Endogeneity, Instrumental Variables

Tan Hong Ming

National University of Singapore

2019

Endogeneity

$$Y = \alpha + \beta X + \epsilon$$

Endogeneity

$$Y = \alpha + \beta X + \epsilon$$

1. X causes Y
2. ϵ causes Y
3. ϵ does not cause X
4. Y does not cause X
5. Nothing which cause ϵ also causes X

Endogeneity

$$Y = \alpha + \beta X + \epsilon$$

1. X causes Y
2. ϵ causes Y
3. ϵ does not cause X
4. Y does not cause X
5. Nothing which cause ϵ also causes X

If X is correlated with ϵ, X is said to be an endogenous explanatory variable.

Within model

Common Types of Endogeneity

Error term

1. - Omitted variables
    - $Y = \alpha + \beta X + \gamma Z + u = \alpha + \beta X + \epsilon$

If it is not observable, then

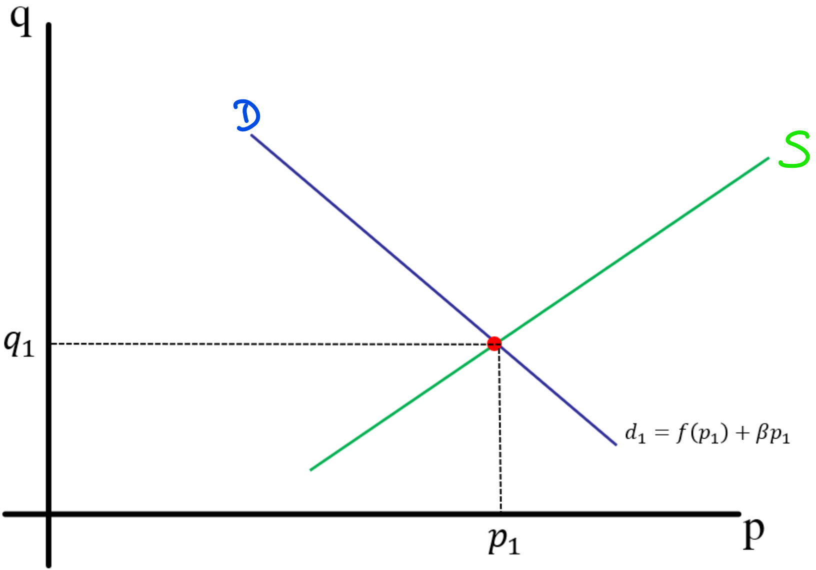
2. - Measurement error
    - we observe $X = X^* + u$
    - $Y = \alpha + \beta X^* + \nu = \alpha + \beta X - \beta u + \nu = \alpha + \beta X + \epsilon$
3. - Self-selection
    - Participation is not determined randomly
    - $E[\epsilon | participation]$

No longer part of randomised control trial

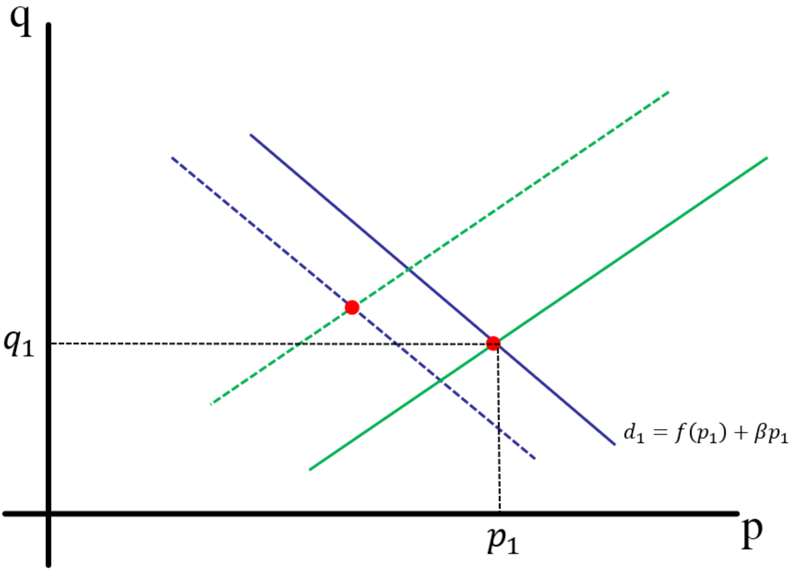
4. - Simultaneity
    - e.g. demand and supply

In project !!

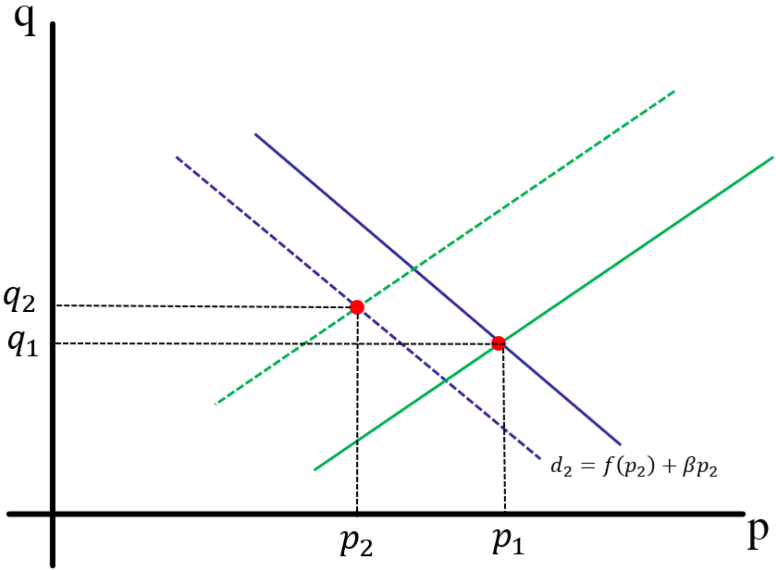
Simultaneity



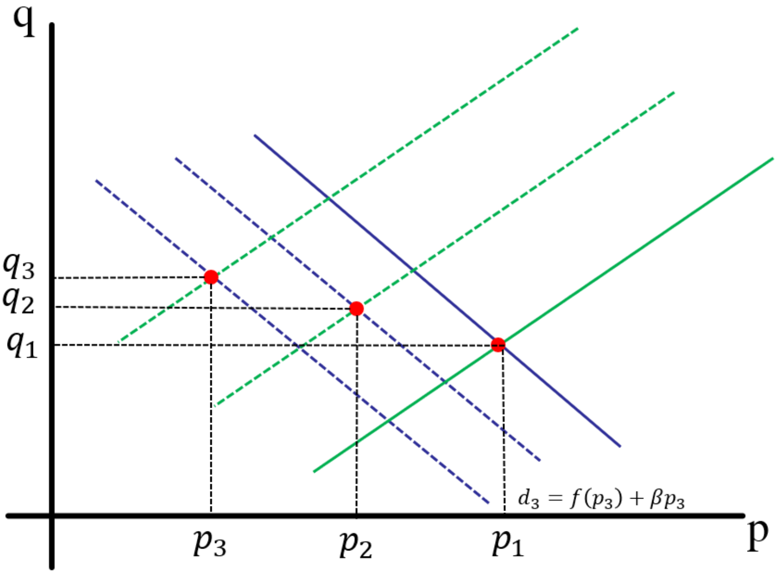
Simultaneity



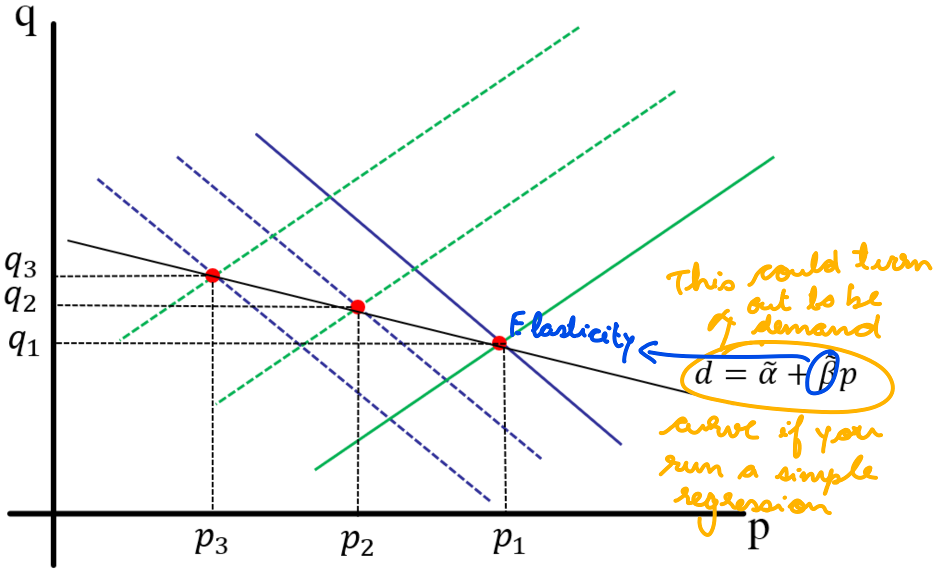
Simultaneity



Simultaneity



Simultaneity



Simultaneity

$$Q = \alpha + \beta P + \epsilon \qquad \text{(D)}$$
$$Q = \gamma + \delta P + \nu \qquad \text{(S)}$$

- Q and P are jointly determined
- Shock cause demand curve to shift
- Due to equilibrium, price P will also change
- Correlation between P and ϵ

Instrumental Variables

because correlatⁿ, $\rho = \frac{cov(X, \varepsilon)}{\sigma_X \sigma_\varepsilon}$

implies some correlation

$$Y = \alpha + \beta X + \epsilon, \text{ where } \boxed{Cov(X, \epsilon) \neq 0} \quad (1)$$

Suppose that we have an observable variable z that satisfies these two assumptions:

1. z is uncorrelated with ϵ: $Cov(z, \epsilon) = 0$ → z is exogenous
2. z is correlated with X: $Cov(z, X) \neq 0$

Then we call z an instrumental variable for X

Problem: ε is unobservable can't be tested mathematically.

- (1) is called the structural equation

$Cov(z, X) \neq 0$

- z is correlated with X can be tested:

$$X = \gamma + \delta z + \nu \tag{2}$$

Reduced form equat^{n}:

- If δ is significant, then we can be fairly confident that $Cov(z, X) \neq 0$
- (2) is an example of a reduced form equation
  - write an endogenous variable in terms of exogenous variables

Example: estimating the causal effect of skipping classes on final exam score

$$score = \beta_0 + \beta_1 skipped + u$$

- *skipped* might be correlated with other factors in u
  - more able, highly motivated students might miss fewer classes
- IV candidate: *distance* between living quarters and campus
  - Bad weather, oversleeping, etc can cause students to miss classes
    - ★ $skipped = \gamma_0 + \gamma_1 distance + \nu$
  - Is the sign of γ_1 important? Yes (check heuristics)
  - Is *distance* uncorrelated with u?

Instrumented Variable

↳ Could be. (cannot be IV then)

Multiple regression model

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u_1 \tag{3}$$

- y_i: endogenous, z_i: exogenous
- z_k: exogenous variable not in (3)

$$y_2 = \pi_0 + \pi_1 z_1 + \cdots + \pi_{k-1} z_{k-1} + \pi_k z_k + v_1 \tag{4}$$

Structural Equal^n

- we require $\pi_k \neq 0$

Two stage least squares

A simple example

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \qquad \text{(structural)}$$

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_1 \qquad \text{(reduced form)}$$

IV

- Think of reduced form as breaking y_2 into two parts:

  $\pi_0 + \pi_1 z_1 + \pi_2 z_2$ uncorrelated with u_1,

  v_1 correlated with u_1
- (1st stage) Estimate the reduced form by OLS and obtain the fitted values:

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2$$

- (2nd stage) Estimate the structural equation using OLS and the fitted values

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + u_1$$

Intuition of 2SLS

- $\hat{y}_2$ is the estimate of $y^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2$
- y^* is uncorrelated with u_1
- $\boxed{\text{2SLS}}$ first "purges" y_2 of its correlation with u_1 before doing OLS
- Recall $y_2 = y^* + v_1$

2-stage Least Squares

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \\
&= \beta_0 + \beta_1 y^* + \beta_2 z_1 + u_1 + \beta_1 v_1 \\
&= \beta_0 + \beta_1 y^* + \beta_2 z_1 + \epsilon
\end{aligned}
$$

Tests

Weak Instruments Low correlation between z and y

H_0 : the IV is weak (reject)

Hausman Test of endogeneity of y

H_0 : y is not endogenous (reject)

Sargan Only if you have more IVs than y

H_0 : all IVs are exogenous (do not reject)

OLS: $Y = \alpha + \beta y + \varepsilon$

2SLS: $y = \alpha + \hat{\beta}\hat{Y} + \varepsilon$

If endogeneity does not exist, then $\beta \approx \hat{\beta}$.

Simultaneous Equations

A simple example

Supply curve; because 'p' is raw material to supply.

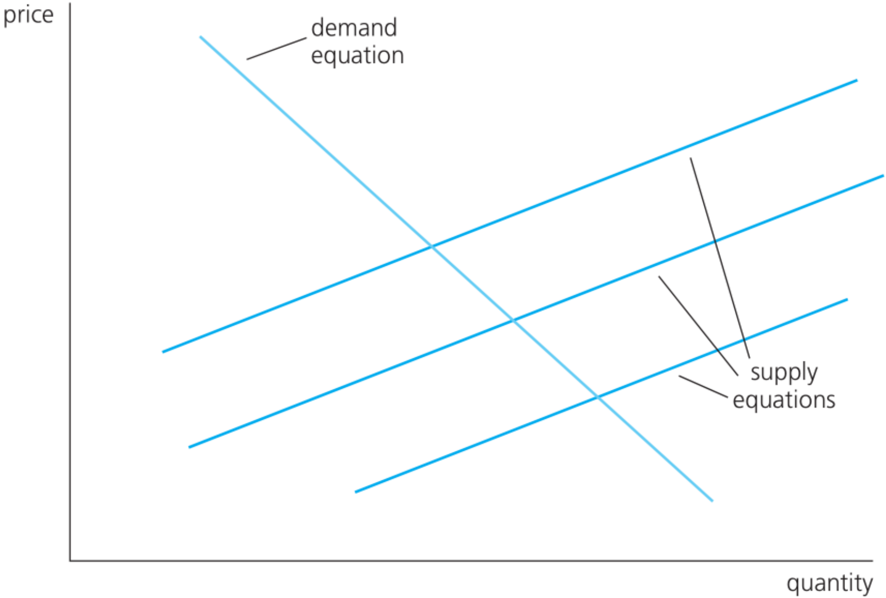
$$
\left\{\begin{aligned}
q &= \alpha_1 p + \beta_1 z_1 + u_1 \qquad (5)\\
q &= \alpha_2 p + u_2 \qquad (6)
\end{aligned}\right.
$$

Both are structural equat^{n}:

- q per capita milk consumption
- p average price per gallon
- z_1 price of cattle feed (exogenous to (5) and (6))

Which is the demand/supply curve? Which equation can be estimated? → Demand ($\because$ z_1 can be used as IV)

Intuition



In general

$$y_1 = \beta_{10} + \alpha_1 y_2 + \beta_{11} z_{11} + \beta_{12} z_{12} + \cdots + \beta_{1n} z_{1n} + u_1 \tag{7}$$
$$y_2 = \beta_{20} + \alpha_2 y_1 + \beta_{21} z_{21} + \beta_{22} z_{22} + \cdots + \beta_{2m} z_{2m} + u_2 \tag{8}$$

- z_i's can overlap in the two equations
- What if all z_i's are the same in both equations?
- What assumption do we need to identify both equations?

Can't do anything. Will not be able to distinguish supply/demand curve.