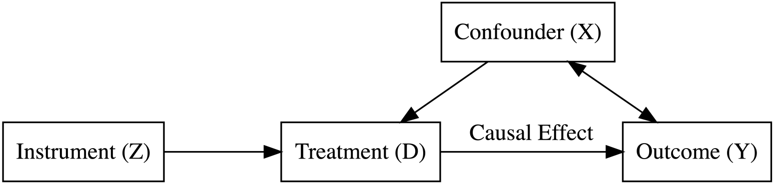
Instrumental Variables and 2SLS

Lecture 7 - Introduction to Causal Inference

Kevin Li

Setup

Issue: We want to find the effect of treatment D on outcome Y, but there is a confounder X. We have an extra variable Z:

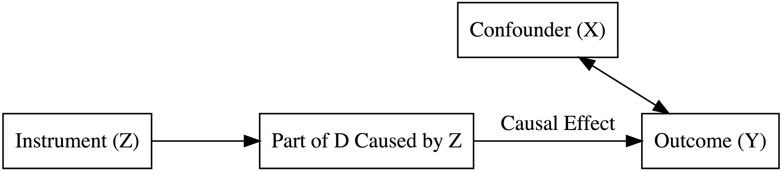


This **instrumental variable** Z has a few assumptions:

1. Z is correlated with the treatment D (**Relevance**).
2. Z is uncorrelated with any confounder X (**Exogeneity**).
3. Z has no direct effect on the outcome Y, only indirectly through D (**Exclusions Restriction**).
4. Z is exogenous to Y and D (also **Exogeneity**).

Inducing Exogeneity in Treatment

Instead of using our original treatment variable D, let us instead only use the **part of D caused by Z**. Let us call this $\hat{D}$.



The confounder X does not cause $\hat{D}$. Because $\hat{D}$ is the part of D caused by Z.

- Thus, X is no longer a confounder - since it doesn't affect selection in $\hat{D}$.

Since $\hat{D}$ is exogenous and there are no more confounders between $\hat{D}$ and Y, we can calculate the causal effect of $\hat{D}$ on Y.

Two-Stage Least Squares

The way to estimate the causal effect in IV is with 2-stage least squares (2SLS) estimator.

1st stage: Regress D on Z to estimate $\hat{D}$.

$$D_i = \delta + Z_i\beta + \varepsilon_i$$

2nd stage: Regression Y on $\hat{D}$ to estimate the causal effect:

$$Y_i = \alpha + \hat{D}_i\tau_{\textsf{LATE}} + u_i$$

The OLS estimate for τ_{LATE} is our causal estimate.

- Note: Standard errors will be incorrect if you manually run the two stages. Use software like **R** with **fixest** package.

Local Average Treatment Effect

The calculated treatment effect of $\hat{D}$ on Y is called the local average treatment effect (LATE).

- Substantively, it is the causal effect of D on Y for the part of D explained by Z
- This is also called the causal effect for **compliers**. Compliers are the units whose treatment D that "comply" (are influenced/caused) by the exogenous Z.

As noted before, this might not be equal to the total average treatment effect (ATE) between D and Y.

(LATE is sometimes caused the Average Causal Response if D is continuous).

Reduced Form of 2SLS

The reduced form regression is the regression of Y on Z:

$$Y_i = \gamma_0 + Z_i\gamma_1 + \epsilon_i$$

Assuming Z is exogenous (add controls if neccesary), that means the estimate γ_1 is the causal effect of Z on Y.

- This effect is also called the **intent-to-treat effect** τ_{ITT}.

If Z and D are both binary (quite common in causal inference), our LATE can also be estimated as:

$$\tau_{\text{LATE}} = \frac{\tau_{\text{ITT}}}{Pr(\text{compliers})} = \frac{\tau_{\text{ITT}}}{\widehat{Cov}(D_i, Z_i)}$$

- This is where the interpretation of LATE as the causal effect of compliers comes from.

Relevance and Weak Instruments

The relevance assumption is that Z must be correlated with D.

$$\boxed{\text{Instrument (Z)}} \xrightarrow{\text{Relevance}} \boxed{\text{Part of D Explained by Z}} \xrightarrow{\text{Causal Effect}} \boxed{\text{Outcome (Y)}}$$

We can test relevance by running a regression of D on Z (which is the 1st stage of 2SLS), and see if β is significant:

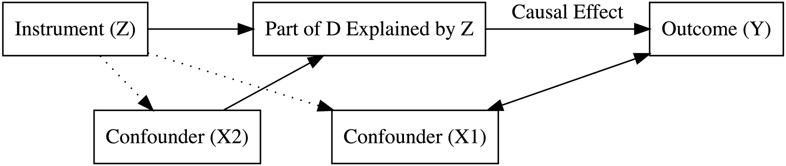
$$D_i = \delta + Z_i\beta + \varepsilon_i$$

If the correlation between Z and D is significant but weak, we have a weak instrument.

- If F test statistic in 1st stage is lower than 10, Z is weak.
- Weak instruments can have very biased τ_{LATE} estimates, especially in small samples.

Exogeneity Assumption

Z must be exogenous/randomly assigned in respect to both D and Y. The dotted lines below shows violations to exogeneity:



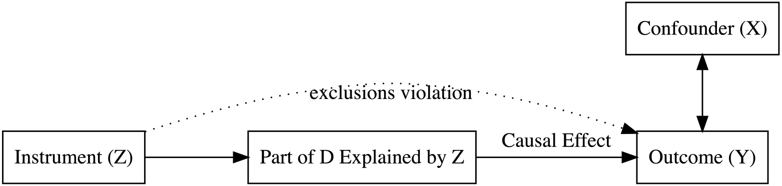
We can solve exogeniety violations by **controlling**/**accounting** for the confounders. We include confounders in both stages of 2SLS:

- 1st stage: $D_i = \delta_0 + Z_i\delta_1 + X_i\delta_2 + \varepsilon_i$
- 2nd stage: $Y_i = \alpha + \hat{D}_i\tau_{\text{LATE}} + X_i\beta + u_i$

Note: panel data - include fixed effects in both stages.

Exclusions Restriction

The exclusions restriction states Z must not have a direct effect on Y. It can only have an indirect effect through D.



Why? Well if Z has an independent effect on Y outside of D, then Z is a confounder between $\hat{D}$ and Y, and $\hat{D}$ will no longer be exogeneous.

There is no way to test the exclusions restriction. You can only justify it through your own understanding of the research topic in question.

Finding Valid Instruments

It is difficult finding an instrument that plausibly satisfies relevance, exogeneity, and exclusions.

- In the econometrics literature, a lot of attention is put on trying to find an instrument that doesn't violate exclusions.
- However, **Exogeneity** is actually probably the more difficult assumption to meet - it is hard to find a Z that is truly randomly assigned in terms of both D and Y.
- The most reliable way to find instruments is with a non-compliance or fuzzy regression discontinuity, which we cover in lectures 9 and 11.
- Other common instruments are often random by nature: Lotteries, rainfall, natural disasters, random selection of beneficiaries for policy pilots, etc.

Extension: Multiple Instruments

You do not need to stick with just one valid instrument. Including more instruments has a few advantages:

1. Having multiple valid instruments provides more variation in $\hat{D}$, allowing for more precise (and less variance) estimates.
2. Multiple instruments can also be a solution for weak-instruments.

First Stage (p number of instruments):

$$D_i = \delta_0 + \delta_1 Z_{1i} + \delta_2 Z_{2i} + \cdots + \delta_p Z_{pi} + \varepsilon_i$$

Second Stage: remains the same.

$$Y_i = \alpha + \hat{D}_i \tau_{\mathsf{LATE}} + u_i$$