Problem Set 3, Question 2

Mitch Downey
Econometrics I

February 18, 2024

1 Original question

1. *FWL, OV'B', and LATE's.* You are interested in the causal effect of parental income on children's outcomes. You will simulate the data, so you know the truth and can compare that with regression results. Simulate 500 observations according to the following data generating process (DGP):

   - Earnings $\sim N(19, 1)$. Note that if I later refer to "labor market earnings," I am referring to this variable.
   - Capital gains $\sim N(1, 1)$
   - $u \sim N(0, 1)$
   - $e \sim N(0, 1)$
   - Occupational status = Earnings + u
   - Child Outcomes = Earnings - Capital gains + e. Note that this means that labor market earnings are good for children (improve their outcomes), while capital gains are actually bad for them (perhaps because they are unearned and send a bad signal to them about the value of hard work.
   - Income $\equiv$ Earnings + Capital gains

   (a) On average, across all individuals in your simulated sample, what fraction of income comes from labor market earnings?

   (b) Note that the average respondent will have income of 20. Given that you simulated the data, what would you say is the true causal effect on child outcomes of increasing the average person's income by 10%, from 20 to 22?

   (c) Regress child outcomes on earnings and capital gains and verify that OLS recovers the correct coefficients. Verify that controlling for occupational status (which is not part of the DGP) does not affect these coefficients.

   (d) A researcher is not interested in the potential distinction between earnings and capital gains, and she pools both into income. Regress child outcomes on income, and compare the coefficient to your answer to (b) above. Is income endogenous?

(e) Within your sample, what is the correlation between occupational status and income?

(f) The researcher is concerned that she should be controlling for occupational status: it is highly correlated with income and excluding it might cause omitted variable bias (OVB). In this univariate context, OVB is a function of three terms, and since you simulated the data, you know what all three terms are. Analytically (by hand, without a computer) calculate the OVB that results from excluding occupational status.

(g) Regress child outcomes on income, controlling for occupational status, and compare the coefficient to your answer to (f) above. Does the researcher conclude that there is OVB? Has she now recovered the causal effect of income on child outcomes? Compare your answer to your answer to (b) above, and discuss the role of endogeneity.

2 Original solution

1. *FWL, OV'B', and LATE's.*

   (a) Since E(earnings) = 19 and E(capital gains) = 1, it will be close to 95% of income.

   (b) There's no right answer to this question, and that's kind of the point. Given that 95% of income comes from earnings, many will answer that the effects of an increase in income will be a weighted average of the effects of earnings and capital gains, with earnings getting a weight of .95 (and that was the point of priming you to think about the share of income that comes from earnings in part (a)). This would imply $2 \times (.95 \times 1 + .05 \times -1) = 1.8$. But regressions don't care about averages, they care about variance. So one might instead expect a variance-weighted average. Since both types of income have the same variance, this would imply an effect of 0. But the point is: It depends. Why is income increasing? Which type of income is increasing? "The true causal effect" is a misleading term because any answer between -2 and 2 is completely reasonable, and all of them would be true causal effects, depending on the source of the income increase.

   (c) Coefficients are close to 1 and -1, regardless of whether or not you control for occupation status.

   (d) The effect is zero, but income is not endogenous.

   (e) The correlation is $\frac{cov(inc,occ)}{\sigma_{inc}\sigma_{occ}} = \frac{E[(cg+e-0)(e+u-0)]}{(\sqrt{v(e)+v(cg)})(\sqrt{v(e)+v(u)})} = \frac{E(e^2)}{(\sqrt{1+1})(\sqrt{1+1})} = \frac{1}{2}$. So your sample should be close to that. Note that this is considered a pretty high correlation.

   (f) The bias is given by $\frac{cov(income,occstat)}{var(income)}\beta_{occstat}$. But $\beta_{occstat} = 0$, so the bias is zero.

   (g) Controlling for occupational status, the coefficient on income goes to -.34 and becomes significant. For most researchers, when you control for a variable correlated with your main x variable of interest and you see the coefficient change, you conclude that there was Omitted Variable Bias, and that your answer in the "longer" regression (i.e., the one that includes more controls) is closer to "the" true causal effect of income. Note that the answer controlling for occupational status isn't *wrong*: She

has recovered a causal effect of income. But it only corresponds to one specific type of income increase (one that is predominantly driven by increased capital gains) and given that capital gains only account for 5% of income, that's probably not what she was expecting or how she's likely to interpret it. The point is that even without omitted variable bias, because of FWL, controls shift which part of the variation is identifying the coefficient, and if there are heterogeneous effects (whether heterogeneous across different people/observations in the sample, or heterogeneous across different sources of variation in the x variable of interest) then that can change which of the effects your regressions are picking up. There's nothing statistically wrong with this, but it makes statistical results very hard to interpret. One solution (the focus of our final lecture together) is to adopt the potential outcomes framework where we are a little more explicit and intentional about what we're trying to estimate, and to use an identification strategy that corresponds to the estimand we're interested in.

3 Further elaboration on part d

A question that has come up a fair bit in whether income is endogenous. This seems like a good opportunity to use this question to explain and reemphasize some core ideas.

3.1 What "should" β be?

As we discussed in the lecture, when you pool over multiple heterogenerous treatment effects, OLS gives you a variance weighted average of those treatment effects. There, the calculation was: Suppose $x = \sum_j x_j$ is a collection of J different components, each with its own effect. For simplicity, let's assume each x_j is mean zero, but that doesn't actually matter. Let's also assume that they are all independent of each other, which matters a lot.

If we estimated a fully disaggregated model of the form:

$$y = \beta_0 + \sum_j \beta_j x_j + \varepsilon \tag{1}$$

then each $\hat{\beta}_j$ would be a consistent estimator for the population projection coefficient β_j, which would be the causal effect of x_j if x_j is independent of potential outcomes. To avoid confusion, let's assume that each x_j is uncorrelated with ε, which is mechanically always true if we think about the β's as the population projection coefficients, but since there's some confusion about causality, let's also assume that it's true about the structural error term (in the terminology of the traditional linear model), which is true in the simulation and allows us to interpret this β_j as the causal effect of x_j. But remember what happens if you estimate this instead:

$$y = \gamma_0 + \gamma_1 x + \nu \tag{2}$$

Well then:

$$\begin{aligned}\gamma_1 &= \frac{cov(x,y)}{var(x)} \\ &= \frac{E\big[(x_1+...+x_J)(\beta_0+\beta_1+...+\beta_J x_J+\varepsilon)\big]}{var(\sum_j x_j)} \\ &= \frac{E(x\beta_0)+E\big[x_1^2\beta_1+...+x_J^2\beta_J\big]+E(x\varepsilon)}{\sum_j var(x_j)} \quad \text{(because of independence)}\end{aligned}$$

So then

$$\gamma_1 = \frac{\sum_j var(x_j)\beta_j}{\sum_j var(x_j)} \tag{3}$$

This says that the coefficient on the pooled variable is a variance-weighted average of the individual coefficients. This is the most important formula for understanding heterogeneous treatment effects, which is a real-world issue that is relevant for all data analysis (including when the treatment variable of interest is randomly assigned). Recognizing this formula and its practical relevance is a key point of this problem set question. This is an important lesson, and that's why I emphasized it in class. Because it's important.

In the case of this problem set example, the two variables have equal variance, and so the coefficient on the pooled variable is the simple mean of the two underlying coefficients (the mean of -1 and 1 is zero). Thus, zero is the "correct" coefficient. To the extent that this doesn't line up with your intuition that about the causal effect that you should recover, it is because you had the wrong intuition. Many plausibly had the intuition that they should focus on the mean of the different income sources, not the variance.

But imagine for a second that the variance of earnings was zero. Well then income is just adding a constant to capital gains, and no matter how big that constant is (i.e., no matter how big are mean earnings), that's not going to affect the coefficient you estimate. Nor should it: Regressions are about using the variance in X to explain the variance in Y. If you don't have any variation in X, then you can't learn anything about its effects.

This insight applies to causal questions, too. If you ever work with non-quantitative people who design programs (e.g., drug treatment programs), they often say "We have collected a lot of data on our participants, you should use it to evaluate the effects of our program." But if you have no control group, then you cannot evaluate something based only on treatment group outcomes, and the statistical reason (which just formalizes your intuition) is because there's no variation in treatment status, and if there's no variation in a variable, then you cannot identify (using the econometric, not the applied micro, terminology) its coefficient because any possible number for that coefficient would be consistent with the observed data, it's just that each coefficient on the variable that doesn't actually vary would imply a different value for the constant term in the regression.

So clearly, if there was no variation in earnings, then it would "feel" correct for the coefficient to simply reflect the effects of capital gains. That intuition is right, but it's from an edge case. Extending it, you quickly arrive at the conclusion that OLS "should" variance-weight. In that sense, OLS recovering a zero effect of income is correct, and the fact that 95% of income comes from earnings and not capital gains was an intentional effort to distract you.

3.2 Does exogeneity require equal variance?

What about this proof circulating that income is only exogenous because the variance weights are equal? I disagree with it. That proof goes like this: Let's say the true DGP is:

$$y = E - C + \nu \tag{4}$$

but we estimate

$$y = \alpha + \beta I + \varepsilon \tag{5}$$

Is it the case that $E(I\varepsilon) = 0$?

$$\begin{aligned}
E(I\varepsilon) &= E\big(I(Y - \alpha - \beta I)\big) \\
&= E(IY) - \alpha - \beta E(I^2) \\
&= E(IY) - \beta\sigma_I^2 \quad \text{(because of mean zero)} \\
&= E\big((E + C)(E - C + \nu)\big) - \beta\sigma_I^2 \\
&= E(E^2 - C^2) - \beta\sigma_I^2 \\
&= \sigma_E^2 - \sigma_C^2 - \beta\sigma_I^2 \\
&= \sigma_E^2 - \sigma_C^2 - 2\beta \\
&= -2\beta
\end{aligned}$$

According to this proof, income is only exogenous when the variance weights are equal because that's the only time that σ_E^2 and σ_C^2 cancel out and $\beta = 0$.

The proof is correct, in that the steps are correct and don't have errors, but I think it's been wrongly interpreted.

This result was interpreted by several as suggesting that exogeneity ($E(I\varepsilon) = 0$) only holds when the variances are equal. But note that if the variances weren't equal, then we'd have a different β So in other words, we're actually using the assumptions that $\sigma_E^2 = \sigma_C^2$ earlier in the proof to cancel terms out and arrive at the conclusion $E(I\varepsilon) = -2\beta$, and then only reevaluating that assumption when looking at that result. So let's write the more general problem, in which both E and C are mean zero and independent of one another, but have generic variances and

each has the coefficient β_E and β_C, respectively. Then:

$$\begin{aligned}
E(I\varepsilon) &= E\big(I(Y-\alpha-\beta I)\big) \\
&= E(IY) - \alpha - \beta E(I^2) \\
&= E(IY) - \beta\sigma_I^2 \quad \text{(because of mean zero)} \\
&= E\big((E+C)(\beta_E E + \beta_C C + \nu)\big) - \beta\sigma_I^2 \\
&= \beta_E E(E^2) + \beta_E E(EC) + \beta_C E(EC) + \beta_C E(C^2) + E\big((E+C)\nu\big) - \beta\sigma_I^2 \\
&= \beta_E\sigma_E^2 + \beta_C\sigma_C^2 - \beta\sigma_I^2 \\
&= \beta_E\sigma_E^2 + \beta_C\sigma_C^2 - \left(\frac{\sigma_C^2\beta_C + \sigma_E^2\beta_E}{\sigma_C^2+\sigma_E^2}\right)(\sigma_C^2+\sigma_E^2) \\
&= \beta_E\sigma_E^2 + \beta_C\sigma_C^2 - (\sigma_C^2\beta_C + \sigma_E^2\beta_E) \\
&= 0 \quad \forall \beta_E, \beta_C, \sigma_E^2, \sigma_C^2
\end{aligned}$$

So it's not a knife edge result at all, but it holds for any values. That's the sense in which the interpretation is wrong. Thus, I is exogenous.

But it's worth commenting on the fact that there was ever any question about this at all. After all, you simulated the data and know that E and C are independent of everything else. How was it plausible that their sum wasn't, since $E\big((E+C)\varepsilon\big) = E(E\varepsilon) + E(C\varepsilon) = 0+0 = 0$?

I think the answer is that the traditional model's definition of exogeneity is inherently confusing. X is exogenous if it is uncorrelated with the structural error term. WTF is that? It's "all other determinants of Y." But of course, $\varepsilon = Y - X\beta$, and so ε inherits its properties from β. And so whenever some misspecification leads to a "wrong" value of β,[1] then it's easy to think that this induces some different ε which might be correlated with X. You can find derivations of the measurement error formula we covered in class which say things like "Because of measurement error, $\tilde{x}$ becomes endogeneous because it becomes correlated with the error term." If that's how you want to define endogeneity then fine, but the important thing is that this can happen no matter what the data generating process is for x, including if x is randomly assigned (or any quasi-experimental identification strategy, no matter how well done it is). To me, calling that endogeneity is like being wedded to a strict definition that doesn't make so much sense, and be willing to throw out all of your useful and practical intuition (e.g., "a randomly assigned x is exogenous") to defend it.

Why do this when there's a simple alternative? That simple alternative is the potential outcomes framework. There, we say that x is exogenous if it is uncorrelated with the potential outcomes. Is that true in this case? Yes.

In this case, the potential outcomes of individual i are a function of the two-dimensional state space: earnings and capital gains. We can write this as: $Y_i(E = x, C = z)$ where x and z are constants. This is different from our examples in class in that the state spaces is two dimensional (which isn't unusual; many experiments have two cross-cutting treatments) and the state space is continuous instead of discrete (which also isn't a conceptual problem). We can say that Income I is exogenous if it is uncorrelated with potential outcomes. In this case, because we generated the data, we know what the true potential outcomes are. They are a

[1]Recall: In this example, the misspecification does not lead to the wrong value of β.

function of the individual-level error term realization e_i (the thing that we simulated in the DGP) and are given by:

$$Y_i(E = x, C = z) = x - z + e_i \quad \forall x, z \tag{6}$$

What does it mean for income to be independent of potential outcomes? It means that:

$$E\big[Y_i(E = x, C = z)|I_i\big] = E\big[x - z + e_i|I_i\big] = E\big[x - z + e_i\big] \quad \forall i, x, z \tag{7}$$

In this case, this is clearly true. We simulated e_i independently of E_i and C_i, and so e_i is independent of $I_i = E_i + C_i$. This tells use that income is exogenous: It is independent of potential outcomes. That means that knowing income about someone tells you nothing about what their outcomes would be under various different income levels. Of course, knowing income about someone tells you a lot about what their outcomes likely will be *under whichever was the income level that was actually realized*, but that's not what potential outcomes are about.[2] Potential outcomes are about causality, and causality is about what would happen under other states of the world, including ones that didn't happen. And in this example, knowing a person's income doesn't tell you anything about what their outcomes would be if their income were different, because "what someone's outcomes would be if their income were different" is only about e_i, and e_i is independent of I_i.

4 Some further examples with potential outcomes

Above, we said that income is independent of the potential outcomes, for all i, x, z. This is actually a pretty strong restriction (which we know to be true because we simulated the data). In other cases, we might only want to assume a weaker version of this.

For instance, you might have income be exogenous only for a subset of i. As an example, half of French tax audits are allocated by random assignment, and half are allocated based on inspectors' expectations of wrongdoing. For the tax filings i where an audit was chosen by random assignment, we can assume that the realized treatment (i.e., the decision to audit) was independent of the potential outcomes (what would have happened if they were audited, and what would have happened if they weren't audited). But for the tax filings i where the audit was chosen by investigators, we expect that the reason why these i were treated (i.e., were audited) was related to their potential outcomes. So we have exogeneity only for a subset of the full sample. We could represent this as:

$$\begin{aligned} &E\big[Y_i(A = 1)|A_i = 1\big] = E\big[Y_i(A = 1)|A_i = 0\big] \text{ and} \\ &E\big[Y_i(A = 0)|A_i = 1\big] = E\big[Y_i(A = 0)|A_i = 0\big] \text{ if } i \in \mathcal{A}_R \end{aligned}$$

where $A_i = 1$ denotes that tax filing i was audited, and $\mathcal{A}_R$ denotes the universe of tax filings

[2] To say "knowing income about someone tells you a lot about what their outcomes likely will be under whichever was the income level that was actually realized" means that for $I_1 \neq I_2$, then $E\big[Y_i(I = I_1)|I_i = I_1\big] \neq E\big[Y_i(I = I_2)|I_i = I_2\big]$. To say "knowing income about someone tells you nothing about what their outcomes would be under various different income levels" means that for any I_3 then $E\big[Y_i(I = I_3)|I_i = I_1\big] = E\big[Y_i(I = I_3)|I_i = I_2\big]$.

among which the random selection was done (i.e., includes treatment group and control group for those, but excludes filings selected for the targeted audit). So that's the relevance of $\forall i$.

What about the meaning of $\forall x, z$? Suppose you're interested in the effect of some level of market competition on firm behavior. Conditional on competition being high enough (the Hirfindahl being low enough), you think publicly listed and private firms would adopt the same pricing strategy, but if the environment wasn't competitive, then they would behave differently. Letting Y_i denote the pricing of firm i and h be the Hirfindahl index, this could be written as:

$$E\big[Y_i(h)|own = public\big] = E\big[Y_i(h)|own = private\big] \text{ if } h < \underline{\text{h}}$$
$$E\big[Y_i(h)|own = public\big] \neq E\big[Y_i(h)|own = private\big] \text{ if } h \geq \underline{\text{h}}$$

The first one says that for any Hirfindahl below $\underline{\text{h}}$, the expected behavior of public and private firms is the same, but for higher Hirfindahl it's not. Under this assumption, you could use privately owned firms as a control group for publicly owned firms to evaluate the effects of some policy that affects only publicly listed firms, but you could only do that when competition is high.

Note that this is different than a version in which potential outcomes are independent of treatment across all states of the world, but only for a subset of the observable levels of treatment. For instance, perhaps it is endogenous whether someone purchases a lottery ticket, but conditional on purchasing a ticket, then it is exogenous how much they won. Suppose you don't observe how many tickets they purchased, you only observe how much they won. Well then perhaps a good assumption is that it is endogenous whether they won zero or greater than zero, but conditional on winning more than zero, it's exogenous how much they won.

Letting w_i be the winnings of individual i, this could be written as:

$$E\big[Y_i(w)|w_i = w_1\big] = E\big[Y_i(w)|w_i = w_2\big] \quad \forall w_1, w_2 > 0 \text{ and } \forall w$$
$$E\big[Y_i(w)|w_i > 0\big] \neq E\big[Y_i(w)|w_i = 0\big]$$

The first one tells you that if you see any two people who have positive winnings ($w_1, w_2 > 0$), then those two people would have the same expected outcomes under the same winnings, no matter what those winnings were ($\forall w$), while the second one says that whenever one person won something and another won nothing, their expected outcomes would not necessarily have been the same even if they had won the same amount. This sort of thing happens a lot in applied work, and you usually hear someone say something like "There might be selection into treatment, but conditional on treatment, we assume that the level of exposure is exogenous" or "The extensive margin might be endogenous, but we assume that the intensive margin variation is exogenous."

Finally, for completeness, it's worth pointing out that potential outcomes can be used to express that treatment is only conditionally exogenous. This is common: In a randomized control trial, often treatment is only exogenous conditional on strata used in the assignment protocol to ensure balance across regions or things like that. So letting X_i be the conditioning

variable(s) and x be some specific value for that variable, then you might say:

$$
\begin{aligned}
&E\big[Y_i(T=1)|T_i=1, X_i=x\big] = E\big[Y_i(T=1)|T_i=0, X_i=x\big] \quad \forall x \\
&E\big[Y_i(T=0)|T_i=1, X_i=x\big] = E\big[Y_i(T=0)|T_i=0, X_i=x\big] \quad \forall x \\
&E\big[Y_i(T=1)|T_i=1\big] \neq E\big[Y_i(T=1)|T_i=0\big] \\
&E\big[Y_i(T=0)|T_i=1\big] \neq E\big[Y_i(T=0)|T_i=0\big]
\end{aligned}
$$

The first one says that the potential outcomes under treatment are the same for the actually treated i's and the actually control i's, conditional on X_i being the same. The second one says that the potential outcomes in the absence of treatment are the same for the actually treated i's and the actually control i's, conditional on X_i being the same. But the third and fourth ones say that if you don't condition on X_i being the same, then there can be systematic differences between the actually treatment i's and the actually control i's in terms of what would have happened under the same treatment state.

What's the key difference between defining exogeneity in the traditional model vs. the potential outcomes case? Well definining it in the traditional model requires you to specify some idealized regression equation. This mixes together definitions/assumptions and estimation, which inveitably creates confusion. Instead of exogeneity being defined solely by the relationships between variables, because the structural error term ε inherits some properties from β, then exogeneity partly depends on estimation. And it's hard to think through because ε doesn't have a positive, constructive definition, but it's just defined as a negative: "And everything else." In my opinion, this is strange and confusing! Instead, within potential outcomes, the definition of exogeneity really is about whether two variables (the X and the latent variable that is the potential outcome) are correlated.

With potential outcomes, not only is it more rigorous and clear, but as these examples show, it's also easy to be flexible.