Econometric Methods for Program Evaluation, ARE, Abadie and Cattaneo (2018)

Introduction

- Using economic model of agent behaviour to extrapolate outside of the support of the data ex-ante is known as structural analysis. Otherwise ex-post analysis of policies usually falls under reduced form analysis.
- This paper focuses on discussing program evaluation for ex post reduced form analysis.

Causality and Potential Outcome Model

- POM writes the realized outcome in terms of the potential outcome and treatment

$\hookrightarrow Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$, $\begin{cases} D_i \in \{0,1\} \text{ is binary treatment} \\ (Y_{0i}, Y_{1i}) \text{ are potential outcomes} \end{cases}$

- SUTVA rules out interference between units. This is violated in educational setting with peer effects as realized outcome depends on treatment status of there units.
- The unit level causal effect is not identified from the data.

$\hookrightarrow Y_{1i} - Y_{0i}$ where (Y_{0i}, Y_{1i}) are potential outcomes

- The causal estimands are generally joint distributions of potential outcomes, treatment, and covariates. The estimand is causal when it just depends on the distribution of the potential outcomes.
- The ATE and ATT only depend on the distribution of the potential outcomes, and hence they are causal estimands.

$$\hookrightarrow \begin{cases} ATE = E(Y_{1i} - Y_{0i}) \\ ATT = E(Y_{1i} - Y_{0i} \mid D_i = 1) \end{cases}, \text{ where } D_i \in \{0,1\}$$

- If potential outcomes are not independent of the treatment, than neither the ATE or ATT is identified from the mean outcome differences across control and treatment.
- Confounding arises when treatment selection is dependant on potential outcome.
- Let us consider the difference in mean outcome between control and treatment:

$$\tau = E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) \implies$$

$$\text{(1)} \quad \tau = E(Y_{1i} - Y_{0i} \mid D_i = 1) + [E(Y_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)]$$
$$= ATT + Bias_{ATT}$$

$$\text{(2)} \quad \tau = E(Y_{1i} - Y_{0i}) + E(Y_{0i}) - E(Y_{1i}) + E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)$$
$$= ATE + Pr(D_i = 1)\, ATU + (1 - Pr(D_i = 1))\, ATT$$
$$= ATE + Bias_{ATE}$$

- In general, neither the ATE and ATT are identified in observational data as shown above.

- A DAG is a visual summary of all the causal effects of interest, nodes are random variables, directed edges represent causal effects.

DAG Ex. $U \to W \to Y \Rightarrow$ { W does not cause Y; W correlated with Y }

- Above DAG is a good example of the cliche "correlation does not imply causation". W does not cause Y, but W is correlated to Y because the confounder U effects both W and Y.

Randomized Experiments

- Randomization of treatment implies treatment is independent of potential outcomes.

↳ $D_i \perp (Y_{0i}, Y_{1i})$

- With randomization, the bias terms for ATT and ATE are both 0 and hence the ATE is identified from comparing mean outcomes across control and treatment.
- Note that the ATT = ATE under randomization of the treatment.

↳ $ATE = ATT = E(Y_i|D_i=1) - E(Y_i|D_i=0) = E(Y_{1i} - Y_{0i})$, with $D_i \perp (Y_{0i}, Y_{1i})$

- The ATE is a function of the marginal distribution of the potential outcomes, they are both identified under randomization.

↳ $\begin{cases} F_{Y_1}(y) = Pr(Y_1 \le y) = Pr(Y \le y | D=1) \\ F_{Y_0}(y) = Pr(Y_0 \le y) = Pr(Y \le y | D=0) \end{cases}$

Identified since D is randomized

- Since the potential outcomes marginal distributions are identified from the data, the quantile treatment effect is also identified.

Identified quantile dbn $\begin{cases} Q_{Y_1}(\theta) = Q_{Y|D=1}(\theta) \\ Q_{Y_0}(\theta) = Q_{Y|D=0}(\theta) \end{cases}$, $\theta \in (0,1)$ quantile of Y

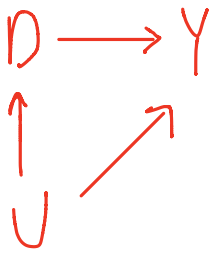
$\Rightarrow$ Quantile Treat. Effect = $Q_{Y_1}(\theta) - Q_{Y_0}(\theta)$

Selection on Observables

- Suppose we are interested in learning the impact of receiving a scholarship on college grades. This is difficult to learn as those who receive scholarships will be of higher ability than those who don't and may receive higher grades even if they had not gotten the scholarship.

- The following DAG represents the above problem:

$$\begin{cases} Y = \text{college grades} \\ D = \text{Recived scholarship} \\ U = \text{Ability} \end{cases} \Rightarrow$$



- The direct causal effect we are interested in is from D to Y, but it is confounded by U.

↳ Paths $\begin{cases} \text{Direct: } D \to Y \\ \text{Backdoor: } D \leftarrow U \to Y \end{cases}$

- We can block the backdoor path by conditioning on U. We can only do this if U is observed.
- Above example motivates the conditional independence assumption (CIA):

↳ $CIA: (Y_{1i}, Y_{0i}) \perp D_i \mid X_i$

- Under the CIA the conditional average treatment effect is identified:

$$E(Y_{1i} - Y_{0i} \mid X_i) \underset{\text{CIA}}{=} E(Y_i \mid X_i, D_i = 1) - E(Y_i \mid X_i, D_i = 0)$$

- CIA along the common support assumption can identify the ATT and ATE:

Assume $\begin{cases} CIA: (Y_{0i}, Y_{1i}) \perp D_i \mid X_i \\ \text{common support}: P(D_i = 1 \mid X_i) \in (0,1) \end{cases} \Rightarrow$

① $ATT = E_{X \mid D=1}\left[E(Y_i \mid X_i, D_i = 1) - E(Y_i \mid X_i, D_i = 0)\right]$

② $ATE = E_X\left[E(Y_i \mid X_i, D_i = 1) - E(Y_i \mid X_i, D_i = 0)\right]$

- Note that both the ATT and ATE are just weighted averages of the within covariate cell treatment effects. We can see this more clearly by letting X take on a finite number of values X1, ..., Xm:

① $ATT = \sum_{R=1}^{m}\left[E(Y_i \mid X_i = X_R, D_i = 1) - E(Y_i \mid X_i = X_R, D_i = 0)\right] P(X_i = X_R \mid D_i = 1)$

$$② \; ATE = \sum_{R=1}^{m} \left[E(Y_i \mid X_i = X_R, D_i = 1) - E(Y_i \mid X_i = X_R, D_i = 0) \right] P(X_i = X_R)$$

- It is noteworthy to mention that regressing Y on D and X and assuming CIA will allow us to recover either the ATT or ATE.

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i \Rightarrow \tau \in \{ATE, ATT\}$$

- The coefficient on D will represent the average of the within covariate cell treatment effects weighted by the variance of the treatment.

$$\tau = \sum_{R=1}^{m} \left[E(Y_i \mid X_i = X_R, D_i = 1) - E(Y_i \mid X_i = X_R, D_i = 0) \right] w_R$$

$$\text{where} \quad w_R = \frac{V(D_i \mid X_i = X_R) P(X_i = X_R)}{\sum_{r=1}^{m} V(D_i \mid X_i = X_r) P(X_i = X_r)}$$

$$\text{with} \quad V(D_i \mid X_i = X_R) = P(D_i = 1 \mid X_i = X_R)\left(1 - P(D_i = 1 \mid X_i = X_R)\right)$$

- The OLS will associate covariate cells that have the most common support across control and treatment group with the largest weights. These weights lower the variance of estimator.
- We can show the ATT within covariate cell effects are weighted by the probability of treatment:

$$ATT = \sum_{R=1}^{m} \left[E(Y_i \mid X_i = X_R, D_i = 1) - E(Y_i \mid X_i = X_R, D_i = 0) \right] w_R^{ATT}$$

$$w_R^{ATT} = P(X_i = X_R \mid D_i = 1) \overset{\text{Bayes}}{=} \frac{P(D_i = 1 \mid X_i = X_R) P(X_i = X_R)}{\sum_{r=1}^{m} P(D_i = 1 \mid X_i = X_r) P(X_i = X_r)}$$

Matching Estimators

- Although OLS cannot be used to estimate the ATT or ATE under CIA, matching estimators can estimate these causal parameters of interest.
- Matching involves pairing each observation i with another observation j(i) that is of the opposite treatment group but with very similar covariates.

$$\text{For each } i, \text{ find } j(i) \text{ such that } \begin{cases} D_{j(i)} = 1 - D_i \\ X_{j(i)} \approx X_i \end{cases}$$

- The matching pairs can be used to identify both the ATT and ATE:

① $ATT = \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - Y_{j(i)})$

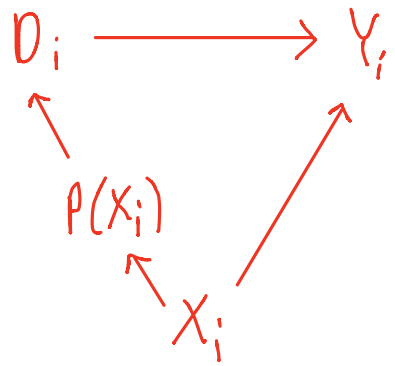
② $ATE = \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - Y_{j(i)}) - \frac{1}{n}\sum_{i=1}^{n}(1-D_i)(Y_i - Y_{j(i)})$

- Another type of matching is known as "prosperity score matching". The propensity score is:
  - ↳ $P(X_i) = P(D_i = 1 \mid X_i)$
- The prosperity score essentially summarizes the information contained in the covariates. That is if CIA is attained from conditioning on X then CIA is also obtained from conditioning on P(X):

$$\text{If } (Y_{1i}, Y_{0i}) \perp D_i \mid X_i \Rightarrow (Y_{1i}, Y_{0i}) \perp D_i \mid P(X_i)$$

- The DAG below illustrates the above implication:

$\begin{cases} Y_i = \text{Outcome} \\ D_i = \text{Binary treatment} \\ X_i = \text{Covariate} \end{cases} \Rightarrow$



- The ATE can be identified using propensity score matching:

$$ATE = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{P(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{(1-P(X_i))}$$

- Observations with large P(X) are overrepresented in the treatment group, so they are weighted down when the treatment group, and weighted up when in control group. The vice-versa is true for small P(X).

Differences-in-Differences

- DD model is identified after making certain assumptions on how unobserved confounders effects the outcome over time.
- Consider the standard panel data model:

$$Y_{it} = D_{it}\tau_{it} + \mu_i + \delta_t + \varepsilon_{it}, \quad D_{it} \in \{0,1\}$$

- The potential outcomes in this panel data setting are:
  - ↳ $\begin{cases} Y_{0it} = \mu_i + \delta_t + \varepsilon_{it} \\ Y_{1it} = \tau_{it} + \mu_i + \delta_t + \varepsilon_{it} \end{cases} \Rightarrow \tau_{it} = Y_{1it} - Y_{0it}$

$$\Rightarrow \text{POM: } Y_{it} = D_{it}Y_{1it} + (1-D_{it})Y_{0it}$$

- For simplicity assume there are two periods: (t = 0, pre-treatment) and (t = 1, post-treatment).

↳ $t \in \{0,1\}$ and $D_{i0} = 1$ for all i

- The unit FE is the confounders and is related to the treatment, and assume changes in unobservable are mean independent of treatment status.

$$\text{Assume} \begin{cases} Cov(\mu_i, D_{i1}) \neq 0 \\ E(\varepsilon_{it} | D_{it}) = E(\varepsilon_{it}) \text{ or } E(\Delta\varepsilon_{i1} | D_{it}) = E(\Delta\varepsilon_{i1}) \end{cases} \quad \varepsilon_{i1} - \varepsilon_{i0} = \Delta\varepsilon_{i1}$$

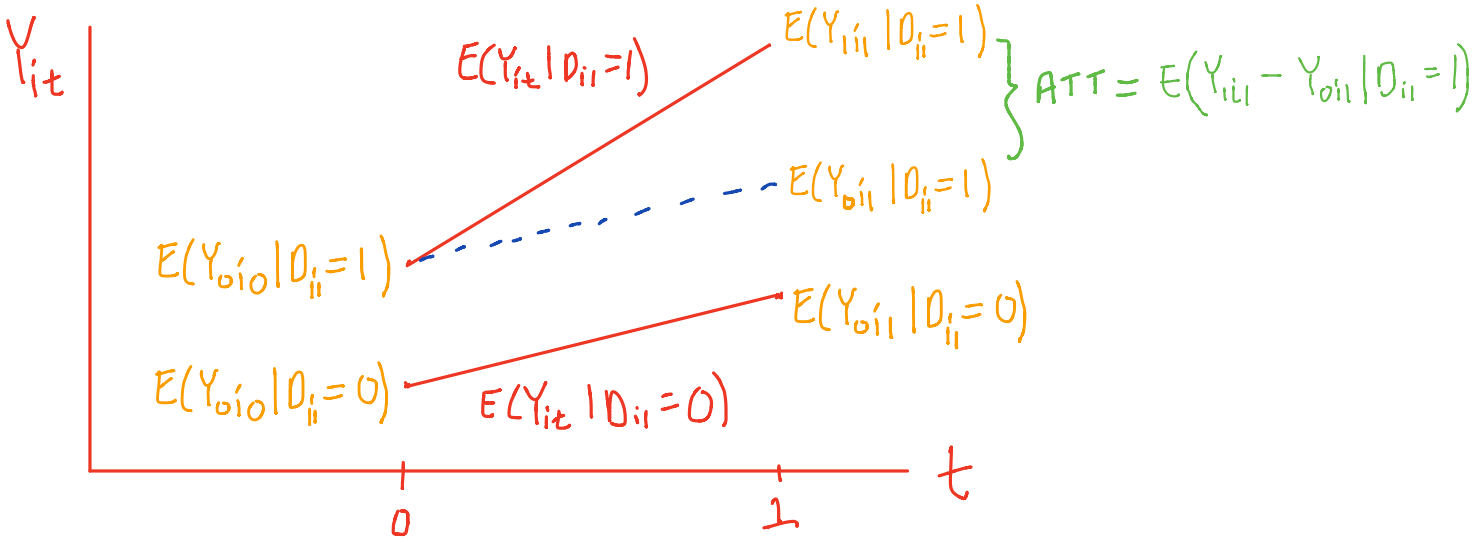
- Under the above assumptions, the ATT is identified using a diff-in-diff estimator:

$$\text{Post Treat} \begin{cases} E(Y_{i1} | D_{i1}=1) = E(\tau_{i1} | D_{i1}=1) + E(\mu_i | D_{i1}=1) + \delta_1 + E(\varepsilon_{i1} | D_{i1}=1) \\ E(Y_{i1} | D_{i1}=0) = E(\mu_i | D_{i1}=0) + \delta_1 + E(\varepsilon_{i1} | D_{i0}=0) \end{cases}$$

$$\text{Pre Treat} \begin{cases} E(Y_{0i} | D_{i1}=1) = E(\mu_i | D_{i1}=1) + \delta_0 + E(\varepsilon_{i0} | D_{i1}=1) \\ E(Y_{0i} | D_{i1}=0) = E(\mu_i | D_{i1}=0) + \delta_0 + E(\varepsilon_{i0} | D_{i1}=0) \end{cases}$$

$$\Rightarrow DD = E(\Delta Y_{i1} | D_{i1}=1) - E(\Delta Y_{i1} | D_{i1}=1)$$
$$= E(\tau_{i1} | D_{i1}=1) + E(\Delta\varepsilon_{i1} | D_{i1}=1) - E(\Delta\varepsilon_{i1} | D_{i1}=0)$$
$$= E(Y_{1i} - Y_{0i} | D_{i1}=1) = ATT$$

- The differences-in-differences intuition is illustrated in the figure below:



- We can use the above figure to define the common trends assumption:

Common trends: $E(Y_{0i1} - Y_{0i0} \mid D_{i1} = 0) = E(Y_{0i1} - Y_{0i0} \mid D_{i1} = 1)$

↳ Implied by $E(\Delta \varepsilon_{i1} \mid D_{it}) = E(\Delta \varepsilon_{i1})$

- Under absence of the treatment, the average change in the outcome across the control and treatment group is the same over time.
- The common trends assumption is not invariant in nonlinear transformation of the outcome.
- That is, the identification of differences-in-differences is dependant on the functional form of the outcome. The DD identified in Y may not be identified in log(Y).
- Plausibility of common trends assumption can be evaluated using pre-treatment periods by checking for identical trends across control and treatment group.

Synthetic Control

- Let j denote the unit, and t be the time period. Assume first unit is exposed to intervention at T0, rest of the units are control group. Also suppose all units have k pre-treat characteristics.

X_1 = pre-treat charac. for treated unit $(K \times 1)$

X_0 = pre-treat charac. for control unit $(K \times J)$

- Synthetic control is essentially a weighted average of untreated units such that they match up with the treated unit prior to the treatment.

$$\min_w \|X_1 - X_0 w\| = \min_w \sqrt{(X_1 - X_0 w)' V (X_1 - X_0 w)'}$$

↳ V weights matches X_1 & $X_0 w$

- The ATT for T > T0 is estimated as follows (compare observed outcome to synthetic control):

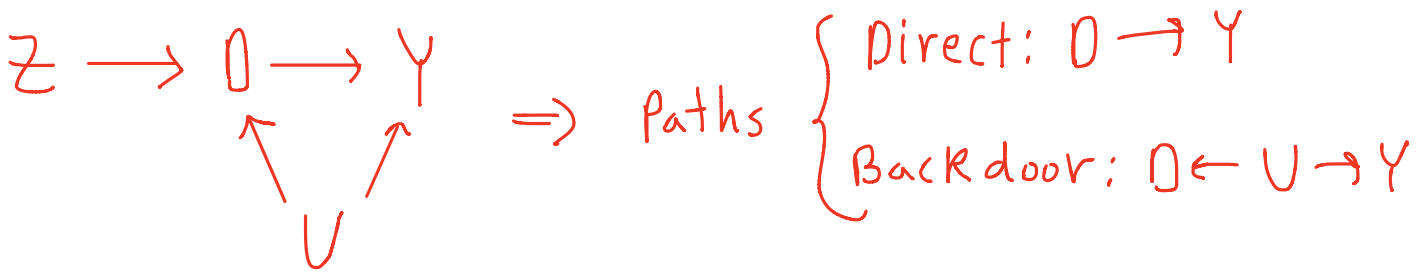
$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \quad \text{for} \quad t \geq T_0$$

Instrumental Variables

- IVs are best motivated for experiments without perfect compliance. Although the treatment assignment is randomized, treatment participation is still endogenous.
- Assignment effects by themselves are known as Intent-to-treat (ITT) effects:

$$ITT = E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)$$

- The ITT are relevant for interventions that are focused on evaluating the effect of providing units access to the treatment.
- The following DAG describes the standard IV setting:



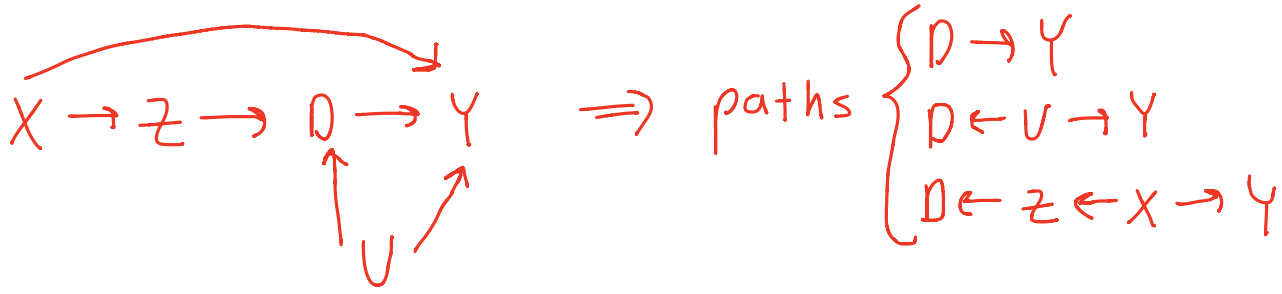
- In the above DAG U is confounder that makes it difficult to measure the direct causal effect of the treatment on the outcome. Z is a instrument effects the outcome only through the treatment.
- The IV uses exogenous variation in D that is induced by Z to estimate the direct effect of D on Y.
- The binary treatment can be expressed using a potential treatment model:
  - $D_i = Z_i D_{1i} + (1 - Z_i) D_{0i}$, where (D_{0i}, D_{1i}) are potential treat.
- One sided compliance is the control group cannot get the treatment, but not everyone in the treatment group participates in the assigned treatment.
  - Perf. Compliance: $Pr(D_{1i} = 1) = Pr(D_{0i} = 0) = 1$
  - One sided: $Pr(D_{0i} = 0) = 1$ but $Pr(D_{1i} = 1) < 1$
- In general, the population can be partitioned into the following groups:
  - pop^n
    - Compliers: $D_{1i} > D_{0i}$ $(D_{1i} = 1 \text{ & } D_{0i} = 0)$
    - Always-takers: $D_{1i} = D_{0i} = 1$
    - Never-takers: $D_{1i} = D_{0i} = 0$
    - Defiers: $D_{1i} < D_{0i}$ $(D_{1i} = 0 \text{ & } D_{0i} = 1)$
- The exclusion restriction says the potential outcome don't directly depend on the treatment assignment. The IV only effects the outcome through its impact on the treatment.
- The independence assumption is implied by the randomization of the IV. The exclusion and independence assumption implies:
  - Exclusion, Indep. $\Rightarrow (Y_{0i}, Y_{1i}, D_{1i}, D_{0i}) \perp Z_i$
- The IV parameter is:

$$\tau_{IV} = \frac{Cov(Y, Z)}{Cov(D, Z)} = \frac{E[(Y_{1i} - Y_{0i})(D_{1i} - D_{0i})]}{E(D_{1i} - D_{0i})}$$

- Under heterogeneous treatment effects and additionally imposing monotonicity (assignment to treatment makes unit more likely to be treated) identifies the average treatment effects for compliers.

$$\text{Monotonicity}: Pr(D_{1i} \geq D_{0i}) \Rightarrow LATE = E(Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i})$$

- Under one sided non-compliance, the LATE is equivalent to the ATT.
- There may be observed variables that effect the instrument and outcome. This model is represented in the DAG below.



- The IV Z above is not valid unless we condition on X. The LATE after condition on X is:

$$LATE = \frac{E_X\left[E(Y_i \mid X_i, Z_i = 1) - E(Y_i \mid X_i, Z_i = 0)\right]}{E_X\left[Pr(D_i = 1 \mid X_i, Z_i = 1) - Pr(D_i = 1 \mid X_i, Z_i = 1)\right]}$$

Regression Discontinuity Design

- The RD design relies on their being discontinuous change in the treatment after the value of the running variable crosses some threshold.

$$\begin{cases} X = \text{running variable} \\ c = \text{threshold} \end{cases} \Rightarrow D_i = I(X_i \geq c) = \begin{cases} 1, \text{ if } X_i \geq c \\ 0, \text{ if } X_i < c \end{cases}$$

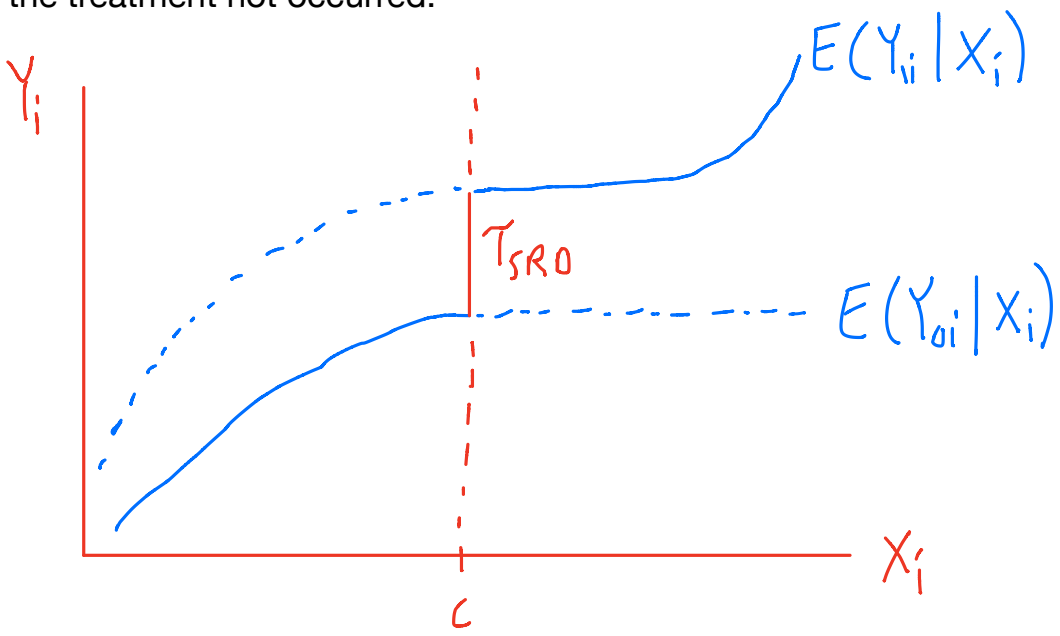
- There are three types of RD designs 1) Sharp RD, 2) Fuzzy RD, and 3) Kink RD. The first two are the most popular.
- The idea behind the RD design is that the treatment status is essential random around the threshold at which the treatment status changes.
- The sharp RD parameter of interest is:

$$\tau_{SRD} = E(Y_{1i} - Y_{0i} \mid X_i = c) = \lim_{x \to c^-} E(Y_i \mid X_i = x) - \lim_{x \to c^+} E(Y_i \mid X_i = x)$$

- The above sharp RD parameter is nonparametrically identified if the conditional expectation of the potential outcomes are continuous at cutoff.

$$\text{Assume} \begin{cases} E(Y_{1i} \mid X_i = x) \\ E(Y_{0i} \mid X_i = x) \end{cases} \text{ are cont. at } x = c$$

- The requirement of the continuity assumption becomes clear in the figure below. It allows us to compare observed mean outcome post-treatment (just to the right of c) to the counterfactual had the treatment not occurred:



- They key RD identifying assumptions can be supported by empirical tests. A common falsification test is to show there is no discontinuity in covariates at the cutoff.
- RD assumes there is no manipulation of individuals at the threshold (individuals can't control which side they end up near the cutoff). This assumption can be supported by showing there is no discontinuity in the distribution of the running variable at the cutoff.
- The most common estimation strategy is a weighted local-linear regression. This involves 1) using data points only near the cutoff and 2) regression that weights points closer to the cutoff.

$$\min_{\beta_0,\beta_1,\beta_2,\tau} \sum_{i=1}^{n} \left(Y_i - \beta_0 - D_i\tau - (X_i - c)\beta_2 - (X_i - c)D_i\beta_3\right)^2 k\left(\frac{X_i - \bar{X}}{h}\right)$$

$$\text{subject to } X_i \in [c-h, c+h],\ h = \text{bandwidth}$$

Conclusion

- Paper discusses the most common empirical tools for program evaluation: 1) Experiments, 2) Conditional Independence, 2) Instrumental Variables, 3) Differences-in-differences, and 4) Regression discontinuity.
- However paper is not exhaustive of all policy tools, see Athey and Imbens (2017) as a nice complement to this paper.