Applied Economics

Quasi-experiments: Instrumental Variables and Sources of Endogeneity

Department of Economics
Universidad Carlos III de Madrid

Policy evaluation with quasi-experiments

- In a quasi-experiment or natural experiment there is a source of randomization that is “as if” randomly assigned, but this variation was not the result of an explicit randomized treatment and control design. We distinguish two types of quasi experiments:
- A case in which treatment (D) is “as if” randomly assigned (perhaps conditional on some control variables X).
- A case in which a variable (Z) that influences treatment (D) is “as if” randomly assigned (perhaps conditional on X), then Z can be used as an instrumental variable for D in an IV regression that includes the control variables X.

  - The article by Angrist is an example of this case.

Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records, Angrist, AER(1990)

- Did military service in Vietnam have a negative effect on earnings?
- A negative relationship between earnings and veteran status does not imply that veteran status causes lower earnings.
- Simple comparisons of earnings by veteran status give a biased measure of the effect of treatment on the treated (unless veteran status is independent of potential earnings).
- Comparisons of earnings controlling for observed characteristics make sense if veteran status is independent of potential earnings after these observed variables are taken into account.

Effect of veteran status on earnings

Let y represent earnings, D_i denote Vietnam-era veteran status, and X_i a set of controls:

Consider estimating the following conditional expectation by OLS

$$E\left[Y_i | D_i, X_i\right] = \beta_0 + \alpha D_i + \sum_k \gamma_k X_{ki}$$

- There is probably some unobserved difference that made some men choose the military and others not, and this difference could be correlated with earning potential.
- If D_i is correlated with unobserved variables that belong to the equation, OLS estimates are inconsistent.
- A possible solution is to find a valid instrumental variable.

An instrument for veteran status

- Concerns about the fairness of the U.S. conscription policy led to the institution of a draft lottery in 1970.
- This lottery was conducted annually during 1970-1972. It assigned random numbers (from 1 to 365) to dates of birth in cohorts of 19-year-olds. Men with the lottery numbers below a cutoff were called to serve (the cutoff was determined every year by the Department of Defense).
- Veteran status was not completely determined by randomized draft eligibility: some volunteered, while others avoided enrollment due to health conditions or other reasons. So, draft eligibility is simply correlated with Vietnam-era veteran status.

Draft eligibility as an instrument 1/2

- Let Z_i indicate draft eligibility (takes the value one if i got a number below the cutoff).
- In order to identify the causal effect of D_i on earnings it is crucial that the only reason for $E(Y_i|Z_i)$ to change when Z_i changes is the variation in $E(D_i|Z_i)$. Draft eligibility affects earnings only through its effect on veteran status.
- A simple check on this is to look for an association between Z_i and personal characteristics that should not be affected by D_i, for example race or sex. Another check is to look for an association between Z_i and Y_i for samples in which there is no relationship between D_i and Z_i.

Draft eligibility as an instrument 2/2

- Angrist looks for instance at 1969 earnings, since 1969 earnings predate the 1970 draft lottery. He finds no effect of the draft eligibility (row 69 in Table 1).
- With the same goal, he also looks at the cohort of men born in 1953. Although there was a lottery drawing that assigned a random number to the 1953 birth cohort in 1972, no one from that cohort was actually drafted. So Z and D are unrelated for this cohort. Angrist finds no significant relationship between earnings and draft eligibility status for men born in 1953 (using the 1972 cutoff).
- These results support the claim that the only reason for draft eligibility to affect earnings is through its impact on veteran status.

Differences in Earnings by Draft Eligibility - Regressions

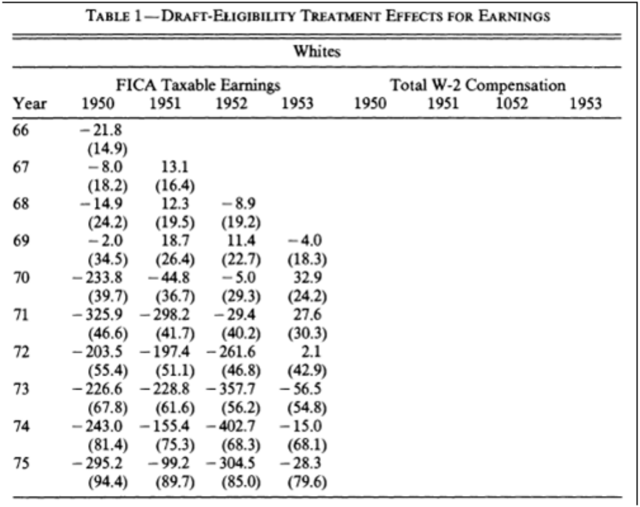
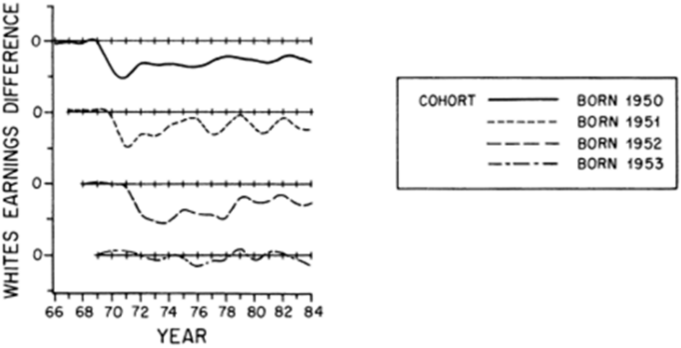


TABLE 1—DRAFT-ELIGIBILITY TREATMENT EFFECTS FOR EARNINGS

	Whites							
	FICA Taxable Earnings				Total W-2 Compensation			
Year	1950	1951	1952	1953	1950	1951	1052	1953
66	−21.8 (14.9)							
67	−8.0 (18.2)	13.1 (16.4)						
68	−14.9 (24.2)	12.3 (19.5)	−8.9 (19.2)					
69	−2.0 (34.5)	18.7 (26.4)	11.4 (22.7)	−4.0 (18.3)				
70	−233.8 (39.7)	−44.8 (36.7)	−5.0 (29.3)	32.9 (24.2)				
71	−325.9 (46.6)	−298.2 (41.7)	−29.4 (40.2)	27.6 (30.3)				
72	−203.5 (55.4)	−197.4 (51.1)	−261.6 (46.8)	2.1 (42.9)				
73	−226.6 (67.8)	−228.8 (61.6)	−357.7 (56.2)	−56.5 (54.8)				
74	−243.0 (81.4)	−155.4 (75.3)	−402.7 (68.3)	−15.0 (68.1)				
75	−295.2 (94.4)	−99.2 (89.7)	−304.5 (85.0)	−28.3 (79.6)				

Differences in Earnings by Draft Eligibility - A Graph



Notes: The figure plots the difference in FICA taxable earnings by draft-eligibility status for the four cohorts born 1950–53. Each tick on the vertical axis represents $500 real (1978) dollars.

Wald Estimator 1/2

- In a simple model with only D as a control: $Y_i = \beta_0 + \alpha D_i + \varepsilon_i$,
- With Z a valid IV we can write $\alpha = Cov(Y_i, Z_i)/Cov(D_i, Z_i)$
- If Z is a dummy variable taking the value one with probability p, for any W we can write:

$$\begin{aligned} Cov(W_i, Z_i) &= E[W_i Z_i] - E[W_i]E[Z_i] \\ &= \{E[W_i|Z_i = 1] - E[W_i|Z_i = 0]\}\, p * (1 - p) \end{aligned}$$

- Then:

$$\alpha = \frac{Cov(Y_i,Z_i)}{Cov(D_i,Z_i)} = \frac{E[Y_i|Z_i{=}1]-E[Y_i|Z_i{=}0]}{E[D_i|Z_i{=}1]-E[D_i|Z_i{=}0]}$$

Wald Estimator 2/2

- If D is also a dummy, for instance representing the treatment group:
  - $E[D_i|Z_i=1]$ is the probability of $D=1$ when $Z=1$, or the proportion of treated among those with $Z=1$
  - $E[D_i|Z_i=0]$ is the probability of $D=1$ when $Z=0$, or the proportion of treated among those with $Z=0$
  - The denominator captures the impact of the instrument on the probability of receiving treatment.
- The sample analogue of α is known as the Wald estimator.

The Wald Estimator (conditioning on X):

$$\hat{\alpha}_W(X) = \frac{\overline{Y}(X,Z=1)-\overline{Y}(X,Z=0)}{\overline{P}_{D=1}(X,Z=1)-\overline{P}_{D=1}(X,Z=0)}$$

Wald Estimator in this case

- Numerator:
  - ○ $\overline{Y}(X, Z = 1)$: average earnings for drafted individuals
  - ○ $\overline{Y}(X, Z = 0)$: average earnings for non-drafted individuals
  - ○ Interpret the coefficients in $Y_i = \beta_0 + \beta_1 Z_i + u_i$
- Denominator:
  - ○ $\overline{P}_{D=1}(X, Z = 1)$: participation rate among those drafted: the proportion of veterans $(D = 1)$ among those drafted $(Z = 1)$
  - ○ $\overline{P}_{D=1}(X, Z = 0)$: participation rate among those not drafted: the proportion of veterans $(D = 1)$ among those not drafted $(Z = 0)$
  - ○ Interpret the coefficients in $D_i = \delta_0 + \delta_1 Z_i + u_i$

Results

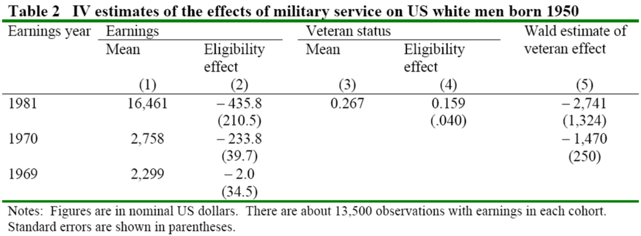


Table 2 IV estimates of the effects of military service on US white men born 1950

Earnings year	Earnings		Veteran status		Wald estimate of veteran effect
	Mean	Eligibility effect	Mean	Eligibility effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	– 435.8	0.267	0.159	– 2,741
		(210.5)		(.040)	(1,324)
1970	2,758	– 233.8			– 1,470
		(39.7)			(250)
1969	2,299	– 2.0			
		(34.5)			

Notes: Figures are in nominal US dollars. There are about 13,500 observations with earnings in each cohort. Standard errors are shown in parentheses.

Table taken from Angrist and Pischke, Mostly Harmless Econometrics.

- For men born in 1950, there are significant negative effects of eligibility status on earnings in 1970, when these men were beginning their military service and in 1981, ten years later.
- In contrast, there is no evidence of an association between eligibility status and earnings in 1969, the year the lottery drawing for men born in 1950 was held but before anyone born in 1950 was actually drafted.
- Since eligibility status was randomly assigned, estimates in column (2) represent the effect of draft eligibility on earnings.

Wald Estimator

- To go from draft-eligibility effects to veteran-status effects we need the denominator of the Wald estimator, which is the effect of draft-eligibility on the probability of serving in the military: $\overline{P}_{D=1}(X,Z=1)-\overline{P}_{D=1}(X,Z=0)$.
- This information is reported in column (4): draft-eligible men were 0.16 more likely to have served in the Vietnam era.
- For earnings in 1981, long after most Vietnam-era servicemen were discharged from the military, the Wald estimate of the effect of military service is about 17 percent of the mean.
- Effects were even larger in percentage terms in 1970, when affected soldiers were still in the army.

Sources of Endogeneity

Sources of Endogeneity: Motivation

- Until now, we have justified endogeneity of the controls as a problem of omitted variables
- Today, we are going to study two alternative sources of endogeneity:
  - errors in variables
  - simultaneity

Errors in variables

Measurement error in dependent variable

- Suppose that the true model is $Y^* = \beta_0 + \beta_1 X + u^*$
- Instead of Y^*, we observe $Y = Y^* + e$ so that

$$Y = \beta_0 + \beta_1 X + (u^* + e)$$

- If $\mathrm{E}(e) \neq 0$, the OLS estimate of β_0 would be inconsistent
- If $\mathrm{cov}(x, e) \neq 0$, the OLS estimate of β_1 would be inconsistent

Measurement error in control

- Suppose that the true model is

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

- Instead of X^*, we observe $X = X^* + e$ with $\mathrm{E}(e) = 0$ so that

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

When e is uncorrelated with X and with u^*

Assume that measurement error is uncorrelated with the observed control and the structural error term:

$$\text{cov}(X, e) = 0 \text{ and } \text{cov}(e, u^*) = 0$$

$$\Rightarrow \text{cov}(X, u) = \text{cov}(X, u^* - \beta_1 e) = \text{cov}(X, u^*) - \beta_1 \text{cov}(X, e) = 0$$

- OLS is consistent
- The estimates will have larger standard deviations: $\text{Var}(u) = \text{Var}(u^*) + \beta_1^2 \text{Var}(e)$

When e is uncorrelated with X^* and with u^*

If, more realistically, the measurement error is uncorrelated with the true control and the structural error term:

$$\text{cov}(X, e) = 0 \text{ and } \text{cov}(e, u^*) = 0$$

$$\Rightarrow \mathrm{cov}(X, u) = \mathrm{cov}(X^* + e, u^* - \beta_1 e) = -\beta_1 \mathrm{Var}(e) \neq 0$$

- OLS is inconsistent: plim $\hat{\beta}_{OLS} = \beta \left(1 - \frac{\text{Var}(e)}{\text{Var}(x)}\right) < \beta$ (attenuation bias)
- With several controls, all their estimators will be inconsistent (although the direction of the bias for the other controls is not clear)

Example: Savings equations

- Suppose that you want to estimate the marginal propensity to save
- Savings equation: $sav = \beta_0 + \beta\, inc^* + u$
- Observed income: $inc = inc^* + e$
- Really estimating: $sav = \beta_0 + \beta\, inc + (u - \beta\, e)$

Savings and instruments

- If measurement error is uncorrelated with true income,

$$\mathrm{cov}\left(inc, e\right) = \mathrm{Var}\left(e\right) \neq 0 \Rightarrow \mathrm{cov}\left(inc, u - \beta e\right) = -\beta \mathrm{var}\left(e\right)$$

- OLS inconsistent: plim $\hat{\beta}_{OLS} = \beta \left(1 - \frac{\mathrm{Var}(e)}{\mathrm{Var}(inc)}\right) < \beta$ (attenuation bias)
- We can use IV: we need a variable
  - correlated with true income (relevance)
  - uncorrelated with the measurement error in observed income
- Any second measure of the income (from the firm, a spouse,...) would be a good instrument
- Alternative, another proxy of the income, such as house size...

Simultaneity: Estimating a demand function

A supply and demand system of equations

- Supply function: $q^s = \gamma_0 + \beta^s p + \gamma x^s + u^s$
- Demand function: $q^d = \alpha_0 + \beta^d p + \alpha x^d + u^d$

- q^s is quantity supplied, q^d is demand, and p is the price
- x^s is an exogenous factor that is observed by the econometrician and affects only the supply
- x^d is an observed exogenous factor that affects only the demand
- u^s are the effects of unobservable factors that affect the supply curve
- u^d are the effects of unobservable factors that affect the demand curve

Equilibrium prices

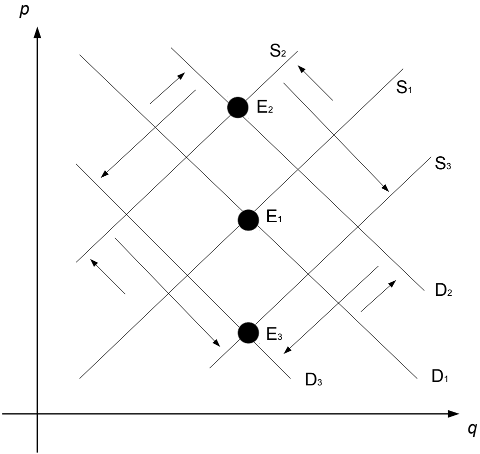
At equilibrium, $q^s = q^d = q$

- $q^s = q^d \Rightarrow \gamma_0 + \beta^s p + \gamma x^s + u^s = \alpha_0 + \beta^d p + \alpha x^d + u^d$
- Hence, at equilibrium,
  $p = \left(\frac{1}{\beta^s - \beta^d}\right)\left[(\alpha_0 - \gamma_0) + (\alpha x^d - \gamma x^s) + (u^d - u^s)\right]$
- At equilibrium, prices depend on demand shifters (x^d and u^d) and supply shifters (x^s and u^s)

Simultaneity

- Observed prices and quantities are simultaneously determined
- Both depend on demand shifters (x^d and u^d) and supply shifters (x^s and u^s)
- if we regress q on p and x^s by OLS, $\hat{\beta}^s$ is inconsistent because $\operatorname{cov}(p, u^s) \neq 0$
- if we regress q on p and x^d by OLS, $\hat{\beta}^d$ is inconsistent because $\operatorname{cov}\left(p, u^d\right) \neq 0$

A graphical interpretation



direct observations don't reveal the negative relation between demand and price

Instruments for the demand equation

- We want to estimate the demand equation: $q = \alpha_0 + \beta^d p + \alpha x^d + u^d$
- At equilibrium, $p = f\left(x^d, x^s, u^d, u^s\right)$
- An instrument is any variable that, independently of the other controls, is correlated with p (relevance) and is not correlated with u^d
- x^d are already controls in the demand equation, so they cannot be instruments
- x^s are potentially good instruments:
  - $\operatorname{cov}(x^s, p) \neq 0$ (relevance) (because p is a function of x^s)
  - $\operatorname{cov}\left(x^s, u^d\right) = 0$ (exogeneity) (otherwise x^s is not really exogenous)

Summary

- If one regressor is measured with error, then it may be endogenous. If we have additional variables which act as proxies for the regressor, we could implement 2SLS
- When two variables are simultaneously determined, then they are both endogenous
- If we want to estimate a demand equation, we need 2SLS and supply shifters