AMM Tutorial 6

Poisonous IV-y

Xinran Hu

August 28, 2024

1 Omitted Variable Bias, Re-explained

Keep in mind that this week's topic is actually a natural progression of what we discussed in the previous weeks - What we should do when the error term is correlated with our key X variable.

The reason why this is a problem goes back to the omitted variable bias. As we know, omitted variables are, well, unobserved. This means that when estimating the regression they end up becoming part of the error term.

Of course we shouldn't be expecting the unobserved terms being mean zero. This is not a big problem when the unobserved term is not correlated with any variable X's, as the mean of the omitted variable will be picked up by the intercept.

Situation becomes a bit different when the **omitted variable is correlated with X**. When this is the case, the mean of the omitted variable will be picked up by the coefficient(s) of the included variables instead. In this situation, the slope coefficient estimates become **a sum of the actual coefficient and the omitted variable bias**, making the estimates biased.

Now let's step back a bit from the intuitions. Given that the omitted variable is part of the error term, it being correlated with one of the regressors is equivalent to the **regressor being correlated with the error term**. This immediately violates assumption 4 for the OLS model, which requires the error term to be unrelated to any explanatory variables.

We kinda know already how to deal with omitted variables if we have panel data and the omitted term does not vary across time. But what about more general scenarios?

Well, that's what instrumental variables are for!

2 What Makes a Good IV?

Remember that the intuition for why we need an IV is that there's some factor that's unavailable for our data, which just so happen to be correlated with both X and Y. In a very sketchy manner, you could think of the IV as a "better" version of X that is not correlated with the unobserved characteristic and hence u (Which is the actual unobserved variable, say a, plus the "good" error term ε). Let's call this candidate Z. Then clearly we would like Z to have the following properties:

$$\begin{aligned} Cov(X,Z) \neq 0 \quad &\Leftrightarrow \quad Corr(X,Z) \neq 0\,; \\ Cov(u,Z) = 0 \quad &\Leftrightarrow \quad Corr(u,Z) = 0\,. \end{aligned}$$

The first property is called **relevance**, and the second property is called **exogeneity**. Relevance should be more self-explanatory - We want to "replace" X with Z so they need to be related for this replacement to be relevant to our goal. Keep in mind that if there are *extra control variables*, we want X and Z to be correlated *after controlling for all other variables.*

What about exogeneity? Well, first as we already know, Z is "better" than X because it's not correlated with u. However, this **DOES NOT mean that Z is not correlated with Y at all**. As a matter of fact, Z and Y are correlated, albeit in a very intricate manner.

For Z to be a good IV of X, it should be **correlated with Y only via X**. In other words, Z is correlated with Y because Y is correlated with X and X is correlated with Z. And to make sure that Z is only correlated with Y via X, Z *should not have any leftover correlation* with Y *in a model with* X as one of the explanatory variables.

Relevance can be tested as it only involves known variables. Exogeneity, however, can't really be tested and requires some creativity to convince the audience.

One last thing: We are dealing with math, not magic. Having an IV doesn't magically solve all the problems as the estimates we get will be less efficient than the OLS estimator. When the correlation between X and Z is **small**, the IV estimates might be **very imprecise** and we call such situation as having a **weak IV**. More on this in the next section as some math is needed to explain this clearly.

3 IV and 2SLS Estimators

When there is only one regressor X and we need an IV Z for it, the estimate using IV is pretty straightforward:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ Cov(Z,Y) &= Cov(Z, \beta_0 + \beta_1 X + u) \\ &= Cov(Z,\beta_0) + Cov(Z,\beta_1 X) + Cov(Z,u) \\ &= 0 + \beta_1 Cov(Z,X) + 0 \\ \Rightarrow \beta_1 &= \frac{Cov(Z,Y)}{Cov(Z,X)} \end{aligned}$$

where once we replace population statistics $Cov(Z,Y)$ and $Cov(Z,X)$ with their sample counterparts, we have an estimate:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}\,.$$

This can be extended to a regression with extra control variables, but β_1 will not have as concise an expression.[1]

One quirk of IV is that you can actually use multiple variables to IV a "bad" regressor if you must. In this situation clearly the IV estimator is no longer applicable as it assumes that Z and X have the same dimensions. (Recall that "good" regressors will serve as their own IV to make the calculation feasible.)

This is where 2SLS comes in. The steps of 2SLS are actually surprisingly straightforward: Imagine the "bad" regressor is still X, but we have two instruments, Z_1 and Z_2. Then

1. Both Z_1 and Z_2 need to satisfy relevance and exogeneity.
2. The regression in the OLS world would be
$$Y_i = \beta_0 + \beta_1 X_i + u_i \,.$$
3. We first estimate
$$X_i = \pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i} + v_i \,;$$
4. Generate $\hat{X}_i$ using the coefficients from the previous step, Z_1, and Z_2;
5. Then estimate
$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i + \beta_1 v_i \,;$$
6. The 2SLS estimate for β_1 would be the IV estimator when $\hat{X}_i$ is the instrument of X_i.

Any extra "good" regressors should be included in both step 3 and step 5.

When there are as many IVs as there are "bad" regressors, the 2SLS estimator will coincide with the IV estimator.

3.1 Weak IV

Now that we know
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})} \,,$$
I can finally explain in detail why the IV estimate might not be desirable when $Corr(X, Z)$ is small. Brace yourself as I'll engage in some math magic that breaks down pages 16 and 17 in lecture note 5:
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})} = \frac{\sum_{i=1}^n (Z_i - \bar{Z})(\beta_0 + \beta_1 X_i + u_i - (\beta_0 + \beta_1 \bar{X} + \bar{u}))}{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})}$$

[1] Well, actually you can still use the same expression if you embrace the matrix notation of regressions, where X will now be a matrix containing all observations for all regressors, and Z will be a matrix containing the instrumental variable and all other "good" regressors. The quotient will now be a vector of all slope coefficients. This is beyond the scope of this course, though, and mathematically identical to the system of equations on slide 13 of lecture note 6.

$$
\begin{aligned}
&= \frac{\sum_{i=1}^{n}(Z_i-\bar{Z})(\beta_1 X_i + u_i - \beta_1\bar{X} - \bar{u})}{\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})} = \frac{\sum_{i=1}^{n}(Z_i-\bar{Z})(\beta_1(X_i-\bar{X}) + u_i)}{\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})} \qquad \Leftarrow \bar{u} = 0 \\
&= \frac{\beta_1\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})}{\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})} + \frac{\sum_{i=1}^{n}(Z_i-\bar{Z})u_i}{\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(Z_i-\bar{Z})u_i}{\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})} \\
\operatorname{plim}\hat{\beta}_1 &= \beta_1 + \operatorname{plim}\frac{\sum_{i=1}^{n}(Z_i-\bar{Z})u_i}{\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})} \\
&= \beta_1 + \frac{\operatorname{plim}\sum_{i=1}^{n}(Z_i-\bar{Z})u_i}{\operatorname{plim}\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})} \\
&= \beta_1 + \frac{\operatorname{plim}\frac{\sum_{i=1}^{n}(Z_i-\bar{Z})u_i}{n}}{\operatorname{plim}\frac{\sum_{i=1}^{n}(Z_i-\bar{Z})(X_i-\bar{X})}{n}} \\
&= \beta_1 + \frac{Cov(Z,u)}{Cov(Z,X)}
\end{aligned}
$$

We already know that under the assumptions relevance and exogeneity, $\operatorname{plim}\hat{\beta}_1 = \beta_1$, i.e., when the sample size is **infinitely large** our estimate behaves like the true coefficient statistically.

There are two kinks in the problem:

1. Our sample size cannot be infinitely large;
2. $Cov(Z,u)$ might not be exactly zero.

Both implies that we cannot perfectly eradicate the bias, but as long as the IV estimator has a smaller bias than OLS, it should be preferred. And at least $Cov(Z,u)$ should be pretty small, so the ratio $\frac{Cov(Z,u)}{Cov(Z,X)}$ would be close to 0, right...? Unless when $Cov(Z,X)$ is also not big! In this scenario, the IV estimator might be worse than the OLS estimator, as in the example of tutorial 6 question 2.

Generally, we call IVs that have a small $Cov(Z,X)$ weak IVs, and try to avoid using them because it's more likely for such an IV to have a large bias (Given that we have little control over $Cov(Z,u)$)!

4 New Stata Commands

4.1 Estimating IVs in Stata

Estimating an IV model in stata requires a different command, `ivregress`. This command requires you to specify how the model would be estimated, but it does not distinguish between the vanilla IV estimator and 2SLS estimator (As IV estimator is a special case of 2SLS). Therefore, we will be using `ivregress 2sls` in this course.

Just as `reg` and `xtreg`, we start with specifying the Y and then key in every "good" control variable. Lastly, we need to include the "bad" regressor and its IV, using the format `(badvar=iv)`. You can ask Stata to report robust standard errors by adding `, robust`.

4.2 Generating Predicted Values

`ivregress 2sls` implicitly does steps 4 and 5 for us. But what if we are interested in the predicted values?

Well, immediately after running a regression (Just as when we want to access `_b[var]`), we can use the `predict new_variable` command to generate a variable called *new_variable* that holds $\hat{y}$ for each observation.

4.3 Correlation Between Variables

The `correlate` command will show you the correlation between two variables. If you want pairwise correlation between more than 2 variables, you'll need to use `pwcorr`, as when more than 2 variables are specified, `correlate` displays the variance-covariance matrix instead.