Lectures 11-13

Instrumental Variables

Oscar Volpe

11/1/2021, 11/3/2021, & 11/10/2021

Three Steps of Causal Inference

Step 1: Write Down a Model

- Define the causal relationship of interest. This requires you, the researcher, to specify a counterfactual question ("What if...?"). No data needed here.
- Under your model, *causal effects* become target parameters.

Step 2: Identification

- Given your model, what can you learn about the target parameters using observed data? *Identification* maps the model and data to information about target parameters. Essentially, what can you recover from data?
- We say that a parameter is *identified* if, under the model assumptions, alternative values of the parameter imply different distributions of the data.

Step 3: Estimation

- In practice, we see finite samples drawn from the population distribution.
- How can we use these samples to estimate the target parameters?

Setup

Let $Y, U \in \mathbb{R}$ and $X \in \mathbb{R}^{k+1}$ with $X_0 = 1$. Consider the linear model:

$$Y = X'\beta + U$$

Suppose that the parameters β are given a causal interpretation.

- Generally, we can always normalize β_0 so that $E(U) = 0$.
- However, we cannot always assume that $E(XU) = 0$.

Definition (Exogenous, Endogenous)

Consider the linear model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U$. Then:

- X_j is *exogenous* if $E(X_j U) = 0$.
- X_j is *endogenous* if $E(X_j U) \neq 0$.

When X_j is *endogenous*, running OLS will not give us an estimate $\hat{\beta}_j$ that is consistent, unbiased, or efficient for β_j. Our BLP assumptions will fail!

Example 1: Omitted Variable Bias

Suppose that an organization implements a high-quality preschool program for children in under-resourced households. Define the variables:

- X_1: a dummy variable for participation in the program
- X_2: the child's socio-economic status
- Y: earnings in adulthood

Assume eligibility for the program is negatively correlated with X_2, but you do not observe X_2. Therefore, you can only estimate (2) among:

$$\begin{aligned} &(1) \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \\ &(2) \quad Y = \tilde{\beta}_0 + \beta_1 X_1 + \tilde{U} \end{aligned}$$

Assume X_1 is exogenous in (1), i.e. $E(X_1 U) = 0$. We cannot estimate β_1 by running OLS on (2). Why? X_1 is endogenous in (2), i.e. $E(X_1 \tilde{U}) \neq 0$!

- As long as $\text{Cov}(X_1, X_2) \neq 0$ and $\beta_2 \neq 0$, we have endogeneity bias.

Example 2: Measurement Error

Suppose Y is earnings, X is ability, and $\tilde{X}$ is some "proxy" used to measure ability, e.g. a test score. Suppose that the *true* model is:

$$Y = \beta_0 + \beta_1 X + U,$$

and here X is exogenous, i.e. $E(XU) = 0$. You only observe $\tilde{X}$, where:

$$\tilde{X} = X + V, \quad \text{where: } E(V) = E(VU) = E(XV) = 0$$

If we wanted to estimate β_1 from $Y = \beta_0 + \beta_1 \tilde{X} + \tilde{U}$, then we cannot use ordinary least squares. Why? Because $\tilde{X}$ is going to be endogenous:

$$E(\tilde{X}\tilde{U}) = E([X + V][U - \beta_1 V]) = \text{Var}(V)\beta_1$$

As the variance of V rises, the degree of endogeneity bias increases.

Example 3: Simultaneity

Consider a classic linear model of supply and demand:

$$Q^d = \beta_0^d + \beta_1^d P + U^d \quad \textit{(Demand)}$$
$$Q^s = \beta_0^s + \beta_1^s P + U^s \quad \textit{(Supply)}$$

In equilibrium, $Q_d = Q_s$, which means that the equilibrium price P^* equals:

$$P^* = \frac{(\beta_0^s - \beta_0^d) + (U^s - U^d)}{\beta_1^d - \beta_1^s}$$

Clearly, price is endogenous in both models of demand and supply.

- If we ran OLS, we could not estimate the elasticity of supply (β_1^s) or the elasticity of demand $(-\beta_1^d)$. Our BLP assumptions don't apply!
- These common issues lead us to use instrumental variables.

Introduction to IV

Consider a simple linear regression model $Y = \beta_0 + \beta_1 X + U$.

- We wish to interpret β_1 as the causal effect of X on Y.
- *Problem:* X is endogenous, i.e. $E(XU) \neq 0$. Cannot run OLS!

Suppose there exists an *instrument* $Z \in \mathbb{R}$ that satisfies:

(1) *Relevance:* $\text{Cov}(Z, X) \neq 0$

(2) *Validity:* $\text{Cov}(Z, U) = 0$

(3) *Exclusion:* Z only affects Y through X.

If we can find an instrument Z for X, then we can use it to solve for β_1.

$$\text{Cov}(Z, Y) = \beta_1 \text{Cov}(Z, X) \quad \Longrightarrow \beta_1 = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}$$

The IV Estimator

Given an *i.i.d.* sample $\{Y_i, X_i, Z_i\}_{i=1}^n$, we construct the IV estimator for β_1.

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

Importantly, the instrumental variables estimator is consistent: $\hat{\beta}_1^{IV} \xrightarrow{p} \beta_1$.

(i) Use the WLLN and CMT to show $\widehat{\text{Cov}(Z_i, Y_i)} \xrightarrow{p} \text{Cov}(Z_i, Y_i)$.

(ii) Use the WLLN and CMT to show $\widehat{\text{Cov}(Z_i, X_i)} \xrightarrow{p} \text{Cov}(Z_i, X_i)$.

(iii) Since $f(a, b) = a/b$ is continuous for all $b \neq 0$, the CMT implies:

$$\hat{\beta}_1^{IV} = \frac{\widehat{\text{Cov}(Z_i, Y_i)}}{\widehat{\text{Cov}(Z_i, X_i)}} \xrightarrow{p} \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, X_i)} = \beta_1$$

Note: we require that $\widehat{\text{Cov}(Z_i, X_i)}$ is nonzero, i.e. *relevance* in the sample.

Testing the IV Assumptions

Question. Can we test for instrument validity and exclusion?

- In general, these conditions are not testable. You must argue why Z is uncorrelated with the error term and why Z is not an omitted variable.
- *Thought Experiment:* Is Z plausibly exogenous in the model? Is there any realistic way Z can affect Y other than through its effect on X?

Question. Can we test for instrument relevance?

- Yes. Run an OLS regression of X on Z, i.e. $X = \gamma_0 + \gamma_1 Z + \varepsilon$. Then, test whether the coefficient γ_1 is significantly different from zero.
  - Compute $\hat{\gamma}_1 = \frac{\widehat{\text{Cov}(Z_i, X_i)}}{\widehat{\text{Var}(Z_i)}}$ and se($\hat{\gamma}_1$).
  - Test $H_0 : \gamma_1 = 0$ against $H_1 : \gamma_1 \neq 0$, e.g. using a t-test.

Strong vs. Weak Instruments

Suppose that $\text{Cov}(Z, X)$ is small, i.e. there is low instrument relevance. In this case, IV can perform poorly even for large samples. To see why, write:

$$\hat{\beta}_1^{IV} = \frac{\widehat{\text{Cov}(Z, Y)}}{\widehat{\text{Cov}(Z, X)}} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)U_i}{\widehat{\text{Cov}(Z, X)}}$$

By instrument exogeneity, $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)U_i \xrightarrow{p} 0$. However, if $\text{Cov}(Z, X)$ is close to zero, then there can be a great deal of bias even for large n.

- This dilemma is known as the *weak instruments* problem.
- If Z is not too relevant, then the IV can be worse than running OLS.

One solution to this problem might be to run an Anderson-Rubin Test. Note that this test statistic does not suffer from weak instrument issues.

Setup: Multiple Linear Regression

Suppose you have data about Y and explanatory variables $X_1, \ldots, X_k$. You decide to write down the following causal model relating Y to X.

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U \\ &= X'\beta + U, \end{aligned}$$

You suspect that some of your X_j's are *endogenous* in this model. Hence, if you were to run OLS, your estimator $\hat{\beta}_n^{\text{OLS}}$ would be inconsistent:

$$\hat{\beta}_n^{\text{OLS}} \xrightarrow{p} E(XX')E(XY) = \beta + E(XX')^{-1}E(XU) \neq \beta$$

IV Assumptions

Suppose that there exists a random vector $Z \in \mathbb{R}^{\ell+1}$ satisfying:

(1) *Validity:* $E(ZU) = \mathbf{0}$

(2) *Relevance/Rank:* $E(ZX') \in \mathbb{R}^{(\ell+1)\times(k+1)}$ has rank $k+1$.

(3) $E(ZZ')$ and $E(ZX')$ exist

(4) No perfect collinearity in Z (i.e. $E(ZZ')$ is invertible)

The components of Z are called *instrumental variables*. Note that any exogenous component of X (including $X_0 = 1$) should be included in Z.

Interpreting the Relevance/Rank Condition

- The *rank* of a matrix is the number of linearly independent columns.
- If $E(ZX')$ has rank $k+1$, then there must be at least as many valid, relevant instruments as there are endogenous regressors.
- A necessary condition for *(2)* is therefore that $\ell \geq k$ *(order condition)*.

Exactly Identified vs. Over-Identified

We say that β is *exactly identified* whenever $\ell = k$.

- In this context, # instruments equals # variables in the model
- Here, $E(ZX')$ is a square, full-rank matrix (i.e. invertible).
- When β is exactly identified, we can use the IV estimator!

We say that β is *over-identified* whenever $\ell > k$.

- In this context, # instruments exceeds # variables in the model
- Here, $E(ZX')$ is not a square matrix (*so:* not invertible).
- When β is over-identified, we can use the TSLS estimator!

Deriving the IV Estimator

Let $Y = X'\beta + U$, where $X \in \mathbb{R}^{k+1}$. Assume $Z \in \mathbb{R}^{\ell+1}$ satisfies *(1)-(4)*.

$$E(ZY) = E(ZX')\beta + E(ZU) = E(ZX')\beta$$

If $\ell = k$, then $E(ZX')$ is invertible. Solving for β, we obtain:

$$\beta = E(ZX')^{-1}E(ZY)$$

The IV estimator $\hat{\beta}_n^{IV}$ can be solved for via the sample analogue principle.

$$\frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - X_i'\hat{\beta}_n^{IV}) = 0 \quad \implies \hat{\beta}_n^{IV} = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i\right)$$

Note: we can use the WLLN and the CMT to prove that $\hat{\beta}_n^{IV} \xrightarrow{p} \beta$.

- In general, the IV estimator is always going to be biased.

Limiting Distribution of $\hat{\beta}_n^{IV}$

Assume that $\text{Var}(ZU)$ exists. Then we can prove that:

$$\sqrt{n}(\hat{\beta}_n^{IV} - \beta) \xrightarrow{d} N(0, \Omega), \quad \text{where } \Omega = E(ZX')^{-1}\text{Var}(ZU)E(ZX')^{-1}$$

Also, we can consistently estimate Ω with $\hat{\Omega}_n = \hat{A}_n \hat{B}_n \hat{A}_n$, where:

- $\hat{A}_n = \left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)^{-1}$, which $\xrightarrow{p} E(ZX')^{-1}$
- $\hat{B}_n = \left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'(\hat{U}_i)^2\right)^{-1}$, which $\xrightarrow{p} \text{Var}(ZU)$

Use this approximation $\hat{\Omega}_n$ to compute test statistics and confidence intervals, e.g. test whether effects are significant using the IV estimator.

Motivating Two-Stage Least Squares

What do we do when $\ell > k$? With more instruments than we need, the matrix $E(ZX')$ is not square. So, we can no longer invert it to solve for β.

- *Goal:* use Z in some "optimal" way to extract as much information about the endogenous X as possible (minimize $\text{Var}(\hat{\beta}_n^{IV} | \{Z_i, X_i\}_{i=1}^n)$.
- *Strategy:* run a least squares regression in two separate stages.
  - *First Stage:* predict X_j (endogenous variable) from Z (instruments)
  - *Second Stage:* regress Y on X using the predicted X_j's instead

Intuition: you are "extracting" the exogenous components of X_j that come from Z, while retaining as much information about X_j as possible. Then, regress Y on the fitted values of X_j, i.e. X_j predicted from (exogenous) Z.

How TSLS Works

Suppose $Y = X'\beta + U$, where X_j is endogenous in the model. You collect data about Z, which is a valid instrument. For TSLS, do the following:

First Stage

- Regress endogenous X_j on Z.
- Collect fitted values $\{\hat{X}_{ji}\}_{i=1}^{n}$ from this regression

Second Stage

- Regress Y on X, replacing X_j with $\hat{X}_j$.
- The coefficient estimates are the TSLS estimators

Important: your exogenous components of X must be included in Z. So you should put your controls in the first stage, as well as the second stage.

Deriving the TSLS Estimand

Define Π so that $\text{BLP}(X|Z) = \Pi' Z$. Thus, $\Pi = E(ZZ')^{-1}E(ZX')$. Write:

$$E(ZY) = E(ZX')\beta \quad \Longrightarrow \Pi' E(ZY) = \Pi' E(ZX')\beta$$

- *Note:* $\Pi' E(ZX') \in \mathbb{R}^{(k+1)\times(k+1)}$ is a square matrix with rank $k+1$.
- Hence, under our IV assumptions, $\Pi' E(ZX')$ will always be invertible.

When running TSLS, we are estimating the β, which equals:

$$\begin{aligned} \beta &= [\Pi' E(ZX')]^{-1}\Pi' E(ZY) \\ &= [\Pi' E(ZZ')\Pi]^{-1}\Pi' E(ZY) \end{aligned}$$

Notice that these two expressions for β are equivalent.

Deriving the TSLS Estimator

Our TSLS estimator has two equivalent representations. We write:

$$\hat{\beta}_n^{\text{TSLS}} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Pi}_n' Z_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Pi}_n' Z_i Y_i\right)$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Pi}_n' Z_i Z_i' \hat{\Pi}_n\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Pi}_n' Z_i Y_i\right),$$

where the estimator $\hat{\Pi}_n$ is equal to $\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i X_i'\right)$.

- *Interpretation:* regress X_i on Z_i to obtain $\hat{\Pi}_n' Z_i$, then regress Y_i on $\hat{\Pi}_n' Z_i$.
- Whenever $\ell = k$, the IV and TSLS estimators are the same.

Properties of $\hat{\beta}_n^{\text{TSLS}}$

Consistency

Just as before, the WLLN and CMT can be used to show $\hat{\beta}_n^{\text{TSLS}} \xrightarrow{p} \beta$.

- In general, the TSLS estimator is *not* unbiased.

Limiting Distribution

Assume $\text{Var}(ZU)$ exists. In this case, we can use the CLT to prove:

$$\sqrt{n}(\hat{\beta}_n^{IV} - \beta) \xrightarrow{d} N(0, \Omega),$$

where the variance is $\Omega = [\Pi' E(ZZ')\Pi]^{-1}\Pi' \text{Var}(ZU)\Pi[\Pi' E(ZZ')\Pi]^{-1}$.

- A natural estimator for Ω is $\hat{\Omega}_n = \hat{A}_n \hat{B}_n \hat{A}_n'$, where:
  - $\hat{A}_n = \left(\frac{1}{n}\sum_{i=1}^n \hat{\Pi}_n' Z_i Z_i' \hat{\Pi}_n\right)^{-1} \hat{\Pi}_n'$, which $\xrightarrow{p} [\Pi' E(ZZ')\Pi]^{-1}\Pi'$
  - $\hat{B}_n = \left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i' (\hat{U}_i)^2\right)^{-1}$, which $\xrightarrow{p} \text{Var}(ZU)$
- We use $\hat{\Omega}_n$ when computing test statistics and confidence intervals.