ECON-UN 3412 - Econometrics Recitation 3

Matthew Alampay Davis
October 5, 2021

Hi all, thought I'd get in the habit of making one of these sheets every once in a while to expand on points from the previous week's lectures that I think are worth emphasizing or have some nuances that might be easy to overlook. These take some time to make and aren't actually part of the job so I can't commit to a weekly just yet.

Logistical notes

Will you have to learn Stata?

I've heard from a couple of students that it was mentioned in lecture that R users would have to learn Stata as well. I think this may be why the R recitation attendance is smaller than the Stata recitations this semester.

It's not really true that you'll have to learn Stata. Lectures will occasionally mention Stata commands or post Stata output and problem-set solutions will be in Stata. These aren't really big problems: you'll be learning to code through recitations, Stata output has the same information that R output does, and I'll be writing R solutions to any empirical problems in the problem sets and practice problems, which will be released at the same time in my recitation folder. This excludes PS1, which did not have an empirical component.

Recitation recordings

My plan this week is to do a trial run with recording my screen (where I'll have my iPad whiteboard and R demonstrations). So I'll record it if I can figure out a good setup in the classroom but won't live broadcast it. If the audio and video quality is all right, I'll post it to my recitation folder at a later date (not sure when yet) so it can be useful for later revision or for students who can't make it to recitation.

Will see how that goes this week before deciding how to manage future recordings (e.g. offer a hybrid option or post the recordings on delay). I know the professors want to encourage in-person attendance as much as possible (especially since this class is small) so it might depend on that as well.

Group chat for R users

Last year, my R students took the initiative of starting a WhatsApp group so they could communicate and help out one another on R-specific issues. This came about during the grad student strike so they had to navigate the problem sets without my help. If that sounds like a good idea, we (me and Tushar, the other TA who teaches R) would be happy to help facilitate it. It's a bit more difficult for students to self-coordinate this year since there are fewer students, the three sections have their own Ed sites (as opposed to one universal site), and there are now two TAs teaching R so if one of you starts a group, we'd be happy to share a link on all three Ed sites and/or send out an email to all three sections with a link. The TAs would not be a part of the chat since we can only reply to questions that all students can see) but we'll of course answer any R-specific questions we see on Ed. If we don't answer the question soon enough (usually one TA is assigned to answer questions per week), just send us an email as a nudge.

1 "Identifying variation"

Remember in our ideal experiment, to get an accurate estimate of the effect of an explanatory variable X_1 on an outcome variable Y, we want to create variation *only* in X_1 so that any change in Y that results can be attributed only to changes X_1. This is possible because we have all other potential factors have been "controlled for" (i.e. are held constant/fixed/unchanging).

In addition, **we generally want to have as much variation in our explanatory variables as possible** (see iPad notes for a demonstration why). We want our sample data to have observations from a wide range of values of X_1 to give us more "identifying variation." This observation makes our effect estimates more precise as measured by the variance of the estimator. This can be seen in an expression we saw in Lecture 5 when talking about the sampling distribution of $\hat{\beta}_1$ in the univariate regression:

$$\hat{\beta}_{\text{OLS}} \sim N\left(\beta, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right) \tag{1}$$

The denominator on the variance term here contains the square of the variance of our explanatory variable (multiplied by the sample size). When this increases, the variance of our estimator is smaller, making our estimate more precise in the same way that increasing our sample size n, also in the denominator intuitively also makes our estimates more precise. One way to visualize this is to think of a see-saw. If we wanted to see which of two people are heavier, we'd have them sit on opposite ends of the see-saw an equal distance from the middle. But the answer is immediately evident if we have them sit on the ends of the see-saw compared to having them sit towards where the pivot is and near each other.

2 Multiple linear regression model

Imagine we're interested in investigating how temperature T_i (in degrees Celsius) affects crop yields Y_i (in kilograms per hectare) for a sample of farms indexed i. We might run a simple regression like

$$Y_i = \alpha + \beta_1 T_i + u_i \tag{2}$$

Suppose we have the relevant data and our regression gives us an estimate $\hat{\beta}_1 = -2.5$ meaning that an increase in a farm's tem-

perature by 1 degree is associated with a decrease in crop yields of 2.5 units. The negative relationship makes some sense: if temperatures are too hot, crops are likely to die out.

But we might also think that temperature isn't the only relevant variable. It seems intuitive that precipitation P_i should also matter: the more rain crops get, the better fed they are. This leads us to propose a multiple linear regression model like

$$Y_i = \alpha + \beta_1 T_i + \beta_2 P_i + e_i \tag{3}$$

We can compare these two models to explore a couple of different topics we've touched on in lecture and explore a few implications.

2.1 Omitted variable bias

Let's be precise. What did we mean above when we said precipitation might also be "relevant"? Intuitively, we can imagine that precipitation also affects crop yields, our outcome variable. But so do any number of other factors: elevation, fertilizer, soil, etc. and we're generally happy to keep those unaccounted for in the error term u_i. Precipitation might be different because **we suspect it also has a relationship with our explanatory variable of interest**, the temperature: when it rains, we expect the air temperature to also cool down.

This presents a problem for the ideal experiment we mentioned at the beginning of 1. The change in Y that results from variation in X_1 (temperature) can no longer be attributed only to X_1 because we cannot assure that all other factors are kept the same. We haven't controlled for precipitation: it is not being kept constant/fixed/unchanging as temperature varies. In other words, we suspect omitted variable bias because we don't know how much of the change in Y (-2.5 kilograms per hectare for every degree Celsius) is due to the one degree change in T_i or changes in P_i that tend to accompany those changes in T_i. Most likely, both are having an effect but the model cannot disentangle them.

So if there is another variable X_2 that tends to change when X_1 changes (i.e. is correlated) *and* that has a plausible effect on Y, the variation we observed in X_1 will also tend to come with variation in X_2. That is to say, the estimator we get from the univariate regression 2 will not represent our ideal experiment because the changes in Y that are associated with changes in X_1 (or in that case T) could also be measuring effects from changes in X_2. We are not controlling for all relevant variables. Mathematically, since we do not have P_i as a regressor, it must exist in the error term u_i along with all the other determinants of Y_i. P_i is particularly concerning because of its correlation with T_i, which makes u_i also correlated with T_i. This violates one of our LSA assumptions and introduces omitted variable bias.

Recall this expression for omitted variable bias from Lecture 6:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_x}\right)\rho_{xu} \tag{4}$$

The term being added to β_1 on the RHS is a measure of omitted variable bias. First look at the term in parentheses. The bias can decrease in magnitude in two ways: by decreasing the standard deviation of the error term—which would mean trying to include as many determinants of Y in our regression as possible—or by increasing the standard deviation of the explanatory variable of interest, which relates to our discussion in Subsection 1 on identifying variation. Variance in your explanatory variable is good for estimation while variance in your error is bad.

The last term here, ρ_{xu} is what introduces OVB in the first place. So we've just said that variance in the error term is bad, but it's unavoidable: you can't possibly include every variable that contributes to crop yields. But it only causes omitted variable bias if it's correlated with your explanatory variable of interest, like precipitation is with temperature.

One last thing on OVB. We know the term in parentheses is always going to be positive since standard deviations are always positive. This means then that the sign of the bias will be the same as the sign of the correlation between the error term and our regressor of interest. In the temperature-precipitation case, we suspect that an increase in precipitation makes temperatures lower so their correlation should be negative. Thus, we would argue that the estimate of $\hat{\beta}_1$ (which we said was -2.5) that we get from the regression model 2 should be biased to be smaller than the true value of β_1.

This is easy to think about when we only consider one omitted variable and one regressor. If we have multiple suspected omitted variables, then it is much more difficult to try to figure out the direction of the omitted variable bias without data. Similarly, the relationship between temperature and precipitation may work in multiple channels: an increase in temperature may increase the moisture content of the air, which may make precipitation more likely or more severe. If so, this positive relationship works in the opposite direction of the more intuitive relationship between the variables and may complicate our idea for the sign of ρ_{xu}. Econometrics is hard sometimes!

2.2 Control variables

We know by now that going with regression model 3 is a way of unbiasing the OVB caused by having precipitation in the error term. In effect, precipitation's inclusion makes it a control variable and the resulting estimate $\hat{\beta}_1$ can be said to have "controlled for" precipitation. I want to focus a bit on intuition for what that means, what that costs, and when it fails.

Once again recall our discussion of identifying variation and its relationship to the ideal experiment. We want X_1 to take on a wide range of values holding all other determinants of Y fixed to more precisely estimate β_1. We now know we also want to include a variable that we suspect would cause significant bias if we did not include it, i.e. X_2 or P in our example. For it to cause OVB, it must be correlated with X_1, whether positively or negatively.

What does this mean for the identifying variation that we argued was needed for precise estimation? If there is indeed correlation between the variables, then by controlling for X_2, we have indeed unbiased our estimator but also reduced the identifying variation. We want X_1 to take on a large range of values controlling for all other variables. But if X_2 is very positively correlated with X_1 then we'll have very few observations where X_1 is high and X_2 is low or where X_1 is low and X_2 is high. IF they're very negatively corre-

lated, then we'll have very few observations where X_1 and X_2 are both high or both low. As a result, our estimator will be less precise (have larger standard errors) since X_2 will be harder to control for. The more covariates we add, the more identifying variation we sacrifice in order to control for the new covariates.

In the extreme, $\rho_{xu} = 1$ or -1 and we get multicollinearity. In that case, we have no identifying variation in X_1 since we can never control for X_2. We can never disentangle the effect X_1 is having on Y from the effect X_2 is having on Y and so the inclusion of X_2 as a control variable actually does not unbias our estimate. This is why R and Stata will automatically remove covariates that are perfectly collinear with another covariate. Even if $\rho_{xu} \neq 1$ or -1 exactly, correlations close enough to -1 or 1 can present a problem from insufficient identifying variation, usually through very large standard errors that make our estimators very imprecise.

In the other extreme, if X_1 and X_2 are independent, then $\rho_{xu} = 0$ (the converse is not necessarily true). In such a case, the bias in Expression 4 becomes zero and the estimates of β_1 are exactly the same in both models 2 and 3. And since none of the identifying variation in X_1 is needed to control for X_2, they will be equally efficient and have the same standard errors.

In practice, adding additional regressors to most models we look at in this course won't be particularly problematic for identification until we get to the machine learning topic in the second half of the course where consider models with very many regressors. Still, I wanted to make these points to show that adding control variables does not come free. We'll also see when we get to panel data that in some contexts, there are more efficient ways to remove bias than just adding a ton of control variables.

3 Standard errors

Finally, just a quick note here on heteroskedasticity-robust standard errors versus standard errors that assume homoskedasticity. In this course, we'll generally prefer to use heteroskedasticity-robust standard errors unless we have reason to do otherwise. The important point here is that assuming homoskedasticity is a strong assumption and imposes a restriction on our model. If that assumption is valid, then our standard errors will be smaller. Instead, we will tend to be conservative or agnostic by not imposing this restriction and use heteroskedasticity-robust standard errors which will always be larger and thus make rejection of the null hypothesis a bit less likely.