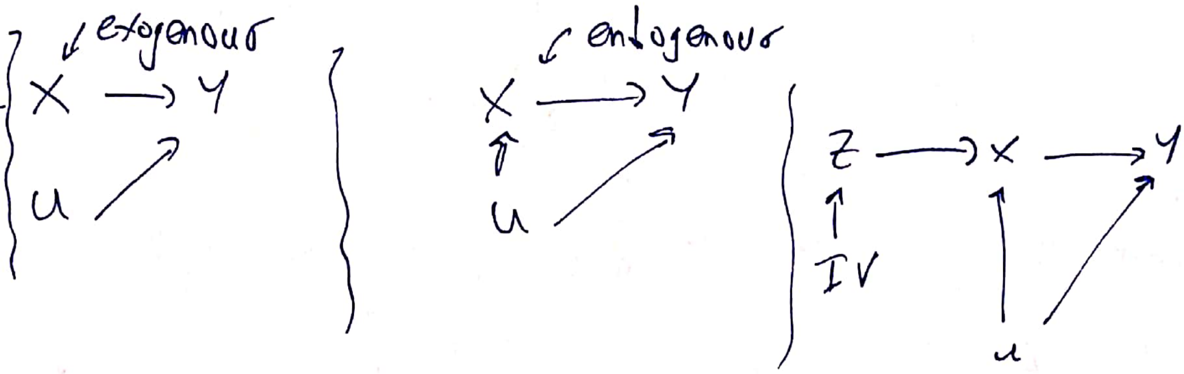
Instrumental Variables

$y = \alpha + bx + u$



Causes of endogeneity:

- measurement error
- omitted variable bias (OVB)

e.g. OVB

True model $y_i = a + b_1 x_i + b_2 z_i + u_i$ 1)

Our model $y_i = a + b_1 x_i + \varepsilon_i$ (missing z_i!)

In our model $Cov(x_i, \varepsilon_i) = Cov(x_i, b_2 z_i + u_i)$

$= Cov(x_i, b_2 z_i) + Cov(x_i, u_i)$

where $Cov(x_i, u_i) = 0$ (x_i is exogenous on 1))

but $Cov(x_i, z_i) \neq 0$ if x_i and z_i are correlated

so x_i is endogenous.

Side note

"Convergence in Probability"

$$\lim P(|X_n - X| > \varepsilon) = 0$$

Let $\hat{\theta}$ be an estimator for θ. Then

If ~~lim~~ $\hat{\theta}$ converges in probability to θ we call $\hat{\theta}$ a consistent estimator.

If $E[\hat{\theta}] = \theta$ we call $\hat{\theta}$ an unbiased estimator.

If all the OLS assumptions hold (including $Cov(x_i, u_i) = 0 \ \forall x_i$.) Then OLS is both consistent and unbiased.

An instrumental variable z_i should satisfy:

1) $Cov(z_i, x_i) \neq 0$ (the higher $Cov(x_i, z_i)$ is the "stronger" the instrument).

2) $Cov(z_i, u_i) = 0$

2 stage linear regression (2SLS)

Stage I.

Regress endogenous variables on instrument,

$x_i = Z_i \delta + \nu_i$

get $\hat{\delta} = (Z^T Z)^{-1} Z^T X$ $\xrightarrow{\text{get estimate for } X}$ $\hat{X} = Z\hat{\delta} = \underbrace{Z(Z^T Z)^{-1} Z^T}_{P_Z} X =$

$= P_Z X$

Stage 2.

Regress Y on estimated $\hat{X}$

$Y = \hat{X}\beta + \varepsilon_i = P_Z X \beta + \varepsilon_i$

$\hat{\beta} = ((P_Z X)^T P_Z X)^{-1} (P_Z X)^T Y$

⊛ $= (X^T P_Z X)^{-1} X^T P_Z Y$

⊛ Side note
you can show
$P_Z^2 = P_Z$
$P_Z^T = P_Z$

↑ substitute P_Z and compute...

So our $\hat{\beta} = (Z^T X)^{-1} Z^T Y = (Z^T X)^{-1} Z^T (X\beta + \varepsilon) =$

↗ substitute $Y = P_Z X \beta + \varepsilon_i$

$= (Z^T X)^{-1} Z^T X \beta + (Z^T X)^{-1} Z^T \varepsilon = \beta + \left(\frac{Z^T X}{n}\right)^{-1} \left(\frac{Z^T \varepsilon}{n}\right)$

from central limit theorem

last term $\frac{1}{n} Z^T \varepsilon = \frac{1}{n} \sum Z_i \varepsilon_i$ $\xrightarrow{n \to \infty}$ $E[Z_i \varepsilon_i] = E(Z_i)E(\varepsilon_i) = 0$

Since Z_i and ε_i are uncorrelated.

Summary

IV estimator is consistent, which intuitively means that as we get more data points our estimates converge in probability to the ~~true~~ true ones.

Disclaimer: I used naively some theorems for convergence etc. This is not a formal proof but gives us an idea.