Causal inference, in a nutshell

1 Causal inference: Background

To start with, we consider a linear structural outcome equation, and homogeneous effects:

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

where y is some outcome, d is an explanatory (or "treatment") variable of interest, and ϵ is an unobservable which represent unobserved determinants of y not accounted for in d. Here β measures the *causal effect* of a unitary change in d on the outcome y. Examples: (y is wages, d is yrs of schooling), (y is quantity demanded, d is price), (y is price, d is market concentration), (y is test scores, d is class size), etc. You want to estimate β. But if d is endogenous (in the sense that $E(\epsilon \cdot d) \neq 0$) then OLS estimate is biased.

A classic solution to the endogeneity problem is to use an instrumental variable z, which should be correlated with d, uncorrelated with ϵ, and excluded from the equation of interest (7).

Heuristically, rewrite the structural model as:

$$y = \beta' d(z, x) + \epsilon$$

where the notation $d(x, z)$ makes explicit that the treatment d depends on both the instrument z (the "exogenous" variation) and other factors x (which are correlated with ϵ, leading to "endogenous" variation). We derive the causal effect of d on y indirectly, by considering an exogenous change in z, which in turn affects d and then affects y. Formally, we have

$$\frac{dy}{dz} = \frac{\partial y}{\partial d}\frac{\partial d}{\partial z} \Longrightarrow \beta = \frac{dy}{dz} \Big/ \frac{\partial d}{\partial z}$$

In the special case when we have a binary auxiliary variable $Z \in \{0, 1\}$, we obtain the following estimator:

$$\frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}.$$

This is the classical *Wald estimator*. A number of the treatment effect estimators we consider below take this form, for different choices of the auxiliary variable Z.

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": $\Delta \equiv Y_1 - Y_0$.
- Goal of inference: moments of Δ.
  - Average Treatment Effect: $E[\Delta]$
  - Average TE on the treated: $E[\Delta|D = 1]$
  - Local ATE: $E[\Delta|Z = z]$ for some auxiliary variable Z (depends on setting)
  - Local ATT, &etc...
  - Note that if Δ is a nondegenerate random variable, it implies that the treatment effect differs across individuals in an arbitrary way. (In a linear model, this is consistent with the model $y_i = \beta_i d_i + \epsilon_i$, so 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 auxiliary 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). Otherwise, typically $E[Y|D = 0] = E[Y_0|D = 0] \neq E[Y_0]$, and similarly to $E[Y|D = 1]$.

2.2 Selection on observables: propensity score weighting and matching

- **Unconfounded assumption**: $Y_0, Y_1 \perp D|Z$, where Z denotes variables observed for each individual. This is *selection on observables*, as the interpretation is that treatments are exogenous once the additional observables Z are controlled for.
- Let F_Z denote the joint distribution of the Z variables. With this assumption, we have that

$$\begin{aligned}\int \{E[Y|D = 1, Z] - E[Y|D = 0, Z]\}\, dF_Z &= \int \{E[Y_1|D = 1, Z] - E[Y_0|D = 0, Z]\}\, dF_Z \\ &= \int \{E[Y_1|Z] - E[Y_0|Z]\}\, dF_Z \\ &= E[Y_1 - Y_0]\end{aligned}$$

which is the average treatment effect.

- But if Z is large dimensional, then inplementing this is not feasible. Therefore we consider some dimension-reducing approaches.

- Define the *propensity score*:

$$Q = Prob(D = 1|Z).$$

This can be estimated for each individual in the sample. Hence we assume that we observe (Y, D, Z, Q) for everyone in the sample. Remember that Q is just a function of Z.

- Rosenbaum and Rubin (1983) theorem: under the selection on observables assumption, we also have $(Y_0, Y_1) \perp D|Q$.

Proof: We want to show that $P(D, Y_1, Y_0|Q) = P(D|Q)P(Y_1, Y_0|Q)$. Starting with the Law of Total Probability, we have $P(D, Y_1, Y_0|Q) = P(D|Y_1, Y_0, Q)P(Y_1, Y_0|Q)$. So it suffices to show $P(D|Y_1, Y_0, Q) = P(D|Q)$. Since D is binary, we can focus on showing this for $P[D = 1|Y_1, Y_0, Q) = P(D = 1|Q)$. Note that

$$\begin{aligned} P[D = 1|Y_0, Y_1, Q] &= E\left\{E[D|Y_1, Y_0, Z]|Y_1, Y_0, Q\right\} \\ &= E\left\{E[D|Z]|Y_1, Y_0, Q\right\} \\ &= E\left\{Q|Y_0, Y_1, Q\right\} = Q. \end{aligned}$$

which does not depend on (Y_1, Y_0). ■

2.2.1 Inverse PS weighting

- Main result: $E(Y_1) = E\left[\frac{D*Y}{Q}\right]$ (Horvitz-Thompson estimator)

Proof:

$$E\left[\frac{D*Y}{Q}\right] = EE\left[\frac{D*Y}{Q}|Z\right] = E\frac{1}{Q}E\left[D*Y_1|Z\right]$$

$$\begin{aligned} &= E\frac{1}{Q}E\left[E\left(D*Y_1|Z, D\right)|Z\right] \\ &= E\frac{1}{Q}E\left[\left(D*E(Y_1|Z, D\right)|Z\right] \\ &= E\frac{1}{Q}E\left[\left(D*E(Y_1|Z)\right)|Z\right] \\ &= E\frac{E(Y_1|Z)}{Q}E\left[D|Z\right] \\ &= E\frac{E(Y_1|Z)}{Q}Q = E(Y_1). \end{aligned}$$

■

Similarly, $E(Y_0) = E\left[\frac{(1-D)*Y}{1-Q}\right]$.

- This is inverse propensity score weighting. Intuitively, in the case of $E(Y_1)$, you weight each individual in the treated sample by the probability of that individual being in the treated sample, which is Q.
- Since we divide by the propensity score Q above, we need that:

$$0 < Q(Z) < 1, \quad \forall Z.$$

This is known as the *overlap* assumption. Practically, it implies that for any Z, individuals with those covariates have a nonzero chance of being treated. Obviously, if there is any set of Z with positive probability for which $Q = 0$, then this set must be excluded from the expectation above, and so it is invalid to interpret it as the unconditional mean of Y_0.

2.2.2 PS matching

This is just dimension reduction. Let F_Q denote the distribution of propensity scores. We have that

$$\begin{aligned}\int \{E[Y|D=1,Q] - E[Y|D=0,Q]\}\, dF_Q &= \int \{E[Y_1|D=1,Q] - E[Y_0|D=0,Q]\}\, dF_Q \\ &= \int \{E[Y_1|Q] - E[Y_0|Q]\}\, dF_Q \\ &= E[Y_1 - Y_0]\end{aligned}$$

which is the average treatment effect. The penultimate equality uses the Rosenbaum-Rubin theorem.

This is "matching" in the sense that for each value of Q, you compare the average outcome of treated vs. untreated with this Q. Many variants on this based on how you match individuals in the treated vs. untreated samples.

2.3 Regression Discontinuity design

2.3.1 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: $Y_0, Y_1 \perp D|Z$ for Z in a neighborhood of $\bar{Z}$. 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:

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

■

Example: Angrist and Lavy (1999): y is test scores, d 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.

2.3.2 "Fuzzy" design

- 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

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

  a Wald-type estimator.[1]

  Proof: We have

$$\begin{aligned} E[Y|\bar{Z}^+] &= (1-P^+)E[Y_0|\bar{Z}^+] + P^+E[Y_1|\bar{Z}^+] \\ &= (1-P^+)E[Y_0|\bar{Z}] + P^+E[Y_1|\bar{Z}] \\ &= E[Y_0|\bar{Z}] + P^+ \cdot \left\{E[Y_1|\bar{Z}] - E[Y_0|\bar{Z}]\right\}. \end{aligned}$$

  Similarly $E[Y|\bar{Z}^-] = E[Y_0|\bar{Z}] + P^- \cdot \left\{E[Y_1|\bar{Z}] - E[Y_0|\bar{Z}]\right\}$. Hence numerator of Wald estimator is $(P^+ - P^-) \cdot \left\{E[Y_1|\bar{Z}] - E[Y_0|\bar{Z}]\right\}$. Denominator is $(P^+ - P^-)$. ■.

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 $\bar{Z}$.

[1] See Hahn, Todd, and van der Klaauw (2001).

2.4 Instrumental variables: LATE

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

- Binary IV: $Z \in \{0, 1\}$.
- Potential treatments (binary) $D_Z \in \{0, 1\}$
- Potential outcomes $Y_{DZ} = y(D, Z)$
- Assumption A1 (Exclusion): $Y_{D,0} = Y_{D,1} \equiv Y_D$ for $D = 0, 1$.
- Assumption A2 (Independence): $Y_1, Y_0, D_1, D_0 \perp Z$
- A3 ("rank"): $E[D_1 - D_0] \neq 0$.
- A4 (Monotonicity): $D_1 \geq D_0$ with probability 1.
- "Full" (latent) sample is (Y_0, Y_1, D_0, D_1, Z). We observe a sample of (Y, D, Z):
  - $D = (1 - Z)D_0 + ZD_1$
  - $Y = (1 - D)Y_0 + DY_1$ (by exclusion restriction, Z doesn't enter)
- Main result: 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: using independence and exclusion 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

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

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

Here, the Wald estimator measures the *average* effect of d on y for those for whom a change in z from 0 to 1 would have affected the treatment d. This insight is known by several terms, including *local IV* and *local average treatment effect (LATE)* (see Angrist and Imbens (1994) for more details).

The LATE framework allows for *selection on observables*. For instance, Y_1, Y_0, D_1, D_0 may all be influenced by some unobservable variable ϵ. This is not ruled out by the independence assumption A2. However, the presence of such an ϵ will typically violate the unconfoundedness assumption.

Examples:

Angrist and Krueger (1991) y is wages, d is yrs of schooling, z is quarter of birth (1=Jan-Aug; 0=Sept-Dec). Exploits two institutional features: (i) can only enter school (kindergarten) when you are 5 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.

Hence, for all kids born in say 2000, those born before 9/1/2000 (tagged $z = 1$) started school a year earlier, and will be in tenth grade when they are allowed to drop out. Those born after 9/1/2000 (tagged $z = 0$) started school a year later, and will only be in ninth grade when they are allowed to drop out.[2]

For this case, the LATE measures the 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.

Angrist (1990) y is lifetime income, d 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, d is price, and z is weather variable.

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

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. For example, d is often a policy change which affects some states but not others.

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.

Abstractly, consider outcome variables indexed by the triple (i, t, d), with $i, t, d \in \{0, 1\}$ (all binary). Here i denotes a subsample, with $i = 1$ being the treated subsample. t denotes time period, with $t = 1$ denoting the period when individuals in subsample $i = 1$ are treated. d is the treatment variable, as before. Of the eight possible combinations, we only observe $Y_{000}, Y_{010}, Y_{100}, Y_{111}$.

- Common trend: $E[Y_{110} - Y_{010}] = E[Y_{100} - Y_{000}] = \alpha$.

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

- The DID estimator is:

$$DID = E[Y_{111} - Y_{100}] - E[Y_{010} - Y_{000}]$$

- Under the common trend assumption,

$$\begin{aligned} DID &= E[Y_{111}] - (E[Y_{000}] + \alpha) - (E[Y_{110}] - \alpha) + E[Y_{000}] \\ &= E[Y_{111} - Y_{110}] \end{aligned}$$

  which is the treatment effect on the treated (see Fig. 1).

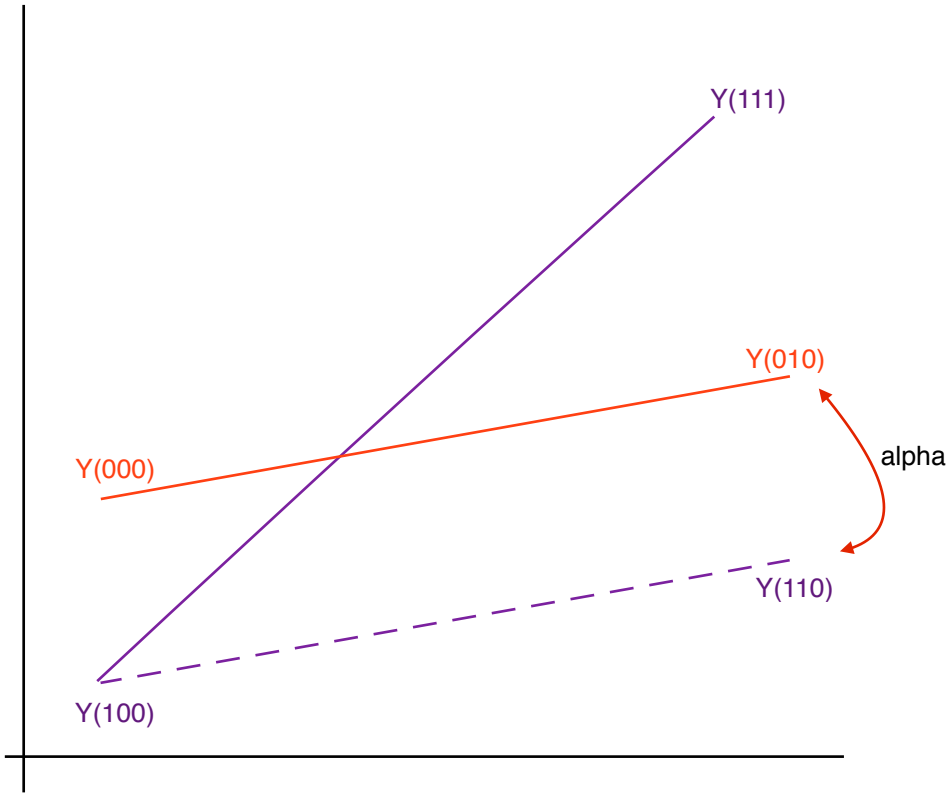

Figure 1: Difference in difference: illustration

The DID is typically obtained by linear regression. Consider the following linear model:

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

with $\epsilon \perp d$. In first differences, this is:

$$\Delta y_i = \beta \Delta d_i + (\gamma_1 - \gamma_0) + \eta_i$$

with $\eta \perp \Delta d_i$. By running this regression, the estimated $\hat{\beta}$ is an estimate of the DID.

In the regression context, it is easy to control for additional variables Z_{it} which also affect outcomes.

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

Card and Krueger (1994) y is employment, d 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, d 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.

Rosenbaum, P. R., and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.