EC 508: Instrumental Variables

Liang Zhong

Boston University [1]

samzl@bu.edu

March 28, 2022

[1] Thanks for Richard, Dr.D and Prof.Paserman

1. Motivation
2. Sources of Endogeneity
3. What is an intrumental variable
4. Applications
5. Things to be aware of

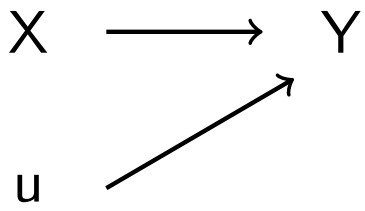
Motivation

- Need $E(\epsilon_i|x_i) = 0$ for unbiasedness of the OLS estimates. (called exogeneity condition)
  - This is restrictive and untestable.
- By large sample theory, $E(x_i\epsilon_i) = 0$ is enough for consistency. (called orthogonality condition)
  - Also needs i.i.d.-ness of observations and a finite nonsingular second-moment matrix of the regressors.
  - Still restrictive and untestable.
- Unfortunately, in many scenarios, even this is too much to ask for.

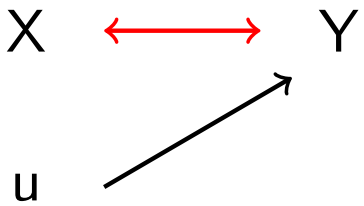
IV: Endogeneity

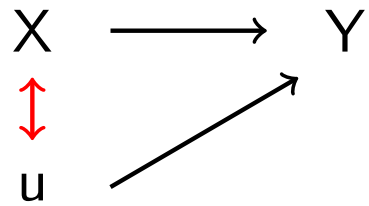
$$Y = \beta_0 + \beta_1 X + u$$

We can use OLS to obtain consistent estimate of the causal effect if:



We **cannot** use OLS to obtain consistent estimate for causal effect if:





Causes of Endogeneity

- Omitted Variable Bias
  - Taught in first half
- Simultaneous Equations
  - Taught by JJ
- Errors in Variable
- Dynamic Models with Autocorrelated Errors
- Sample selection bias

Errors in Variable

- When we measure variables for running a regression, we cannot always expect to measure them accurately.
  - The question is does this affect the consistency of OLS estimates?
- Suppose the true model is:

$$Y^* = \beta_0 + \beta_1 X + u$$

  - Suppose $E(uX) = 0$, but we do not have data on Y^*
  - All we have is Y, where $Y = Y^* + \epsilon$
- Now the true model of our regression is:

$$Y = \beta_0 + \beta_1 X + (u + \epsilon)$$

- As long as $E(\epsilon X) = 0$, the estimator is still consistent. $(E((\epsilon + u)X) = 0)$
- Consistency remains when the lhs variable is measured with error.

Errors in Variable

- Suppose the true model is:

$$Y = \beta_0 + \beta_1 X^* + u$$

  - Suppose $E(uX^*) = 0$, but we do not have data on X^*
  - All we have is X, where $X = X^* + \epsilon$, $E(\epsilon X^*) = 0$
- Now the true model of our regression is:

$$Y = \beta_0 + \beta_1 X + (u - \beta_1 * \epsilon)$$

- Let $\theta = u - \beta 1 * \epsilon$ as the error term,

$$E(\theta * X) = E(u - \beta_1 * \epsilon) * (X^* + \epsilon) = -\beta_1 * Var(\epsilon)$$

- It is nonzero when β_1 is nonzero. Hence, we have endogeneity and inconsistency of OLS estimates

Errors in Variable

- For the true model:

$$Y = \beta_0 + \beta_1 X + (u - \beta_1 * \epsilon)$$

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\text{Cov}(Y, X)}{\text{Var}(X)} \\
&= \frac{\text{Cov}(\beta_0 + \beta_1 X + (u - \beta_1 * \epsilon), X)}{\text{Var}(X)} \\
&= \beta_1 + \frac{\text{Cov}(u - \beta_1 * \epsilon, X^* + \epsilon)}{\text{Var}(X^* + \epsilon)} \\
&= \beta_1 - \beta_1 \frac{\text{Var}(\epsilon)}{\text{Var}(X^*) + \text{Var}(\epsilon)}
\end{aligned}$$

- If $\beta_1 > 0$, the (probability limit of the) coefficient still remains positive but moves closer towards 0.
- If $\beta_1 < 0$, it still remains negative and moves closer towards 0.
- This phenomenon is known as 'attenuation bias' for the errors in variable model.

Dynamic Models with Autocorrelated Errors

- Suppose, we wish to estimate the expectation-augmented Phillips curve equation:

$$\pi_t = \pi_t^e + X_t'\beta + u_t$$

  - π_t is inflation, π_t^e is expected inflation, X_t is a vector of other variables, β is a parameter vector and u_t is an error term.
- Assume that we have AR(1) errors, i.e. $u_t = \rho u_{t-1} + \epsilon_t$
- If one assumes an adaptive expectations framework, where $\pi_t^e = \pi_{t-1} + \alpha(\pi_{t-1} - \pi_{t-2})$
- Then one can rewrite the Phillips curve equation as:

$$y_t = \alpha y_{t-1} + X_t'\beta + u_t$$

  - $y_t = \pi_t - \pi_{t-1}$
- Now we are in trouble because the error term is correlated with y_{t-1} (why?) - resulting in an endogeneity problem
- Don't put lagged values of the lhs as a regressor if you suspect the error term is autocorrelated!!

Sample selection bias

- Example: the effect of veteran status on earnings
- Potential outcomes:

$$Y_{0i} = \beta_0 + \eta_i$$

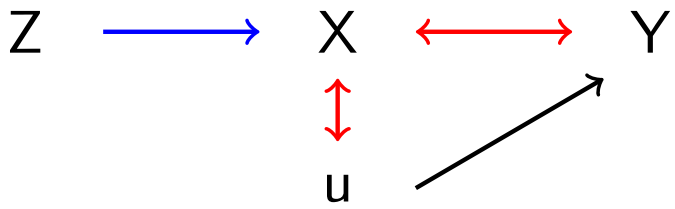
$$Y_{1i} = Y_{0i} + \delta$$

- Observed outcome

$$Y_i = \beta_0 + D_i\delta + \eta_i$$

  - $D_i = \{0, 1\}$, δ is the parameter of interest
- D_i not randomly assigned, so OLS is biased.
  - For example, people with low potential civilian earnings more likely to serve in the military

Instrumental Variables



Assumptions:

1. Exclusion restriction: Z is uncorrelated with u, $\text{Cov}(Z, u) = 0$
2. Relevance: Z is correlated with X, $\text{Cov}(X, Z) \neq 0$

Intuition: Because X is endogenous, β_1 cannot be interpreted as a causal effect. Instead, we look for an exogenous (i.e. determined outside of the model) variable, Z. We need this variable to be related to Y only through its association with X. This lets us tease out the causal effect of X on Y.

Exogeneity of Instrument Z

Structural equations:

$$Y = \beta_0 + \beta_1 X + u$$

$$X = \alpha_0 + \alpha_1 Z + v$$

Reduced form:

$$Y = \delta_0 + \delta_1 Z + \varepsilon$$

- It is a challenge to determine whether Z is exogeneous
- You may be tempted to regress $\widehat{u}$ on Z, but
- ...with only one instrument, Z, we cannot "test" for exogeneity

Why?

Structural equations:

$$Y = \beta_0 + \beta_1 X + u$$

$$X = \alpha_0 + \alpha_1 Z + v$$

Reduced form:

$$Y = \delta_0 + \delta_1 Z + \varepsilon$$

$$\begin{aligned}
\frac{\text{Cov}(u, Z)}{\text{Var}(Z)} &= \frac{\text{Cov}(Y - \beta_0 - \beta_1 X, Z)}{\text{Var}(Z)} \\
&= \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)} - \beta_1 \frac{\text{Cov}(X, Z)}{\text{Var}(Z)} \\
&= \delta_1 - \beta_1 \alpha_1
\end{aligned}$$

Note that reduced form coefficient, $\delta_1 = \beta_1 \alpha_1$ so $\frac{\text{Cov}(u,Z)}{\text{Var}(Z)} = 0$.

Single Regressor Case

Suppose that

$$Y = \beta_0 + \beta_1 X + u \tag{1}$$

where X is endogenous

Suppose Z is a "good" instrument for X and

$$X = \pi_0 + \pi_1^{FS} Z + v \tag{2}$$

Substituting (2) into (1)

$$\begin{aligned} Y &= \beta_0 + \beta_1[\pi_0 + \pi_1^{FS} Z + v] + u \\ &= \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1^{FS} Z + \beta_1 v + u \\ &= \delta_0 + \delta_1^{RF} Z + \epsilon \end{aligned}$$

Deriving IV Estimator

Note that

$$\delta_1^{RF} = \beta_1 \pi_1^{FS}$$

$$\beta_1^{IV} = \beta_1 = \frac{\delta_1^{RF}}{\pi_1^{FS}}$$

$$= \frac{\text{Cov}(Z, Y)/\text{Var}(Z)}{\text{Cov}(Z, X)/\text{Var}(Z)}$$

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

Where do assumptions enter into this formulation?

1. If Z is not exogenous, then δ_1^{RF} is inconsistent
2. If instrument relevance does not hold, then $\pi_1^{FS} = 0$

When either assumption fails, β_1^{IV} fails to capture the causal effect

TSLS

We can make this operational via TSLS

1. Regress X on Z
2. Collect predicted values $\widehat{X}$
3. Regress Y on $\widehat{X}$

The resulting estimator is consistent since sample covariances are consistent estimators of population covariances

Back to sample selection case

- Suppose researcher has access to instrument Z that is randomly assigned
- Example: draft lottery number (Angrist 1990)
  - Between 1970 and 1972, random sequence numbers were associated to each birth date in cohorts of 19-year olds.
  - Those with low lottery numbers were drafted into the military
  - Draft eligibility and veteran status not perfectly correlated. People with low numbers could still obtain deferments, people with high numbers could still volunteer
- When Z_i is binary, we can simplify the estimator to the Wald estimator. (purely technical)

$$\hat{\beta}_{wald} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}$$

Estimation results

Table 5
IV estimates of the effects of military service on white men[a]

Earnings year	Earnings		Veteran status		Wald estimate of veteran effect
	Mean (1)	Eligibility effect (2)	Mean (3)	Eligibility effect (4)	(5)
A. Men born 1950					
1981	16461	−435.8 (210.5)	0.267	0.159 (0.040)	−2741 (1324)
1970	2758	−233.8 (39.7)			−1470 (250)
1969	2299	−2.0 (34.5)			
B. Men born 1951					
1981	16049	−358.3 (203.6)	0.197	0.136 (0.043)	−2635 (1497)
1971	2947	−298.2 (41.7)			−2193 (307)
1970	2379	−44.8 (36.7)			
C. Men born 1953 (no one drafted)					
1981	14762	34.3 (199.0)	0.130	0.043 (0.037)	No first stage
1972	3989	−56.5 (54.8)			
1971	2803	2.1 (42.9)			

[a] Note: Adapted from Angrist (1990, Tables 2 and 3), and unpublished author tabulations. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 observations with earnings in each cohort.

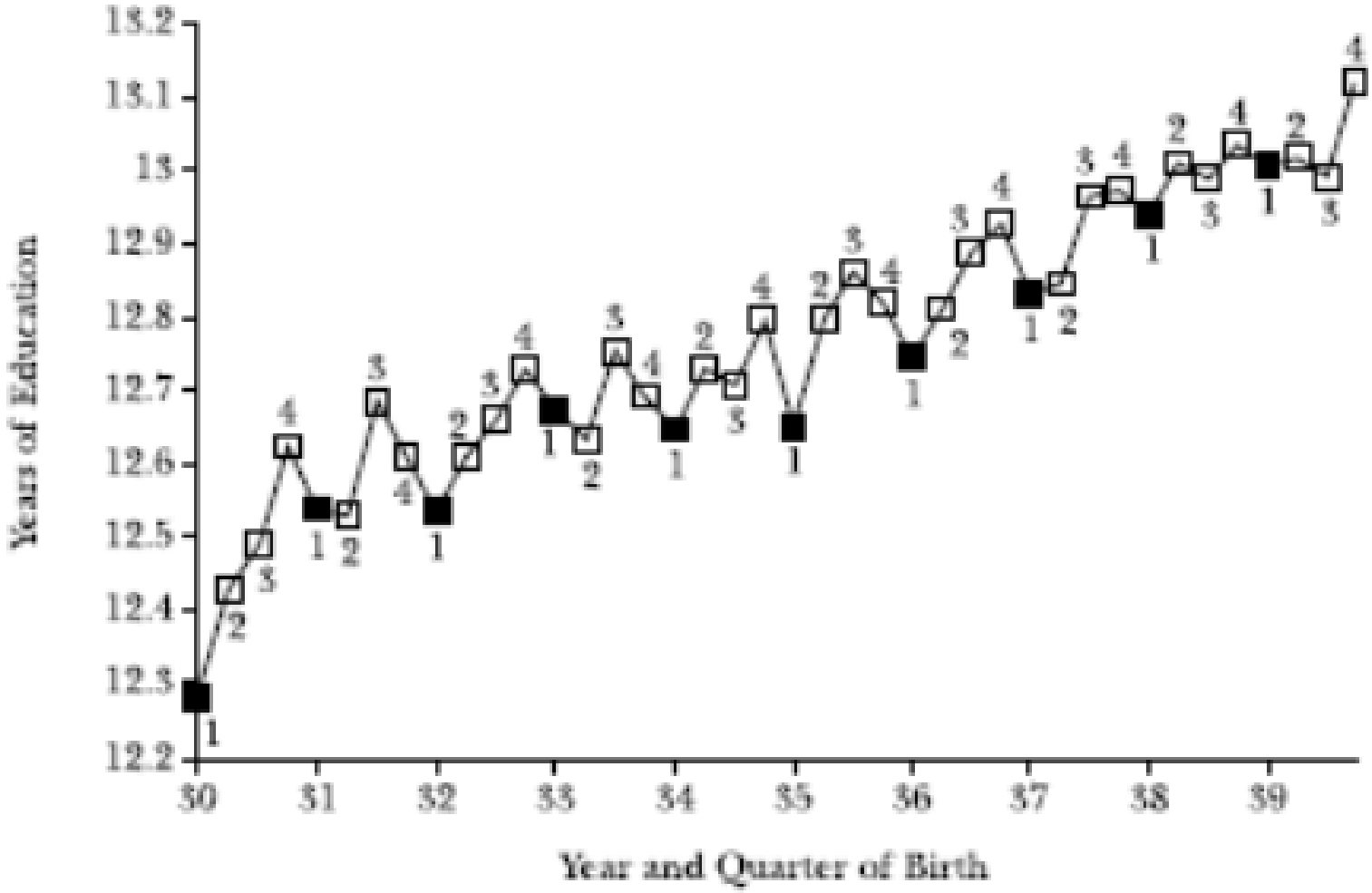
A well-known study

- Angrist and Krueger (1991) study the returns to education.
  - It is a classical endogeneity problem because the ability is an obvious omitted variable but can not observe.
- They use a novel instrument: the individual's quarter of birth.
  - Due to compulsory schooling laws, children enter school the **year** they turn six, and they are required to remain in school till their sixteenth birthday.
  - This allows children born in the first quarter of the year to 'get away' with less education than those born in the third and fourth quarter.
  - Consequently, the birth-quarter dummies are correlated with years of schooling, but clearly, they have nothing to do with the omitted ability variable.
- The following diagrams show that there indeed is a strong relationship between earnings, birth quarter, and educational achievement.

QoB and Education

Figure 1

Mean Years of Completed Education, by Quarter of Birth

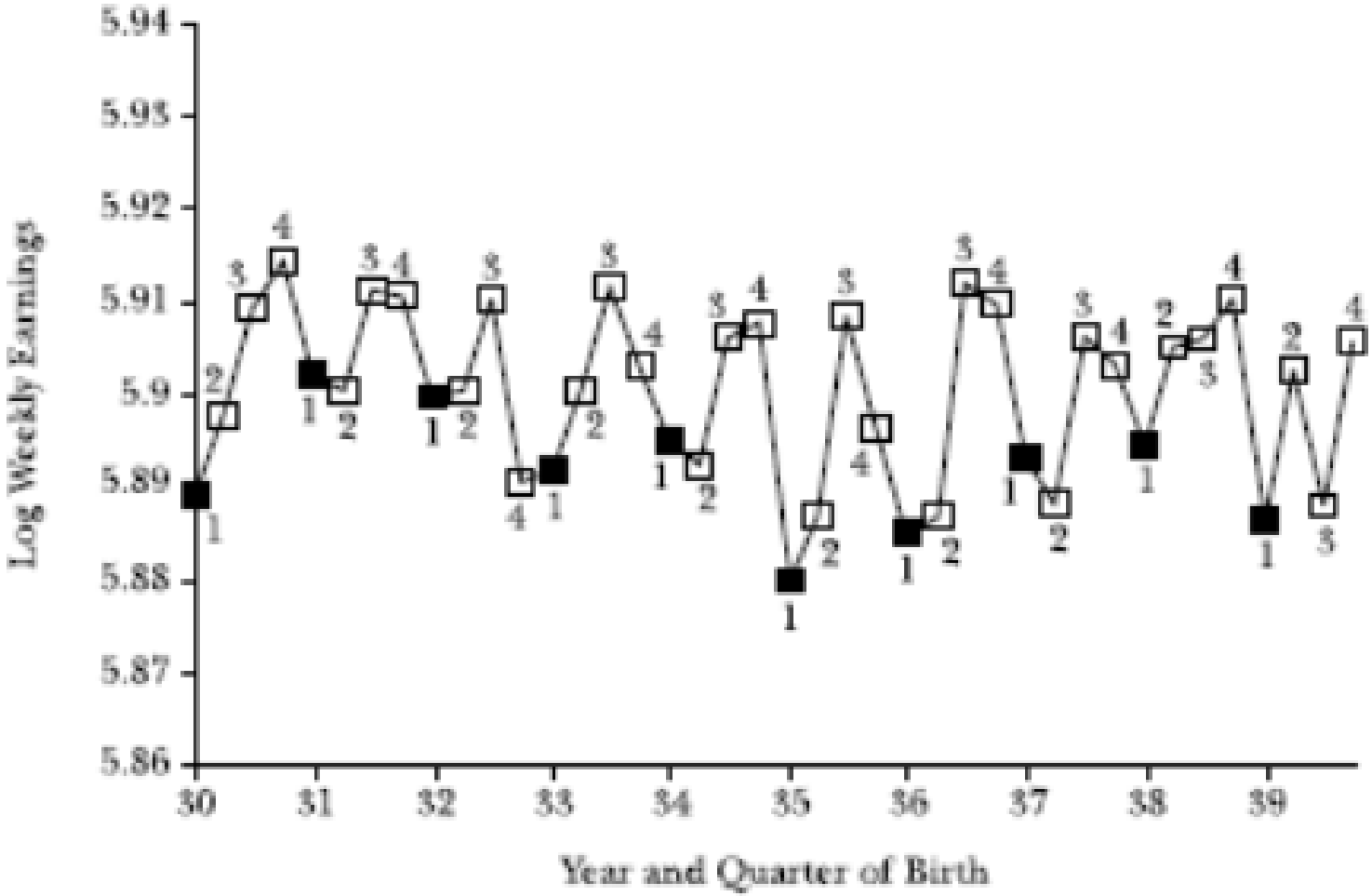


Source: Authors' calculations from the 1980 Census.

QoB and Earning

Figure 2

Mean Log Weekly Earnings, by Quarter of Birth



Source: Authors' calculations from the 1980 Census.

Weaknesses of IV framework

- Key identifying assumption is not testable. Must rely on economic theory, institutional knowledge
- Finite sample bias: weak instruments, too many instruments, small samples $\rightarrow$ all lead to the biased inference, IV in finite samples is biased towards OLS
- If the treatment effect is heterogeneous, IV can still estimate the treatment effect for a particular subpopulation, but not necessarily something we care about

Checking Assumptions

How to check for instrument relevance and exogeneity?

- Relevance: First stage F-statistics
  - $F > 10$
- Exogeneity: More difficult. In many cases it comes down to making a plausible case for $\text{Cov}(Z, u) = 0$
  - Just identified: Rely on theory, knowledge of the empirical issue at hand
  - If over identified, can test whether a subset of instruments is exogenous
- J-test
  - Regress residuals of TSLS on instruments
  - Scale resulting F-statistic by m (number of overidentifying restrictions)
  - This is the J-stat $\sim \chi^2_{m-k}$
  - H_0: overidentifying restrictions are valid