Empirical Methods in Labour Economics

An Introduction

EC317

Lent Term 2023 — Class 1

The Evaluation Problem
Randomised Experiments
Regression
Matching
Instrumental Variables
Fixed Effects
Difference-in-Differences
Regression Discontinuity Design
References

The Evaluation Problem

*"The path to causal understanding is rough and shadowed as it snakes around the boulders of **selection bias**. And yet, masters of 'metrics walk this path with confidence as well as humility, successfully linking cause and effect."*

Angrist & Pischke (2015), Intro.

The Evaluation Problem

- Labour economists are typically interested in using data to answer *causal* questions of the form
  - What are the wage returns to education?
  - How do minimum wages affect wages and employment?
  - Do unions raise wages?
  - Does education reduce crime?
- These are all *counterfactual* questions: How much would outcome Y_i change if the variable of interest D_i were to increase?

The Evaluation Problem

- To make matters simple, suppose that the variable of interest is binary, i.e., $D_i \in \{0, 1\}$.
- Then, we can think of two potential states of the world concerning individual i:
  0. $D_i = 0$ and i's realised outcome is Y_{0i}
  1. $D_i = 1$ and i's realised outcome is Y_{1i}
- We call Y_{0i} and Y_{1i} the *potential outcomes* corresponding to states $D_i = 0$ and $D_i = 1$, respectively.

The Evaluation Problem

- Notice that it is impossible for the same individual to have $D_i = 0$ and $D_i = 1$ at the same time.
- We do not know the individual causal effect $Y_{1i} - Y_{0i}$ since we only observe

$$\begin{aligned} Y_i &= (1 - D_i)\, Y_{0i} + D_i\, Y_{1i} \\ &= Y_{0i} + D_i(\, Y_{1i} - Y_{0i}). \end{aligned}$$

- Similarly, we cannot directly estimate $\mathbb{E}\,[\, Y_{0i} \mid D_i = 1]$ and $\mathbb{E}\,[\, Y_{1i} \mid D_i = 0]$ from observed data $(\, Y_i, D_i)$.

The Evaluation Problem

- What we can directly estimate is

$$\mathbb{E}\left[Y_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i \mid D_i = 0\right] = \mathbb{E}\left[Y_{1i} \mid D_i = 1\right] - \mathbb{E}\left[Y_{0i} \mid D_i = 0\right]$$

- Yet, the answers to our causal questions of interest lie in parameters such as
  - Average treatment effect (ATE): $\mathbb{E}\left[Y_{1i} - Y_{0i}\right]$
  - Average treatment effect on the treated (ATT): $\mathbb{E}\left[Y_{1i} - Y_{0i} \mid D_i = 1\right]$
  - Average treatment effect on the untreated (ATU): $\mathbb{E}\left[Y_{1i} - Y_{0i} \mid D_i = 0\right]$

The Evaluation Problem

- Simple comparisons of observed outcomes are subject to *selection bias*:

$$\mathbb{E}\left[Y_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i \mid D_i = 0\right] = \mathbb{E}\left[Y_{1i} \mid D_i = 1\right] - \mathbb{E}\left[Y_{0i} \mid D_i = 0\right]$$

$$= \underbrace{\mathbb{E}\left[Y_{1i} - Y_{0i} \mid D_i = 1\right]}_{\text{average treatment effect on the trated (ATT)}}$$

$$+ \underbrace{\left(\mathbb{E}\left[Y_{0i} \mid D_i = 1\right] - \mathbb{E}\left[Y_{0i} \mid D_i = 0\right]\right)}_{\text{selection bias}}$$

- Empirical methods try to solve this problem.

Randomised Experiments

*"Random assignment isn't the same as holding everything else fixed, but it has the same effect. Random manipulation makes **other things equal** hold **on average** across the groups that did and did not experience manipulation."*

Angrist & Pischke (2015), Intro.

RCT: Random Assignment

- **Random assignment:**
  - A framework for answering causal questions.
  - A benchmark by which results from other methods are judged.
- **Randomised Trial:**
  - Experimental samples are created by sampling from a population of interest.
  - Sampled subjects are randomly divided (as if by a coin toss) into treatment and control groups.

RCT: Random Assignment Solves the Selection Problem

- Intuitively:
  - Randomly assigned treatment and control groups come from the same underlying populations.
  - The *Law of Large Numbers* promises that those individuals in these groups will be similar if the samples are large enough.
- Formally:
  - Random assignment makes D_i independent of potential outcomes.
  - Therefore, $\mathbb{E}[Y_{0i} \mid D_i = 1] = \mathbb{E}[Y_{0i} \mid D_i = 0]$, so selection bias is zero.
  - Moreover, ATE $=$ ATT $=$ ATU.

RCT: Estimating ATE

- We can consistently estimate the ATE by simple differences of sample means:

$$\widehat{\rho} = \frac{1}{N_1} \sum_{i \in \mathcal{D}_1} Y_{1i} - \frac{1}{N_0} \sum_{i \in \mathcal{D}_0} Y_{0i}$$

where $\rho \equiv \mathbb{E}\left[Y_{1i} - Y_{0i} \right]$ is the ATE, $\mathcal{D}_1$ represents the set of N_1 individuals in the treatment group, and $\mathcal{D}_0$ represents the set of N_0 individuals in the control group.

RCT: Estimating ATE

- Alternatively, we can run a linear regression to obtain a numerically identical estimate plus its standard error

$$\begin{aligned}
\mathbb{E}\left[Y_i \mid D_i\right] &= \mathbb{E}\left[Y_{i0} \mid D_i\right] + \mathbb{E}\left[Y_{1i} - Y_{i0} \mid D_i\right] D_i \\
&= \mathbb{E}\left[Y_{0i}\right] + \mathbb{E}\left[Y_{1i} - Y_{0i}\right] D_i \\
&\equiv \alpha + \rho D_i \\
\implies Y_i &= \alpha + \rho D_i + \varepsilon_i
\end{aligned}$$

where $\varepsilon_i \equiv Y_i - \mathbb{E}\left[Y_i \mid D_i\right]$.

Regression

*"Regression-based causal inference is predicated on the assumption that when key **observed variables** have been **made equal** across treatment and control groups, selection bias from the things we can't see is also mostly eliminated."*

Angrist & Pischke (2015), Ch. 2

Regression: The CEF

- **Conditional Expectation Function:** The CEF for dependent variable Y_i given $K \times 1$ covariate vector X_i is the population mean of Y_i with X_i held fixed. That is,

$$\mathbb{E}\left[Y_i \mid X_i = x \right] = \begin{cases} \int s \, f_{Y|X}(s \mid X_i = x) \, ds & \text{for continuous } Y_i \\ \sum_s s \, \mathbb{P}(Y_i = s \mid X_i = x) & \text{for discrete } Y_i \end{cases}$$

- It is a function of X_i and, because X_i is random, the CEF is random.

Regression: Properties of the CEF

- **Decomposition property:** Any random variable Y_i can be decomposed into a piece that is "explained by X_i" —the CEF— and an error term that is uncorrelated with any function of X_i:

$$Y_i = \mathbb{E}\left[Y_i \mid X_i\right] + \varepsilon_i$$

- **Prediction property:** The CEF is the best predictor of Y_i given X_i in the minimum mean squared error (MMSE) sense:

$$\mathbb{E}\left[Y_i \mid X_i\right] = \underset{m(X_i)}{\arg\min}\ \mathbb{E}\left[\left(Y_i - m(X_i)\right)^2\right]$$

Regression: Population Linear Regression

- **Population Regression Coefficients:** Solution to the population least squares problem

$$\beta = \arg\min_b \mathbb{E}\left[\left(Y_i - X_i'b\right)^2\right]$$

$$= \mathbb{E}\left[X_i X_i'\right]^{-1} \mathbb{E}\left[X_i Y_i\right]$$

- **Population residual:**

$$e_i \equiv Y_i - X_i'\beta$$

- e_i is uncorrelated with X_i by construction: FOC is $\mathbb{E}\left[X_i\left(Y_i - X_i'\beta\right)\right] = 0$.

Regression: Linear Regression and the CEF

- The CEF is a good summary of the relationship between Y_i and X_i, and regression is tightly linked to the CEF.

- **Linear CEF:** If $\mathbb{E}[Y_i \mid X_i] = X_i'\beta^*$, then the population regression function is the CEF since, by the CEF decomposition,

$$\mathbb{E}\left[X_i\left(Y_i - X_i'\beta^*\right)\right] = 0 \iff \beta^* = \mathbb{E}\left[X_i'X_i\right]^{-1}\mathbb{E}\left[X_i\,Y_i\right] = \beta$$

Regression: Linear Regression and the CEF

- **Best linear predictor:** just as the CEF is the MMSE predictor of Y_i given X_i among **all** functions, $X_i'\beta$ is the MMSE predictor among **linear** functions, since β solves the population least squares problem.

- **Best linear approximation:** It can be shown that, even if the CEF is nonlinear, $X_i'\beta$ provides the MMSE **linear** approximation to $\mathbb{E}[Y_i \mid X_i]$:

$$\beta = \arg\min_b \mathbb{E}\left[\left(\mathbb{E}[Y_i \mid X_i] - X_i'b\right)^2\right]$$

Regression: The OLS Estimator

- The Ordinary Least Squares (OLS) estimator is the sample analog of the population regression coefficients

$$\widehat{\beta}_{OLS} = \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i Y_i \right)$$

- It receives its name because it solves the sample analog of the population least squares problem:

$$\widehat{\beta}_{OLS} = \arg\min_{b} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - X_i' b \right)^2$$

Regression: Causality

- When can we think of a regression coefficient as approximating the causal effect that would be revealed by a randomised trial?
  - A regression is causal when the CEF it approximates is causal.
  - A CEF is causal when it describes differences in average potential outcomes for a fixed reference population.

Regression: Causality

- Consider the regression of Y_i on a constant and our variable of interest D_i.
- We have already discussed the selection bias problem

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \rho_{\text{ATT}} + (\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0])$$

where $\rho_{\text{ATT}} \equiv \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$ is the ATT.

- Without random assignment (as in a randomised trial), we cannot rely on independence between D_i and potential outcomes to eliminate selection bias.

Regression: Causality

- **Conditional Independence Assumption:** The CIA (also known as the *selection on observables* assumption) states that

$$D_i \text{ is independent of } Y_{1i} \text{ and } Y_{0i} \text{ conditional on } X_i$$

- Under the CIA, selection bias disappears when we compare outcomes of people with the same value of X_i

$$\begin{aligned}\mathbb{E}\left[Y_{0i} \mid D_i = 1, X_i\right] - \mathbb{E}\left[Y_{0i} \mid D_i = 0, X_i\right] &= \mathbb{E}\left[Y_{0i} \mid X_i\right] - \mathbb{E}\left[Y_{0i} \mid X_i\right] \\ &= 0\end{aligned}$$

Regression: Causality

- Under the CIA, treatment-control comparisons of average outcomes for people with the same value of X_i have a causal interpretation

$$\mathbb{E}\left[Y_i \mid D_i = 1, X_i\right] - \mathbb{E}\left[Y_i \mid D_i = 0, X_i\right] = \mathbb{E}\left[Y_{1i} - Y_{0i} \mid X_i\right]$$

$$\equiv \rho(X_i)$$

- But this is the causal effect for a specific value of X_i, so there are as many as distinct values of X_i.

- However, the Law of Iterated Expectations implies that

$$\rho = \mathbb{E}\left[Y_{1i} - Y_{0i}\right] = \mathbb{E}\Big[\mathbb{E}\left[Y_{1i} - Y_{0i} \mid X_i\right]\Big] = \mathbb{E}\left[\rho(X_i)\right]$$

Regression: Causality

- To make matters simple, suppose that

$$Y_{1i} = Y_{0i} + \rho$$

$$\mathbb{E}\left[Y_{0i} \mid X_i\right] = \alpha + X_i'\beta$$

- Then,

$$\begin{aligned}\mathbb{E}\left[Y_i \mid D_i, X_i\right] &= \mathbb{E}\left[Y_{0i} \mid D_i, X_i\right] + \mathbb{E}\left[Y_{1i} - Y_{0i} \mid D_i, X_i\right] D_i \\ &= \mathbb{E}\left[Y_{0i} \mid X_i\right] + \mathbb{E}\left[Y_{1i} - Y_{0i} \mid X_i\right] D_i \\ &= \alpha + X_i'\beta + \rho D_i\end{aligned}$$

Regression: Causality

- Since the resulting CEF is linear, the population regression coefficients are the CEF coefficients α, ρ, and β.
- Therefore, the population regression coefficients have a causal interpretation.
- Since $\rho = \operatorname{plim} \widehat{\rho}_{OLS}$ has a causal interpretation, the OLS estimator $\widehat{\rho}_{OLS}$ from regression

$$Y_i = \alpha + X_i'\beta + \rho D_i + \varepsilon_i$$

consistently estimates a causal effect.

Regression: Causality

- The key assumption is that the observable characteristics X_i are the only reason why D_i correlates with $\eta_i \equiv X_i'\beta + \epsilon_i$ in short regression

$$Y_i = \alpha + \rho D_i + \eta_i$$

- This is why the CIA is also known as *selection on observables*.

- This analysis extends (with a bit more work) to the more general case where D_i takes more than 2 values.

Matching

*"The method of matching sorts individuals into groups with the **same values of control variables**. Matched comparisons within these groups are then averaged to get a single overall effect. Regression is an automated matchmaker. The regression estimate of a causal effect is also an average of within-group comparisons."*

Angrist & Pischke (2015), Ch. 2

Matching

- Just as causal regression, matching is motivated/relies on the CIA/selection on observables.
- As such, we won't give much detail here.
- The course will not focus on papers using matching as their identification strategy.

Matching

- Matching amounts to covariate-specific treatment-control comparisons, weighted together to produce a single overall average treatment effect.

- Regression can be motivated as a particular sort of matching estimator, and therefore the differences between regression and matching estimates are unlikely to be of major empirical relevance.

Matching

- As already discussed for regression, the CIA implies that selection bias disappears after conditioning on X_i:

$$\mathbb{E}\left[Y_{0i} \mid D_i = 1, X_i\right] - \mathbb{E}\left[Y_{0i} \mid D_i = 0, X_i\right] = 0$$

- The ATT can be recovered by iterating expectations on X_i:

$$\begin{aligned}
\rho_{\text{ATT}} &= \mathbb{E}\left[Y_{1i} - Y_{0i} \mid D_i = 1\right] \\
&= \mathbb{E}\Big[\mathbb{E}\left[Y_{1i} - Y_{0i} \mid D_i = 1, X_i\right] \Big| D_i = 1\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\left[Y_{1i} - Y_{0i} \mid X_i\right] \Big| D_i = 1\Big] \\
&= \mathbb{E}\left[\rho(X_i) \mid D_i = 1\right]
\end{aligned}$$

Matching

- If covariates X_i are discrete, the matching estimand can be written as

$$\mathbb{E}\left[Y_{1i} - Y_{0i} \mid D_i = 1\right] = \sum_x \rho(x)\, \mathbb{P}\left(X_i = x \mid D_i = 1\right)$$

- The estimator simply replaces $\rho(x)$ by the sample covariate-specific treatment-control outcome differences, and the conditional probability mass function by the empirical distribution of covariates among the treated.

Instrumental Variables

"Just as in randomised trials, the forces of nature, including human nature, sometimes manipulate treatment in a manner that obviates the need for controls. Such forces are rarely the only source of variation in treatment, but this is an obstacle easily surmounted. The instrumental variables method harnesses ***partial or incomplete random assignment****, whether* ***naturally occurring or generated by researchers****."*

Angrist & Pischke (2015), Ch. 3

Instrumental Variables

- To fix ideas, suppose we are interested in the causal effect of schooling D_i on log-wages Y_i.

- Potential outcomes —i.e., the Y_i that would obtain if $D_i = s$ for each possible value s of D_i— are given by

$$Y_{si} = \alpha + \rho s + \eta_i$$

Instrumental Variables

- Suppose that the CIA holds for the causal effect of D_i on Y_i conditional on "ability" variables X_i:

$$\eta_i = X_i'\beta + \varepsilon_i$$

  where β is a vector of population coefficients, so X_i and ε_i are uncorrelated by construction.

- Our *selection on observables* assumption states that X_i are the only reason why D_i correlates with η_i, so

$$\mathbb{E}\left[D_i\varepsilon_i\right] = 0$$

Instrumental Variables

- If we could control for X_i, we would consistently estimate ρ by OLS estimation of the long regression

$$Y_i = \alpha + \rho D_i + X_i'\beta + \varepsilon_i$$

- But X_i is unobserved, so we are left with the regression of Y_i on D_i and a constant.

- Since $\mathbb{E}[D_i \eta_i] \neq 0$, OLS estimation of the short regression does not yield a consistent estimate of ρ.

Instrumental Variables

- Suppose we have access to a variable Z_i that is correlated with D_i but uncorrelated with any other determinants of Y_i, that is

$$\text{Cov}(Z_i, D_i) \neq 0$$

$$\text{Cov}(Z_i, X_i) = \text{Cov}(Z_i, \varepsilon_i) = 0 \iff \text{Cov}(Z_i, \eta_i) = 0$$

- From the long (causal) regression, we see that

$$\begin{aligned}\text{Cov}(Z_i, Y_i) &= \rho \,\text{Cov}(Z_i, D_i) + \underbrace{\text{Cov}(Z_i, X_i'\beta)}_{0 \text{ since } \text{Cov}(Z_i, X_i)=0} + \underbrace{\text{Cov}(Z_i, \varepsilon_i)}_{0} \\ &= \rho \,\text{Cov}(Z_i, D_i)\end{aligned}$$

Instrumental Variables

- Therefore,

$$\rho = \frac{\text{Cov}\left(Z_i,\, Y_i\right)}{\text{Cov}\left(Z_i, D_i\right)}$$

$$= \frac{\frac{\text{Cov}\left(Z_i, Y_i\right)}{\text{Var}\left(Z_i\right)}}{\frac{\text{Cov}\left(Z_i, D_i\right)}{\text{Var}\left(Z_i\right)}}$$

$$\equiv \frac{\gamma}{\pi}$$

where γ and π are the population regression coefficients of Y_i on Z_i (called the *reduced form*) and D_i on Z_i (called the *first stage*), respectively.

Instrumental Variables

- The IV estimator is the sample analog of the ratio of covariances

$$\frac{\text{Cov}\left(Z_i,\, Y_i\right)}{\text{Cov}\left(Z_i,\, D_i\right)}$$

- The assumptions needed for this estimand to equal the causal effect ρ are

1. Z_i is **as good as randomly assigned**, so it is independent of potential outcomes.
2. **Exclusion restriction:** Z_i has no effect on Y_i other than through D_i.
3. **First stage:** Z_i has an effect on D_i.

Instrumental Variables

- We can think of instrumental variables as initiating a causal chain where Z_i affects D_i (through the first stage), which in turn affects Y_i.
- For this chain to capture the causal effect of D_i on Y_i:
  1. First, Z_i must have a clear effect on D_i (the first stage).
  2. Second, the only reason for the relationship between Y_i and Z_i must be the first stage (independence + exclusion restriction).

Instrumental Variables: The Wald Estimator

- Consider the special case where
  - The instrument is binary, i.e. $Z_i \in \{0, 1\}$
  - D_i and η_i may be correlated in the causal linear regression model

$$Y_i = \alpha + \rho D_i + \eta_i$$

Instrumental Variables: The Wald Estimator

- Then,

$$\begin{aligned}
\mathrm{Cov}\left(Z_i,\, Y_i\right) &= \mathbb{E}\left[Z_i\, Y_i\right] - \underbrace{\mathbb{E}\left[Z_i\right]}_{p_z \cdot 1 + (1-p_z)\cdot 0}\ \mathbb{E}\left[Y_i\right] \\
&= \mathbb{E}_z\Big[Z_i\, \mathbb{E}\left[Y_i \mid Z_i\right]\Big] - p_z \mathbb{E}_z\Big[\mathbb{E}\left[Y_i \mid Z_i\right]\Big] \\
&= p_z\, \mathbb{E}\left[Y_i \mid Z_i = 1\right] \\
&\qquad - p_z\Big(p_z\, \mathbb{E}\left[Y_i \mid Z_i = 1\right] + (1-p_z)\, \mathbb{E}\left[Y_i \mid Z_i = 0\right]\Big) \\
&= p_z(1-p_z)\Big(\mathbb{E}\left[Y_i \mid Z_i = 1\right] - \mathbb{E}\left[Y_i \mid Z_i = 0\right]\Big)
\end{aligned}$$

where $p_z \equiv \mathbb{P}(Z_i = 1)$.

Instrumental Variables: The Wald Estimator

- Similarly,

$$\operatorname{Cov}(Z_i, D_i) = p_z(1-p_z)\Big(\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]\Big)$$

- Therefore,

$$\rho = \frac{\operatorname{Cov}(Z_i, Y_i)}{\operatorname{Cov}(Z_i, D_i)}$$

$$= \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]}$$

- The sample analog, i.e. the IV estimator, is the Wald estimator.

Angrist (1990)

- Angrist (1990) uses the Wald estimator in his study of the effects of veteran status on civilian earnings.
- Concerns about the fairness of the U.S. conscription policy during the Vietnam War era led to the institution of a draft lottery in 1970 that was sed to determine priority for conscription.
- In each year from 1970 to 1972, random sequence numbers were randomly assigned to each birth date in cohorts of 19-year-olds. Men with lottery numbers below a cutoff were eligible for the draft, while men with numbers above the cutoff could not be drafted.

Angrist (1990)

- In practice, many draft-eligible men were still exempted for health or other reasons, while many men who were draft-exempt nevertheless volunteered for service. So veteran status was not completely determined by ransomised draft eligibility, but the correlation is strong.
- While veteran status was not completely determined by randomised draft eligibility, they a
- Draft eligibility is a binary instrument for Vietnam veteran status, since it was determined by a lottery over birthdays and is highly correlated with veteran status.

Angrist (1990)

- Angrist (1990) uses instrument

$$Z_i = \mathbb{I}\left[i \text{ had a lottery number above the cutoff}\right]$$

to estimate the effect of veteran status D_i on earnings after the war Y_i.

- The Wald/IV estimator is

$$\widehat{\rho}_{\mathsf{IV}} = \frac{\widehat{\mathbb{E}}\left[Y_i \mid Z_i = 1\right] - \widehat{\mathbb{E}}\left[Y_i \mid Z_i = 0\right]}{\widehat{\mathbb{E}}\left[D_i \mid Z_i = 1\right] - \widehat{\mathbb{E}}\left[D_i \mid Z_i = 0\right]}$$

where $\widehat{\mathbb{E}}\left[\cdot \mid Z_i = z\right]$ represents the sample mean or average over the subsample with $Z_i = z$.

Angrist (1990)

- An important feature of the Wald/IV estimator is that the identifying assumptions are easy to assess and interpret.
- The fundamental claim justifying the interpretation of the Wald estimator as capturing the causal effect of D_i is that the only reason why $\mathbb{E}[Y_i \mid Z_i]$ changes as Z_i changes is the variation in $\mathbb{E}[D_i \mid Z_i]$.
- A simple check is looking at the correlation of Z_i and individual characteristics that should not be affected by Z_i such as race, sex, or any other predetermined characteristic (relative to D_i).

Angrist (1990)

- Another check is looking for correlation between Z_i and Y_i in other samples —samples from other populations— where there is no relationship between Z_i and D_i.
- If the only reason for draft eligibility effects on earnings is veteran status, then draft eligibility effects on earnings should be zero in samples where draft eligibility status is unrelated to veteran status.
- Angrist (1990) looks at earnings in 1969, finding a zero effect of draft eligibility. This is comforting, since 1969 earnings predate the 1970 draft lottery.

Fixed Effects

"Fixed effects estimators are based on the presumption of ***time-invariant*** *(or group-invariant)* ***omitted variables****."*

Angrist & Pischke (2009), Ch. 5

Fixed Effects

- Controlling for *fixed effects* is a strategy that uses data with a time/cohort dimension and repeated individual/group observations along this dimension to control for **unobserved** but **fixed** (within group) omitted variables.
- We will work with an individual i and a time t dimension to fix ideas, but the method is more general in the sense that i and t could be different dimensions of the unit of observation, e.g., twin $j \in \{1, 2\}$ of family/twin-pair f.
- This type of data is typically called *panel* or *longitudinal* data.

Fixed Effects: Model

- Suppose we are interested in the causal effect of union membership status $D_{it} \in \{0, 1\}$ on log-earnings Y_{it} of individual i at time t.

- The observed Y_{it} is potential outcome Y_{1it} when $D_{it} = 1$, and Y_{0it} when $D_{it} = 0$.

Fixed Effects: Model

- Suppose that

$$\mathbb{E}\left[Y_{0it} \mid D_{it}, A_i, X_{it}, t \right] = \mathbb{E}\left[Y_{0it} \mid A_i, X_{it}, t \right]$$

  where X_{it} is a vector of observed time-varying covariates and A_i is a vector of **unobserved but fixed** "ability" confounders.

- This is just a version of the CIA: union status D_{it} is as good as randomly assigned conditional on A_i and X_{it}. The problem is that we do not observe A_i.

Fixed Effects: Model

- The key assumption is that A_i appears without a t-subscript in the linear CEF

$$\mathbb{E}\left[Y_{0it} \mid A_i, X_{it}, t\right] = \alpha + \lambda_t + A_i'\gamma + X_{it}'\beta$$

- We also assume an additive, constant causal effect of D_{it} on Y_{it}:

$$\mathbb{E}\left[Y_{1it} \mid A_i, X_{it}, t\right] = \mathbb{E}\left[Y_{0it} \mid A_i, X_{it}, t\right] + \rho$$

Fixed Effects: Model

- Combining our assumptions, we obtain

$$\mathbb{E}\left[Y_{it} \mid D_{it}, A_i, X_{it}, t\right] = \alpha + \lambda_t + \rho D_{it} + A_i'\gamma + X_{it}'\beta$$

where ρ is the causal effect of interest.

- Finally, using the CEF decomposition property, we obtain our causal linear fixed-effects regression model

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\beta + \varepsilon_{it}$$

with $\varepsilon_{it} \equiv Y_{0it} - \mathbb{E}\left[Y_{0it} \mid A_i, X_{it}, t\right]$ and $\alpha_i \equiv \alpha + A_i'\gamma$.

Fixed Effects: Model

- While we cannot directly control for unobserved A_i, given panel data (repeated observations on individuals), ρ can be consistently estimated by treating the **fixed effect** α_i as an additional parameter to be estimated.
- The time effect λ_t is also treated as a parameter to be estimated.
- But how? ⌣̈

Fixed Effects: Model

- Let $d_{ji} = \mathbb{I}[i = j]$ for $j = 1, \ldots, N$ be a set of individual dummies, one for each worker i.
- Notice that for any given worker i, only one of the N dummies, d_{ii}, equals 1, and the other $N - 1$ equal 0.
- Hence,

$$\sum_{j=1}^{N} \alpha_j d_{ji} = \alpha_i$$

Fixed Effects: Model

- An analogous argument establishes that

$$\sum_{s=1}^{T} \lambda_s d_{st} = \lambda_t$$

where $d_{st} = \mathbb{I}[t = s]$ are time dummies for $s = 1, \ldots, T$ and T is the number of periods (which, for simplicity, we assume equal for all workers).

Fixed Effects: Model

- Combining these two results, our causal linear fixed-effects model can be equivalently written as

$$Y_{it} = \sum_{j=1}^{N} \alpha_j \, d_{ji} + \sum_{s=1}^{T} \lambda_s \, d_{st} + \rho D_{it} + X_{it}' \beta + \varepsilon_{it}$$

- That is, the unobserved individual effects α_i are coefficients on dummies for each individual, while the time effects are coefficients on time dummies.

Fixed Effects: Estimation

- A natural and direct way to proceed is to estimate the dummy version of the model by OLS, including one dummy for each individual and one time dummy for each period (excluding one of each if we include an intercept: *beware the dummy variable trap* ö).
- However, this means adding $N + T$ (or $N + T - 2$ if we include a common intercept) variables and their corresponding coefficients to estimate.
- It turns out that dummy-OLS is algebraically equivalent to OLS in deviations from means.

Fixed Effects: Estimation

- Consider the individual average (which we can compute directly)

$$\overline{Y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} Y_{it}$$

$$= \alpha_i + \overline{\lambda} + \rho\overline{D}_i + \overline{X}'_{it}\beta + \overline{\varepsilon}_i$$

- Deviation from means "kills" the unobserved individual effects:

$$Y_{it} - \overline{Y}_i = \underbrace{(\alpha_i - \alpha_i)}_{0} + (\lambda_t - \overline{\lambda}) + \rho(D_{it} - \overline{D}_i) + (X_{it} - \overline{X}_{it})'\beta + (\varepsilon_{it} - \overline{\varepsilon}_i)$$

Fixed Effects: Estimation

- An alternative to deviations from means is first differences, which amounts to OLS estimation of

$$\Delta Y_{it} = \underbrace{\Delta \alpha_i}_{0} + \Delta \lambda_t + \rho \Delta D_{it} + \Delta X_{it}' \beta + \Delta \varepsilon_{it}$$

where $\Delta(\cdot)_{it} \equiv (\cdot)_{it} - (\cdot)_{it-1}$.

- First-differencing also “kills” the unobserved individual effects.

Fixed Effects: Estimation

- With two periods ($T = 2$), first-differencing is algebraically equivalent to deviations from means, but not otherwise.
- With more than two periods, first-differenced standard errors should be adjusted for the fact that the differenced residuals are serially correlated.
- Similarly, the standard errors of the deviations from means estimator need to be adjusted for the degrees of freedom lost in the estimation of the N individual means.
- The deviations from means estimator has many names, including the "*within estimator*", "*analysis of covariance*", and "*absorbing*" the fixed effects.

Fixed Effects: Estimation

- **Incidental parameters problem:** The fixed effects —i.e., the α_i coefficients— are not estimated consistently in a panel with T fixed and $N \to \infty$ since the number of parameters grows with N.
- However, the other parameters in the model —in particular ρ, the causal effect of interest—are consistently estimated.
- **Attenuation bias:** Attenuation bias from measurement error is exacerbated in the case of fixed-effect estimators.
- A possible solution is to use instrumental variable methods.

Difference-in-Differences

"The difference-in-differences method recognises that, in the absence of random assignment, treatment and control groups are likely to differ for many reasons. Sometimes, however, treatment and control outcomes ***move in parallel in the absence of treatment****. When they do, the divergence of a post-treatment path from the trend established by a comparison group may signal a treatment effect."*

Angrist & Pischke (2015), Ch. 5

Difference-in-Differences

- The fixed effects strategy requires panel data: repeated observations on the same individual (i) over time (t).
- Often, however, the regressor of interest D_{st} varies only at a more aggregate level s.
- This is the case, for example, when:
  - Regional labour market policies vary over time but are fixed across workers/firms within regions.
  - A sizeable common shock affects a subpopulation of individuals.

Difference-in-Differences

- For instance:
  - Changes to or introduction of minimum wages

(Card & Krueger, 1994; Machin, Manning & Rahman, 2003; Draca, Machin & Van Reenen, 2011)

  - Changes in maternity leave entitlements

(Dustmann & Schönberg, 2012; Carneiro, Løken & Salvanes, 2015; Danzer & Lavy, 2018)

  - Job loss during mass layoffs

(Gathmann, Helm & Schönberg, 2020; Britto, Pinotti & Sampaio, 2022)

  - Increased police deployment following terror attacks

(Di Tella & Schargrodsky, 2004; Klick & Tabarrok, 2005; Draca, Machin & Witt, 2011)

Difference-in-Differences

- The sources of omitted variable bias when evaluating these policies must therefore be unobserved variables at the group (s) and time (t) level.
- The difference-in-differences (DiD) identification strategy is based on the simple idea that, in some cases, group-level omitted variables can be captured by group-level fixed effects.
- To make matters concrete, suppose we are interested in the effect of the minimum wage on employment.

Card & Krueger (1994)

- Card & Krueger (1994) exploit a natural experiment whereby, on April 1, 1992, New Jersey (NJ), raised the state minimum wage from \$4.25 to \$5.05.
- They collected data on employment in fast food restaurants in NJ in February 1992 and again in November 1992.
- They also collected data on the same type of restaurants from eastern Pennsylvania (PA), just across the river, where the minimum wage stayed constant at \$4.25 throughout the period.

Card & Krueger (1994)

- Card & Krueger used their data to compute DiD estimates of the employment effect of the NJ wage increase by comparing the February-to-November change in employment in NJ to the change in PA over the same period.

- DiD is a version of fixed effects estimation using aggregate data. We will use the Card & Krueger (1994) example to introduce the model and notation.

DiD: Model

- In this simple setting, we observe two states $s \in \{\text{NJ}, \text{PA}\}$ at two points in time $t \in \{\text{Feb}, \text{Nov}\}$.

- Our variable of interest D_{st} is a dummy for high-minimum-wage states and periods. That is,

$$D_{st} = \begin{cases} 1 & \text{if } s = \text{NJ and } t = \text{Nov} \\ 0 & \text{otherwise} \end{cases}$$

DiD: Model

- Potential outcomes of restaurant i in state s and period t are Y_{1ist} if there is a high state minimum wage in period t, i.e., if $D_{st} = 1$, and Y_{0ist} otherwise.
- The key assumption is that the structure of potential outcomes in the no-treatment state is additive:

$$\mathbb{E}\left[Y_{0ist} \mid s, t\right] = \alpha_s + \lambda_t$$

- The time-invariant, additive state effect α_s plays the role of the unobserved individual effect α_i discussed in the previous section.

DiD: Parallel Trends Assumption

- In words: In the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect and a time effect that is common across states.

- A more intuitive way to understand this key assumption is as a **common trends assumption**:

  Employment trends would be the same in both states in the absence of treatment.

DiD: Parallel Trends Assumption

- In this simple two period context, the trends absent treatment are simply

$$\mathbb{E}\left[Y_{0ist} \mid s, t = \text{Nov}\right] - \mathbb{E}\left[Y_{0ist} \mid s, t = \text{Feb}\right]$$

- Because of our additivity assumption, they are the same and equal

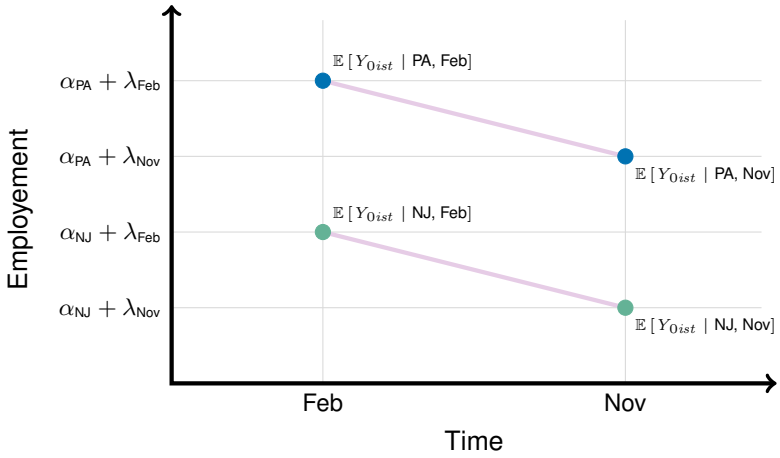
$$\lambda_{\text{Nov}} - \lambda_{\text{Feb}}$$

(the state effects cancel out and the time effects are the same across states).

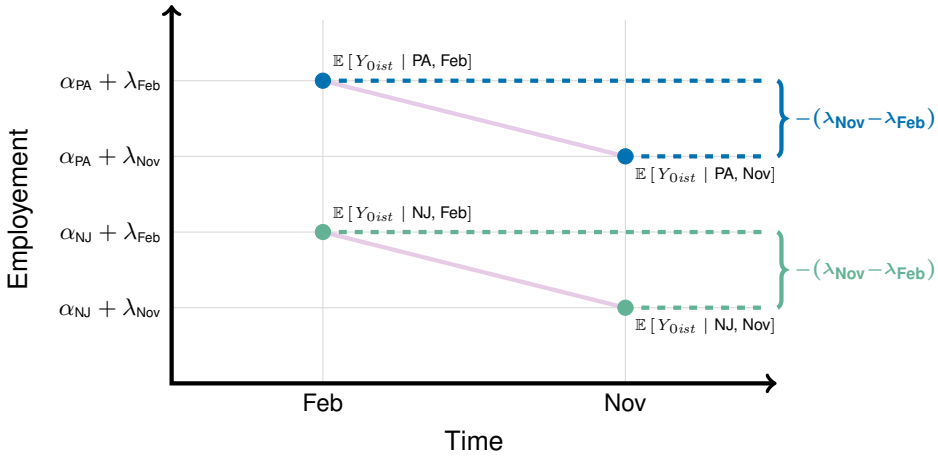
DiD: Parallel Trends Assumption

- The treatment and control states can differ.
- The difference is meant to be captured by the state fixed effect α_s.
- It plays the same role as the unobserved individual effect α_i in fixed effects models.

DiD: Parallel Trends Assumption



DiD: Parallel Trends Assumption



DiD: Identifying the Treatment Effect

- Assuming

$$\mathbb{E}\left[Y_{1ist} - Y_{0ist} \mid s, t \right] = \rho$$

observed employment is given by

$$Y_{ist} = \alpha_s + \lambda_t + \rho D_{st} + \varepsilon_{ist}$$

with $\mathbb{E}\left[\varepsilon_{ist} \mid s, t \right] = 0$.

DiD: Identifying the Treatment Effect

- Therefore, the February-to-November change in employment in NJ is

$$\mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Nov}\right] - \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Feb}\right] = \lambda_{\text{Nov}} - \lambda_{\text{Feb}} + \rho$$

- The analogous time-difference for PA is

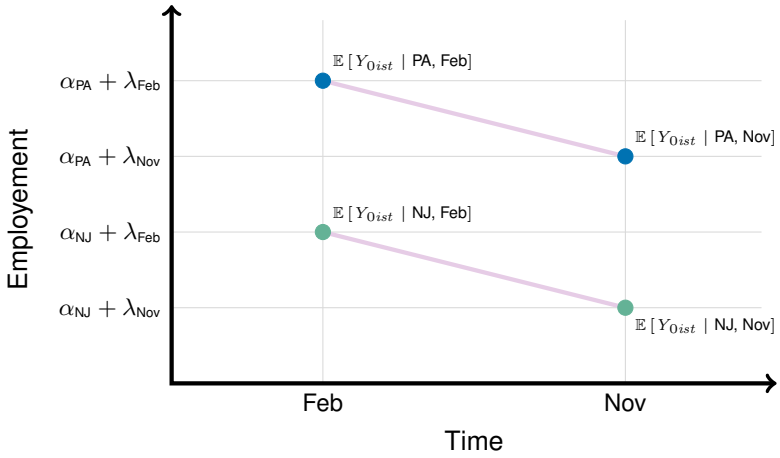
$$\mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Nov}\right] - \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Feb}\right] = \lambda_{\text{Nov}} - \lambda_{\text{Feb}}$$

DiD: Identifying the Treatment Effect

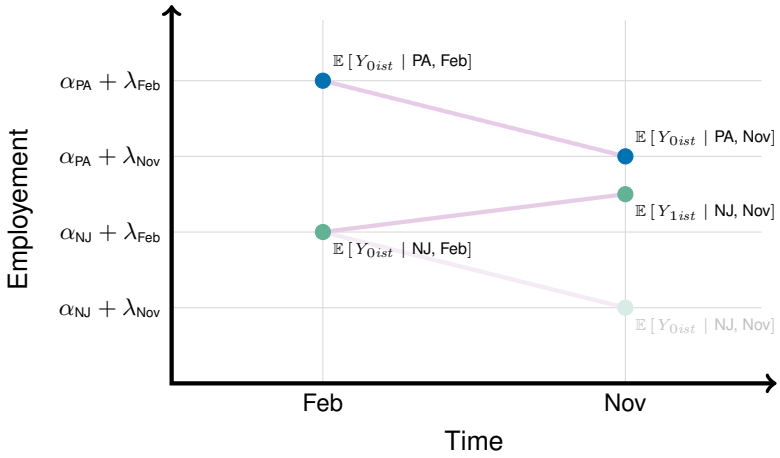
- Treatment induces a deviation from the common trend.
- The population *difference-in-differences* captures the causal effect of interest

$$\Big\{ \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Nov} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Feb} \right] \Big\}$$
$$- \Big\{ \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Nov} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Feb} \right] \Big\} = \rho$$

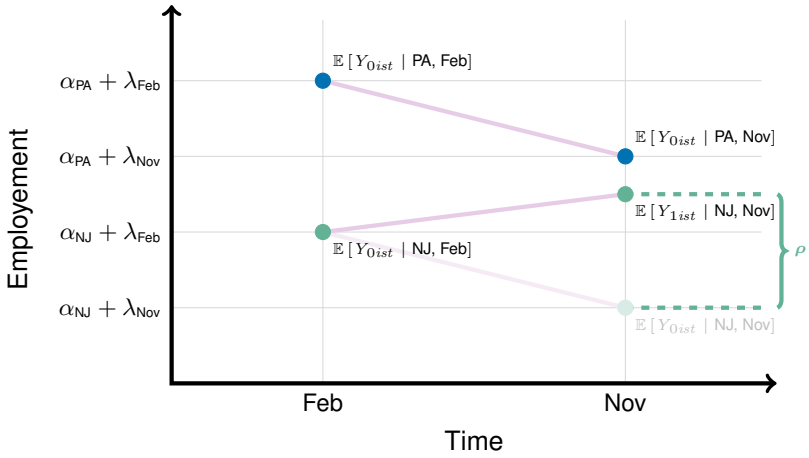
DiD: Identifying the Treatment Effect



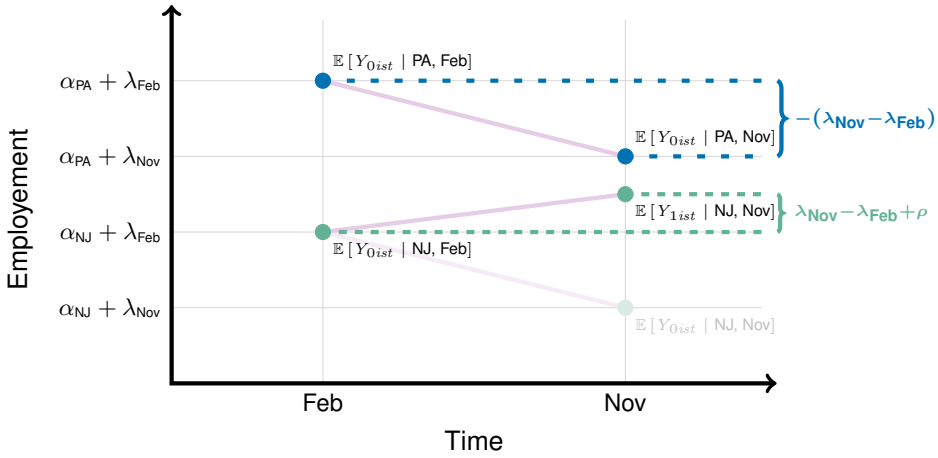
DiD: Identifying the Treatment Effect



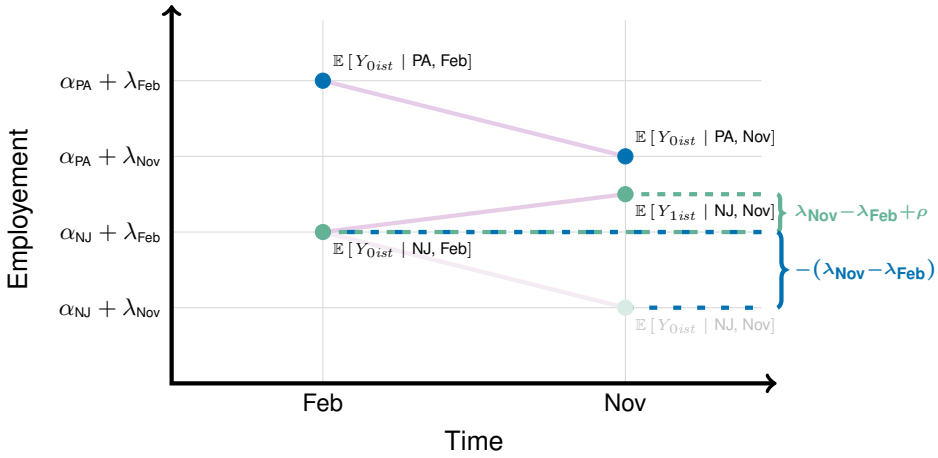
DiD: Identifying the Treatment Effect



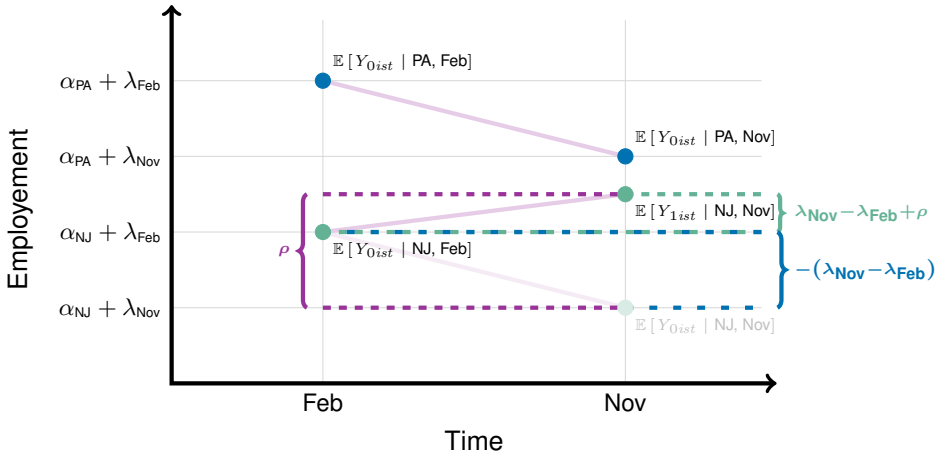
DiD: Identifying the Treatment Effect



DiD: Identifying the Treatment Effect



DiD: Identifying the Treatment Effect



DiD: Estimation

- The DiD estimator is the sample analog of the population DiD

$$\widehat{\rho}_{\text{DiD}} = \left\{ \overline{Y}_{\text{NJ,Nov}} - \overline{Y}_{\text{NJ,Feb}} \right\} - \left\{ \overline{Y}_{\text{PA,Nov}} - \overline{Y}_{\text{PA,Feb}} \right\}$$

where $\overline{Y}_{st}$ denotes the sample average of Y_{ist}.

Population DiD the Other Way Around

- We motivated the population DiD as the cross-state difference of within-state time differences, but it is algebraically equivalent to the time difference of the cross-state differences

$$\begin{aligned}
&\Big\{ \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Nov} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Feb} \right] \Big\} \\
&\qquad - \Big\{ \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Nov} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Feb} \right] \Big\} \\
&= \Big\{ \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Nov} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Nov} \right] \Big\} \\
&\qquad - \Big\{ \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Feb} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Feb} \right] \Big\}
\end{aligned}$$

DiD Estimator the Other Way Around

- The same is true for the DiD estimator

$$\left\{\overline{Y}_{\text{NJ,Nov}} - \overline{Y}_{\text{NJ,Feb}}\right\} - \left\{\overline{Y}_{\text{PA,Nov}} - \overline{Y}_{\text{PA,Feb}}\right\} = \left\{\overline{Y}_{\text{NJ,Nov}} - \overline{Y}_{\text{PA,Nov}}\right\} - \left\{\overline{Y}_{\text{NJ,Feb}} - \overline{Y}_{\text{PA,Feb}}\right\}$$

Regression DiD

- We can conveniently recast our model as a population regression of observed employment Y_{ist} on

1. a treatment-state dummy $d_s^{\text{NJ}} \equiv \mathbb{I}[s = \text{NJ}]$,
2. an after-treatment dummy $d_t^{\text{Nov}} \equiv \mathbb{I}[t = \text{Nov}]$,
3. and their interaction $d_s^{\text{NJ}} \times d_t^{\text{Nov}} = \mathbb{I}[s = \text{NJ and } t = \text{Nov}]$, which is equivalent to our binary regressor of interest D_{st}.

Regression DiD

- To see this, notice that

$$\lambda_t = d_t^{\text{Nov}}\,\lambda_{\text{Nov}} + (1 - d_t^{\text{Nov}})\,\lambda_{\text{Feb}}$$

$$= \lambda_{\text{Feb}} + \underbrace{(\lambda_{\text{Nov}} - \lambda_{\text{Feb}})}_{\equiv\lambda}\, d_t^{\text{Nov}}$$

$$\alpha_s = d_s^{\text{NJ}}\,\alpha_{\text{NJ}} + (1 - d_s^{\text{NJ}})\,\alpha_{\text{PA}}$$

$$= \alpha_{\text{PA}} + \underbrace{(\alpha_{\text{NJ}} - \alpha_{\text{PA}})}_{\equiv\alpha}\, d_s^{\text{NJ}}$$

Regression DiD

- Finally, let $\gamma \equiv \alpha_{\text{PA}} + \lambda_{\text{Feb}}$ and rewrite

$$\begin{aligned} Y_{ist} &= \alpha_s + \lambda_t + \rho D_{st} + \varepsilon_{ist} \\ &= (\alpha_{\text{PA}} + \alpha\, d_s^{\text{NJ}}) + (\lambda_{\text{Feb}} + \lambda\, d_t^{\text{Nov}}) + (\rho\, d_s^{\text{NJ}}\, d_t^{\text{Nov}}) + \varepsilon_{ist} \\ &= \gamma + \alpha\, d_s^{\text{NJ}} + \lambda\, d_t^{\text{Nov}} + \rho\, d_s^{\text{NJ}}\, d_t^{\text{Nov}} + \varepsilon_{ist} \end{aligned}$$

- This is a population regression *in levels*, where we regress the observed outcome on the treatment-state and after-period dummies and their interaction.

Regression DiD

- The intercept

$$\gamma = \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Feb} \right]$$

captures the average of the observed outcome in in the control state before the treatment.

- The coefficient on the treated-state dummy

$$\alpha = \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Feb} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Feb} \right]$$

captures the cross-state difference in average outcomes before the treatment.

Regression DiD

- The coefficient on the after-treatment dummy

$$\lambda = \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Nov} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Feb} \right]$$

  captures the common time trend.

- The coefficient on the interaction of the treatment-state and after-treatment dummies captures the population difference-indifferences

$$\begin{aligned} \rho = & \left\{ \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Nov} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{NJ}, t = \text{Feb} \right] \right\} \\ & - \left\{ \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Nov} \right] - \mathbb{E}\left[Y_{ist} \mid s = \text{PA}, t = \text{Feb} \right] \right\} \end{aligned}$$

  which is our causal effect of interest.

Regression DiD in Changes

- In this treatment/control–before/after setting, the regression *in levels* is equivalent to the following regression *in changes*

$$\begin{aligned}\Delta Y_{is} &\equiv Y_{is\text{Nov}} - Y_{is\text{Feb}} \\ &= \underbrace{(\gamma - \gamma)}_{=0} + \alpha \underbrace{(d_s^{\text{NJ}} - d_s^{\text{NJ}})}_{=0} + \underbrace{\lambda (1 - 0)}_{=\lambda} + \underbrace{\rho\, d_s^{\text{NJ}} (1 - 0)}_{=\rho\, d_s^{\text{NJ}}} + \underbrace{\varepsilon_{is\text{Nov}} - \varepsilon_{is\text{Feb}}}_{\equiv \Delta\varepsilon_{is}} \\ &= \lambda + \rho\, d_s^{\text{NJ}} + \Delta\varepsilon_{is}\end{aligned}$$

- That is, the population DiD can also be reformulated as a population regression of the changes in employment (the time differences for each restaurant) on a dummy indicating the treatment state (NJ) and a constant capturing the common time trend.

Advantages of the Regression DiD Formulation

- We can estimate any of these regressions DiD by OLS and easily obtain the corresponding standard errors.
- Regression DiD offers a few more advantages:
  - We can easily expand our sample by adding additional control states and pre-treatment periods. We would just need to include additional dummies.
  - It facilitates the study of policies that cannot be described by dummy variables. We can consider treatment intensity, e.g., the differences between state and federal minimum wages across all states.
  - We can easily add additional control variables.

Regression Discontinuity Design

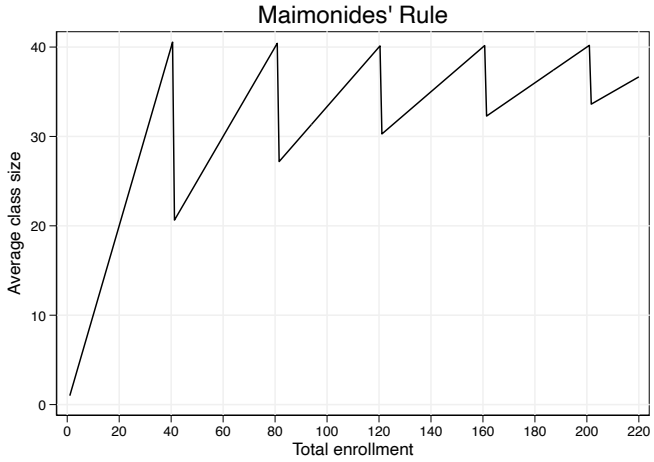
*"RD is based on the seemingly paradoxical idea that **rigid rules** —which at first appear to reduce or even eliminate the scope for randomness— **create valuable experiments**."*

Angrist & Pischke (2015), Ch. 4

RDD: Intuition

- RDD can be used when the treatment is a **discontinuous** function of an underlying **continuous** variable. Examples:
  - A student fails a course if her grade is below the passing grade.
  - Youth gain legal access to alcohol when they reach the minimum legal drinking age.
  - Youth are incapacitated from engaging in crime while attending school. Increases in the legal dropout age alter the crime-age profile for individuals younger than the new dropout age.
  - Maimonides' rule in Israeli public schools: maximum class size of 40.

RDD: Example


Maimonides' Rule
Average class size
Total enrollment
0
10
20
30
40
0
20
40
60
80
100
120
140
160
180
200
220

RDD: Intuition

- Under such a rule, two individuals/cohorts with very **close characteristics** will be exposed to **different treatments**.
- Idea of RDD: compare outcomes of people **just above** and **just below** the discontinuity. For instance,
  - Later labour market outcomes of passing/failing students obtaining grades just above/below the passing grade.
  - Death rates for people just after/before their 18^{th} birthday.
  - Arrest rates for cohorts just above/below the new dropout age at the time of a compulsory schooling law reform. (Bell, Costa & Machin, 2022)
  - Test scores of students from school cohorts with total enrollment just above/below integer multiples of 40. (Angrist & Lavy, 1999)

RDD: Setting

- Individual i's treatment status, D_i is a discontinuous function of the *running variable* (also called forcing variable or score), X_i.
  - D_i is determined (at least in part) by whether X_i crosses the *threshold* (also called cutoff) c.
- Potential outcomes vary smoothly (no discontinuities) with X_i.

RDD: Setting

***Sharp* RD:**

- D_i is a **deterministic** function of X_i such that treatment jumps cleanly as the running variable passes the cutoff.

- For binary treatment $D_i \in \{0, 1\}$, treatment switches cleanly off/on:

$$D_i = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases}$$

RDD: Setting

***Fuzzy* RD:**

- It is the **probability** or the **expected value** of treatment that discontinuously jumps at the threshold.

- For binary treatment $D_i \in \{0, 1\}$, the treatment probability jumps but not from 0 to 1:

$$\mathbb{P}(D_i = 1 \mid X_i) = \begin{cases} g_1(X_i) & \text{if } X_i \geq c \\ g_0(X_i) & \text{if } X_i < c \end{cases}$$

where $g_1(X_i) \neq g_0(X_i)$.

Sharp RD: Specifics

- Notice that there is no value of X_i at which there are both treated and untreated individuals.
- In matching/OLS, we compared treatment and control outcomes at particular values of the control variables.
- Here, we rely on extrapolation across values of X_i in a neighbourhood of c, where D_i switches on.

Sharp RD: Specifics

- To fix ideas, suppose

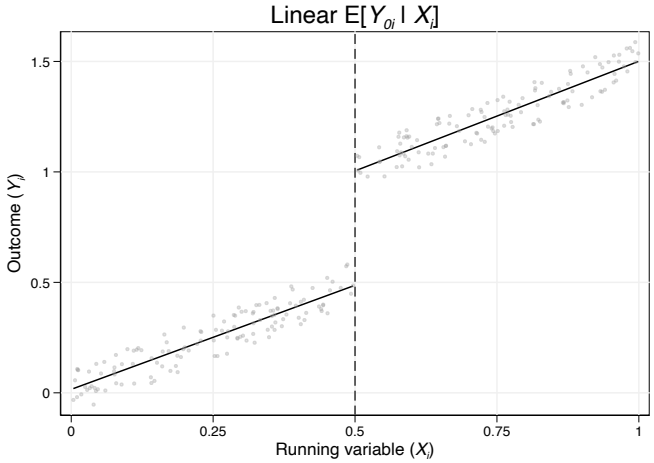
$$\begin{aligned} Y_{1i} &= Y_{0i} + \rho \\ \mathbb{E}\left[Y_{0i} \mid X_i \right] &= \alpha + \beta X_i \end{aligned}$$

- This leads to regression

$$Y_i = \alpha + \beta X_i + \rho D_i + \varepsilon_i$$

- RD captures causal effects by distinguishing the nonlinear, discontinuous function $D_i(X_i) = \mathbb{I}\left[X_i \geq c \right]$ from the linear, smooth function X_i

Sharp RD: Specifics



Sharp RD: Specifics

- Now, suppose that

$$\mathbb{E}\left[Y_{0i} \mid X_i\right] = f(X_i)$$

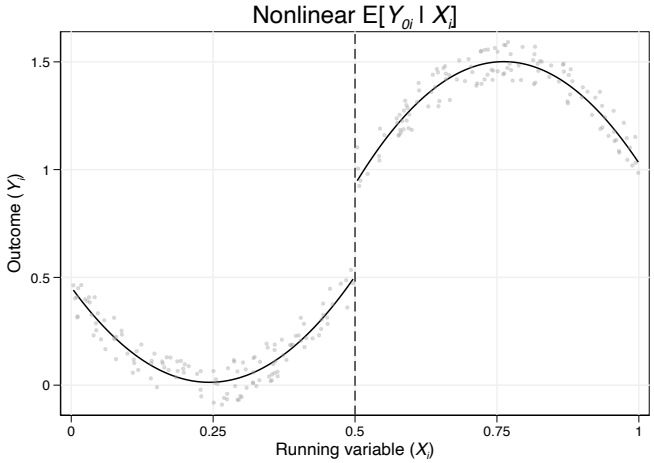
instead, where $f(X_i)$ is some smooth nonlinear function.

- It is still possible to estimate ρ by fitting

$$Y_i = f(X_i) + \rho D_i + \varepsilon_i$$

as long as $f(X_i)$ is continuous in a neighbourhood of c.

Sharp RD: Specifics

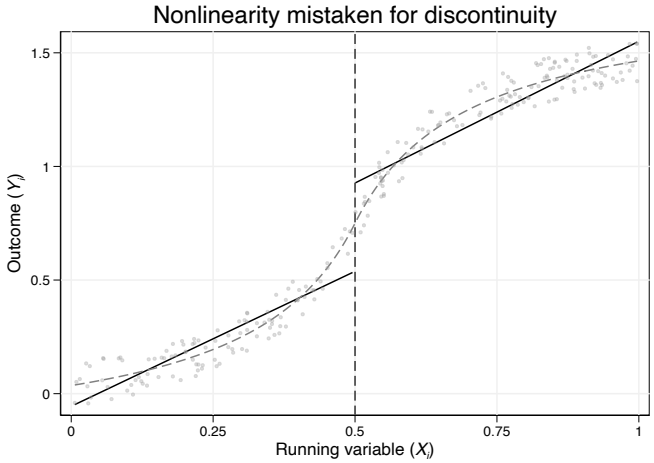


Sharp RD: Specifics

- However, the validity of such estimates of ρ depends on our ability to reasonably approximate $\mathbb{E}\left[Y_{0i} \mid X_i\right]$.

- If not, we can mistake a nonlinearity for a treatment effect.

Sharp RD: Specifics



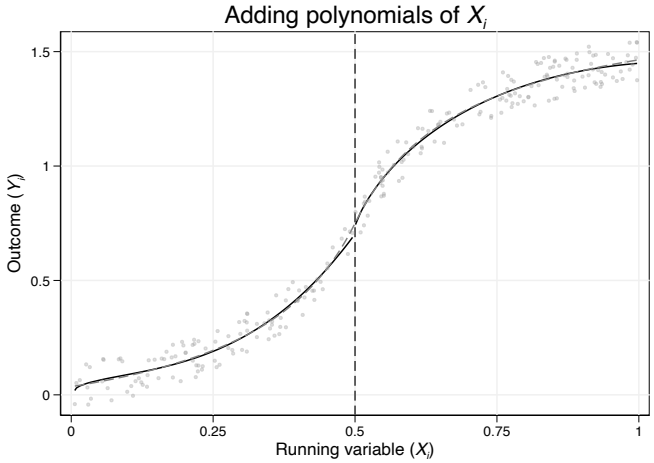
Sharp RD: Specifics

- Two strategies to reduce the likelihood of mistakes:

1. *Parametric* RD: Modelling nonlinearities directly, typically using polynomial functions of the running variable.

2. *Non-parametric* RD: Focusing solely on observations near the cutoff.

Parametric RD



Non-parametric RD

- Non-parametric RD exploits the fact that the problem of distinguishing jumps from nonlinear trends vanishes as the neighbourhood around the cutoff shrinks.

- Using only observations such that $X_i \in [c - \Delta, c + \Delta]$ for some small $\Delta > 0$, we have that

$$\lim_{\Delta \to 0} \mathbb{E}\left[Y_i \mid c \leq X_i < c + \Delta \right] - \mathbb{E}\left[Y_i \mid c - \Delta < X_i < c \right] = \mathbb{E}\left[Y_{1i} - Y_{0i} \mid X_i = c \right]$$

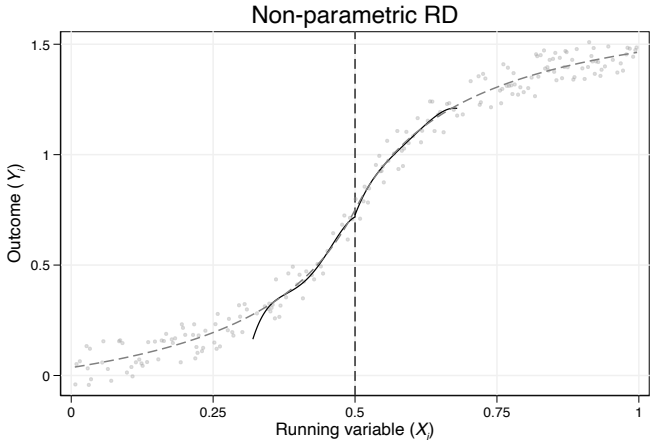
Non-parametric RD

- Non-parametric RD compares averages in a narrow window (defined by the *bandwidth* h) just to the left and just to the right of the cutoff, i.e.,

$$c - h \leq X_i \leq c + h$$

- The mean of Y_i at each side of the cutoff is typically estimated by semi- or non-parametric methods such as local linear regression.
- There is a trade-off between bias near the cutoff and variance (smaller sample size the narrower the window).
- Bandwidth selection is important.

Non-parametric RD



Kernel: Epanechnikov, **Bandwidth:** 0.18.

Fuzzy RD

- To fix ideas, suppose that

$$g_1(c) > g_0(c)$$

so that $X_i \geq c$ makes treatment more likely.

- Let $Z_i \equiv \mathbb{I}[X_i \geq c]$ and note that it indicates the point at which $\mathbb{E}[D_i \mid X_i]$ is discontinuous:

$$\begin{aligned}\mathbb{E}[D_i \mid X_i] &= \mathbb{P}(D_i = 1 \mid X_i) \\ &= g_0(X_i) + [g_1(X_i) - g_0(X_i)] Z_i\end{aligned}$$

Fuzzy RD

- Now, to make matters simple, suppose that

$$g_0(X_i) = \gamma_0 + \gamma_1 X_i$$
$$g_1(X_i) = \delta_0 + \delta_1 X_i$$

- Then,

$$\mathbb{E}\left[D_i \mid X_i\right] = \gamma_0 + \gamma_1 X_i + \left[\pi_0 + \pi_1 X_i\right] Z_i$$

where $\pi_0 = \delta_0 - \gamma_0$ and $\pi_1 = \delta_1 - \gamma_1$.

Fuzzy RD

- Notice that

$$\lim_{\Delta \to 0} \frac{\mathbb{E}\left[Y_i \mid c \leq X_i < c + \Delta\right] - \mathbb{E}\left[Y_i \mid c - \Delta < X_i < c\right]}{\mathbb{E}\left[D_i \mid c \leq X_i < c + \Delta\right] - \mathbb{E}\left[D_i \mid c - \Delta < X_i < c\right]}$$

$$= \frac{\rho\left[\pi_0 + \pi_1 c\right]}{\pi_0 + \pi_1 c}$$

$$= \rho$$

- Fuzzy RD is IV with instrument Z_i for D_i.

Example: Maimonides' rule

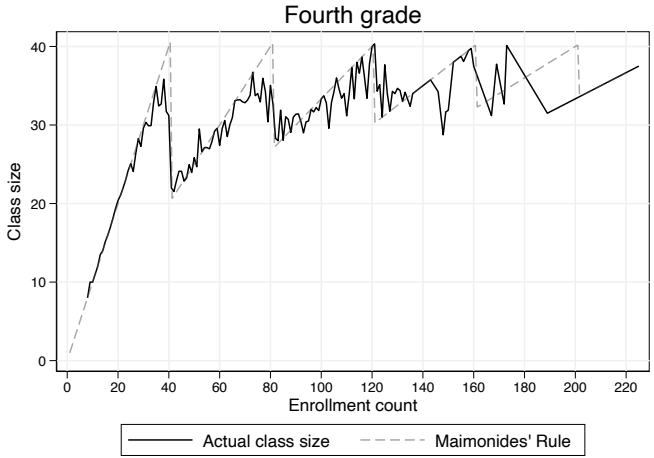
- Angrist & Lavy (1999) use IVs to estimate the causal effect of class size on test scores in Israeli public schools following a fuzzy RDD:

$$\text{Expected class size} = \frac{\text{Enrolment}}{\text{int}\left[\frac{\text{Enrolment}-1}{40}\right] + 1}$$

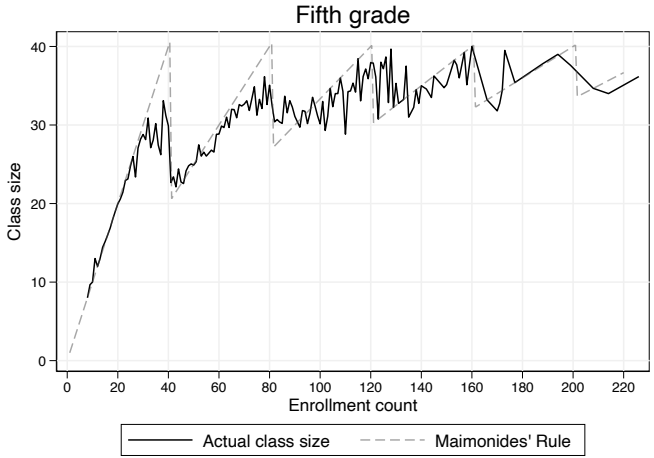
  where int (x) is the integer part (or floor) of x.

- Schools can sometimes afford to add extra classes before reaching the maximum class size of 40 $\implies$ Maimonides' rule does not predict class size perfectly.

Angrist & Lavy (1999): Fuzzy RD



Angrist & Lavy (1999): Fuzzy RD



References

References

Additional references (not on the course reading list):

Angrist, J. D. (1990).
Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records.
American Economic Review, 80(3):313–336.

Angrist, J. D. and Pischke, J.-S. (2009).
Mostly Harmless Econometrics: An Empiricist's Companion.
Princeton University Press.

Carneiro, P., Løken, K. V., and Salvanes, K. G. (2015).
A Flying Start? Maternity Leave Benefits and Long Run Outcomes of Children.
Journal of Political Economy, 123(2):365–412.

Danzer, N. and Lavy, V. (2018).
Paid Parental Leave and Children's Schooling Outcomes.
The Economic Journal, 128(608):81–117.

Di Tella, R. and Schargrodsky, E. (2004).
Do Police Reduce Crime? Estimates Using the Allocation of Police Forces After a Terrorist Attack.
American Economic Review, 94(1):115–133.

Dustmann, C. and Schönberg, U. (2012).
Expansions in Maternity Leave Coverage and Children's Long-Term Outcomes.
American Economic Journal: Applied Economics, 4(3):190–224.

Gathmann, C., Helm, I., and Schönberg, U. (2020).
Spillover Effects of Mass Layoffs.
Journal of the European Economic Association, 18(1):427–468.

Klick, J. and Tabarrok, A. (2005).
Using Terror Alert Levels to Estimate the Effect of Police on Crime.
The Journal of Law and Economics, 48(1):267–279.