ECON-UN 3412 - Points of Emphasis

Matthew Alampay Davis
June 9, 2023

1 "Identifying variation"

Think back to our very simple single-variable regression model:

$$Y = \beta_0 + \beta X + u \tag{1}$$

If the regression model above is true, then any variation in Y we observe in our data must be explaiend by variation in X and/or variation in u, i.e., the sum of all other relevant contributors to Y. To produce an unbiased estimate of β, the expected change in Y attributable to a linear change in X, we at the least need data with enough variation in X that we can compare Y at different values of X. In addition, we need this variation in X to be *independent of variation in all other contributors to* Y, i.e., variation in u, even if we cannot measure it. This is what the OLS assumption $\mathbb{E}[u|X] = 0$ captures.

If the OLS assumption above holds, then these are completely separable sources of variation: X does not co-vary with u and regressing Y on X permits an unbiased measure of the association between them. In this ideal case, all the variation in X can be considered "identifying variation" because it can be used to "identify" β without bias. You can think of our econometrics course as covering the search for valid identifying variation under contexts where the OLS assumptions do not necessarily hold.

A takeaway here is that as much as possible, we generally want to have more variation in our explanatory variable(s) and ideally as little variation as possible in everything else we don't observe that explains Y. Mathematically, we want as much of the variation in Y in our data to come from variation in X rather than variation in u. This can be seen in the expression we saw in lecture for the sampling distribution of $\hat{\beta}$ under OLS assumptions:

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

Concentrating on the variance term here, we see that the numerator contains the variance of u and the denominator contains (the square of) the variance in X. This means that when the variance of our explanatory variable increases relative to the variance of the error, the variance of our estimator decreases and our estimate becomes more precise. Otherwise, we are increasingly vulnerable to the exact same estimation problem that we face when we have a small sample size n (which also appears in the denominator): our estimates are too imprecise to be useful.

As a sidenote, this also recalls the question in the first problem set which asked how a regression coefficient can be significant when the R^2 is low. While $\frac{\sigma_u^2}{(\sigma_X^2)^2}$ is not the formula for R^2, which is the proportion of variation explained by the covariates, they are obviously inversely related so that an increase in $\frac{\sigma_u^2}{(\sigma_X^2)^2}$ corresponds to a decrease in R^2. Thus we can see that a smaller R^2 implies larger standard errors and thus makes it more difficult to find evidene against the null hypothesis. But as long as there is some positive variance in X, meaning there are multiple values of X in the dataset, we can produce precise estimates as long as n is large enough to offset the low R^2. Indeed, this is very common in economics research where studies very rarely have R^2 greater than, say, 0.6: it is rarely the case that a major outcome of interest like a poverty rate can be dominantly explained by a small set of variables we have data on, but this does not prevent us from identifying important contributors.

2 Multiple linear regression model

If the OLS assumption above doesn't hold, then at least some of the variation in X co-moves with some of the variation in u in a way we do not observe. The same simple regression of Y on X would really be measuring the combined effect of X and some component(s) of u but attributing it only to X. This is our familiar omitted variable bias result. The total variation in X is no longer valid for identifying β. To unbias our estimate, we'd need to use only a component of the variation in X which does not co-vary with u. We've encountered one way to do this: control variables. By including some relevant omitted variable(s) W, our estimate of β comes not from comparing the Y in observations with low X to high X but from comparisons of observations with low X to high X with the same value of W. Our identifying variation comes from the component of the variation in X which does not vary in W. Problem solved so long as this identifying variation is great enough. We get in trouble if there is very little variation in X that does not also co-vary with W, i.e., when X and W are highly correlated. High enough correlation between X and W means that for a given level of W, there is very little variation in X. Going back to the equation above, the variation in X which isn't already explained by W that we're using for identification is so small that the variance in our estimator can prevent precise estimation unless we have large enough n to compensate for it. Otherwise, there is simply too much correlation between X and W that we cannot separate the influence of one from the influence of the other.

2.1 Omitted variable bias

Recall this expression for omitted variable bias from Lecture 6:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_x}\right)\rho_{xu} \tag{3}$$

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 in practice might mean getting

as much data on omitted variables as possible to use as controls—or by increasing the standard deviation of the explanatory variable of interest, which clearly relates to our discussion above on identifying variation.

The same problematic correlation mentioned above appears here as ρ_{xu}, the source of omitted variable bias 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 recitation, my omitted-variable example was using precipitation as a control in a regression estimating the effect of temperature on crop yields. We suspect that rainy days tend to be cooler days so the correlation between precipitation and temperature should be negative. Thus, we would argue that the estimate of $\hat{\beta}_{temperature}$ that we get from the regression of crop yields on temperature without including a precipitation control should be biased to be smaller than the 'true' value of $\beta_{temperature}$.

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 looking at 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.

2.2 Control variables

Re-iterating what was said above about control variables and identifying information using the example just introduced. We want our data to contain days where $Temperature$ takes on a wide range of values holding all other determinants of Y fixed. We know we also want to include a variable that we suspect would cause significant bias if we did not include it, i.e. $Precipitation$ in our example. For omitted variable to be present, $Temperature$ and $Precipitation$ must be correlated, whether positively or negatively. But if $Precipitation$ is very negatively correlated with $Temperature$ then we'll have very few observations in our data where $Temperature$ is high and $Precipitation$ is low or where $Temperature$ is low and $Precipitation$ is high. For a given precipitation level, the range of possible temperatures is much smaller than if we didn't control for precipitation and our identifying variation is smaller and our estimate of the effect of $Temperature$ on $Yields$ will be less precise. Even more, we might think that other variables like $Humidity$ and $Elevation$ are relevant. The more covariates we add, the more identifying variation we sacrifice to guard against omitted variable bias. If a setting is complicated enough, there may no satisfactory way to estimate the desired effect even with data on as many control variables as we want.

In the extreme, $\rho_{xu} = 1$ or -1 and we get multicollinearity. In that case, we have no identifying variation in X since we can never control for W. We can never disentangle the effect X is having on Y from the effect W is having on Y and so the inclusion of W as a control variable actually does not reduce any bias in our estimate. This is why R and Stata will automatically remove covariates that are perfectly collinear with another covariate. We either have to drop one of these covariates or include the covariates but drop the intercept which, since it is a constant, is collinear with a linear combination of the covariates.

In the other extreme, if X and W are independent, then $\rho_{xu} = 0$ (though note the converse is not necessarily true). In such a case, the bias becomes zero and the estimates of β are unbiased in the regression of Y on X regardless of whether W is included as a control.

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 Joint significance

Above, I noted that if two covariates have high enough correlation, then it becomes impossible to disentangle the two effects and we can not confidently estimate the effect of one covariate on the outcome variable. This can be related to the topic of joint significance tests: sometimes, we don't care whether a specific variable is relevant but whether *at least one* of several variables jointly have a significant effect, regardless of which it is.

A couple of exam-type questions to think about on your own: if two covariates are jointly significant at a given level, does this mean that at least one of them has a t-statistic that is also significant at that level? And conversely, if one covariate has a t-statistic that is individually significant, does that mean that the joint significance test of that covariate and another covariate must be jointly significant?

General problem set and exam advice

- Read and answer the entire question/subquestion *especially* if you're feeling the pressure of a timed exam. They often ask you for multiple things or to explain your answer and people often lose easy points by only providing the first thing they ask for.
- Interpreting coefficients
  - A subquestion that comes up in problem sets and past

exams a lot asks you to "interpret" a regression coefficient. If so, it is not enough to say it is significant or what its sign is. You should try to translate it into a claim along the lines of

> *"An increase in X by one [unit/percent/percentage point]* ***is associated with*** *an increase in Y of [$\hat{\beta}$]."*

– We never make definite claims like "X **causes** Y to increase" or "An increase in X by 1 **will** increase Y by [$\hat{\beta}$]" or "If there is zero X, then **Y will be** $[\hat{\beta}_0]$. We are producing estimates with uncertainty from a simple model built on strong assumptions and that's how we should talk about them