Econometrics Lecture 13

Instrument Variable

Hang Miao

Rutgers University

April 13, 2021

Overview

Introduction

Introduction

Threats Internal Validity

Three important threats to internal validity are:

- Omitted Variable Bias from a variable that is correlated with X but is unobserved (so cannot be included in the regression) and for which there are inadequate control variables;
- Simultaneous Causality Bias ($\boldsymbol{x}$ causes y, y causes $\boldsymbol{x}$);
- Errors-in-Variables Bias ($\boldsymbol{x}$ is measured with error)

Solution

- All three problems result in $\mathrm{E}[u|\boldsymbol{X}] \neq 0$.
- Instrumental Variables Regression can eliminate bias when $\mathrm{E}[u|\boldsymbol{X}] \neq 0$ by using an instrumental variable (IV), z.

Underlying Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + u$$

Suppose x is correlated with u: $\mathrm{Cov}[x, u] \neq 0$, then Strict Exogeneity Assumption 3: $\mathrm{E}[u|\boldsymbol{X}] \neq 0$ is violated. As a result, $\hat{\boldsymbol{\beta}}_{ols}$ is biased and inconsistent.

Endogeneity and Exogeneity

- An **Endogenous Variable** is one that is correlated with u
- An **Exogenous Variable** is one that is uncorrelated with u
- "Endogenous" literally means "determined within the system." If x is jointly determined with y, then a regression of y on x is subject to simultaneous causality bias. IV is first proposed to address such endougeneity problem. But this definition of endogeneity is too narrow because IV regression can be used to address OVB and Errors-in-Variable Bias. Thus we use the broader definition above.

Definition

Definition (Instrument Variable)

An valid Instrumental Variable, z, must satisfies the following two conditions:

- Instrument Relevance: $\mathrm{Cov}[z, x] \neq 0$
- Instrument Exogeneity: $\mathrm{Cov}[z, u] = 0$

Key Idea

Instrumental Variables (IV) regression breaks x into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate $\hat{\beta}_{ols}$.

Estimation

Approach 1: Two Stage Least Square (TSLS)

As it sounds, TSLS has two stages – two regressions:

First Stage

- Isolate the part of X that is uncorrelated with u by regressing x on z using OLS:

$$x_i = \pi_0 + \pi_1 z_i + v_i$$

- Compute the predicted values of $\hat{x}_i$:

$$\hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$$

- Since $\mathrm{Cov}[z_i, u_i] = 0$, $\mathrm{Cov}[\hat{x}_i, u_i] = 0$

Second Stage

Replace x_i by $\hat{x}_i$ and regress y on $\hat{x}_i$ using OLS:

$$y = \beta_0 + \beta_1 \hat{x} + u$$

Approach 2: Sample Covariance IV Estimator

The IV estimator could also be estimated by directly exploiting the information from IV exogeneity $\mathrm{Cov}[z, u] = 0$

$$
\begin{aligned}
\mathrm{Cov}\left(y_i, z_i\right) &= \mathrm{Cov}\left(\beta_0 + \beta_1 x_i + u_i,\ z_i\right) \\
&= \mathrm{Cov}\left(\beta_0, z_i\right) + \mathrm{Cov}\left(\beta_1 x_i, z_i\right) + \mathrm{Cov}\left(u_i, z_i\right) \\
&= 0 \quad + \mathrm{Cov}\left(\beta_1 x_i, z_i\right) + 0 \\
&= \beta_1 \mathrm{Cov}\left(x_i, z_i\right) \\
\beta_1 &= \frac{\mathrm{Cov}\left(y_i, z_i\right)}{\mathrm{Cov}\left(x_i, z_i\right)}
\end{aligned}
$$

The IV estimator replaces these population covariances with sample covariances:

The IV estimator replaces these population covariances with sample covariances:

$$\hat{\beta}_1 = \frac{s_{yz}}{s_{xz}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z})}$$

Approach 3: "Reduced Form"

The "reduced form" is more of providing another perspective to interpret the IV coefficient rather an estimation technique.

Reduced Form

$$x_i = \pi_0 + \pi_1 z_i + v_i$$
$$y_i = \gamma_0 + \gamma_1 z_i + w_i$$

Intuition

- A unit change in z_i results in a change in x_i of π_1 and a change in y_i of γ_1.
- Because that change in x_i arises from the exogenous change in z_i, that change in Xi is exogenous.
- Thus an exogenous change in x_i of π_1 units is associated with a change in y_i of γ_1 units so the effect on Y of an exogenous change in X is $\beta_1 = \gamma_1/\pi_1$.

Unbiased and Consistency of the TSLS Estimator

From the Second Estimation Approach we have Sample Covariance IV Estimator as follow

$$\hat{\beta}_1^{TSLS} = \frac{s_{yz}}{s_{xz}}$$

Unbiasedness

From the derivation we have $\mathrm{Cov}\left(y_i, z_i\right) = \beta_1 \mathrm{Cov}\left(x_i, z_i\right)$

$$\mathrm{E}[\hat{\beta}_1^{TSLS}|\boldsymbol{X}] = \frac{\mathrm{E}[s_{yz}|\boldsymbol{X}]}{\mathrm{E}[s_{xz}|\boldsymbol{X}]} = \frac{\mathrm{Cov}\left(y_i, z_i\right)}{\mathrm{Cov}\left(x_i, z_i\right)} = \frac{\beta_1 \mathrm{Cov}\left(x_i, z_i\right)}{\mathrm{Cov}\left(x_i, z_i\right)} = \beta_1$$

Consistency

Since $s_{yz} \xrightarrow{p} \mathrm{Cov}\left(y_i, z_i\right)$ and $s_{xz} \xrightarrow{p} \mathrm{Cov}\left(x_i, z_i\right)$

$$\hat{\beta}_1^{TSLS} = \frac{s_{yz}}{s_{xz}} \xrightarrow{p} \frac{\mathrm{Cov}\left(y_i, z_i\right)}{\mathrm{Cov}\left(x_i, z_i\right)} = \frac{\beta_1 \mathrm{Cov}\left(x_i, z_i\right)}{\mathrm{Cov}\left(x_i, z_i\right)} = \beta_1$$

Examples

Example 1: Effect of Studying on Grades

Stinebrickner, Ralph and Stinebrickner, Todd R. (2008) The Causal Effect of Studying on Academic Performance,

Formulation

What is the effect on grades of studying for an additional hour per day?

$x = \pi_0 + \pi_1 z + v_{\mathrm{i}}$

$y = \gamma_0 + \gamma_1 z + w_{\mathrm{i}}$

$y = GPA$(4 point scale)

$x =$ time spent studying (hours per day)

$z = 1$ if roommate brought video game, $= 0$ otherwise

Estimation Result

$\hat{\pi}_1 = -.668$

$\hat{\gamma}_1 = -.241$

$$\hat{\beta}_1^{IV} = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{-.241}{-.668} = 0.360$$

Example 2: Supply and Demand for Butter

Philip, Wright and Sewall, Wright (1928) The tariff on Animal and Vegetable Oils , Appendix B

Formulation

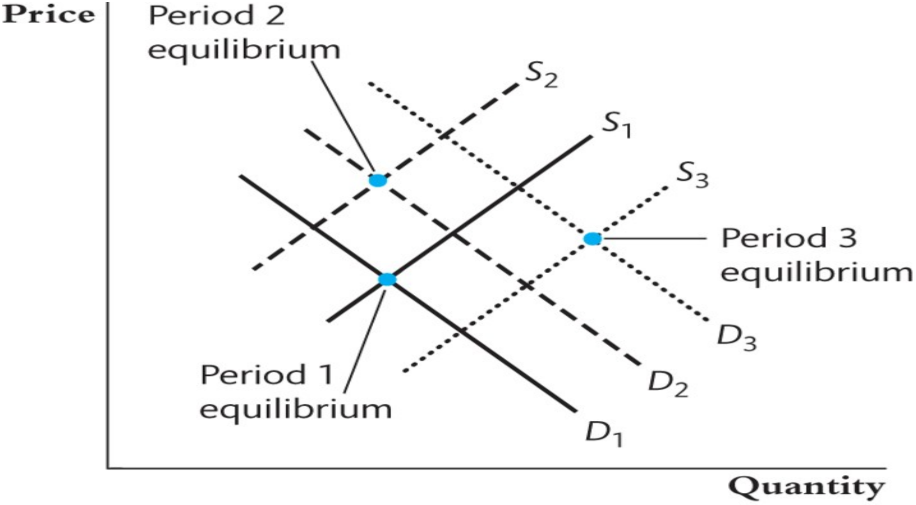
$$\ln\left(Q_i^{buttar}\right) = \beta_0 + \beta_1 \ln\left(P_i^{butter}\right) + u_i$$

Algebraic Derivation for Simultaneous Causality

$$\begin{array}{ll} \text{Demand:} & q_i = \beta_0 + \beta_1 p_i + u_i \\ \text{Supply:} & q_i = \alpha_0 + \alpha_1 p_i + v_i \end{array} \implies \begin{cases} p_i = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{v_i - u_i}{\beta_1 - \alpha_1} \\ q_i = \frac{\alpha_0\beta_1 - \alpha_1\beta_0}{\beta_1 - \alpha_1} + \frac{\beta_1 v_i - \alpha_1 u_i}{\beta_1 - \alpha_1} \end{cases}$$

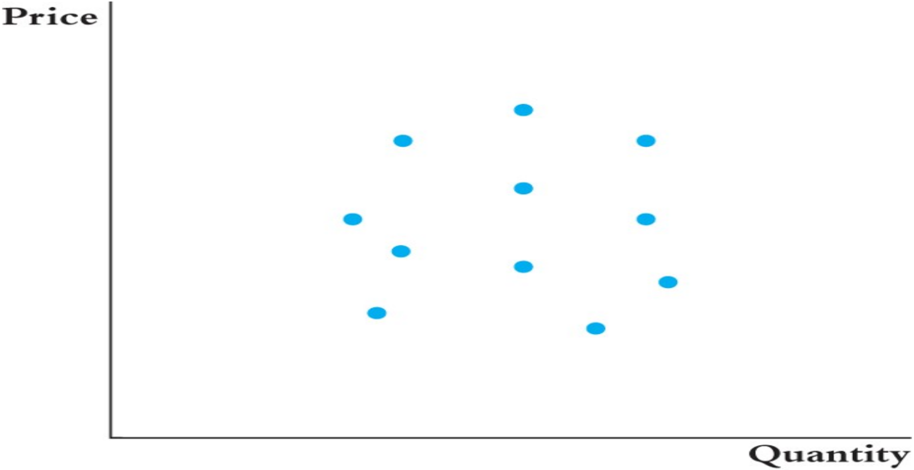
So we have $\mathrm{Cov}[p_i, u_i] = -\frac{Var[u_i]}{\beta_1 - \alpha_1} > 0$, price p_i is an endogeneous variable. Assumption 3 doesn't hold.

(q_i, p_i) is Jointed Determined by Demand and Supply



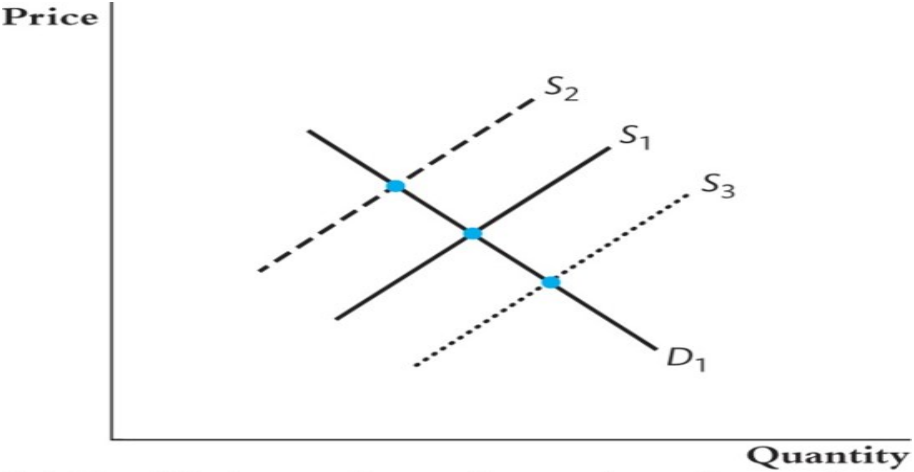
(a) Demand and supply in three time periods

(q_i, p_i) is Jointed Determined by Demand and Supply



(b) Equilibrium price and quantity for 11 time periods

Trace out Demand Curve by Isolating the Supply Shock



(c) Equilibrium price and quantity when only the supply curve shifts

Using Supply Shifter as an Instrument Variable to Estimate Demand Elasticity

Instrument Variable: Supply Shifter z

From the previous graph, we could trace out the quantity and price (p_i, q_i) tuple on demand curve by shifting the supply curve. Let z be a supply shifter like rainfall in dairy-producing regions. Then z satisfy the two requirement for instrument variable:

- Relevance $\mathrm{Cov}[z_i, p_i] \neq 0$: insufficient rainfall means less grazing means less butter means higher prices
- Exogenous $\mathrm{Cov}[z_i, u_i] = 0$: whether it rains in dairy-producing regions should not affect demand side uncertainty u_i (which normally depends on the taste, income etc)

'Reduced Form' Estimation Approach

Simultaneous Causality Bias Eliminated by Instrument Variable

Demand: $q_i = \beta_0 + \beta_1 p_i + u_i$

Supply: $q_i = \alpha_0 + \alpha_1 p_i + z_i + \eta_i$

where v_i is replace by $z_i + \eta_i$

$$\Longrightarrow \begin{cases} p_i = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{v_i - u_i}{\beta_1 - \alpha_1} & = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{1}{\beta_1 - \alpha_1} z_i + \frac{\eta_i - u_i}{\beta_1 - \alpha_1} \\ q_i = \frac{\alpha_0 \beta_1 - \alpha_1 \beta_0}{\beta_1 - \alpha_1} + \frac{\beta_1 v_i - \alpha_1 u_i}{\beta_1 - \alpha_1} & = \frac{\alpha_0 \beta_1 - \alpha_1 \beta_0}{\beta_1 - \alpha_1} + \frac{\beta_1}{\beta_1 - \alpha_1} z_i + \frac{\beta_1 \eta_i - \alpha_1 u_i}{\beta_1 - \alpha_1} \end{cases}$$

'Reduced Form' Estimation Approach

$$p_i = x_i = \pi_0 + \pi_1 z_i + v_i$$

$$q_i = y_i = \gamma_0 + \gamma_1 z_i + w_i$$

- Run OLS seperately on the above two equations. $\hat{\pi}_1$ is an estimation of $\frac{1}{\beta_1 - \alpha_1}$, $\hat{\gamma}_1$ is an estimation of $\frac{\beta_1}{\beta_1 - \alpha_1}$.
- If we devide $\hat{\gamma}_1$ over $\hat{\pi}_1$, we got $\hat{\beta}_1 = \hat{\gamma}_1 / \hat{\pi}_1$ which is the estimation β_1

TSLS Estimation Approach

Stage 1

- Regress $\ln(p_i)$ on instrument variable rain z_i to get the OLS estimator $\hat{\pi}_0$, $\hat{\pi}_1$.
- Estimate the predicted price level $\widehat{\ln(p_i)} = \hat{\pi}_0 + \hat{\pi}_1 z_i$
- In this stage, we isolate the price change arise only from the supply shifter z_i

Stage 2

- Regress $\ln(q_i)$ on estimated price $\widehat{\ln(p_i)}$ to get the OLS estimator $\hat{\beta}_0$, $\hat{\beta}_1$.

$$\ln(q_i) = \hat{\beta}_0 + \hat{\beta}_1 \widehat{\ln(p_i)} + u_i$$

- This step is the regression counterpart of using supply shifter to trace out the demand curve.

General IV Regression Model

General IV Regression Model

Underlying Model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + u$$

Instrument Variables are $z_1, z_2, \cdots, z_m$

- y is the dependent variable
- $x_1, \cdots, x_k$ are the **endogenous regressors** (potentially correlated with u)
- $w_1, , w_r$ are the included **exogenous regressors** (uncorrelated with ui) or **control variables** (included so that z_k is uncorrelated with u, once the w's are included)
- $z_1, , z_m$ are the m **instrumental variables** (the excluded exogenous variables)
- The coefficients are **overidentified** if $m > k$; **exactly identified** if $m = k$; and **underidentified** if $m < k$.

TSLS Estimation with a Single Endogenous Variable

Underlying Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 w_1 + \ldots + \beta_{1+r} w_r + u$$

Stage 1

- Regress x_1 on all the exogenous regressors: regress x_1 on $w_1, \cdots, w_r, z_1, \cdots, z_m$ and an intercept, by OLS
- Compute the predicted $\hat{x}_1$

Stage 2

- Regress y on $\hat{x}_1, w_1, \cdots, w_r$ and an intercept, by OLS
- The estimated coefficient from the second stage are the TSLS estimators

Example: Effect of Studying Time on Grades

Formulation

What is the effect on grades of studying for an additional hour per day?

$y = \beta_0 + \beta_1 x + u$

$y =$ GPA (4 point scale)

$x =$ time spent studying (hours per day)

$z = 1$ if roommate brought video game, $= 0$ otherwise

Roommates were randomly assigned

Possible Drawbacks

- Can you think of a reason that z might be correlated with u even though it is randomly assigned?
- What else enters the error term – what are other determinants of grades, beyond time spent studying?

Example: Effect of Studying Time on Grades

Why might z be correlated with u?

Heres a hypothetical possibility: the students sex.

- Although roommates are randomly assigned, normally men assigned with men and women assigned with women.
- Suppose: women get better grades than men, holding constant hour spent studying
- Suppose: men are more likely to bring a video game than women
- Then $\mathrm{Cov}[z, u] < 0$ (males are more likely to have a [male] roommate who brings a video game but males also tend to have lower grades, holding constant the amount of studying).

Solution

This is the IV version of OVB. We could use the extended general IV model by adding a sex binary control variable w to eliminate the bias: $y = \beta_0 + \beta_1 x + \gamma w + u$

Instrument Validity Issues

Instrument Validity Issues: Inadequate Relevance

Definition (Instrument Variable)

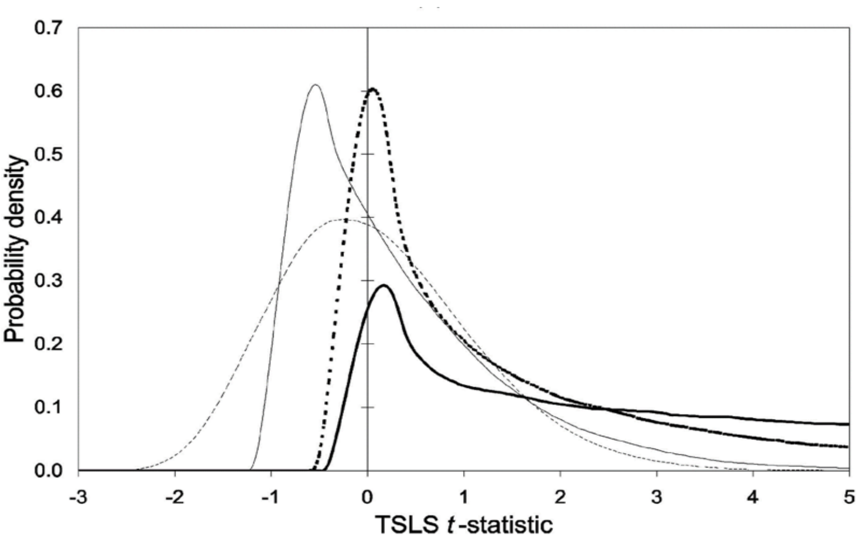
An valid Instrumental Variable, z, must satisfies the following two conditions:

- Instrument Relevance: $\mathrm{Cov}[z, x] \neq 0$
- Instrument Exogeneity: $\mathrm{Cov}[z, u] = 0$

Relevance

- If IV z weakly correlated with endogenous variable x i.e $\mathrm{Cov}[x, z]$ near zero, then we call it **weak instrument**
- We could check for weak instruments by the first-stage F test, where H_0 : all instrument coefficient is 0.
- If $F > 10$ we reject the null, indicating that there is strong correlation between x and z. Thus z is not weak instrument.
- If $F < 10$ we can't reject the null. There is possible weak instrument concern need to address

Consequence of Weak Instrument



Instrument Validity Issues: Inadequate Exogeneity

Exogeneity

- If IV z correlated with uncertainty u i.e $\mathrm{Cov}[u, z] \neq 0$, then we can not isolate the uncorrelated component $\hat{x}$ in the first stage of TSLS i.e $\mathrm{Cov}[u, \hat{x}] = \mathrm{Cov}[u, z] \neq 0$. Thus $\hat{\beta}_1$ is still biased and inconsistency
- If the coefficients are **overidentified**, i.e if there are more instruments than endogenous regressors, it is possible to test for instrument exogeneity by **J-test**

Algorithm for J-test

- Run TSLS and compute the estimated $\hat{y}_i$.
- Compute the resudual $\hat{u}_i = y_i - \hat{y}_i$.
- Regress $\hat{u}_i$ on all exogenous control variables and instrument variables $w_1, \cdots, w_r, z_1, \cdots, z_m$.
- Compute the F-statistic testing with Null Hypothesis that the coefficients on instrument $z_1, \cdots, z_m$ are all zero;
- The J-Statstics $= m \cdot$ F-statistic
- J-Statstics $\sim \chi^2(m-k)$, k is the number of endogenous variable, k is the number of instrument variable.