Econometrics II

Fabian Waldinger (LMU Munich)

Motivating Example Omitted Variable Bias: Ability Bias

- One of the most famous examples of omitted variable bias in economics is the ability bias when estimating returns to education
- More able people could get more schooling and at the same time earn more not because of the additional schooling but just because they are more able:

$$\log(y) = \beta_1 + \beta_2 S + \beta_3 A + \varepsilon$$

  - $y =$ earnings
  - $S =$ years of schooling
  - $A =$ ability
- Most datasets do not contain measures of ability; we therefore estimate

$$\log(y) = \beta_1 + \beta_2 S + \varepsilon$$

Ability Bias

- What is the expected value of $\tilde{\beta}_2$?

$$E(\tilde{\beta}_2|\boldsymbol{X}) = \beta_2 + \beta_3 \frac{Cov(S, A)}{Var(S)}$$

- Is this likely positive or negative?
- Under which circumstances would the ability bias be 0?

How IV Can be Used to Obtain Unbiased Estimates?

- How can we estimate the true β_2 if ability is unobserved?
- Randomly assigning S will be difficult
- Solution 2: Use an instrumental variable (IV): Z
- 2 important conditions for a valid IV:
  1. $Cov(S, Z) \neq 0$ (first stage exists)
  2. $Cov(Z, \varepsilon) = 0$ (exclusion restriction: Z is uncorrelated with any other determinants of the dependent variable)
- While we can test whether the first condition is satisfied the second condition cannot be tested. As a researcher you have to try to convince your audience that it is satisfied
- What do the two conditions mean for the ability bias example?
- Is a proxy variable for ability (e.g. IQ) a good IV?

The IV Estimator

- The model of interest is:

$$y = \beta_1 + \beta_2 x + \varepsilon$$

- But here $Cov(x, \varepsilon) \neq 0$ because of a violation of GM3 and we can therefore not obtain an unbiased/consistent estimate of β_2 by OLS
- We therefore find an IV (z) that satisfies the following conditions:
  1. $Cov(x, z) \neq 0$ (first stage exists)
  2. $Cov(z, \varepsilon) = 0$ (exclusion restriction: z is uncorrelated with any other determinants of the dependent variable)

The IV Estimator

- The IV estimator can be derived as follows:

$$\begin{aligned} Cov(z, y) &= Cov(z, \beta_1 + \beta_2 x + \varepsilon) \\ &= Cov(z, \beta_1) + \beta_2 Cov(z, x) + Cov(z, \varepsilon) \\ &= \beta_2 Cov(z, x) + Cov(z, \varepsilon) \end{aligned}$$

- Now under assumption 2: $Cov(z, \varepsilon) = 0$ and hence the last term drops out
- We now solve for β_2 (under the assumption that $Cov(x, z) \neq 0$) and get:

$$\beta_2 = \frac{Cov(z, y)}{Cov(z, x)}$$

The IV estimator

- Given a random sample, we can estimate the population quantities by the sample analogs:

$$\hat{\beta}_2^{IV} = \frac{\sum_{i=1}^{N} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N} (z_i - \bar{z})(x_i - \bar{x})}$$

- When $z = x$ (i.e. x is exogenous and can be used at its own instrument) we would get the OLS estimator
- Because thinking in regression coefficients can sometimes be easier we can divide both denominator and numerator by Var(Z) to get:

$$\hat{\beta}_2^{IV} = \frac{Cov(y, z)/Var(z)}{Cov(z, x)/Var(z)}$$

- The coefficient of interest is the ratio of the population regression of y on z (reduced form) to the population regression of x on z (first stage)

IV is Biased in Small Samples

- Can we prove that the IV estimator is unbiased?

$$\hat{\beta}_2^{IV} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})}$$

- Now substitute the true model for y:

$$\begin{aligned}
\hat{\beta}_2^{IV} &= \frac{\sum(z_i - \bar{z})([\beta_1 + \beta_2 x_i + \varepsilon_i] - [\beta_1 + \beta_2 \bar{x} + \bar{\varepsilon}])}{\sum(z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\sum(z_i - \bar{z})(\beta_2[x_i + \bar{x}] + [\varepsilon_i - \bar{\varepsilon}])}{\sum(z_i - \bar{z})(x_i - \bar{x})} \\
&= \beta_2 + \frac{\sum(z_i - \bar{z})[\varepsilon_i - \bar{\varepsilon}])}{\sum(z_i - \bar{z})(x_i - \bar{x})}
\end{aligned}$$

- Because x and ε are correlated we cannot take expectations of the last term -> IV is not unbiased

IV is Consistent

- However, we can show that the IV estimator is consistent if the IV assumptions are satisfied
- Taking plims of the the last term:

$$plim\left(\frac{\sum(z_i-\bar{z})[\varepsilon_i-\bar{\varepsilon}])}{\sum(z_i-\bar{z})(x_i-\bar{x})}\right)=plim\left(\frac{\frac{1}{N}\sum(z_i-\bar{z})[\varepsilon_i-\bar{\varepsilon}])}{\frac{1}{N}\sum(z_i-\bar{z})(x_i-\bar{x})}\right)$$

- The plim of $\frac{1}{N}\sum(z_i-\bar{z})[\varepsilon_i-\bar{\varepsilon}])=Cov(z,\varepsilon)$ which is 0 by the exclusion restriction
- The plim of $\frac{1}{N}\sum(z_i-\bar{z})(x_i-\bar{x})=Cov(z,x)=\sigma_{zx}$ which is $\neq 0$ by the first stage
- And hence:

$$plim\,\hat{\beta}_2^{IV}=\beta_2+\frac{0}{\sigma_{zx}}=\beta_2$$

IV Jargon

- Causal relationship of interest:

$$y = \beta_1 + \beta_2 x + \varepsilon$$

- First-Stage regression:

$$x = \gamma_1 + \gamma_2 z + \mu$$

- Second-Stage regression:

$$y = \beta_1 + \beta_2 \hat{x} + \upsilon$$

- Reduced form:

$$y = \lambda_1 + \lambda_2 z + \upsilon$$

Inference with the IV Estimator

- If the IV assumptions 1 and 2 are satisfied and if we add the additional assumption that

$$E(\varepsilon^2|z) = \sigma^2$$

- We can then show that the asymptotic variance of $\hat{\beta}_2^{IV}$ is

$$\frac{\sigma^2}{n\sigma_x^2\rho_{xz}^2}$$

- The asymptotic standard error is therefore:

$$\frac{\hat{\sigma}^2}{TSS_x R_{xz}^2}$$

  - $\hat{\sigma}^2$ is the consistent estimator for σ^2 : $\hat{\sigma}^2 = \frac{1}{N-k}\sum \varepsilon_i^2$
  - TSS_x is the total sum of squares of the x_i
  - R_{xz}^2 is the R-squared of the regression of $x_i$on z_i (i.e. the first stage regression)

IV versus OLS

- The variance of the OLS estimator is:

$$\frac{\sigma^2}{TSS_x}$$

- while the comparable formula for the IV estimator is:

$$\frac{\sigma^2}{TSS_x R^2_{xz}}$$

- As R^2_{xz} is always lower than 1 the IV variance is always larger than the OLS variance

Two-Stage Least Squares

- If we have multiple IVs for one endogenous variable we can use two-stage least squares to estimate unbiased coefficients.
- Suppose you have two exogenous variables: z_1 and z_2
- As before the IVs have to satisfy the exclusion restriction and there has to be a first stage
- We can then estimate the following first stage regression:

$$x = \gamma_1 + \gamma_2 z_1 + \gamma_3 z_2 + \mu$$

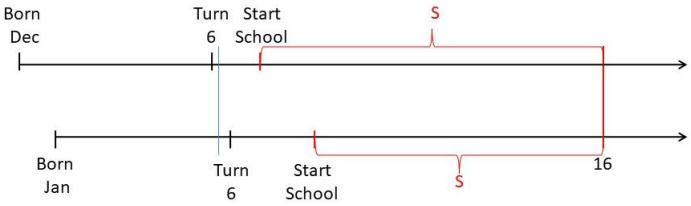
- Obtain $\hat{x}$ and then estimate the following second stage regression:

$$y = \beta_1 + \beta_2 \hat{x} + \upsilon$$

- (One would have to adjust the standard errors)
- This would estimate a consistent β_2

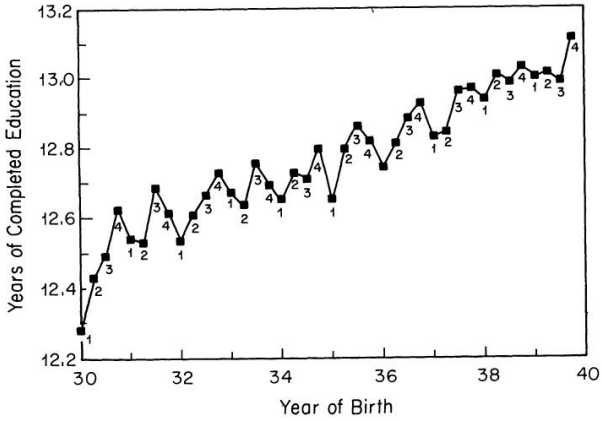
Instrumenting Using Compulsory Schooling Laws

- In practice it is often difficult to find convincing instruments (in particular because many potential IVs do not satisfy the exclusion restriction)
- In the returns to education literature Angrist and Krueger (1991) had a very influential study where they used quarter of birth as an instrumental variable for schooling
- In the US you could drop out of school once you turned 16
- Children have different ages when they start school and thus different lengths of schooling at the time they turn 16 when they can potentially drop out



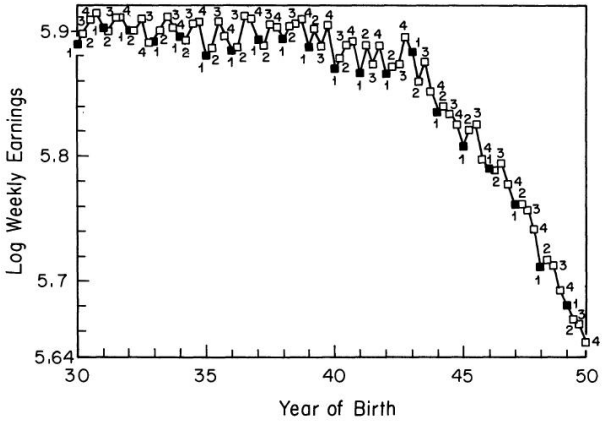
First Stages

- Men born earlier in the year have lower schooling. This indicates that there is a first stage



Reduced Form

- Do differences in schooling due to different quarter of birth translate into different earnings?



Two Stage Least Squares

- The first stage regression is:

$$S = \gamma_1 + \gamma_2 Z + \gamma_3 X + \mu$$

- The reduced form regression is:

$$Y = \lambda_1 + \lambda_2 Z + \lambda_3 X + \upsilon$$

- The covariate adjusted IV estimator is the sample analog of the ratio $\frac{\lambda_3}{\gamma_3}$
- Here we would estimate Two-stage-least squares (2SLS)
- It is called 2SLS because you could estimate it as follows (see above):
  1. Obtain the first stage fitted values:

$$\hat{S} = \hat{\gamma}_1 + \hat{\gamma}_2 Z + \hat{\gamma}_3 X$$

     where $\hat{\gamma}$ are the OLS estimates of the first stage regression.
  2. Plug the first stage fitted values into the "second-stage equation":

$$Y = \beta_1 + \beta_2 \hat{S} + \beta_3 X + \upsilon$$

Two Stage Least Squares

- Despite the name the estimation is usually not done in two steps (if you would do that the standard errors would be wrong)
- The standard regression packages usually do the job for you (and get the standard errors right)
- The intuition of 2SLS, however, is very useful: 2SLS only retains the variation in S that is generated by quasi-experimental variation (and thus hopefully exogenous)
- Angrist and Krueger use more than one instrumental variable to instrument for schooling: they include a dummy for each quarter of birth
- Their estimated first-stage regression is therefore:

$$S = \gamma_1 + \gamma_2 Z_1 + \gamma_3 Z_2 + \gamma_4 Z_3 + \gamma_5 X + \mu$$

- The second stage is the same as before but the fitted values are from the new first stage

First Stage Regressions in Angrist & Krueger (1991)

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect[a] I	II	III	F-test[b] [P-value]
Total years of education	1930–1939	12.79	−0.124	−0.086	−0.015	24.9
			(0.017)	(0.017)	(0.016)	[0.0001]
	1940–1949	13.56	−0.085	−0.035	−0.017	18.6
			(0.012)	(0.012)	(0.011)	[0.0001]
High school graduate	1930–1939	0.77	−0.019	−0.020	−0.004	46.4
			(0.002)	(0.002)	(0.002)	[0.0001]
	1940–1949	0.86	−0.015	−0.012	−0.002	54.4
			(0.001)	(0.001)	(0.001)	[0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004	0.051	0.012	5.9
			(0.014)	(0.014)	(0.014)	[0.0006]
	1940–1949	14.28	0.005	0.043	−0.003	7.8
			(0.011)	(0.011)	(0.010)	[0.0017]
College graduate	1930–1939	0.24	−0.005	0.003	0.002	5.0
			(0.002)	(0.002)	(0.002)	[0.0021]
	1940–1949	0.30	−0.003	0.004	0.000	5.0
			(0.002)	(0.002)	(0.002)	[0.0018]

First Stage Regressions in Angrist & Krueger (1991)

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect[a] I	II	III	F-test[b] [P-value]
Total years of education	1930–1939	12.79	−0.124	−0.086	−0.015	24.9
			(0.017)	(0.017)	(0.016)	[0.0001]
	1940–1949	13.56	−0.085	−0.035	−0.017	18.6
			(0.012)	(0.012)	(0.011)	[0.0001]
High school graduate	1930–1939	0.77	−0.019	−0.020	−0.004	46.4
			(0.002)	(0.002)	(0.002)	[0.0001]
	1940–1949	0.86	−0.015	−0.012	−0.002	54.4
			(0.001)	(0.001)	(0.001)	[0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004	0.051	0.012	5.9
			(0.014)	(0.014)	(0.014)	[0.0006]
	1940–1949	14.28	0.005	0.043	−0.003	7.8
			(0.011)	(0.011)	(0.010)	[0.0017]
College graduate	1930–1939	0.24	−0.005	0.003	0.002	5.0
			(0.002)	(0.002)	(0.002)	[0.0021]
	1940–1949	0.30	−0.003	0.004	0.000	5.0
			(0.002)	(0.002)	(0.002)	[0.0018]

- Quarter of birth is a strong predictor of total years of education

First Stage Regressions in Angrist & Krueger (1991)

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect[a] I	II	III	F-test[b] [P-value]
Total years of education	1930–1939	12.79	−0.124	−0.086	−0.015	24.9
			(0.017)	(0.017)	(0.016)	[0.0001]
	1940–1949	13.56	−0.085	−0.035	−0.017	18.6
			(0.012)	(0.012)	(0.011)	[0.0001]
High school graduate	1930–1939	0.77	−0.019	−0.020	−0.004	46.4
			(0.002)	(0.002)	(0.002)	[0.0001]
	1940–1949	0.86	−0.015	−0.012	−0.002	54.4
			(0.001)	(0.001)	(0.001)	[0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004	0.051	0.012	5.9
			(0.014)	(0.014)	(0.014)	[0.0006]
	1940–1949	14.28	0.005	0.043	−0.003	7.8
			(0.011)	(0.011)	(0.010)	[0.0017]
College graduate	1930–1939	0.24	−0.005	0.003	0.002	5.0
			(0.002)	(0.002)	(0.002)	[0.0021]
	1940–1949	0.30	−0.003	0.004	0.000	5.0
			(0.002)	(0.002)	(0.002)	[0.0018]

- Reassuringly quarter of birth does not affect the probability of graduating from college

IV Results

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS
Years of education	0.0711	0.0891	0.0711	0.0760
	(0.0003)	(0.0161)	(0.0003)	(0.0290)
Race (1 = black)	—	—	—	—
SMSA (1 = center city)	—	—	—	—
Married (1 = married)	—	—	—	—
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No
Age	—	—	−0.0772	−0.0801
			(0.0621)	(0.0645)
Age-squared	—	—	0.0008	0.0008
			(0.0007)	(0.0007)
χ^2 [dof]	—	25.4 [29]	—	23.1 [27]

IV Results - Including More Covariates and Interacting Quarter of Birth

- They also include specifications using 30 (quarter of birth x year) dummies and 150 (quarter of birth x state) dummies as IVs (the effect of quarter of birth may vary by birth year or state)
- This reduces standard errors
- But also comes at the cost of potentially having a weak instruments problem (see below)

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS
Years of education	0.0672	0.0635	0.0671	0.0555
	(0.0013)	(0.0185)	(0.0003)	(0.0199)
SMSA (1 = center city)	—	—	—	—
Married (1 = married)	—	—	—	—
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No
49 State-of-birth dummies	Yes	Yes	Yes	Yes
Age	—	—	−0.0309	−0.3274
			(0.2538)	(0.2560)