# 10: Instrumental Variables

## Exogeneity

* Most of the time nodes in DAGs will be endogenous with something coming into it.
* Even if we say it’s not, there’s often something else (perhaps an omitted/unknown variable?) that will be influencing it.
* Tricky because we often want to know what the exogenous variation looks like.

## Adjusting for Endogeneity

* RCTs are great because you get to throw away all the arrows going into a node (which essentially makes it exogenous).
* The influence on an input node will always be a mix of endogenous and exogenous factors.
* Could just stick the other variables in a regression model, but this assumes that the relationships are all linear. (Hence a better approach is matching or IPTW).
* The naïve linear model including just the input variable and outcome of interest is incorrect because it includes endogeneity.

## Unmeasurable Confounding

* Often you can’t directly measure or don’t have data on the important confounders.
* So can’t do matching or IPTW etc. (Can do RDD and DiD as these are techniques for handling unmeasured confounding).

## Instrumental Variables

* Would be great if we could somehow remove the endogenous part of the policy/program/input variable and include it in the error term. Just left with the exogenous part’s influence on the outcome variable.
* Using an instrument achieves this.
  + Very useful when you have unmeasurable/unmeasured backdoors.
* Exclusion criterion: doesn’t directly cause the outcome (or no plausible explanation for how it could do so), but causes the outcome through the original policy/program/input variable.
* Exogeneity: for example, in the education-earnings scenario, the instrument can’t be connected to ability. Otherwise it would just be another confounding variable.
* Instrument has to be totally separate from the rest of the DAG.

## General Terms

* Policy/program/intervention node is X.
* Outcome variable is Y.
* Unmeasured confounders are U.
* Instrument is Z.
* Relevance: Strong connection between Z and X.
  + Can be tested directly with statistics.
  + Run a model to check the strength of the relationship between X and Z.
* Excludability: Z can only get to Y via X. i.e. the direct correlation between Z and Y should be 0.
  + Can half-test with a model checking the strength of the relationship between Z and Y.
  + The idea that Z can only work through X is a convincing/plausible story/argument that can’t be directly tested.
* Exogeneity:
  + Can’t be done statistically at all (as there is no data on U).
  + Relies entirely on a convincing argument/story.

## Relevance

* Social security number is a common instrument but likely implausible in the example of education causing earnings.
* 3rd grade test scores could be a plausible instrument.
* Similarly for father’s education.

## Excludability

* No obvious connection between the instrument and the outcome variable.

## Exogeneity

* Proving or providing a plausible story about how a proposed instrument meets the excludability and exogeneity criteria can often be difficult as easy instruments tend to be related to other things we care about in the study.
* Basically, a valid instrument is going to have to be weird and unobvious/non-intuitive to satisfy the exclusion restriction.

## Examples

* Father’s education
  + Used to be popular but it doesn’t satisfy all the criteria for a valid instrument.
* Distance to college
  + The closer you live to a university, the more likely it is you will attend university.
  + Distance to college only causes better wages because you go to school more.
  + But this can break because towns/cities with universities generally have more start-ups/entrepreneurs etc due to the presence of business schools; or there’s a whole community of higher wage earners so those neighbourhoods will have higher wages and better opportunities in general.
  + So you’re probably going to be better off anyway if you live in a town with a university even if you don’t attend.
* Military draft
  + Selection based randomly on social security number.
  + But could also break because there is probably other reasons why being drafted would affect wages beyond education e.g. PTSD.

### Tobacco Taxes

* Using tobacco taxes to analyse the claim that smoking causes health issues.
* There are a whole other set of behaviours that cause you to smoke and negatively impact health.
* Increases in taxes will lead to a decrease in smoking due to the increased cost of cigarettes.
* No reasonable direct path between taxes and health issues.

### Policing

* Using election cycles to analyse the claim that patrol hours cause a decrease in the crime rate. Can’t directly estimate the number of criminals in a city.
* Using overcrowding lawsuits to analyse the claim that the incarceration rate affects crime.
  + The one thing instrumental variables let you do is deal with simultaneous causality i.e. chicken-and-egg problem.
  + Instrumental variables mean you’re looking at one direction of that causal relationship because they’re only going into one of the two nodes.

### Americanisation

* Are non-natives who assimilate into American society more successful in the labour market?
* Ability is the biggest unmeasured confounder (there are others).
* Used the scrabble score of the person’s name as the instrument. Works because the scores attached to each scrabble letter are derived from as the frequency of letters in a text corpus of 1920s/1930s English newspapers.
* Hence, more common names like John had a fairly low scrabble score and names with less common letters i.e. more foreign names would have higher scores.
* Found that immigrants with more American-sounding names had higher rates of success in the labour market.

### Weather

* Super-common instrument used is the weather because we have no control over it.
* For example, often weather is used as the instrument in studies of the effect of economic growth on conflict (i.e. civil war).
* As it turns out, rainfall isn’t weird enough to be valid in this case.
* In the entire list of previous examples, the only solid instrument was the scrabble score.
* The reasoning about weather is that sufficient rainfall allows for economic growth (droughts mean crops etc can’t be grown). Droughts then change the propensity for conflict and strife.
* Paper in 2020 fond 137 ways to get from weather to the outcome not through economic growth.

### Covid-19

* Theory is that covid causes school attendance to changes, which in turn affects grades or earnings.
* But this doesn’t quite work because covid-19 can cause changes to grades/earnings only through attendance. i.e. there’s no other way… which is absolute nonsense.

## Introduction to Using Instruments

* Using an instrument splits the intervention into its endogenous and exogenous components.
* Means something statistically and numerically.
* If we don’t adjust for anything, we find that education has a 13.1 unit increase in wages.
* If we adjust for ability (if we could measure it), we find that the true effect is around 7.8 units.
* Need to get the special version of education containing just the exogenous part.
* First need to satisfy ourselves that the instrumental variable requirements are met.

## Relevancy

* Goal here is to see if there’s a relationship between the input variable and the instrument.
* Can do so simply by running regression model.
* More formal approach is to use the F-statistic, which outputs the joint significance of all parts of the model.
  + Should be looking at a statistic greater than 104-ish.

## Exclusion

* Can do some statistics e.g. look at the relationship between the outcome variable and the instrument.

## Exogeneity

* Need to check that the assignment of the instrument value is random.

## Two-Stage Least Squares (2SLS) Regression

* Stage 1: predict the input variable using the instrument. This gives you the exogenous part of the input variable.
  + Any differences between the predicted and actual values are due to the unobserved confounders.
* Stage 2: use those predictions to estimate the outcome.
  + Can talk about the effect in causal terms as we removed the endogenous component in Stage 1.
  + Purely observational data… no treatment and control groups etc.

## Multiple Instruments

* No reason why you can’t use more than one instrument to explain the endogeneity in the input node.
* Each of the instruments still need to meet all of the assumptions.
* Both instruments included in the first stage regression.

## Other Control Variables

* Can also include other control variables in the regressions… but they need to be added to both stages.

## Faster, More Accurate Ways to Run 2SLS

* R can run both stages automatically and correct the standard errors (which affect the statistical significance of the coefficients.)
* ivreg package won’t give you an F-statistic, but does provide an equivalent weak instruments test. Insignificance indicates a weak instrument.
* estimatr package allows you to do fancy things with the standard errors (but same syntax as ivreg).