FD.1 is identical to FE.1, and FD.2 is similar in spirit to FE.2.

(10.64)

Page 317: problem if some variable always increases by 1 each period for *all* individuals, like work experience possibly. Then perfect multicollinearity; e.g., if $T = 2$, then $\Delta d2 = \Delta\text{experience} = 1$.

Discussion Question 10.9. If we just drop the year dummies, then can we estimate the coefficient on experience consistently? Why/not?

Page 318 near top: relative efficiency of FE and FD depends on whether FE.3 or FD.3 is closer to being true. Basically depends on amount of serial correlation in idiosyncratic error terms: if lots, then FD better, if little or none, then FE. But, also possible bias difference; see Section 10.7, like (10.76),(10.77): with contemporaneous but not strict exog, and some time series weak dependence assumptions, FE bias is $O(1/T)$ while FD bias is $O(1)$; former maybe small if T big enough, but not latter.

Section 10.6.4

More on difference-in-differences in a bit. For now, just a question.

Discussion Question 10.10. Imagine program targets individuals with low Y_{i1} (first-period outcome). Let $T = 2$, model for FD estimation is just $Y_{it} = \beta_0 + \beta_1 d2_t + \beta_2 X_{it} + C_i + U_{it}$, where $X_{it} = 1$ if individual i participated in the program *and* $t = 2$, and o/w $X_{it} = 0$; and $d2_t = \mathbb{1}\{t = 2\}$. Imagine in reality $\beta_1 = \beta_2 = C_i = 0$, so the true DGP is just $Y_{it} = U_{it}$, where $U_{it} \overset{iid}{\sim} F_U$. 1) Compute whether $\mathrm{E}(\Delta X \Delta U)$ is $= 0$, > 0, or < 0; or draw a picture to figure it out. Hint: use binary like $Y = \mathbb{1}\{\text{employed}\}$ if it makes it easier. 2) Will our FD estimator of β_2 be correct, biased up, or biased down? 3) How will (2) affect our policy decision about whether to continue the program or not: will we be overly likely to continue the program, not likely enough, or just right?

Section 10.7.1

skip

Section 10.7.2

Ehh maybe skip.

Also page 334: "seemingly large differences between the RE and FE estimates but, due to large standard errors, the Hausman statistic fails to reject. What should be done in this case? A typical response is to conclude that the random effects assumptions hold and to focus on the RE estimates. Unfortunately, we may be committing a Type II error: failing to reject Assumption RE.1b when it is false."

Diff-in-diff

[Following mostly copied from Intro text.]

Imagine two groups of individuals, two time periods. The $t = 1$ is pre-treatment, $t = 2$ is post-treat. One group is never treated. The other group is "treated" b/w observation periods. Ex: one city that increased minimum wage, another that didn't, year before and year after increase.

Discussion Question 10.11. In min wage example, imagine we look at $t = 2$ after the min wage has taken effect, and compare employment (Y) in the treated city vs. untreated city. Explain a scenario where we clearly shouldn't interpret this as a causal effect of min wage.

Discussion Question 10.12. In min wage example, imagine we look at how employment rate changes from $t = 1$ to $t = 2$ in treated city. Explain a scenario where we clearly shouldn't interpret this as a causal effect of min wage.

Consider CEF model w/ $T = 2$: let $X_i \equiv \mathbb{1}\{\text{treated}_i\}$,

$$\mathrm{E}(Y_{it} \mid X_i, t) = \beta_0 + \beta_1 X_i + \beta_2 \mathbb{1}\{t = 2\} + \beta_3 X_i \mathbb{1}\{t = 2\}. \tag{10.3}$$

Let $m(a,b) \equiv \mathrm{E}(Y_{it} \mid X_i = a, \mathbb{1}\{t=2\} = b)$, so

$$m(0,0) = \beta_0 + (\beta_1)(0) + (\beta_2)(0) + (\beta_3)(0)(0) = \beta_0, \tag{10.4}$$
$$m(0,1) = \beta_0 + (\beta_1)(0) + (\beta_2)(1) + (\beta_3)(0)(1) = \beta_0 + \beta_2, \tag{10.5}$$
$$m(1,0) = \beta_0 + (\beta_1)(1) + (\beta_2)(0) + (\beta_3)(1)(0) = \beta_0 + \beta_1, \tag{10.6}$$
$$m(1,1) = \beta_0 + (\beta_1)(1) + (\beta_2)(1) + (\beta_3)(1)(1) = \beta_0 + \beta_1 + \beta_2 + \beta_3. \tag{10.7}$$

From (10.4)–(10.6) and their differences,

$$\overbrace{\beta_0 = m(0,0)}^{(10.4)}, \tag{10.8}$$
$$\beta_1 = \overbrace{(\beta_0 + \beta_1) - \beta_0 = m(1,0) - m(0,0)}^{\text{(10.6) minus (10.4)}}, \tag{10.9}$$
$$\beta_2 = \overbrace{(\beta_0 + \beta_2) - \beta_0 = m(0,1) - m(0,0)}^{\text{(10.5) minus (10.4)}}, \tag{10.10}$$
$$\begin{aligned}\beta_3 = [\beta_2 + \beta_3] - [\beta_2] &= \overbrace{[(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1)]}^{\text{(10.7) minus (10.6)}} - \overbrace{[(\beta_0 + \beta_2) - (\beta_0)]}^{\text{(10.5) minus (10.4)}} \\ &= \overbrace{\overbrace{[m(1,1) - m(1,0)]}^{\text{difference}} - \overbrace{[m(0,1) - m(0,0)]}^{\text{difference}}}^{\text{difference-in-differences}}.\end{aligned} \tag{10.11}$$

Before interpreting each coefficient economically, note that it is mathematically (though not "economically") equivalent to write

$$\begin{aligned}\beta_3 &= \overbrace{[(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2)]}^{\text{(10.7) minus (10.5)}} - \overbrace{[(\beta_0 + \beta_1) - (\beta_0)]}^{\text{(10.6) minus (10.4)}} \\ &= [m(1,1) - m(0,1)] - [m(1,0) - m(0,0)].\end{aligned} \tag{10.12}$$

This structure is behind the name **difference-in-differences**, or diff-in-diff, or DiD (or DD).

Using the results in (10.8)–(10.12), the four β_j have the following interpretations.

- $\beta_0 = m(0,0)$ is the mean wage among low-education, low-experience individuals; more generally, the mean Y in the subpopulation with $X_1 = 0$ and $X_2 = 0$. Note: generally $\beta_0 \neq \mathrm{E}(Y)$, the unconditional mean.

- $\beta_1 = m(1,0) - m(0,0)$ is the mean wage difference between high-education and low-education individuals within the low-experience subpopulation; more generally, the mean Y difference between $X_1 = 1$ and $X_1 = 0$ individuals within the $X_2 = 0$ subpopulation. Note: generally $\beta_1 \neq \mathrm{E}(Y \mid X_1 = 1) - \mathrm{E}(Y \mid X_1 = 0)$; it additionally conditions on $X_2 = 0$.

- $\beta_2 = m(0,1) - m(0,0)$ is the mean wage difference between high-experience and low-experience individuals within the low-education subpopulation; more generally, the mean Y difference between $X_2 = 1$ and $X_2 = 0$ individuals within the $X_1 = 0$ subpopulation. Note: generally $\beta_2 \neq \mathrm{E}(Y \mid X_2 = 1) - \mathrm{E}(Y \mid X_2 = 0)$; it additionally conditions on $X_1 = 0$.

- $\beta_3 = [m(1,1) - m(1,0)] - [m(0,1) - m(0,0)]$ is the difference between the (predicted or causal) "effect" on mean wage of experience (high vs. low) in the high-education subpopulation and the "effect" in the low-education subpopulation. More generally, β_3 is the difference between the "effect" of X_2 in the $X_1 = 1$ subpopulation and its "effect" in the $X_1 = 0$ subpopulation.

- Equivalently, $\beta_3 = [m(1,1) - m(0,1)] - [m(1,0) - m(0,0)]$ is is the difference between the (predicted or causal) "effect" on mean wage of education (college vs. not) in the high-experience subpopulation and the "effect" in the low-experience subpopulation. More generally, β_3 is the difference between the "effect" of X_1 in the $X_2 = 1$ subpopulation and its "effect" in the $X_2 = 0$ subpopulation.

The β_j interpretations can also be seen by considering the regression of Y on X_1 separately when $X_2 = 0$ and $X_2 = 1$. That is, plugging in (conditioning on) $X_2 = 0$, the CEF becomes

$$m(x_1, 0) = \beta_0 + \beta_1 x_1 + (\beta_2)(0) + (\beta_3)(x_1)(0) = \beta_0 + \beta_1 x_1. \tag{10.13}$$

Plugging in $X_2 = 1$, the CEF is instead

$$m(x_1, 1) = \beta_0 + \beta_1 x_1 + (\beta_2)(1) + (\beta_3)(x_1)(1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1. \tag{10.14}$$

That is, when changing from $X_2 = 0$ to $X_2 = 1$, the intercept changes by β_2 and the slope changes by β_3. These changes could be positive or negative, or zero. Thinking of the slope on X_1 as the (predicted or causal) "effect" of X_1 on Y, the interaction coefficient β_3 describes how this effect of X_1 differs when $X_2 = 1$ versus $X_2 = 0$. Equivalently, we could switch all the X_1 and X_2 and interpret β_3 as the difference between the "effect" of X_2 when $X_1 = 1$ versus when $X_1 = 0$:

$$m(0, x_2) = \beta_0 + (\beta_1)(0) + \beta_2 x_2 + (\beta_3)(0)(x_2) = \beta_0 + \beta_2 x_2, \tag{10.15}$$
$$m(1, x_2) = \beta_0 + (\beta_1)(1) + \beta_2 x_2 + (\beta_3)(1)(x_2) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x_2. \tag{10.16}$$

The difference-in-differences idea is to try to combine the before vs. after comparison with the our city vs. another city comparison. That is, maybe we can control for macroeconomic changes over time by seeing how incomes in another city change. Or, maybe we can control for persistent differences between the treated and untreated cities by looking at how different they were before the minimum wage increase took effect.

Conceptually, the goal is to construct a **counterfactual**: what would our city's average income have been if there were not a minimum wage increase? We observe the actual average income, but we want to compare it to this counterfactual parallel universe where the minimum wage did not pass, like the parallel universes in the potential outcomes framework. Without additional assumptions, we cannot do this.

The key for causal identification is called the **parallel trends** assumption. The cool thing is that this assumption can replace independence, which we may not often believe outside randomized experiments. That is, parallel trends can let us learn causal effects from observational (non-randomized) data, which is important for the many economic cases where randomization is not possible (e.g., we cannot tell cities to randomize their minimum wage laws). The SUTVA assumption (Chapter 21, page 905) can be relaxed a little bit

(e.g., allowing for wage effect spillovers within our city, since we're looking at the effect of "treating" the entire city), but it is still necessary to preclude spillovers between the treated and untreated units. For example, our city's minimum wage law would probably affect wages in all bordering cities, so those would not be a good choice for the comparison city.

Parallel trends can be understood in the following way. Conceptually, parallel trends says that without the minimum wage law, our city's average income would have increased by exactly the same amount as the other city's average income increased. Mathematically, the other city's average income increase is

$$\mathrm{E}(Y \mid X_1 = 0, X_2 = 1) - \mathrm{E}(Y \mid X_1 = 0, X_2 = 0). \tag{10.17}$$

Parallel trends assumes that adding this increase to the baseline average income in our city, $\mathrm{E}(Y \mid X_1 = 1, X_2 = 0)$, gives us the counterfactual income for our city in the later time period. That is, it assumes we can learn a causal effect by comparing

$$\begin{aligned} &\overbrace{\mathrm{E}(Y \mid X_1 = 1, X_2 = 1)}^{\text{actual (our city, after)}} \quad \text{vs.} \\ &\underbrace{\mathrm{E}(Y \mid X_1 = 1, X_2 = 0)}_{\text{our city, before}} + \overbrace{\underbrace{\mathrm{E}(Y \mid X_1 = 0, X_2 = 1) - \mathrm{E}(Y \mid X_1 = 0, X_2 = 0)}_{\text{increase in other city over time}}}^{\text{counterfactual}}. \end{aligned} \tag{10.18}$$

We can draw (10.18) after rearranging. With $m(x_1, x_2) \equiv \mathrm{E}(Y \mid X_1 = x_1, X_2 = x_2)$, the causal effect (we'll be more specific about what type of causal effect later) is

$$\begin{aligned} &\overbrace{m(1,1)}^{\text{actual}} - \overbrace{\{m(1,0) + [m(0,1) - m(0,0)]\}}^{\text{counterfactual}} \\ &= [m(1,1) - m(1,0)] - [m(0,1) - m(0,0)] = \beta_3, \end{aligned}$$

the interaction term coefficient. Figure 2 visualizes this treatment effect. We can think of constructing the counterfactual outcome, and then subtracting it from the actual outcome $m(1,1)$, or we can think of taking the before/after difference for our city, $m(1,1) - m(1,0)$, and subtracting off the before/after difference in the other city, $m(0,1) - m(0,0)$.

0.5.1 Identification

Population Objects of Interest

Most fundamentally, the difference-in-differences approach only learns the average treatment effect for the group that was actually treated (in our universe). This is called the **average treatment effect on the treated** (ATT) (or sometimes ATTE or ATET). Recall that mathematically, ATE meant $\mathrm{E}(Y^1 - Y^0)$, where Y^1 is the treated potential outcome (somebody's outcome in the parallel universe where they're treated) and Y^0 is the untreated potential outcome (somebody's outcome in the parallel universe where they're not treated). ATT is the same comparison, but for the subpopulation who was actually treated in our universe. In our diff-in-diff example, $X_1 = 1$ if somebody is in the treated group (i.e., in our city where there was a minimum wage increase). Thus, the ATT is

$$\mathrm{ATT} \equiv \mathrm{E}(Y^1 - Y^0 \mid X_1 = 1). \tag{10.19}$$

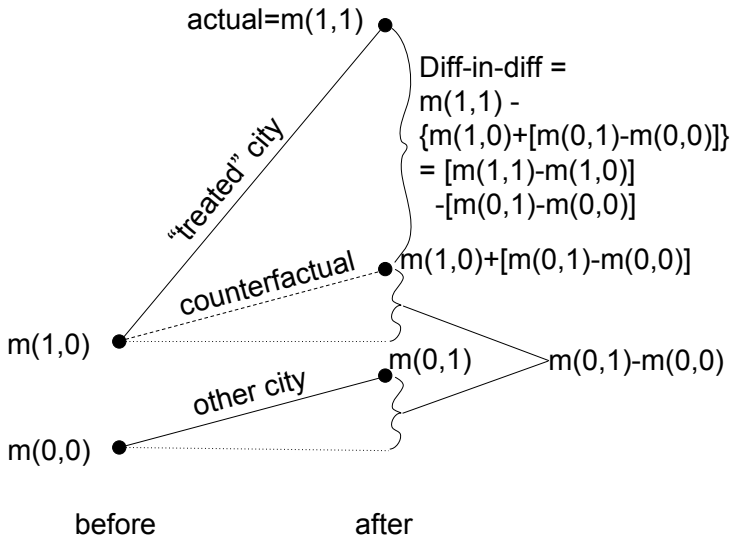

Figure 2: Difference-in-differences.

If we assume that the average effect is the same in the $X_1 = 1$ and $X_1 = 0$ groups, then the ATE and ATT are the same, but sometimes we might think the effect is different. For example, maybe there are different demographics in our city than the comparison city, or different levels of unionization, or different other labor laws, or different industry mix. This is essentially a question of "external validity."

Identification of ATT

Why does diff-in-diff identify the ATT? The key identifying assumption is "parallel trends":

$$\begin{aligned} &\mathrm{E}(Y^0 \mid X_1 = 1, X_2 = 1) - \mathrm{E}(Y^0 \mid X_1 = 1, X_2 = 0) \\ &= \mathrm{E}(Y^0 \mid X_1 = 0, X_2 = 1) - \mathrm{E}(Y^0 \mid X_1 = 0, X_2 = 0). \end{aligned} \tag{10.20}$$

That is, the mean untreated potential outcome changes over time ($X_2 = 0$ to $X_2 = 1$) by the same amount in the treated ($X_1 = 1$) and untreated ($X_1 = 0$) groups. Since the observed Y is $Y = Y^1 T + Y^0(1 - T)$, where $T \equiv \mathbb{1}\{X_1 = 1, X_2 = 1\}$ (or equivalently $T = X_1 X_2$), three of the four terms in (10.20) are observable, i.e., other than the first term, $Y^0 = Y$ since $T = 0$. It is only $\mathrm{E}(Y^0 \mid X_1 = 1, X_2 = 1)$ that we cannot observe data for. Indeed, this is the counterfactual: what would the average wage have been in our city ($X_1 = 1$) in the later time period ($X_2 = 1$) if we had not passed the minimum wage law (Y^0, not Y^1)? Rearranging (10.20) thus gives us the counterfactual in terms of things we can learn about:

$$\begin{aligned} &\mathrm{E}(Y^0 \mid X_1 = 1, X_2 = 1) \\ &= \mathrm{E}(Y^0 \mid X_1 = 1, X_2 = 0) + [\mathrm{E}(Y^0 \mid X_1 = 0, X_2 = 1) - \mathrm{E}(Y^0 \mid X_1 = 0, X_2 = 0)] \\ &= \mathrm{E}(Y \mid X_1 = 1, X_2 = 0) + [\mathrm{E}(Y \mid X_1 = 0, X_2 = 1) - \mathrm{E}(Y \mid X_1 = 0, X_2 = 0)] \\ &= m(1,0) + [m(0,1) - m(0,0)]. \end{aligned} \tag{10.21}$$

The ATT is thus

$$\begin{aligned}
\text{ATT} &= \text{E}(Y^1 - Y^0 \mid X_1 = 1, X_2 = 1) \\
&= \overbrace{\text{E}(Y^1 \mid X_1 = 1, X_2 = 1)}^{Y^1 = Y \text{ since } X_1 = 1, X_2 = 1} - \overbrace{\text{E}(Y^0 \mid X_1 = 1, X_2 = 1)}^{\text{use counterfactual from (10.21)}} \\
&= \text{E}(Y \mid X_1 = 1, X_2 = 1) - \overbrace{\{m(1,0) + [m(0,1) - m(0,0)]\}}^{\text{counterfactual}} \\
&= m(1,1) - \{m(1,0) + [m(0,1) - m(0,0)]\} \\
&= [m(1,1) - m(1,0)] - [m(0,1) - m(0,0)] \\
&= \beta_3.
\end{aligned}$$

Discussion Question 10.13. Imagine $T = 11$. The treatment is between $t = 10$ and $t = 11$. You compare/graph $\text{E}(Y_{it} \mid X_i = 1, t = s) - \text{E}(Y_{it} \mid X_i = 0, t = s)$ for $s = 1, 2, \ldots, T - 1$, and the values are all extremely similar. 1) Why is this related to parallel trends? 2) Why does this not prove that parallel trends assumption holds?

Skepticism: parallel trends may not hold, and we cannot directly test it since it has to do with unobservable (counterfactual). But, if pre-treatment trends look parallel, somewhat reassuring. But note parallel trends are fragile, e.g., if trend in Y parallel, then trend in $\ln(Y)$ can't be, and vice-versa. Ways to relax to at least conditional parallel trends.

Chapter 13

I covered this in the past but might skip this year (partly b/c Zack covered MLE in 9472, I heard). But, maybe still worth showing the KLIC stuff to understand you don't need the full joint distribution, just marginals are fine.

Section 13.1

Page 469 "efficiency usually comes at the price of nonrobustness"

Page 470 "there *are* cases in which MLE turns ou to be robust to failure of certain assumptions, but these must be examined on a case-by-case basis"

Page 470 "many problems for which it is indispensable"

Section 13.2

Page 471 has margin notes/examples in my book, but maybe skip this year.

Section 13.3

For notational simplicity, imagine scalar, continuous Y with support $\mathbb{R}$. True PDF of Y evaluated at $Y = y$ is $p_o(y)$. Expectations are integrals against the true PDF. E.g., $\mathrm{E}(Y) = \int_{\mathbb{R}} y p_o(y)\,\mathrm{d}y$.

Let $f(\cdot)$ denote another PDF, the wrong one. Then,

$$\begin{aligned}\mathrm{E}[f(Y)/p_o(Y)] &= \int_{\mathbb{R}} \frac{f(y)}{p_o(y)} p_o(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}} f(y)\,\mathrm{d}y = 1\end{aligned}$$

since f is a PDF. Since $\ln(1) = 0$, then

$$\ln\{\mathrm{E}[f(Y)/p_o(Y)]\} = 0. \tag{13.1}$$

By Jensen's inequality, since $\ln(\cdot)$ is concave,

$$\mathrm{E}[\ln\{f(Y)/p_o(Y)\}] \leq 0. \tag{13.2}$$

Now imagine a family of PDFs with parameter $\boldsymbol{\theta}$: $f(\cdot;\boldsymbol{\theta})$. For example, these could be normal PDFs with $\boldsymbol{\theta} = (\mu, \sigma^2)$. Imagine our model is **properly specified** so that there exists $\boldsymbol{\theta}_o$ such that $p_o(\cdot) = f(\cdot;\boldsymbol{\theta}_o)$. Then,

$$\mathrm{E}\{\ln[f(Y;\boldsymbol{\theta}_o)/p_o(Y)]\} = \mathrm{E}\{\ln[1]\} = 0. \tag{13.3}$$

Let $\boldsymbol{\theta}$ be any possible value, not necessarily the true one. Then,

$$\begin{aligned}
\mathrm{E}\{\ln[f(Y;\boldsymbol{\theta}_o)]\} - \mathrm{E}\{\ln[p_o(Y)]\} &= \mathrm{E}\{\ln[f(Y;\boldsymbol{\theta}_o)] - \ln[p_o(Y)]\} \\
&= \mathrm{E}\{\ln[\overbrace{f(Y;\boldsymbol{\theta}_o)/p_o(Y)}^{=1}]\} \\
&= 0 \\
&\geq \mathrm{E}[\ln\{f(Y;\boldsymbol{\theta})/p_o(Y)\}] \\
&= \mathrm{E}\{\ln[f(Y;\boldsymbol{\theta})] - \ln[p_o(Y)]\} \\
&= \mathrm{E}\{\ln[f(Y;\boldsymbol{\theta})]\} - \mathrm{E}\{\ln[p_o(Y)]\}.
\end{aligned}$$

After adding $\mathrm{E}\{\ln[p_o(Y)]\}$ to both sides,

$$\mathrm{E}\{\ln[f(Y;\boldsymbol{\theta}_o)]\} \geq \mathrm{E}\{\ln[f(Y;\boldsymbol{\theta})]\}. \tag{13.4}$$

So, we can think of the true parameter value $\boldsymbol{\theta}_o$ as the value that maximizes $\mathrm{E}\{\ln[f(Y;\boldsymbol{\theta})]\}$:

$$\boldsymbol{\theta}_o = \arg\max_{\boldsymbol{\theta}} \mathrm{E}\{\ln[f(Y;\boldsymbol{\theta})]\}. \tag{13.5}$$

We could also write

$$\ell_i(\boldsymbol{\theta}) \equiv \ln[f(Y_i;\boldsymbol{\theta})] \tag{13.6}$$

and say $\boldsymbol{\theta}_o$ maximizes $\mathrm{E}[\ell_i(\boldsymbol{\theta})]$. This $\ell_i(\cdot)$ is called the **log-likelihood** for observation i.

In the sample, the analogy principle suggests the estimator

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \hat{\mathrm{E}}[\ell_i(\boldsymbol{\theta})] = \arg\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ln[f(Y_i;\boldsymbol{\theta})]. \tag{13.7}$$

Discussion Question 13.1. Does this result require independence of the Y_i, like $Y_i \perp\!\!\!\perp Y_k$? If so, where/which step?

Page 475, top half. Coincidentally, since $\ln(A) + \ln(B) = \ln(AB)$, and since $\ln(\cdot)$ is a strictly increasing function, this happens to be mathematically equivalent to maximizing

$$\prod_{i=1}^{n} f(Y_i;\boldsymbol{\theta}).$$

Again coincidentally, this is the joint PDF for all $Y_1, \ldots, Y_n$ *if* they are all independent. But we did not need independence before. It's also not clear why maximizing the joint likelihood should be a good estimator of the true parameter; e.g., there's no analogy principle argument.

Now, conditional, basically just condition on $\mathbf{X} = \mathbf{x}$ everywhere. True conditional PDF of Y given $\mathbf{X}$, evaluated at $Y = y$ and $\mathbf{X} = \mathbf{x}$, is $p_o(y \mid \mathbf{x})$. Estimator solves (13.15). Called **conditional maximum likelihood estimator** (CMLE).

Section 13.8.1

Example of usefulness of only needing marginals: panel data. Don't need to specify time series dependence structure, just use partial/pooled CMLE.

Chapter 15

Section 15.1

Binary $Y \in \{0, 1\}$, $\mathrm{P}(Y = 1) = p$, so $\mathrm{P}(Y = 0) = 1 - p$.

Discussion Question 15.1. What's $\mathrm{E}(Y)$? $\mathrm{Var}(Y)$?

[Side note: redundant to report mean and variance of binary variable.]
Thus CEF w/ binary Y is equal to cond prob fn (CPF) in (15.1),

$$p(\mathbf{x}) \equiv \mathrm{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \mathrm{E}(Y \mid \mathbf{X} = \mathbf{x}). \tag{15.1}$$

So, could do usual CEF stuff, just interpret in terms of prob instead of mean.

Section 15.2

LPM: **linear probability model**. Ambiguous: here it's linear-in-variables, but people also mean linear-in-parameters (e.g., include X^2).

As w/ CEF, true function may not be linear; then we have a LP/BLA/BLP interpretation.

Discussion Question 15.2. Let $Y = \mathbb{1}\{\text{employed}_i\}$, X years of education. Imagine estimate $\hat{\beta}_0 = 0.56$, $\hat{\beta}_1 = 0.02$ in $\hat{p}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. 1) Would these estimates ever predict employment probability $> 100\%$? Or $< 0\%$? If so, when? 2) Does (1) imply our model is bad? Why/not?

Pros: 1) can use OLS, 2) easy interpretation if linear-in-variables, 3) LP/BLA/BLP interp if misspecified.

Cons: 1) can have fitted or predicted $\hat{p}(\mathbf{x}) < 0$ or > 1 for certain $\mathbf{x}$; especially problematic for more "extreme" $\mathbf{x}$; 2) constant marginal effects.

But: 1) maybe we don't care about extreme $\mathbf{x}$, like education of 30 years; 2) can relax to nonlinear-in-variables but still linear-in-parameters; 3) maybe linear is good approx over the values of $\mathbf{x}$ that we care about.

Discussion Question 15.3. Say $Y = \mathbb{1}\{\text{employed}_i\}$, $X = \mathbb{1}\{\text{male}_i\}$, $\mathrm{P}(Y \mid x) = \beta_0 + \beta_1 x$. What's interpretation of β_0 and β_1? What are sample analogs $\hat{\beta}_0$ and $\hat{\beta}_1$? Is $\hat{p}(x) < 0$ or > 1 possible? Why/not? Hint: can treat as CEF if more comfortable; or, try plugging in x values.

Section 15.3

A type of (single) **index model**:

$$p(\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta}), \tag{15.2}$$

for known (nonlinear) function $G(\cdot)$. The $\mathbf{x}\boldsymbol{\beta}$ part is called the "index." A cynical view of the model is that it just takes the linear model and pushes the too-high parts down below 1, and pulls up the too-low parts above 0. That is, for *any* r, $0 \leq G(r) \leq 1$, so for *any* $\hat{\boldsymbol{\beta}}$, $\hat{p}(\mathbf{x}) = G(\mathbf{x}\hat{\boldsymbol{\beta}}) \in [0,1]$, too. A more aspirational view of the model is that $G(\cdot)$ is the CDF of some latent structural error term.

Discussion Question 15.4. The "probit" model has $G(\cdot)$ as the normal CDF, $\Phi(\cdot)$. Consider simple model $\Phi(\beta_0 + \beta_1 x)$. 1) Why is derivative wrt x equal $\phi(\beta_0 + \beta_1 x)\beta_1$, where $\phi(\cdot)$ is normal PDF? 2) Can derivative depend on x (unlike LPM)? 3) Can deriv be constant? 4) Can deriv have different sign ($\pm$) at different x?

Don't mistake "nonlinear" for "flexible" (or "robust to misspecification"). Although nonlinear, the index model is not any more "flexible" than the LPM. There are still the same number of parameters. The flexibility just goes in different directions. Just as the model $\beta_0 + \beta_1 x$ cannot yield nonlinear estimates (no matter how nonlinear the data look), the model $G(\beta_0 + \beta_1 x)$ cannot yield linear estimates (unless $\beta_1 = 0$). If $G(\cdot)$ is a Gaussian or logistic CDF, then $\hat{p}(\cdot)$ always has an S shape even if the true $p(\cdot)$ does not.

A structural view of the index model comes from a linear-in-parameters structural model for a latent (unobserved) outcome variable Y^*. This Y^* is often something like utility. The observed binary Y is $Y = \mathbb{1}\{Y^* > 0\}$. For example, maybe Y^* is the utility gain (which could be negative) from getting married, and Y is somebody's decision to get married; or serve in the military, go to college, buy organic milk, etc. We imagine

$$Y^* = \mathbf{X}\boldsymbol{\beta} + U, \quad Y = \mathbb{1}\{Y^* > 0\}, \tag{15.3}$$

where the structural error U is continuous, symmetric, and has CDF $G(\cdot)$. "Symmetric" meaning $G(r) = 1 - G(-r)$; w/ $r = 0$ implies $G(0) = 0.5$. But, the critical "exogeneity" assumption is $U \perp\!\!\!\perp \mathbf{X}$. Then (can add detail for 2nd equality and 3rd equality in class)

$$\mathrm{P}(Y = 1 \mid \mathbf{x}) = \mathrm{P}(Y^* > 0 \mid \mathbf{x}) = \overbrace{\mathrm{P}(U > -\mathbf{x}\boldsymbol{\beta} \mid \mathbf{x}) = 1 - G(-\mathbf{x}\boldsymbol{\beta})}^{\text{by exogeneity}} = G(\mathbf{x}\boldsymbol{\beta}), \tag{15.4}$$

where the last equality uses symmetry.

The **probit model** has $G(\cdot) = \Phi(\cdot)$, the standard normal CDF (meaning the structural U is std normal), while **logit model** has $G(\cdot) = \Lambda(\cdot)$, the (standard) logistic CDF. My impression is economists use probit more b/c used to normal errors, while statisticians prefer logit because of interpreting coefficients in terms of log odds ratios and stuff?

Why assume $\mathrm{Var}(U) = 1$? Well, imagine simplest possible model,

$$\mathrm{P}(Y = 1) = \mathrm{P}(Y^* > 0) = \mathrm{P}(U > -\beta_0) = G(\beta_0).$$

Assume $U \sim \mathrm{N}(0, \sigma_U^2)$. Imagine $\mathrm{P}(Y = 1) = 0.95$. Then it's possible $U \sim \mathrm{N}(0,1)$ and $\beta_0 = 1.64$. But, it's also possible $\mathrm{Var}(U) = 4$ and $\beta_0 = 2 \times 1.64$:

$$\mathrm{P}\{\mathrm{N}(0,4) \leq 2 \times 1.64\} = \mathrm{P}\{\mathrm{N}(0,1) \leq 1.64\}.$$

Discussion Question 15.5. In this simple example w/ only β_0, can we even estimate the sign of β_0? Say $\hat{\mathrm{E}}(Y) = 0.4$; can we estimate if $\beta_0 < 0$ or > 0 or $= 0$? Why/not/how? Hint: recall $G(\cdot)$ symmetric.

For this reason, we say the parameter β_0 is only **identified up to scale**, which is also true of all other β_j for the same reason. But, it's hard to interpret the parameters themselves anyway when Y^* is something like utility. Usually focus on partial effects of the X on probabilities. Turns out, these partial effects (derivatives) do *not* depend on scale; phew! Also, statistical significance doesn't depend on scale since just about distinguishing from zero.

As usual, we can look at partial derivatives like (15.13), or just finite differences like (15.14).

Discussion Question 15.6. Let $p(x) = \Phi(\beta_0 + \beta_1 x)$, so partial deriv wrt x is $\phi(\beta_0 + \beta_1 x)\beta_1$; $Y = \mathbb{1}\{\text{employed}_i\}$, X is income of spouse measured in \$1000s (like $x = 50$ means \$50,000). Let $\hat{\beta}_0 = 4$, $\hat{\beta}_1 = -0.1$, SE of $\hat{\beta}_1$ is 0.01. 1) Statistical significance? 2) Economic significance? Hint: are stat sig and econ sig the same at all x? Also, $\phi(0) \approx 0.4$, $\phi(2) \approx 0.05$, $\phi(4) \approx 0.0001$.

Finally (p. 567), if index is nonlinear-in-variables (but linear-in-parameters), then can use chain rule to get partial derivatives; or again just use finite differences.

Section 15.4

skip?

Section 15.6

Things to report.

Page 573: percent correctly predicted. But: what's predicted? Let y actual values, g guessed (predicted). Imagine loss function $L_0(y, g) = 0$ if $y = g$ (correct guess good, zero loss), $L_0(y, g) = 1$ if $y \neq g$ (incorrect guess bad). More succinctly,

$$L_0(y, g) = \mathbb{1}\{y \neq g\} \tag{15.5}$$

is known as **0–1 loss**. Consider long-run average loss after many predictions, which should be $\mathrm{E}[L_0(Y, g)]$, where the expectation is over the distribution of rv Y. This **expected loss** is also called **risk**. Given 0–1 loss, risk is

$$\mathrm{E}[L_0(Y, g)] = \mathrm{E}[\mathbb{1}\{Y \neq g\}] = \mathrm{P}(Y \neq g) = 1 - \mathrm{P}(Y = g). \tag{15.6}$$

Picking g to minimize risk is equivalent to

$$g_0^* \equiv \arg\min_g \mathrm{E}[L_0(Y, g)] = \arg\min_g [1 - \mathrm{P}(Y = g)] = \arg\max_g \mathrm{P}(Y = g). \tag{15.7}$$

That is, g_0^* is the mode, the single most probable value of Y. This is true even w/ non-binary Y. This is also true conditional on $\mathbf{X} = \mathbf{x}$:

$$g_0^*(\mathbf{x}) = \arg\min_g \mathrm{E}[L_0(Y, g) \mid \mathbf{X} = \mathbf{x}] = \arg\max_g \mathrm{P}(Y = g \mid \mathbf{x}). \tag{15.8}$$

So for binary Y, the optimal prediction is 1 if $p(\mathbf{x}) > 0.5$ and is 0 if < 0.5 (and indifferent if $= 0.5$). So, in practice, w/ 0–1 loss, predict 1 if $\hat{p}(\mathbf{x}) > 0.5$, o/w 0; i.e., given the estimated model and some new $\mathbf{x}$ value, predict

$$\hat{y}(\mathbf{x}) = \mathbb{1}\{\hat{p}(\mathbf{x}) > 0.5\}. \tag{15.9}$$

Discussion Question 15.7. I thought the CEF was the best predictor. The CEF is $\mathrm{E}(Y \mid \mathbf{x}) = p(\mathbf{x})$. Is this the same as (15.9)? If not, is it better or worse? Why?

Also, if trying to measure goodness of fit, better to use leave-one-out predictions or something anyway. And also (page 575) better to focus on statistical and economic significance?

Also, sometimes we care much more about accuracy for predicting, say, $Y = 1$ than $Y = 0$, like if Y is an indicator of (future) homelessness or crime etc. In that case, the "loss" from guessing 0 when $y = 1$ is much higher/worse than guessing 1 when $y = 0$ (assuming it's like, trying to help somebody, not pre-emptively arresting them or something!). That is, we'd like $L(y, g)$ to have $L(1, 0) > L(0, 1) > 0 = L(0, 0) = L(1, 1)$; similar to thinking type II errors are worse than type I. This is weighted 0–1 loss. The optimal prediction is then no longer the mode; we may want a lower threshold τ in $\hat{y} = \mathbb{1}\{\hat{p}(\mathbf{x}) > \tau\}$, where τ is like a critical value or significance level α in hypothesis testing.

Discussion Question 15.8. When would lower τ help better predict $Y = 1$? When would lower τ help better predict $Y = 0$?

Discussion Question 15.9. 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.

Page 575: partial effect at the average (PEA). One drawback: like average "male" maybe 0.5, not actually possible. The mean individual may not exist; although the median individual does.

Page 577: average partial effect (APE), (15.32–33).

Section 15.7.5

Pass.

Chapter 16

Skip.

Chapter 19

Probably not enough time, skip whole chapter.

Section 19.3

Terms: **selected sample**, **selection mechanism** (p. 790).
Example 19.2.
Stata default: **complete case analysis** (drop any row with *any* missing value).
Example 19.3, emphasize self-selection.

Section 19.4.1 through top p. 796

.....

Section 19.5

Skip. (Tobit.)

Chapter 21

Section 21.1

Eh.

Section 21.2

Rubin causal model, **potential outcomes**: like parallel universes, numbered 0 and 1. Each individual has (Y_0, Y_1), but we're stuck in our universe so can only observe one, not both. Fundamental problem of causal effect identification is this unobservability.

Page 905: SUTVA. Rules out spillovers, general eqm effects, etc. Often very restrictive in economics (vs. surgery, etc.).

Discussion Question 21.1. Consider a study of 100 people that concluded the average causal effect of getting a college degree on annual salary was a \$15,000/yr increase. If the U.S. government then gave everyone a college degree, do you think average salary would increase by \$15,000, or less, or more? Why?

Discussion Question 21.2. Consider a population with four types of individuals, each with probability 0.25. Each "type" has a different (Y_0, Y_1) potential outcome pair; see Table 1. Compute: $\mathrm{E}(Y_0)$, $\mathrm{E}(Y_1)$, $\mathrm{E}(Y_1) - \mathrm{E}(Y_0)$, $\mathrm{E}(Y_1 - Y_0)$, $\mathrm{Q}_{0.4}(Y_1) - \mathrm{Q}_{0.4}(Y_0)$, and $\mathrm{Q}_{0.4}(Y_1 - Y_0)$.

Note $\mathrm{E}(Y_1) - \mathrm{E}(Y_0) = \mathrm{E}(Y_1 - Y_0)$ by linearity, but different interpretations. For quantiles, nonlinear operator so not generally equal; QTE (former version) is easier to estimate, often more policy-relevant anyway.

Observe $Y = WY_1 + (1 - W)Y_0$, where $W = 1$ if treated, $W = 0$ o/w.

(21.1) ATE is $\mathrm{E}(Y_1 - Y_0)$ by definition; equals $\mathrm{E}(Y_1) - \mathrm{E}(Y_0)$ by linearity.

Table 1: Potential outcomes example.

Y_0	Y_1	$Y_1 - Y_0$	Probability
0	1	1	0.25
1	2	1	0.25
2	4	2	0.25
3	0	-3	0.25

Table 2: Potential outcomes example.

Y_0	Y_1	W	Probability
0	5	1	0.25
3	4	1	0.25
2	2	0	0.25
5	0	0	0.25

Page 906, (21.2) ATT: $\mathrm{E}(Y_1 - Y_0 \mid W = 1)$. LATE on same page but save for later section w/ more detail. For ATT,

$$\begin{aligned}\mathrm{E}(Y_1 - Y_0 \mid W = 1) &= \mathrm{E}(Y_1 \mid W = 1) - \mathrm{E}(Y_0 \mid W = 1) \\ &= \mathrm{E}(Y \mid W = 1) - \mathrm{E}(Y_0 \mid W = 1).\end{aligned}$$

The first term is just a statistical object, and conditional mean involving (always) observable variables, so identified by definition. The second term is the counterfactual: we never observe Y_0 when $W = 1$, so in order to learn ATT, we need identifying assumptions that let us learn about this fundamentally unobservable object.

Discussion Question 21.3. Consider Table 2. Compute: 1) ATT, 2) ATE, 3) $\mathrm{E}(Y \mid W = 1) - \mathrm{E}(Y \mid W = 0)$. If we compute the sample mean difference between treated and untreated groups, is that a consistent estimator for any, all, or none of (1,2,3)?

Conditional: CATE, CATT, same but also condition on $\mathbf{x}$. By LIE,

$$\mathrm{E}[\mathrm{E}(Y_1 - Y_0 \mid X)] = \mathrm{E}[\mathrm{CATE}(X)] = \mathrm{ATE}\,. \tag{21.1}$$

Discussion Question 21.4. Consider expanding Medicaid (health insurance for low-income people). Do we care about the average benefit to current Medicaid recipients, or hypothetical average benefit to non-recipients, or another subpopulation?

Page 907: identification under independence $W \perp\!\!\!\perp (Y_0, Y_1)$, and under mean independence $\mathrm{E}(Y_0 \mid W) = \mathrm{E}(Y_0)$, $\mathrm{E}(Y_1 \mid W) = \mathrm{E}(Y_1)$. Since $Y = Y_1$ when $W = 1$, and $Y = Y_0$ when $W = 0$, then altogether

$$\mathrm{E}(Y_w) = \mathrm{E}(Y_w \mid W = w) = \mathrm{E}(Y \mid W = w), \quad w = 0, 1.$$

The final term is a statistical object, just a feature of the joint distribution of the observable Y and W. Thus we can consistently estimate

$$\hat{\mathrm{E}}(Y \mid W = w) \xrightarrow{p} \mathrm{E}(Y \mid W = w) = \mathrm{E}(Y_w), \quad w = 0, 1. \tag{21.2}$$

So the sample mean difference fundamentally estimates a population mean difference, but the identifying (mean) independence assumption (along with SUTVA!) lets us interpret it with causal meaning. Also, independence implies ATE=ATT since $\mathrm{E}(Y_1 - Y_0 \mid W = 1) = \mathrm{E}(Y_1 - Y_0)$.

However, even when eligibility for "economic" programs (like job training) is randomized, people are usually not *forced* to participate. The fact that they get to choose usually means there is self-selection, that familiar enemy of identification. ATT can still be identified if $W \perp\!\!\!\perp Y_0$, but even that may be suspect.

Discussion Question 21.5. In which direction do you think self-selection would bias ATE estimator if: 1) everyone has same Y_0? 2) everyone has same Y_1? 3) $(Y_1 - Y_0)$ is decreasing in Y_0? Hint: imagine the true ATE is just zero for simplicity; is the sample mean difference positive or negative?

If time, (21.5) is interesting: shows cond mean diff equals ATT up to some bias term that's zero under mean indep assumption.

Section 21.3

Go through bold terms and numbered/display equations.

Section 21.3.1

Show CATE identification under ATE.1'; basically same as before but just condition on $\mathbf{x}$ everywhere:

$$\begin{aligned} \mathrm{E}(Y_1 \mid \mathbf{x}) &\overbrace{= \mathrm{E}(Y_1 \mid \mathbf{x}, W = 1)}^{\text{by ATE.1'}} \\ &= \mathrm{E}(Y \mid \mathbf{x}, W = 1) \equiv m_1(\mathbf{x}), \end{aligned}$$

which is just a statistical object. The same argument holds for $W = 0$, so

$$\mathrm{E}(Y_0 \mid \mathbf{x}) = \mathrm{E}(Y \mid \mathbf{x}, W = 0) \equiv m_0(\mathbf{x}). \tag{21.3}$$

Thus, the CATE is identified by a difference in conditional means, which are both identified as long as ATE.2 (conditional overlap) holds:

$$\mathrm{CATE}(\mathbf{x}) \equiv \mathrm{E}(Y_1 - Y_0 \mid \mathbf{x}) = \mathrm{E}(Y_1 \mid \mathbf{x}) - \mathrm{E}(Y_0 \mid \mathbf{x}) = m_1(\mathbf{x}) - m_0(\mathbf{x}). \tag{21.4}$$

Note that this is nonparametric identification: we have not assumed a functional form for the CEFs. We might assume a functional form to estimate the CEF, but that only affects estimation, not interpretation. In other cases, identification itself depends on the functional form (linear, logit, etc.).

Discussion Question 21.6. In Chapter 4, to estimate $\mathrm{E}(Y \mid \mathbf{x}, w) = \beta_0 + \beta_1 w + \mathbf{x}\boldsymbol{\beta}_2$, with binary $w \in \{0, 1\}$, did we require $0 < \mathrm{P}(W = 1 \mid \mathbf{x}) < 1$ for all $\mathbf{x}$? Why/not?

(21.16): if CATEs identified, then ATE is identified b/c ATE is just average of CATEs, like (21.1).

Section 21.4.3

Reconsider the simple IV model from Chapter 5, with binary X (now W) and Z. We'll call Y the outcome, W the treatment, and Z the offer (of treatment). A common situation has

Table 3: Potential treatments and outcomes example.

Type	Probability	W_0	W_1	Y_0	Y_1
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	2	8

treatment offer Z randomized, but W can still be chosen to some degree, so there is some self-selection.

Recall the "Wald estimator," the ratio of the LPCs in the LP of Y onto $(1, Z)$ and X onto $(1, Z)$. In simple binary LP $Y = \alpha + \lambda Z + V$, $\lambda = \mathrm{E}(Y \mid Z = 1) - \mathrm{E}(Y \mid Z = 0)$, and similarly in $W = \delta_0 + \theta Z + R$, $\theta = \mathrm{E}(W \mid Z = 1) - \mathrm{E}(W \mid Z = 0)$; see (5.2) and (5.4). Thus, our estimator $\hat{\lambda}/\hat{\theta}$ can be written as

$$\hat{\tau}_{LATE} = \frac{\hat{\mathrm{E}}(Y \mid Z = 1) - \hat{\mathrm{E}}(Y \mid Z = 0)}{\hat{\mathrm{E}}(W \mid Z = 1) - \hat{\mathrm{E}}(W \mid Z = 0)}. \tag{21.5}$$

Here, we use potential outcomes to derive a new interpretation of the same IV estimator in (21.5).

In addition to potential outcomes, we also have "potential treatments." Recall the idea of potential outcomes: Y_0 is the outcome (Y) in the untreated universe ($W = 0$), while Y_1 is the outcome in the treated universe ($W = 1$); i.e., Y_w is the outcome when $W = w$. We can also consider two parallel universes defined by the treatment offer Z, and what the individual decides for W in each universe. So, W_z is the treatment status when the offer is $Z = z$: W_0 is the treatment status when not offered, and W_1 is the treatment status when offered. Nesting subscripts, we could in principle write the outcome in the $Z = 1$ universe as Y_{W_1}, etc. Similar to $Y = Y_0(1 - W) + Y_1 W = Y_0 + W(Y_1 - Y_0)$,

$$W = W_0(1 - Z) + W_1 Z = W_0 + Z(W_1 - W_0). \tag{21.6}$$

Consider four "types" of individuals based on their value of (W_0, W_1):

A Always takers: $(W_0, W_1) = (1, 1)$, they always get treated regardless of Z.

N Never takers: $(W_0, W_1) = (0, 0)$, they never get treated regardless of Z.

D Defiers: $(W_0, W_1) = (1, 0)$, they always "defy" the offer and do the opposite.

C Compliers: $(W_0, W_1) = (0, 1)$, they always "comply" with the offer and do whatever it says, getting treated if $Z = 1$ and not treated if $Z = 0$.

Table 3 shows an example of potential treatments for the four types, along with average (conditional) potential outcomes within each type. We could replace Y_0 with $\mathrm{E}(Y_0 \mid \text{type})$, and replace Y_1 with $\mathrm{E}(Y_1 \mid \text{type})$, 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, but it is also often plausible. That is, we could imagine an always-taker who

sneaks into the treatment even if not offered because they benefit so much; we could imagine a never-taker who would be harmed by the treatment; we could imagine a complier who's helped by treatment but can't get it without the offer; but a defier does not seem rational.

Discussion Question 21.7. Using Table 3, compute and interpret: 1) $\mathrm{E}(Y \mid Z = 0)$, 2) $\mathrm{E}(Y \mid Z = 1)$, 3) $\mathrm{P}(W = 1 \mid Z = 0)$, 4) $\mathrm{P}(W = 1 \mid Z = 1)$.

Discussion Question 21.8. Using your previous calculations, compute and interpret

$$\frac{\hat{\mathrm{E}}(Y \mid Z = 1) - \hat{\mathrm{E}}(Y \mid Z = 0)}{\hat{\mathrm{E}}(W \mid Z = 1) - \hat{\mathrm{E}}(W \mid Z = 0)}.$$

The assumption of "no defiers" is in (21.90), called **monotonicity** since it's equivalent to $W_1 \geq W_0$: the only time $W_1 < W_0$ is for defiers.

Also note "compliers" are identified by $W_1 - W_0 = 1$. Defiers have $W_1 - W_0 = -1$. Types A and N have $W_1 - W_0 = 0$.

Then: go through formal identification math on pages 951–953.

For the numerator, we can show (21.91):

$$\mathrm{E}(Y \mid Z = 1) - \mathrm{E}(Y \mid Z = 0) = \mathrm{E}(Y_1 - Y_0 \mid W_1 - W_0 = 1)\,\mathrm{P}(W_1 - W_0 = 1), \tag{21.7}$$

where as a reminder $W_1 - W_0 = 1$ is equivalent to "complier." Plugging (21.6) into $Y = Y_0 + W(Y_1 - Y_0)$,

$$Y = Y_0 + W_0(Y_1 - Y_0) + Z(W_1 - W_0)(Y_1 - Y_0). \tag{21.8}$$

Given independence of Z and everything, conditioning does not change the distribution of any other rv or its mean. Thus,

$$\begin{aligned}
\mathrm{E}(Y \mid Z = 1) &= \mathrm{E}[Y_0 + W_0(Y_1 - Y_0) + (W_1 - W_0)(Y_1 - Y_0) \mid Z = 1] \\
&= \mathrm{E}[Y_0 \mid Z = 1] + \mathrm{E}[W_0(Y_1 - Y_0) \mid Z = 1] + \mathrm{E}[(W_1 - W_0)(Y_1 - Y_0) \mid Z = 1] \\
&= \mathrm{E}[Y_0] + \mathrm{E}[W_0(Y_1 - Y_0)] + \mathrm{E}[(W_1 - W_0)(Y_1 - Y_0)], \\
\mathrm{E}(Y \mid Z = 0) &= \mathrm{E}[Y_0 + W_0(Y_1 - Y_0) \mid Z = 1] \\
&= \mathrm{E}[Y_0 \mid Z = 1] + \mathrm{E}[W_0(Y_1 - Y_0) \mid Z = 1] \\
&= \mathrm{E}[Y_0] + \mathrm{E}[W_0(Y_1 - Y_0)].
\end{aligned}$$

Subtracting,

$$\begin{aligned}
\mathrm{E}(Y \mid Z = 1) - \mathrm{E}(Y \mid Z = 0) &= \mathrm{E}[Y_0] + \mathrm{E}[W_0(Y_1 - Y_0)] + \mathrm{E}[(W_1 - W_0)(Y_1 - Y_0)] \\
&\quad - \{\mathrm{E}[Y_0] + \mathrm{E}[W_0(Y_1 - Y_0)]\} \\
&= \mathrm{E}[(W_1 - W_0)(Y_1 - Y_0)] \\
&= (1)\,\mathrm{E}[Y_1 - Y_0 \mid W_1 - W_0 = 1]\,\mathrm{P}(W_1 - W_0 = 1) \\
&\quad + (0)\,\mathrm{E}[Y_1 - Y_0 \mid W_1 - W_0 = 0]\,\mathrm{P}(W_1 - W_0 = 0) \\
&\quad + (-1)\,\mathrm{E}[Y_1 - Y_0 \mid W_1 - W_0 = -1]\,\mathrm{P}(W_1 - W_0 = -1) \\
&= \mathrm{E}[Y_1 - Y_0 \mid \overbrace{W_1 - W_0 = 1}^{\text{compliers}}]\,\mathrm{P}(W_1 - W_0 = 1) \\
&\quad - \mathrm{E}[Y_1 - Y_0 \mid \overbrace{W_1 - W_0 = -1}^{\text{defiers}}]\,\overbrace{\mathrm{P}(W_1 - W_0 = -1)}^{\text{assume zero}} \\
&= \mathrm{E}[Y_1 - Y_0 \mid W_1 - W_0 = 1]\,\mathrm{P}(W_1 - W_0 = 1).
\end{aligned}$$

This is like the "CATE" for compliers times the proportion of compliers. The denominator will cancel out the proportion, leaving the CATE for compliers.

For the denominator, use the fact that Z is independent of everything, so conditioning on Z has no effect, and the observation equation for $W = W_1$ if $Z = 1$ and $W = W_0$ if $Z = 0$ (and equating probabilities with expectations for binary variables):

$$\begin{aligned}
\mathrm{E}(W \mid Z = 1) - \mathrm{E}(W \mid Z = 0) &= \mathrm{E}(W_1 \mid Z = 1) - \mathrm{E}(W_0 \mid Z = 0) \\
&= \mathrm{E}(W_1) - \mathrm{E}(W_0) \\
&= \mathrm{E}(W_1 - W_0) \\
&= (1)\,\mathrm{P}(W_1 - W_0 = 1) + (0)\,\mathrm{P}(W_1 - W_0 = 0) \\
&\quad + (-1)\,\mathrm{P}(W_1 - W_0 = -1) \\
&= \mathrm{P}(W_1 - W_0 = 1) - \mathrm{P}(W_1 - W_0 = -1).
\end{aligned}$$

This is the probability (population proportion) of compliers minus the probability (population proportion) of defiers. If the latter is assumed zero, then altogether

$$\mathrm{E}(W \mid Z = 1) - \mathrm{E}(W \mid Z = 0) = \mathrm{P}(W_1 - W_0 = 1) = \mathrm{P}(\text{complier}). \tag{21.9}$$

The LHS is a statistical object that can be consistently estimated; the RHS has causal meaning and cannot be observed directly.

Unfortunately, although we can estimate the proportion of compliers (assuming no defiers), we have no way of identifying who these compliers are, even in our sample. We can tell somebody isn't a complier if $(Z, W) = (1, 0)$ or $(0, 1)$, but we don't know if $(Z, W) = (1, 1)$ is a complier or always-taker, and we don't know if $(Z, W) = (0, 0)$ is C or N.

So, we can formally define what we're estimating, but it's hard to interpret, which may make it less helpful for policy.

Further, note that the population LATE itself depends on Z. If we have two different instruments, even if both are valid, we get not only two different estimates, but two different population objects.

Bibliography

Hansen, B. E., 2018. Econometrics, unpublished textbook, available at https://www.ssc.wisc.edu/~bhansen/econometrics/.

Last updated: February 28, 2019

https://faculty.missouri.edu/~kaplandm/personalTeaching.html