Two-stage Least Squares: Compiled QM questions: 2020-2021

Binta Zahra Diop

What is the interpretation of the 2SLS estimate

A few definitions:

- *The first stage*: the regression of X on Z
  $X_i = \pi_0 + \pi_1 Z_i + u_i$
- *The second stage/the structural equation*: the regression of Y on $\widehat{X}$
  $Y_i = \beta_0 + \beta_1 \widehat{X_i} + \epsilon_i$
- *The Reduced form*: the regression of Y on Z
  $Y_i = \gamma_0 + \theta_1 \pi_1 Z_i + \epsilon_i$ (you can write it this way, because we know that $\widehat{X_i} = \pi_1 Z_i$
  In the case of one instrument, $\beta_{2SLS} = \theta$ is the ratio of the reduced form coefficient of the first stage coefficient.

Looking at the reduced form, it is clear that what you are measuring with β_{2SLS} is the the impact of X on Y but *only* via the channel of the exogenous variable. Meaning that what you are measuring, is the impact of X on Y for those who are impacted by the instrument. In the case of the example in class 4, you are looking at the impact of having more children, for those who have had an unexpected increase in the number of children by having twins.

Measurement Error: How attenuation biased is caused (both mathematically and intuitively)

Think of the model: $y = \alpha + \beta x + \epsilon$.
You care about the true link between x and y but you only observe $\tilde{x} = x + u$. Where u is measurement error.
We assume that the measurement error in x is on average 0 and that it's uncorrelated with x and y. If you substitute the second equation into the first you get:

$$y = \alpha + \beta(\tilde{x} - u) + \epsilon = \alpha + \beta\tilde{x} + (\epsilon - \beta u)$$

Because of the correlation between x and the error term, you have endogeneity.
Looking at $\tilde{x} = x + u$ you see that u and x are positively correlated and so $\widehat{\beta}$ will be correlated negatively. You can rewrite:

$$\widehat{\beta} = \frac{cov(\tilde{x}, y)}{var(\tilde{x})} = \frac{cov(x + u, \alpha + \beta x + \epsilon)}{var(x + u)}$$

so from there you can see that:

$$plim\widehat{\beta} = \beta\frac{var(x)}{var(x) + var(u)}$$

And because $0 < \frac{var(x)}{var(x)+var(u)} < 1$, $\widehat{\beta}$ is biased towards 0.

Intuitively: Think about the example in the cheat sheet I uploaded here. If you want to know the impact of height on being good at basketball, but you can't tell for sure how tall somebody is. Even if you know that your errors are on average 0, you might not be able to tell who is 1m90 and who is 1m80. For that reason and if height has any impact on being good at basketball, you will be able to say "meh, it seems like on average people who are taller are better at basketball". Though some of the differences in ability to play the game is attributable to height, you won't be able to tell, because your data is not good at distinguishing people who are close in height.

Basically, because measurement error induces noise in what you can say, it decreases your ability to say the extent to which a relationship is true (this means bias downward).

Exactly how IV overcomes this measurement error

In the case above, any instrument z that is correlated with your true x, but not correlated with your measurement error u will help you identify the true coefficient:

$$\beta_{IV} = \frac{cov(y, z)}{var(\tilde{x}, z)} = \frac{cov(\alpha + \beta x + \epsilon, z)}{var(x + u, z)}$$

note that: $plim\widehat{\beta}_{IV} = \beta\frac{cov(x,z)}{cov(x,z)} = \beta$