Econometrics of causal inference

Throughout, we consider the simplest case of a linear outcome equation, and homogeneous effects:

$$y = \beta x + \epsilon \tag{1}$$

where y is some outcome, x is an explanatory variable, and ϵ is an unobservable which represent unobserved determinants of y not accounted for in x. Consider simplest case: just one explanatory variable. Here β measures the *causal effect* of a unitary change in x[1] on the outcome y. Examples: (y is wages, x is yrs of schooling), (y is quantity demanded, x is price), (y is price, x is market concentration), (y is test scores, x is class size), etc.

You want to estimate β. But if x is endogenous (in the sense that $E(\epsilon x) \neq 0$) then OLS estimate is biased. Two ways out:

1. Find an IV for x: roughly speaking, correlated with x, uncorrelated with ϵ, and (of course) excluded from the equation of interest (1)
2. Find an *experimental* (or quasi-experimental) situation where one could plausibly claim that x is exogenous (in the sense of being uncorrelated with the unobservable ϵ).

Basically speaking, a *natural experiment* is a *discrete* (usually binary) variable z which fulfills one of the functions above.

1 Natural experiments as IV

Estimation of β using **Wald estimator** $(z'x)^{-1}(z'y)$. Justification is because population analog is

$$\frac{cov(y,z)}{cov(x,z)} = \frac{cov(x,z)\beta + cov(\epsilon,z)}{cov(x,z)} = \beta$$

when z is valid instrument.

For a binary z, Wald estimator becomes

$$\frac{\frac{\sum_{i=1}^n y_i z_i}{\sum_{i=1}^n z_i} - \frac{\sum_{i=1}^n y_i(1-z_i)}{\sum_{i=1}^n (1-z_i)}}{\frac{\sum_{i=1}^n x_i z_i}{\sum_{i=1}^n z_i} - \frac{\sum_{i=1}^n x_i(1-z_i)}{\sum_{i=1}^n (1-z_i)}}.$$

Examples:

Angrist and Krueger (1991) y is wages, x is yrs of schooling, z is quarter of birth (1=Jan-Sept; 0=Sept-Dec).[2] Exploits two institutional features: (i) can only enter

[1] If x is a continuous variable, then β is a *marginal* effect.

[2] Angrist and Krueger do the analysis for each birth year separately.

school (first grade) when you are 6 yrs old by Sept. 1; (ii) must remain in school until age 16 $\Longrightarrow$ people with $z = 1$ forced to complete more yrs of schooling before they can drop out.

Ex: Children born in 2000 can drop out on their birthday in 2016. Those with $z = 0$ will be in ninth grade; those with $z = 1$ will be in tenth grade.[3]

Angrist (1990) y is lifetime income, x is years of experience in the (civilian) workforce, and z is draft eligibility. Intuition: that draft eligibility led to exogenous shift in years of experience.

Angrist, Graddy, and Imbens (2000) y is quantity demanded, x is price, and z is weather variable.

Angrist and Evans (1990) y is parents' labor supply, x is number of children, z is indicator of sex composition of children (i.e., whether first two births were females)

Angrist and Lavy (1999) y is test scores, x is class size, z is indicator for whether total enrollment was "just above" a multiple of 40. Maimonides' rules states (roughly) that no class size should exceed forty, so that if enrollment (treated as exogenous) is "just below" 40, class sizes will be bigger, whereas if enrollment is "just above" 40, class sizes will be smaller.

They restrict their sample to all (school-cohorts) where total enrollment was within +/- 5 of a multiple of 40. This analysis represents an example of "regression-discontinuity design".

2 Cross-sectional approaches

Here we consider the situation where each individual in the dataset is only observed *once*. We also restrict attention to the binary treatment case. (Most common case for policy evaluation.)

2.1 Rubin causal framework

- Treatment $D \in 0, 1$
- Potential outcomes Y_D, $D = 0, 1$
- "Treatment effect": $\Delta Y \equiv Y_1 - Y_0$.
- Goal of inference: moments of ΔY.
  - Average Treatment Effect: $E[\Delta Y]$

[3]Note that if compulsory schooling were described in terms of *years of schooling*, then identification strategy fails.

- Average TE on the treated: $E[\Delta Y|D=1]$
- Local ATE: $E[\Delta Y|Z=z]$ for some auxiliary variable Z (depends on setting)
- Local ATT, &etc...
- Note that if ΔY is a nondegenerate random variable, it implies that the treatment effect differs across individuals in an arbitrary way. (In a regression context, this means that the coefficient on the treatment variable is different for every individual.)

In the cross-sectional setting, the crucial data limitation is that each individual can only be observed in one of the possible treatments: that is, defining

$$Y = D * Y_1 + (1 - D) * Y_0$$

the researcher observes a sample of (Y, D, Z) across individuals (Z are auxialary variables).

A naive estimator of ATE is just the difference in conditional means $E[Y|D=1] - E[Y|D=0]$. This is obviously not a good thing to do unless $Y_0, Y_1 \perp D$ – that is, unless treatment is *randomly* assigned (as it would be in a controlled lab setting, or in a tightly controlled field experiment).[4] Otherwise, typically $E[Y|D=0] = E[Y_0|D=0] \neq E[Y_0]$, and similarly to $E[Y|D=1]$.

2.2 Regression Discontinuity design

- Basic setup ("sharp" design):
  - Forcing variable Z: $D=0$ when $Z \leq \bar{Z}$; $D=1$ when $Z > \bar{Z}$. This implies you observe $E[Y_0|Z]$ for $Z \leq \bar{Z}$, and $E[Y_1|Z]$ for $Z > \bar{Z}$.
  - Continuity assumption: $E[Y_D|Z]$ continuous at $Z = \bar{Z}$, for $D = 0, 1$.
  - Local unconfoundedness: $Y_0, Y_1 \perp D|Z$. This means that $P(Y_1, Y_0, D|Z) = P(Y_1, Y_0|Z)P(D|Z)$.
  - $E[Y|D=1, \bar{Z}^+] - E[Y|D=0, \bar{Z}^-]$ estimates $E[Y_1 - Y_0|\bar{Z}]$, the "local" treatment effect for individuals with forcing variable $Z = \bar{Z}$.
    Proof:

$$\begin{aligned} E[Y|D=1, \bar{Z}+] - E[Y|D=0, \bar{Z}-] &= E[Y_1|D=1, \bar{Z}+] - E[Y_0|D=0, \bar{Z}-] \\ &= E[Y_1|\bar{Z}^+] - E[Y_0|\bar{Z}^-] \quad \text{(by cond. independence)} \\ &= E[Y_1|\bar{Z}] - E[Y_0|\bar{Z}] \quad \text{(by continuity)} \end{aligned}$$

■

- "Fuzzy" design (Hahn, Todd, and van der Klaauw (2001))

[4] The terminology "unconfoundedness" is also used for randomized treatment.

- Probability of treatment jumps discontinuously at $\bar{Z}$: that is, $P[D = 1|Z]$ jumps (up) at $Z = \bar{Z}$. Define $P^+ = P(D = 1|\bar{Z}+)$ and analogously P^-.
- Conditional independence: $Y_1, Y_0 \perp D|Z$ in a neighborhood of $\bar{Z}$.
- Continuity: $E[Y_D|\bar{Z}^+] = E[Y_D|\bar{Z}^-]$ for $D = 0, 1$.
- Let $Y = (1 - D)Y_0 + DY_1$. Then

$$E[Y_1 - Y_0|\bar{Z}] \approx \frac{E[Y|\bar{Z}^+] - E[Y|\bar{Z}^-]}{E[D|\bar{Z}^+] - E[D|\bar{Z}^-]}.$$

  Interpretation: above is Wald IV estimator in regression of observed outcome Y on D, using values of the instrument Z close to the jump point $\bar{Z}$.
  Proof: Derive that $E[Y|Z^+] = E[Y_0|Z] + P^+ \cdot \{E[Y_1|Z] - E[Y_0|Z]\}$ and similarly $E[Y|Z^-] = E[Y_0|Z] + P^- \cdot \{E[Y_1|Z] - E[Y_0|Z]\}$. Hence numerator of Wald estimator is $(P^+ - P^-) \cdot \{E[Y_1|Z] - E[Y_0|Z]\}$. Denominator is $(P^+ - P^-)$. ■.

2.3 Instrumental variables: LATE

More formally, the basic binary local average treatment effect ("LATE") setup is the following (cf. Angrist and Pischke (2009)):

- Binary IV: $Z \in \{0, 1\}$.
- Binary treatment $D_Z \in \{0, 1\}$
- Potential outcomes $Y_{DZ} = y(D, Z)$
- Assumption A1 (Independence): $Y_{D_1,1}, Y_{D_0,0}, D_1, D_0 \perp Z$
- A2 (Exclusion): $Y_{D,0} = Y_{D,1} \equiv Y_D$ for $D = 0, 1$.
- A3 ("rank"): $E[D_1 - D_0] \neq 0$.
- A4 (Monotonicity): $D_1 \geq D_0$ with probability 1.
- Then the Wald estimator $\frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]}$ estimates $E[Y_1 - Y_0|D_1 > D_0]$.

  Proof: by exclusion restriction, we have $Y = (1 - D)Y_0 + DY_1$ (Z doesn't enter); then using independence assumptions, we have

$$E[Y|Z = 1] = E[(1 - D)Y_0 + DY_1|Z = 1] = E[(1 - D_1)Y_0 + D_1Y_1].$$

  Similarly, $E[Y|Z = 0] = E[(1 - D_0)Y_0 + D_0Y_1]$, implying that the numerator is

$$\begin{aligned} &E[Y|Z = 1] - E[Y|Z = 0] = E[(Y_1 - Y_0)(D_1 - D_0)] \\ &= E[(Y_1 - Y_0) \cdot 1|D_1 > D_0]P(D_1 > D_0) + E[(Y_1 - Y_0) \cdot 0|D_1 = D_0]P(D_1 = D_0) \\ &\quad + E[(Y_1 - Y_0)(D_1 - D_0)|D_1 < D_0]P(D_1 < D_0) \\ &= E[(Y_1 - Y_0)|D_1 > D_0]P(D_1 > D_0). \end{aligned}$$

  Denominator, by similar argument, equals $P(D_1 > D_0)$.

Evidently, this Wald estimator measures the *average* effect of x on y for those for whom a change in z from 0 to 1 would have affected the treatment x.

For example, in the schooling/wages example, Wald estimator measures effect of an extra year of schooling on those (dropout) students for whom an earlier birth (ie. change z from 0 to 1) would have been forced to complete an extra year of schooling before dropping out. This insight is known by several terms, including *local IV* and *local average treatment effect (LATE)* (see Angrist and Imbens (1994) for more details).

3 Panel data

In panel data, one observes the same individual over several time periods, including (ideally) periods both before and after a policy change. In this richer data environment, one can estimate the effect of the policy change while controlling arbitrarily for individual-specific heterogeneity, as well as for time-specific effects. This is the *difference-in-difference* approach.

For example, x is often a policy change, such as rise in the minimum wage, which affects some states but not others. Estimation of (1) usually proceeds via the regression:

$$y_{it} = \alpha_i + \beta x_{it} + \gamma_t + \epsilon_{it}$$

Let $\hat{\beta}$ denote our estimate of β from this equation. Given the exogeneity and mean-zero assumptions on ϵ, we can interpret $\hat{\beta}$ in the following manner.

For simplicity, assume two periods, t and t'. Assume the policy is enacted in period t', and let x be a binary variable, which turns to 1 when the policy is enacted. Let i denote a cross-sectional unit which experienced the shift in x, and j denote a unit which did not experience such a shift. If we subtract (or "difference") the expected outcome equations for units i and j between periods t and t', we get rid of the unit-specific fixed effects:

$$\begin{aligned} E\left(y_{it'} - y_{it} | x_{it}, x_{it'}\right) &= \hat{\beta} + \left(\hat{\gamma}_t - \hat{\gamma}_{t'}\right) \\ E\left(y_{jt'} - y_{jt} | x_{jt}, x_{jt'}\right) &= \left(\hat{\gamma}_t - \hat{\gamma}_{t'}\right). \end{aligned} \tag{2}$$

Now difference the two equations in (2):

$$E\left[\left(y_{it'} - y_{it}\right) - \left(y_{jt'} - y_{jt}\right) | \cdots\right] = \hat{\beta} \tag{3}$$

so that $\hat{\beta}$ is a **difference-in-difference** estimate of the effect of a change in x on y.

Alternatively, the reasoning can go the other way: you construct the "diff-in-diff" estimate:

$$\tilde{\beta}_n \equiv \left(\bar{y}_{1t'} - \bar{y}_{1t}\right) - \left(\bar{y}_{0t'} - \bar{y}_{1t}\right)$$

where the bars ($\bar{}$) denote sample averages, and the subscript "1" denotes the cross-sectional units which experienced the policy shift, and "0" the cross-sectional units that did not.

Using the above argument backwards, $\tilde{\beta}_n$ is an estimator of the coefficient on x of a panel regression with fixed effects for cross-sectional units as well as time periods.

There are many many examples of this. Two examples are:

Card and Krueger (1994) y is employment, x is minimum wage (look for evidence of general equilibrium effects of minimum wage). Exploit policy shift which resulted in rise of minimum wage in New Jersey, but not in Pennsylvania. Sample is fast food restaurants on the NJ/Pennsylvania border.

Kim and Singal (1993) y is price, x is concentration of particular flight market. Exploit merger of Northwest and Republic airlines, which affected only markets (so we hope) in which Northwest or Republic offered flights.

References

Angrist, J. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313–336.

Angrist, J., and W. Evans (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313–336.

Angrist, J., K. Graddy, and G. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simulataneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67, 499–527.

Angrist, J., and G. Imbens (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476.

Angrist, J., and A. Krueger (1991): "Does Compulsory School Attendance Affect Scholling and Earnings?," *Quarterly Journal of Economics*, 106, 979–1014.

Angrist, J., and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533–575.

Angrist, J., and J. Pischke (2009): *Mostly Harmless Econometrics*. Princeton University Press.

Card, D., and A. Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–93.

Hahn, J., P. Todd, and W. van der Klaauw (2001): "Estimation of Treatment Effectswith a Quasi-Experimental Regression-Discontinuity Design," *Econometrica*, 69, 201–210.

Kim, E., and V. Singal (1993): "Mergers and Market Power: Evidence from the Airline Industry," *American Economic Review*, 83, 549–569.