Instrumental Variable Regression

2013/07/17

Tobias Brandt

Chair for Information Systems Research

UNI FREIBURG

Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1 \dots n$$

Gauss-Markov-Theorem:

OLS is Best Linear Unbiased Estimator (BLUE) **iff** error terms

- have zero mean: $E(\epsilon_i) = 0$
- are uncorrelated: $Cov(x, \epsilon) = 0$
- are homoscedastic: $Var(\epsilon_i) = \sigma^2 \quad \forall \quad i$

Bias

Omitted Variable Bias

- True Model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 r_i + \epsilon_i$ with $Cov(x, r) \neq 0$
- r_i is unobserved and $y_i = \widehat{\beta_0} + \widehat{\beta_1} x_i + \eta_i$ is estimated instead
- $\eta_i = \gamma r_i + \epsilon_i \quad \Rightarrow \quad \boldsymbol{Cov(x, \epsilon) \neq 0}$ (Endogeneity)

Simultaneity

- $y_i = \beta_1 x_i + \beta_2 r_i + \epsilon_i$ and $r_i = \gamma_1 x_i + \gamma_2 y_i + \eta_i$
- $E(r_i \epsilon_i) = \frac{\gamma_2}{1 - \beta_2 \gamma_2} E(\epsilon_i \epsilon_i) \neq 0 \quad \Rightarrow \quad \boldsymbol{Cov(r, \epsilon) \neq 0}$ (Endogeneity)

Instruments

Idea

- Assuming an endogeneity bias in $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$…

- … x_i can be “broken” into two parts – one correlated and one not using an instrumental variable z_i

- Instrument only detects movements in x_i that are uncorrelated with ϵ_i

Instruments

Validity

- A valid instrument must be
  - relevant: $Cov(z, x) \neq 0$
  - exogenous: $Cov(z, \epsilon) = 0$
- $Cov(z, \epsilon) = 0$ cannot be tested (bias) $\Rightarrow$ use common sense & economic theory
- $Cov(z, x) \neq 0$ can be tested by regressing $x_i = \gamma_0 + \gamma_1 z_i + \eta_i$
  First Stage Regression

Two Stage Least Squares (TSLS)

First Stage Regression

- Compute values $\hat{x}_i = \widehat{\gamma_0} + \widehat{\gamma_1} z_i$

Second Stage Regression

- Regress $y_i = \beta_0 + \beta_1 \hat{x}_i + \epsilon_i$
- IV estimator is not unbiased, but consistent (as sample size increases, bias disappears)
- Bias is larger for “weak” instruments (low explanatory power)

Example: Angrist & Krueger (1991)

Estimating return to education

- $\ln(wage) = \beta_0 + \beta_1 schooling + \epsilon$
- Problem: “Able” people generally perform better in school and earn more money, but inherent ability cannot be observed (OVB)

Instrument: Quarter of Birth

- Mandatory schooling laws (e.g. enter school in September of the year they turn six, may drop out after 16^{th} birthday)
- Q1 drops out after 9^{th} grade, Q4 after 10^{th} grade

Example: Angrist & Krueger (1991)

Problem: weak instrument

- F-statistics in 1^{st} stage in some specifications less than 2 (target is 10)
- Despite large sample (360,000), TSLS shows same result when quarters are randomly assigned

Strong instruments

- Better results for specifications with large F-stat
- Estimated return is 8% per year of schooling

Example: Acemoglu et al. (2001)

Estimating impact of institutions on GDP

- $GDP = \beta_0 + \beta_1 APAE + \gamma X + \epsilon$
  (APAE = Average Protection against Expropriation)

- Problem: Richer countries may prefer and can afford better institutions (Simultaneity)

Instrument: Colonial Settler Mortality

- settler mortality ⇒ settlements ⇒ early institutions ⇒ current institutions ⇒ current performance

Example: Acemoglu et al. (2001)

Results

- Strong instrument (25-35% R^2 in 1^{st} stage)
- Shows significant impact of mortality on institution and of institutions on GDP

Assessment

- Overall, strong support for the importance of institutions
- Still problematic due to possible measurements errors

Summary

- IV tries to replicate a natural experiment
- IV is not a silver bullet to solve all endogeneity problems
- IV can provide additional support for hypotheses
- IV fundamentally relies on the strong instruments