Econometrics 2 (Part 1)

Sergey Lychagin

Central European University

Winter 2020

Motivation

Early econometric research on

- Simultaneous Equations Models
- Measurement Error Bias

Today

- Omitted Variables Bias
- Applications:
  - Effect of military service on earnings
  - Effects of family size on female labor supply
  - Returns to education

Omitted variables problem

- Constant Effects Setup

$$\begin{aligned} y_{si} &= f_i(s) \\ f_i(s) &= \alpha + \rho s + \eta_i \\ \eta_i &= A_i'\gamma + v_i,\ E[\eta_i|A_i] = A_i\gamma. \end{aligned}$$

- A_i are the only reason why η_i and s_i may be correlated: $s_i \amalg \eta_i|A_i$.
- γ — population regression coefficients (not necessarily causal)

$$\begin{aligned} E[A_i v_i] &= 0 \\ E[s_i v_i] &= 0 \end{aligned}$$

- If A_i is observed:

$$Y_i = \alpha + \rho s_i + A_i'\gamma + v_i \Rightarrow \text{Long Regression}$$

- Problems:
  - A_i is unobserved, how can we estimate ρ?

Instrumental Variable

$$Y_i = \alpha + \rho s_i + \eta_i$$

- Instrumental Variable
  - z_i correlated with s_i, but uncorrelated with any other determinants of Y_i (instrument relevance)
  - $Cov(\eta_i, z_i) = 0$, or z_i uncorrelated with both A_i and v_i (instrument exogeneity)
- *Exclusion Restriction/Instrument exogeneity*:

$$\rho = \frac{Cov(Y_i, z_i)}{Cov(s_i, z_i)} = \frac{Cov(Y_i, z_i)/V(z_i)}{Cov(s_i, z_i)/V(z_i)}$$

- Ratio of population regression of Y_i on z_i (*reduced form*) and s_i on z_i (*first stage*).

Assumptions

- 2 important assumptions:
  - *Relevance*: z_i has an effect on s_i (can be tested)

$$Cov\left(s_i, z_i\right) \neq 0$$

  - *Exogeneity/exclusion restriction:* z_i affects Y_i only via s_i:

$$Cov\left(\eta_i, z_i\right) = 0$$

- How do we find Instrumental Variables?
  - Institutional knowledge
  - Ideas about the process determining s_i
- Examples:
  - Compulsory schooling law
  - Schooling decision based on costs and benefits
  - College proximity as determinant of schooling decision

General model

- Structural Equation

$$Y_i = X_i'\alpha + \rho s_i + \eta_i,\ E[\eta_i|X_i] = 0.$$

  - First Stage:

$$s_i = X_i^{'}\pi_{10} + \pi_{11} z_i + \xi_{1i},\ E[\xi_{1i}|X_i, z_i] = 0$$

  - Reduced Form:

$$Y_i = X_i^{'}\pi_{20} + \pi_{21} z_i + \xi_{2i}$$

- s_i and Y_i are *endogenous variables*. Fundamental issue: ξ_{1i} correlates with η_i (e.g. schooling and wage are driven by unobserved ability).
- z_i instrumental variable, conditionally independent of η_i: $z_i \amalg \eta_i|X_i$.
- X_i — controls

Indirect Least Squares

Covariate adjusted IV estimator:

$$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{Cov\left(Y_i, \tilde{z}_i\right)}{Cov\left(s_i, \tilde{z}_i\right)}$$

- $\tilde{z}_i$ residual from regressing z_i on x_i (regression anatomy)
- Proof:

$$\begin{aligned} Y_i &= X_i'\alpha + \rho s_i + \eta_i \\ Cov\left(Y_i, \tilde{z}_i\right) &= \rho Cov\left(s_i, \tilde{z}_i\right) \end{aligned}$$

  - $\tilde{z}_i$ uncorrelated with X_i by construction.
  - $\tilde{z}_i$ uncorrelated with η_i by assumption (easy to check)

Alternative Representation

$$Y_i = X_i'\alpha + \rho s_i + \eta_i$$

- Subsitute first stage

$$\begin{aligned} Y_i &= X_i'\alpha + \rho\left[X_i'\pi_{10} + \pi_{11}z_i + \xi_{1i}\right] + \eta_i \\ Y_i &= X_i'\left[\alpha + \rho\pi_{10}\right] + \rho\pi_{11}z_i + \left[\rho\xi_{1i} + \eta_i\right] \end{aligned}$$

- Reduced Form

$$Y_i = X_i'\pi_{20} + \pi_{21}z_i + \xi_{2i}$$

- Compare coefficients

$$\begin{aligned} \pi_{20} &= \alpha + \rho\pi_{10} \\ \rho\pi_{11} &= \pi_{21} \qquad \Rightarrow \rho = \frac{\pi_{21}}{\pi_{11}} \\ \xi_{2i} &= \rho\xi_{1i} + \eta_i \end{aligned}$$

Two Stage Least Squares

Re-write structural equation

$$Y_i = X_i'\alpha + \rho \underbrace{\left[X_i'\pi_{10} + \pi_{11} z_i\right]}_{s_i^*} + \rho\xi_{1i} + \eta_i$$

- s_i^* population fitted value from first stage
- X_i and z_i are uncorrelated with ξ_{1i}
- *second stage* regression coefficient on s^* equals ρ

Two stage least squares

- 2 stage procedure:
  - Fitted First Stage

$$\hat{s}_i = X_i^{'}\hat{\pi}_{10} + \hat{\pi}_{11} z_i$$

  - Second Stage Equation

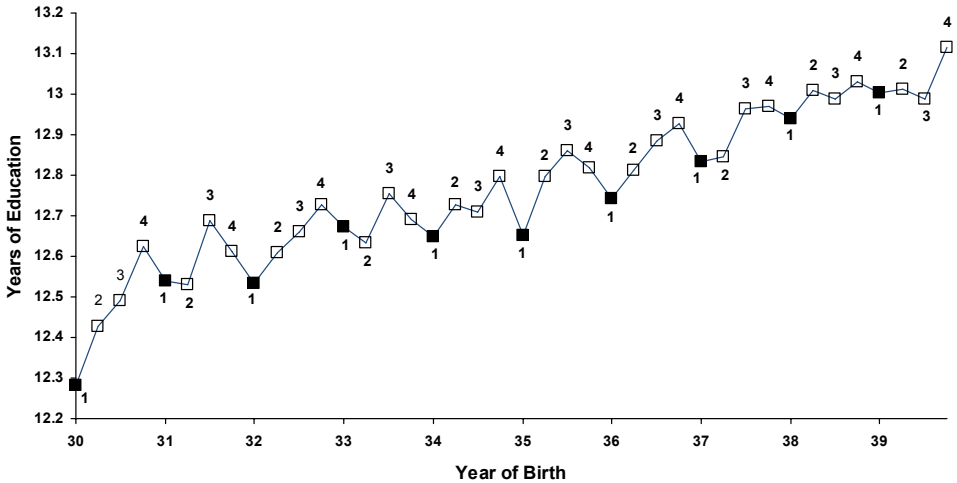
$$Y_i = X_i'\alpha + \rho\hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)]$$

  - Exclusion Restriction: $\hat{s}_i$ not correlated with η_i
  - By construction:$\hat{s}_i$ not correlated with $s_i - \hat{s}_i$
- 2SLS can be performed in two steps, but second stage standard errors are incorrect.
- Better to use STATA procedure!
- In a model with one endogenous variable and a single instrumental variable 2SLS is the same as ILS.

Compulsory schooling law

- School entry date determined by the calendar year when a child turns 6
- Those born later in the year are younger when they start school
- Compulsory schooling law: earliest school leaving date 16th birthday
- Kids born early in the year can leave before finishing 10th grade
- Does this variation in schooling levels influence earnings?

A. Average Education by Quarter of Birth (first stage)



B. Average Weekly Wage by Quarter of Birth (reduced form)

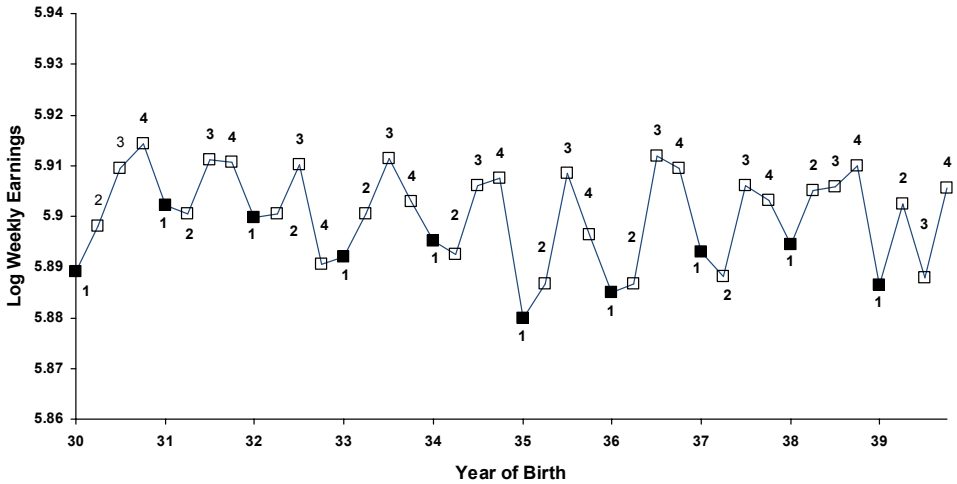


Figure 4.1.1: Graphical depiction of first stage and reduced form for IV estimates of the economic return to schooling using quarter of birth (from Angrist and Krueger 1991).

Multiple Instruments

- z_{1i}, z_{2i}, z_{3i} dummy variables for quarter of birth
- 2 stage least squares estimation
- First stage equation

$$s_i = X_i' \pi_{10} + \pi_{11} z_{1i} + \pi_{12} z_{2i} + \pi_{13} z_{3i} + \xi_{i1}$$

- $\hat{s}_i$ fitted values from first stage regression
- 2SLS "instrument": linear combination of all instrumental variables – increases efficiency.

Table 4.1.1: 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Years of education	0.075	0.072	0.103	0.112	0.106	0.108	0.089	0.061
	(0.0004)	(0.0004)	(0.024)	(0.021)	(0.026)	(0.019)	(0.016)	(0.031)
Covariates:								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year of birth dummies		✓			✓	✓	✓	✓
50 state of birth dummies		✓			✓	✓	✓	✓
Instruments:			dummy for QOB=1	dummy for QOB=1 or QOB=2	dummy for QOB=1	full set of QOB dummies	full set of QOB dummies int. with year of birth dummies	full set of QOB dummies int. with year of birth dummies

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses.

Wald Estimator

- Special case: z_i dummy variable
- Structural model $Y_i = \alpha + \rho s_i + \eta_i$

$$\begin{aligned} E(Y_i|z_i) &= \alpha + \rho E(s_i|z_i) + E(\eta_i|z_i) \\ E(Y_i|z_i = 1) &= \alpha + \rho E(s_i|z_i = 1) + E(\eta_i|z_i = 1) \\ E(Y_i|z_i = 0) &= \alpha + \rho E(s_i|z_i = 0) + E(\eta_i|z_i = 0) \end{aligned}$$

- Wald estimator

$$\begin{aligned} \rho &= \frac{E\left[Y_i|z_i = 1\right] - E\left[Y_i|z_i = 0\right]}{E\left[s_i|z_i = 1\right] - E\left[s_i|z_i = 0\right]} \\ &= \frac{\text{difference in mean earnings by z}}{\text{difference in mean schooling by z}} \end{aligned}$$

Draft lottery

- U.S. conscription during the Vietnam war era
  - Institution of draft lottery in 1970
  - each year 1970-1972 a random sequence of lottery numbers were assigned to each birth date in the cohort of 19-year olds.
  - lottery numbers below a cutoff were eligible to be drafted
  - exceptions for volunteers, school attendance, bad health etc.
- use draft eligibility status as binary instrument for military service
- lottery number positively correlated with veteran status: relevance
- lottery number uncorrelated to other determinants of earnings: exclusion restriction
- discrete instrument: lottery number groups, visual IV

Table 4.1.3: Wald estimates of the effects of military service on the earnings of white men born in 1950

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean	Eligibility Effect	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8	0.267	0.159	-2,741
		(210.5)		(0.040)	(1,324)
1971	3,338	-325.9			-2050
		(46.6)			(293)
1969	2,299	-2.0			
		(34.5)			

Notes: Adapted from Angrist (1990), Tables 2 and 3. 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 individuals in the sample.

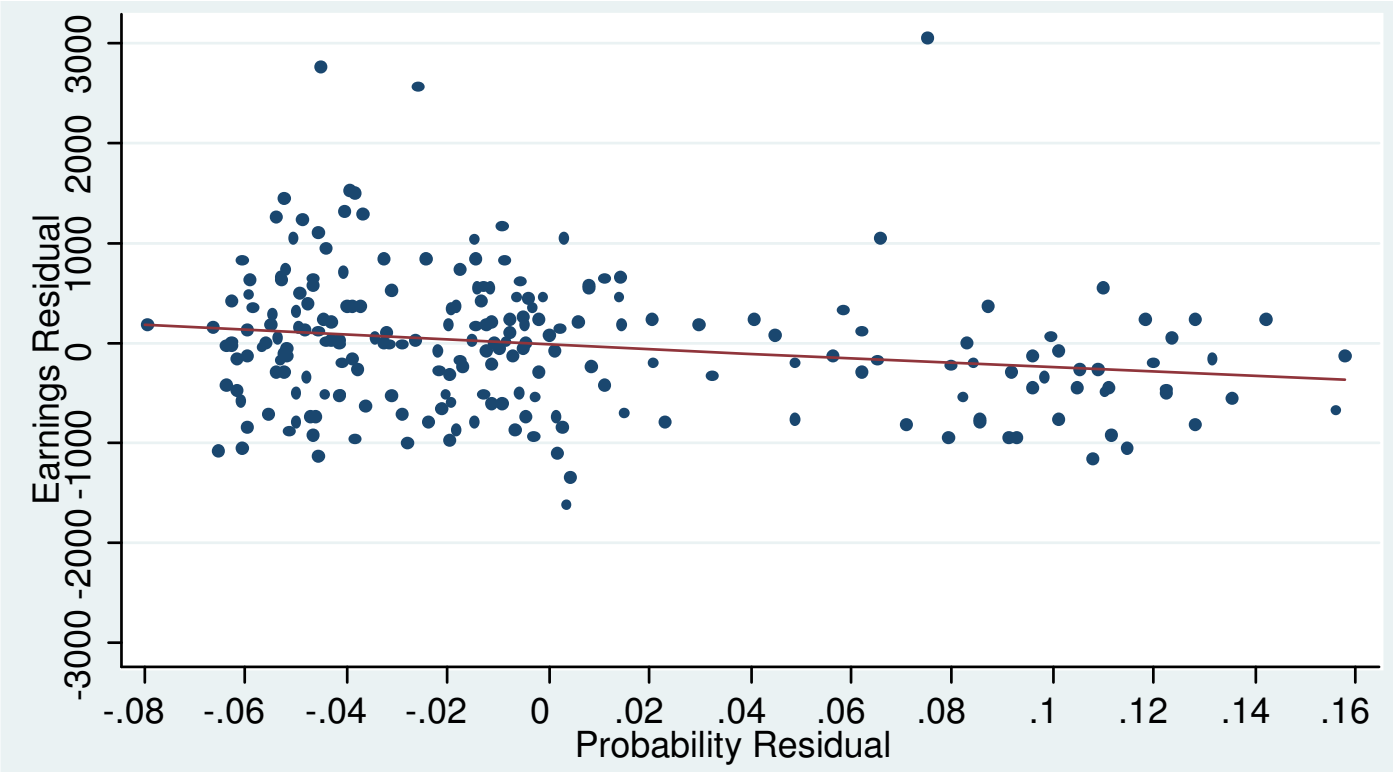
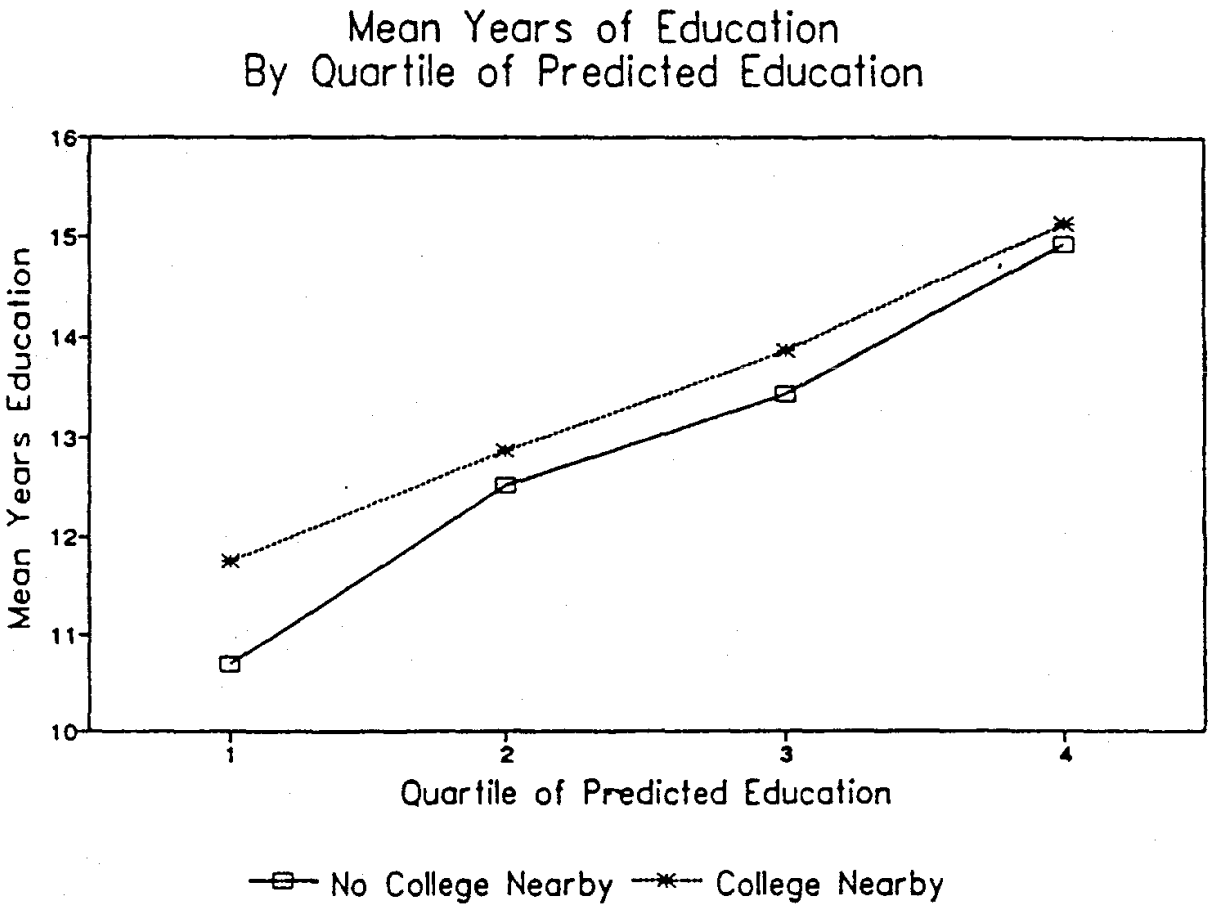


Figure 4.1.2: The relationship between average earnings and the probability of military service (from Angrist 1990). This is a VIV plot of average 1981-84 earnings by cohort and groups of five consecutive draft lottery numbers against conditional probabilities of veteran status in the same cells. The sample includes white men born 1950-53. Plotted points consist of average residuals (over four years of earnings) from regressions on period and cohort effects. The slope of the least-squares regression line drawn through the points is -2,384, with a standard error of 778.

"Using geographic variation in college proximity to estimate the return to education", Card (1993, NBER WP 4483)

- Ability Bias
  - Individual with high test scores have higher schooling upward biased OLS $\hat{\rho}$
- Absence of "pure" random assignment
  - Use the presence of a nearby college as exogenous variation in education
  - Students who grow up in an area without a college face a higher cost of college education, since the option of living at home is precluded.
  - ρ might depend on levels of income

Relevance



Note: prediction equation is fit to subsample with no college nearby

Structural Model Equation

Model

$$Y_i = \alpha + \rho s_i + \gamma_1 exper_i + \gamma_2 exper_i^2 + \eta_i$$

- Potential experience $exper_i = age_i - s_i - 6$
- Additional covariates: parents' education, region of residence, etc
- Proxy for ability: 'knowledge of the world of work' test score
- instrument c_i college proximity
- First stage

$$s_i = \pi_{10} + \pi_{11} c_i + \pi_{12} exper_i + \pi_{13} exper_i^2 + \xi_{1i}$$

Multiple Endogenous Variables

- $experience$, $experience^2$
- Need two additional excluded variables z_2, z_3 correlated with $experience$, $experience^2$
- age, age^2
- **Three** first stage equations:

$$\begin{aligned} s_i &= X_i'\pi_{10} + \pi_{11}z_{1i} + \pi_{12}z_{2i} + \pi_{13}z_{3i} + \xi_{1i} \\ exper_i &= X_i'\pi_{20} + \pi_{21}z_{1i} + \pi_{22}z_{2i} + \pi_{23}z_{3i} + \xi_{2i} \\ exper_i^2 &= X_i'\pi_{30} + \pi_{31}z_{1i} + \pi_{32}z_{2i} + \pi_{33}z_{3i} + \xi_{3i} \end{aligned}$$

- Reduced form equation:

$$Y_i = X_i'\pi_{40} + \pi_{41}z_{1i} + \pi_{42}z_{2i} + \pi_{43}z_{3i} + \xi_{4i}$$

Table 3: Reduced Form and Structural Estimates of Education and Earnings Models

	Reduced Form Models: Education		Reduced Form Models: Earnings		Structural Models of Earnings	
	(1)	(2)	(3)	(4)	(5)	(6)
A: Treat Experience and Experience Squared as Exogenous						
1. Live Near College in 1966	0.320 (0.088)	0.322 (0.083)	0.042 (0.018)	0.045 (0.018)	--	--
2. Education	--	--	--	--	0.132 (0.055)	0.140 (0.055)
3. Family Background Variables[a]	no	yes	no	yes	no	yes
B: Treat Experience and Experience Squared as Endogenous[b/]						
4. Live Near College in 1966	0.382 (0.114)	0.365 (0.105)	0.047 (0.019)	0.048 (0.019)	--	--
5. Education	--	--	--	--	0.122 (0.046)	0.132 (0.049)
6. Family Background Variables[a]	no	yes	no	yes	no	yes

Exclusion restriction

- Exclusion restriction does not allow for a *direct* effect of college proximity on earnings.
  - Better schools in college areas
  - Geographic wage premia
  - Selection of families into college areas

"Forbidden regression"

Common mistake — "forbidden regression"

- Suppose one fits $\widehat{exper}_i$ and $\widehat{s}_i$, runs Y_i on $\widehat{exper}_i$, $\widehat{exper}_i^2$ and $\widehat{s}_i$ rather than on $\widehat{exper}_i$, $\widehat{exper^2}_i$ and $\widehat{s}_i$.
- What the second stage estimates:

$$\begin{aligned} Y_i &= \alpha + \rho s_i + \gamma_1 exper_i + \gamma_2 exper_i^2 + \eta_i \\ &= \alpha + \rho(\widehat{s}_i + \xi_{1i}) + \gamma_1(\widehat{exper}_i + \xi_{2i}) + \gamma_2(\widehat{exper}_i + \xi_{2i})^2 + \eta_i \\ &= \alpha + \rho\widehat{s}_i + \gamma_1\widehat{exper}_i + \gamma_2\widehat{exper}_i^2 \\ &+ \underbrace{(\rho\xi_{1i} + \gamma_1\xi_{2i} + \gamma_2\xi_{2i}^2 + 2\gamma_2\widehat{exper}_i\xi_{2i} + \eta_i)}_{\zeta_i} \end{aligned}$$

- RHS variables are likely to correlate with ζ_i (e.g. if ξ_{2i} is heteroskedastic).

Run three separate 1st stage regressions, don't predict $exper_i$ and $exper_i^2$ using one 1st stage regression.

Testing for Endogeneity

- 2SLS less efficient than linear regression (larger standard errors)

$$Y_i = \alpha X_i^{'} + \rho s_i + \eta_i$$

- z_i exogenous instrument.
- If $Cov(s_i, \eta_i) = 0$, we can use linear regression
  - 2SLS consistent but less efficient
- If $Cov(s_i, \eta_i) \neq 0$, should use 2SLS with instrument z_i
- Idea: Compare OLS and 2SLS estimates

Testing for Endogenity

- First Stage

$$s_i = X_i^{'}\pi_{10} + \pi_{11}z_i + \xi_{1i}$$

- Predict first stage residual $\hat{\xi}_{1i}$ and include it in structural equation.

$$Y_i = X_i^{'}\alpha + \rho s_i + \delta\hat{\xi}_{1i} + error$$

- Hausman Test: Test $H_0 : \delta = 0$

Testing Overidentification Restrictions

$$Y_i = X_i^{'}\alpha + \rho s_i + \eta_i$$

- Two instruments z_1 and z_2
- We could generate 2 IV estimators one using z_1, one using z_2 and compare or check for correlation between IV-residuals with the other instrument.
- Test procedure
  - Estimate 2SLS using z_1 and z_2 and predict residuals $\hat{\eta}_i$
  - Regress $\hat{\eta}_i$ on all exogenous variables (X, z_1, z_2) and obtain R^2
  - H_0 : z_1 and z_2 uncorrelated to η_i
  - Under H_0, $nR^2 \approx \chi^2_q$, $q = 2$, number of instruments

Testing Overidentification Restrictions

- Caveat:
  - IV estimators often imprecise tests don't have much power.
  - Treatment effect heterogeneity

Using IV to address simultaneity issues

Typical application: demand estimation

$$q_i = \alpha - \rho p_i + X'_{di}\beta_d + u_{di} \tag{1}$$

Identification issue: price p_i is driven by demand shocks u_{di}.
Using OLS is a bad idea.

- Typical strategy: use observable supply shifters X_{si} as instruments for p_i. Be careful: X_{si} cannot correlate with u_{di}!
  - Exogenous changes in market structure (entry/exit/mergers)
  - Fluctuations in wages, prices for inputs
  - Disruptions in supply chains
- Estimating supply: symmetric case; use demand shifters X_{di} as instruments for q_i.

Weak instruments

Suppose we estimate ρ using one exogenous IV:

$$Y_i = \alpha + \rho s_i + \eta_i$$

$$\widehat{\rho} = \rho + \frac{\frac{1}{n}\sum_i (\eta_i - \overline{\eta}_i)(z_i - \overline{z}_i)}{\frac{1}{n}\sum_i (s_i - \overline{s}_i)(z_i - \overline{z}_i)}$$

LLN+CLT imply that

$$\frac{1}{n}\sum_i (\eta_i - \overline{\eta}_i)(z_i - \overline{z}_i) = cov(\eta_i, z_i) + \frac{1}{\sqrt{n}} N_1 + o\left(\frac{1}{\sqrt{n}}\right)$$

$$\frac{1}{n}\sum_i (s_i - \overline{s}_i)(z_i - \overline{z}_i) = cov(s_i, z_i) + \frac{1}{\sqrt{n}} N_2 + o\left(\frac{1}{\sqrt{n}}\right)$$

where N_1 and N_2 are two normal random variables.

Weak instruments

$$\widehat{\rho} = \rho + \frac{\frac{1}{\sqrt{n}} N_1 + o\left(\frac{1}{\sqrt{n}}\right)}{cov(s_i, z_i) + \frac{1}{\sqrt{n}} N_2 + o\left(\frac{1}{\sqrt{n}}\right)}$$

- In the limit, the denominator $\approx cov(s_i, z_i)$, and then $\widehat{\rho} \approx \rho + \frac{1}{\sqrt{n} cov(s_i, z_i)} N_1$.
- However, in finite (and not necessarily small!) samples, $\frac{1}{\sqrt{n}} N_2$ may dominate $cov(s_i, z_i)$.
- In this case, $\widehat{\rho} \approx \rho + \frac{N_1}{N_2}$ — biased point estimate, the distribution of $\widehat{\rho}$ far from normal.
- This may happen even if z_i appears significant at the conventional levels in the 1st stage.

There is more on this at the Friday's seminar (how bad the bias can be, how to test properly). Stay tuned!