Instrumental Variables Cheat Sheet

Binta Zahra Diop

Why use an Instrumental Variable

Reverse Causality

X causes Y but Y **also** causes X. There is a feedback loop.

- **Impact**: Your estimate will have a bias.
- **Example**: Think of education (X) and wealth (Y). The more educated household members are, the richer a household will become (X causes Y). However, being richer will lead to higher levels of education (Y causes E) of household members. So any causal link you will try to identify with an OLS will be biased by the feedback of the relationship – that is because you will be unable to identify the true causal link there. Primarily because there is no way to truly say that the βs you'll identify is a true result of X on Y and not biased by of Y on X.

Measurement Error: $X_{i,observed} = X_{i,true} + v_i$

X is not measured precisely, there is measurement error v_i that is added (v can be positive or negative) to the true value.

- **Impact**: Attenuation a bias – meaning that whatever β you are measuring will be biased towards 0.
- **Example**: Think of height and basketball. If you can't measure height well – even if this error has an expectation of 0, whatever association between height and ability to be good at basketball won't be as strong. That is because the perturbation of your height measure introduced by the measurement error will dampen your ability to say as much as you could have with the true height. (stems from not being able to distinguish between somebody who is 2m10 and somebody who is 1m90, admittedly large error) but you can tell that as people as getting taller, on average they are better at basketball.

Endogeneity

X is correlated with ϵ

- **Impact**: Omitted variable a bias.
- **Example**: Think of education and wealth again. Being in an urban areas with a lot of well-paid jobs that require high levels of education could lead people to going to school for longer, but could also lead people to be richer (better salaries). In that case, the β_{OLS} (identifying the impact of education on wealth) would be biased by the correlation of education and the characteristics in the error term (the characteristics of the labor market for example).

What are the necessary characteristics for an IV (Z)

Relevance: $corr(Z, X) > 0$

Z has to be correlated with X to be able to capture enough of the exogenous variations

- **How to test it**: t-test of the first stage, F-test of the first stage (rule of thumb 10)

Exclusion: $cov(Z, \epsilon) = 0$

$cov(Z, \epsilon) = 0$, meaning that the IV must be excluded from the second stage. This is what we mean by "exogenous" variable.

- **How to test it**: if you have more IVs than endogenous variables, the Hansen J test of over identification. Strictly speaking, one can never test for exogeneity.
  We can test for something close to it, which is making sure that the set of instruments used is not redundant. If it were redundant, this may be evidence that the instruments we had intended to use are correlated with the unobserved component (that's the test for over-identifying restrictions). see video here. If your model is just identified, you can only use theory and reasoning to get at this. A good example of that is in the discussion of rainfall as a instrument here or in this blogpost.