Econometrics 2 –Part 1

Arieda Muço

Central European University

Winter 2023

Motivation

Instrumental Variables can help us with:

- OVB
- Measurement error
- Simultaneity

- 2 parts:
  - Restricted model with constant coefficients (Potential outcomes are the same for everybody)
  - Unrestricted model with heterogenous potential outcomes
- Applications:
  - Returns to education
  - Effect of military service on earnings
  - Effects of family size on female labor supply

Omitted variables problem

- Constant Effects Setup

$$\begin{aligned} y_{si} &= f_i(s) \\ f_i(s) &= \alpha + \rho s_i + \eta_i \\ \eta_i &= A_i\gamma + v_i \end{aligned}$$

- A_i is the only reason why η_i and s_i may be correlated.
- γ population regression coefficients.

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

- If A_i is observed:

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

- 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)
  - $Cov\left(\eta_i, z_i\right) = 0$, or z_i uncorrelated with both A_i and v_i (instrument exogeneity)
- *Exclusion Restriction/Instrument exogeneity*:

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

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

Assumptions

- 2 important assumptions:
  - z_i has an effect on s_i. First stage is not zero *Relevance*:

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

  - z_i has an effect on Y_i only through affecting 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

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

  - First Stage:

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

  - Reduced Form:

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

- s_i and Y_i are *endogenous variables*
- X_i and z_i are *exogenous variables*
- z_i instrumental variable
- X_i exogenous covariates

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)
- Indirect Least Squares (ILS) estimator of the causal effect ρ in the model
- Structural Equation:

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

  - $\tilde{z}_i$ uncorrelated with X_i by construction.
  - $\tilde{z}_i$ uncorrelated with η_i by assumption.

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?

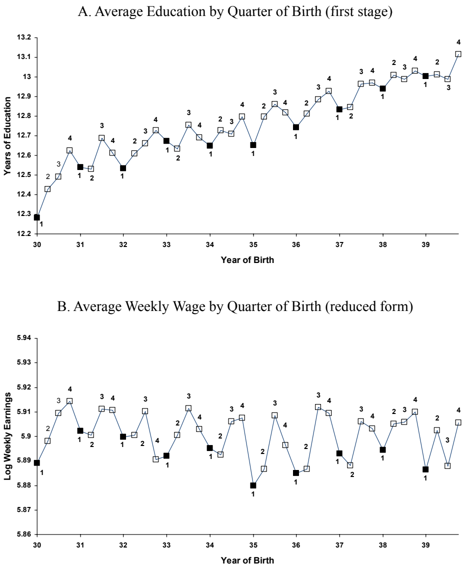


Figure 4.1.1: Graphical depiction of first stage and reduced form for IV estimates of the economic return to schooling using quarter of birth (from Angrist and Krueger 1991).

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": a 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 (0.0004)	0.072 (0.0004)	0.103 (0.024)	0.112 (0.021)	0.106 (0.026)	0.108 (0.019)	0.089 (0.016)	0.061 (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

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

Table 4.1.2: Wald estimates of the returns to schooling using quarter of birth instruments

	(1) Born in the 1st or 2nd quarter of year	(2) Born in the 3rd or 4th quarter of year	(3) Difference (std. error) (1)-(2)
ln (weekly wage)	5.8916	5.9051	-0.01349 (0.00337)
Years of education	12.6881	12.8394	-0.1514 (0.0162)
Wald estimate of return to education			0.0891 (0.0210)
OLS estimate of return to education			0.0703 (0.0005)

Notes: Adapted from a re-analysis of Angrist and Krueger (1991) by Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930-39 birth cohorts in the 1980 Census 5 percent file. The sample size is 329,509.

Applications

- Effects of veteran status on earnings
  - Does serving in the military have an impact on earnings later in life?
  - Instrument: Vietnam war draft lottery (Angrist,1990)
- Effects of family size on female labor supply (Angrist and Evans, 1998)
  - Instrument: Multiple births, sex composition
- Returns to schooling
  - Ability bias or sorting based on returns to schooling?
  - Instrument: Quarter of birth (Angrist and Krueger, 1991)
  - Instrument: Proximity to college (Card, 1993)

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 a 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 (210.5)	0.267	0.159 (0.040)	-2,741 (1,324)
1971	3,338	-325.9 (46.6)			-2050 (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.

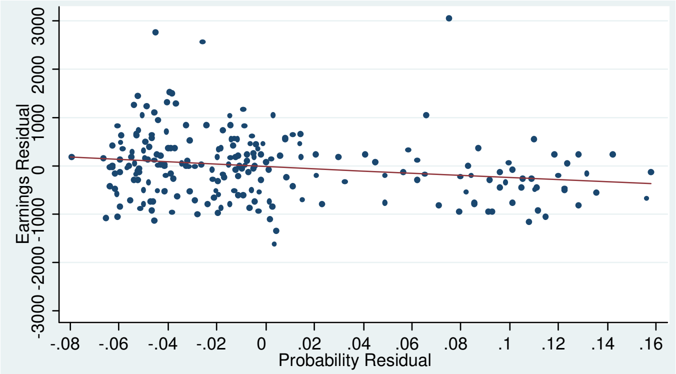


Figure 4.1.2: The relationship between average earnings and the probability of military service (from Angrist 1990). This is a VIV plot of average 1981-84 earnings by cohort and groups of five consecutive draft lottery numbers against conditional probabilities of veteran status in the same cells. The 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.

College Proximity

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

Data

- National Longitudinal Survey of Youths
- Young men age 14-24 in 1966
- 1st survey 1966
  - family composition
  - father's, mother's education
  - characteristics of local labor market e.g. college
- Follow up surveys every 2 years
  - large attrition 20% drop out in first 3 years
  - select 1976 interview for labor market information
  - education, wages

Table 1: Sample Characteristics for Overall Sample and 1976 Subset of National Longitudinal Survey of Young Men

	Overall NLS-YM Sample	Subset Interview in 1976: Valid Education	Valid Wage & Education
1. Age Distribution in 1966:			
Age 14-15 (%)	25.9	25.3	25.5
Age 16-17	24.9	23.8	24.1
Age 18-20	23.1	24.1	24.6
Age 21-24	26.1	26.7	25.8
2. Regional Distribution in 1966:			
Northeast (%)	20.2	20.0	20.7
Midwest	25.4	26.3	26.0
South	41.1	41.3	41.4
West	13.3	12.5	11.9
3. Lived in SMSA 1966 (%)	66.0	64.3	65.0
4. Lived Near 4-year College in 1966 (%)	69.2	67.8	68.2
5. Family Structure at Age 14:			
Mother & Father (%)	76.8	79.2	78.9
Mother Only (%)	11.8	10.0	10.1
6. Average Parental Education			
Mother's Education (yrs)	10.3	10.4	10.3
Father's Education (yrs)	9.4	10.0	10.0
7. Percent Black	27.5	23.0	23.0
8. Average Score on KWW Test	33.0	33.5	33.5
9. Interviewed in 1976 (%)	70.7	100.0	100.0
10. Mean Education in 1976	13.2	13.2	13.3
11. Live in South in 1976 (%)	39.6	40.0	40.3
12. Sample Size	5225	3613	3010

Notes: Means are based on all available valid observations in any

- Linear model

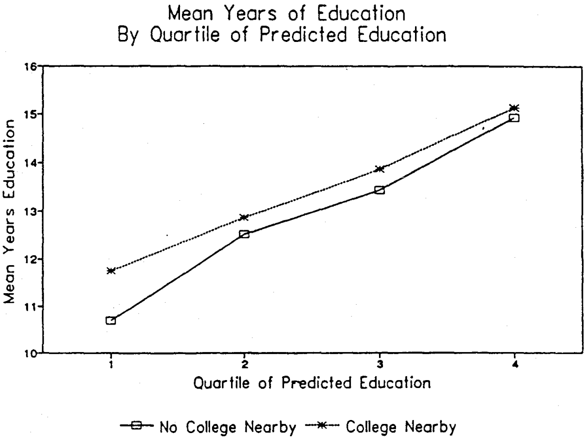
$$Y_i = \alpha + \rho s_i + \eta_i$$

Table 2: Estimated Regression Models for Log Hourly Earnings

	(1)	(2)	(3)	(4)	(5)
1. Education	0.074 (0.004)	0.075 (0.003)	0.073 (0.004)	0.074 (0.004)	0.073 (0.004)
2. Experience	0.084 (0.007)	0.085 (0.007)	0.085 (0.007)	0.085 (0.007)	0.085 (0.007)
3. Experience-Squared /100	-0.224 (0.032)	-0.229 (0.032)	-0.230 (0.032)	-0.226 (0.032)	-0.229 (0.032)
4. Black Indicator	-0.190 (0.017)	-0.199 (0.018)	-0.194 (0.019)	-0.194 (0.019)	-0.189 (0.019)
5. Live in South	-0.125 (0.015)	-0.148 (0.026)	-0.146 (0.026)	-0.145 (0.026)	-0.146 (0.026)
6. Live in SMSA	0.161 (0.015)	0.136 (0.020)	0.136 (0.020)	0.137 (0.020)	0.138 (0.020)
7. Region in 1966 (8 indicators)	no	yes	yes	yes	yes
8. Live in SMSA in 1966	no	yes	yes	yes	yes
9. Parental Education[a] (main effects)	no	no	yes	yes	yes
10. Interacted Parental Education Classes[b]	no	no	no	yes	yes
11. Family Structure[c] (2 indicators)	no	no	no	no	yes
12. R-squared	0.291	0.300	0.301	0.303	0.304
13. P-value for family background effects	--	--	0.235	0.462	0.165

- 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

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

- 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

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

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

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 4: OLS and Instrumental Variables Estimates of the Return to Education: Alternative Specifications

	OLS Estimate	IV Estimate[a]
1. Basic Specification (N=3010)	0.073 (0.004)	0.132 (0.049)
2. Use 1978 Wages and Education (N=2639 with 1978 data)	0.066 (0.006)	0.117 (0.061)
3. Include KWW Test Score (N=2963 with valid KWW)	0.055 (0.004)	0.136 (0.078)
4. Include KWW Test Score, Instrument KWW with IQ[b] (N=2040 with valid KWW and IQ)	0.061 (0.005)	0.089 (0.085)
5. Use Proximity to Public College as instrument for education	as in row 1	0.194 (0.059)
6. Use Proximities to 2-year and 4-year colleges as instruments for education	as in row 1	0.117 (0.047)
7. Use Subsample Age 14-19 in 1966 (N=2037)	0.076 (0.006)	0.094 (0.064)

Multiple Instruments

- z_{1i} proximity to 4 year college
- z_{2i} proximity to 2 year college

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

- $\hat{s}_i$ fitted values from first stage regression
- 2SLS "instrument": Residual from a regression of first stage fitted values on exogenous covariates increases efficiency.

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

Direct effect from college proximity

- Idea: college proximity should have a bigger effect on educational choice of low income families
- Instrument for education: $c_i * p_i$ were p_i is an indicator for low parental background
- First stage

$$s_i \quad = \quad \pi_{10} + \pi_{11} c_i + \pi_{12}(c_i * p_i) + \pi_{13} exper_i + \pi_{14} exper_i^2 + \xi_{1i}$$

- Earnings Equation

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

Table 5: Instrumental Variables Estimates of the Return to Education Based on Interaction of Parental Education and Proximity to College

	Reduced Form Models: Education	Reduced Form Models: Earnings	Structural Models of Earnings	
	(1)	(2)	(3)	(4)
1. Live Near College in 1966	0.154 (0.135)	0.029 (0.024)	0.015 (0.029)	0.013 (0.024)
2. Live Near College * Low Parental Education[a]	0.462 (0.186)	0.043 (0.032)	--	--
3. Education[b]	--	--	0.093 (0.065)	0.097 (0.048)
4. Family Background Variables[c]	yes	yes	yes	yes

Testing for Endogeneity

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

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

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

Testing for Endogenity

- First Stage

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

- $Cov\,(s_i, \xi_{1i}) = 0$
  - 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 whether $\delta = 0$

$$H_0 : \delta = 0$$

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 and obtain R^2
  - $H_0 : z_1$ and z_2 uncorrelated to η_i
  - $nR^2 \approx \chi^2_q$, $q = 2$, number of instrument

Testing Overidentification Restrictions

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