Chapter 3
Causal inference methods

Simon Fraser University
ECON 483
Summer 2023

SFU DEPARTMENT OF ECONOMICS

Disclaimer

These notes are partly based on **Causal inference: The mixtape** by Scott Cunningham. However I am entirely responsible for any error.

I do not allow this content to be published without my consent.

Causal inference

- Economists are often interested in the impact of a policy on a given outcome, and might want to identify its impacts before implementing it
- The same way pharmaceutical companies conduct trials before selling a drug, or medical doctors test treatments on a sample of patients
- Some individuals might be subject to the policy (or the "treatment"), others not. And individuals can differ across various characteristics
- The question is then: how to estimate the pure effect of the policy?
- We want to make sure the difference in the outcome variable is not due to some particular characteristics of the treated individuals

Outline

- **Treatment effects taxonomy**
- **Estimating average treatment effects**
- **Difference-in-differences (DiD) estimators**
- **Regression Discontinuity Design (RDD)**
- **Synthetic control methods**
- Suggested readings: Chapters 4, 6, 7, 9, 10 in **Causal Inference**

Definitions, assumptions and estimation

Treatment effects taxonomy

- Consider a binary variable, $\boldsymbol{D_i}$, that equals 1 if Mr. $\boldsymbol{i}$ is subject to the treatment, and 0 if he is not
- We are interested in the impact of the treatment on an **outcome variable** $\boldsymbol{Y_i}$
- Individuals subject to the treatment are part of the **treatment group**, and individuals who are not are part of the **control group**
- Consider the **potential outcomes** $\boldsymbol{Y_i^1}$ and $\boldsymbol{Y_i^0}$ for individual $\boldsymbol{i}$: Those are the values of the outcome variable for Mr. $\boldsymbol{i}$ if he gets treated vs if he doesn't
- In the data, we only observe one or the other, as Mr. i is either treated or not. What we observe is the **actual world**, what we do not observe is the **counterfactual world**: "**What** would have happened **if**..."
- For each individual $\boldsymbol{i}$, define the unit-specific treatment effect $\boldsymbol{\delta_i = Y_i^1 - Y_i^0}$

Treatment effects definitions

Definition: Treatment effects

We are interested in the following treatment effects:

- **Average Treatment Effect** is defined as

$$\tau_{ATE} \equiv \mathbb{E}\left[Y_i^1\right] - \mathbb{E}\left[Y_i^0\right]$$

- **Average treatment Effect on the treated (ATT)** is defined as

$$\tau_{ATT} \equiv \mathbb{E}\left[Y_i^1 | D_i = 1\right] - \mathbb{E}\left[Y_i^0 | D_i = 1\right]$$

- The **Average Treatment Effect on the untreated (ATU)** is defined as

$$\tau_{ATU} \equiv \mathbb{E}\left[Y_i^1 | D_i = 0\right] - \mathbb{E}\left[Y_i^0 | D_i = 0\right]$$

Treatment effects definitions

- The ***ATT*** is the average treatment effect on the group that was assigned the treatment
- We cannot compute the ***ATE***, ***ATT*** and ***ATU*** as we only observe one outcome (we do not observe the **counterfactual outcomes**)
- But we can estimate them
- Depending on the question, one or all parameters are of interest. The most common ones are ***ATE*** and ***ATT***

Treatment effects decomposition

- Let $\boldsymbol{\pi}$ be the share of observations getting treatment
- We observe a **sample analog** of the following:

$$\mathbb{E}\left[\boldsymbol{Y_i^1}|\boldsymbol{D_i}=\mathbf{1}\right]-\mathbb{E}\left[\boldsymbol{Y_i^0}|\boldsymbol{D_i}=\mathbf{0}\right]$$

  which can be decomposed into 3 terms:

- The $\boldsymbol{ATE}=\boldsymbol{\tau_{ATE}}=\mathbb{E}\left[\boldsymbol{Y_i^1}\right]-\mathbb{E}\left[\boldsymbol{Y_i^0}\right]$
- A **selection bias** $\mathbb{E}\left[\boldsymbol{Y_i^0}|\boldsymbol{D_i}=\mathbf{1}\right]-\mathbb{E}\left[\boldsymbol{Y_i^0}|\boldsymbol{D_i}=\mathbf{0}\right]$
- A **heterogeneous treatment effect bias** $(\mathbf{1}-\boldsymbol{\pi})\left(\boldsymbol{\tau_{ATT}}-\boldsymbol{\tau_{ATU}}\right)$

Treatment effects decomposition

- In words: What we are able to observe is:
  - The pure average effect of the treatment (***ATE***)
  - The average difference in outcomes for treated and non treated **if** no one had been treated (**selection bias**)
  - The weighted difference in the effect of the treatment on the treated and the untreated (if they had been treated): The **hetergoneous treatment effect bias**
- The latter two effects are a problem: They are not due to the treatment itself, but to inherent characteristics that differ between treated and non treated units
- We cannot observe them, so we cannot subtract them to get an estimate of ***ATE***
- But under some assumptions, they are equal to zero!

Causal inference assumptions

Assumption 1 : Stable Unit Treatment Value Assumption (SUTVA)

- Potential outcomes of one individual are not affected by treatment status of any other individual

Assumption 2 : Conditional Independence

$$\boldsymbol{Y_i^0, Y_i^1 \perp D_i | X_i}$$

- In words: The treatment assignment is independent of the potential outcomes, i.e. it **was not assigned based on what could** happen to individuals if they were treated or not.

Meaning of Assumption 2

- This assumption says that selection into treatment does not depend on potential outcomes, i.e. nobody got selected into the treatment or control group based on what could happen to the individual if treated or not
- Example: Give a drug to someone knowing it will benefit them vs another violates independence, as there is another variable, a **confounder**, that affects who the drug is given to
- The treatment must be assigned **"ignoring"** how each individual would respond
- It guarantees that observations across treatment and control groups are comparable
- Hence, there are many other names for that assumption: **Unconfoundedness**, **selection on observables**, **ignorability**

Implications of Assumption 2

- Assumption **2** means that

$$\mathbb{E}\left[\boldsymbol{Y^1}|\boldsymbol{D}=\mathbf{1}\right]-\mathbb{E}\left[\boldsymbol{Y^1}|\boldsymbol{D}=\mathbf{0}\right]=\mathbf{0}$$
$$\mathbb{E}\left[\boldsymbol{Y^0}|\boldsymbol{D}=\mathbf{1}\right]-\mathbb{E}\left[\boldsymbol{Y^0}|\boldsymbol{D}=\mathbf{0}\right]=\mathbf{0}$$

In words:

- The average potential outcomes ($\boldsymbol{Y^1}$ and $\boldsymbol{Y^0}$) are the **same** for either the treatment or the control group
- The average outcome had units been treated is the same when looking at the treatment group vs the control group
- Assumption **2** implies there is **no selection bias** nor **heterogeneous treatment effect bias**
- How to make sure it is satisfied? **Randomize** treatment assignment!

Estimating treatment effects

- If Assumptions **1** and **2** are satisfied, one can use linear regression tools and estimate the following model:

$$\boldsymbol{Y_i = \beta_0 + \tau D_i + u_i}$$

  With that regression (as $\boldsymbol{D_i}$ is uncorrelated with $\boldsymbol{u_i}$, i.e. it is **exogenous**): $\mathbb{E}[\boldsymbol{Y_i}|\boldsymbol{D_i = 1}] - \mathbb{E}[\boldsymbol{Y_i}|\boldsymbol{D_i = 0}] = \boldsymbol{\tau = \tau_{ATE}}$
- So $\boldsymbol{\hat{\tau}}$ estimates $\boldsymbol{\tau_{ATE}}$
- Standard hypothesis testing can be used to test the significance of $\boldsymbol{\tau}$, i.e. to test whether the treatment has a significant effect or not
- One can control for additional covariates $\boldsymbol{X_i}$ as well (as long as they satisfy the standard OLS assumptions). It can increase the precision of the estimates via lower residual variance

Treatment effects: Illustration in **Rstudio**

Treatment effects: Selection bias in **Rstudio**

```r
# If treatment is based on potential outcome
# i.e. if it depends on what could happen if treated
d_pot <- ifelse(y1 > 3, 1, 0)
ate_pot <- lm(y ~ d_pot + x)
summary(ate_pot)
```

```
##
## Call:
## lm(formula = y ~ d_pot + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2258 -1.2439  0.0578  1.2896  5.1940
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.79185    0.08473  21.148  < 2e-16 ***
## d_pot        1.20350    0.21213   5.674 2.38e-08 ***
## x            2.87545    0.07594  37.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.735 on 497 degrees of freedom
## Multiple R-squared:  0.7439, Adjusted R-squared:  0.7428
## F-statistic: 721.7 on 2 and 497 DF,  p-value: < 2.2e-16
```

Treatment effects: Consistent estimation in Rstudio

```r
# If treatment is random
ate <- lm(y ~ d + x)
summary(ate)
```

```
##
## Call:
## lm(formula = y ~ d + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5048 -0.9085  0.0101  0.9475  4.5910
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.97957    0.08884   11.03   <2e-16 ***
## d            2.12921    0.12949   16.44   <2e-16 ***
## x            2.93330    0.06314   46.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.441 on 497 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8227
## F-statistic:  1158 on 2 and 497 DF,  p-value: < 2.2e-16
```

Dealing with confounding variables

- The presence of confounding variables stays the main problem for the estimation of treatment effects
- Example: Cochran (1968) reports data that has a higher death rate among cigars/pipes smokers than among cigarettes smokers
- Strange! Smoking cigars or pipes implies **not inhaling**, so less tar should reach the lungs for cigar/pipes smokers
- Some variables are confounding the effect of smoking cigars or pipes vs cigarettes: Age
- It turns out, those who smoke cigars and pipes are older than those who smoke cigarettes, so their higher death rate is not due to what they smoke, but other factors related to their age
- Result: It looks like more people die from smoking cigars and pipes than cigarettes

Subclassification

- We can reduce the bias due to selection into treatment by weighting averages according to the confounding variables. Let's call them $\boldsymbol{X}$
- In the example above: Divide the data by **age** groups as age is the confounding variable. If the distribution of cigarettes and cigars/pipes smokers differs by age groups, we say the age distribution is **imbalanced**
- Say, 50% of cigarette smokers are under 50 years old (20 deaths recorded per 100,000), and 50% are above (45 deaths recorded per 100,000)
- 20% of cigar smokers are under 50 years old, 80% are above
- Unbalanced mortality rate: $\mathbf{20 \times 50\% + 45 \times 50\% = 32.5}$
- Balanced mortality rate: $\mathbf{20 \times 20\% + 45 \times 80\% = 40}$

Subclassification

- In words: If cigarette smokers had the same age distribution as cigar smokers, the mortality rate would be **40** deaths per 100,000 people
- The method is simple, and allows to make groups comparable with respect to the confounding variables
- But as the number of confounders increases (dimension of $\boldsymbol{X}$), we end up with more and more subgroups, for which we might not have observations from the treatment or control group
- An important assumption is violated: **Common support**

Assumption 3 : Common support

$$0 < \mathbb{P}(\boldsymbol{D}_i = 1|\boldsymbol{X}) < 1 \ \ \forall \boldsymbol{x} \in \boldsymbol{\chi}$$

- In words: The probability of being treated given any value of $\boldsymbol{X}$ must be positive

Matching

- Subclassification is simple, but it becomes quite cumbersome if there are many confounding variables
- If observations between control and treatment groups are somewhat similar regarding the confounding variables, we can **match** them one on one
- Look for observations in both groups that have the same value of $\boldsymbol{X}$ (**exact matching**), or "close" values of $\boldsymbol{X}$ in terms of distance (**approximate matching**)
- If observation $\boldsymbol{i}$ has more than one match, take the average of $\boldsymbol{Y}$ for these matches to make one single counterfactual
- Once every observation is matched, compute the difference between an observation in the treatment group and its match in the control group
- The estimated average treatment effect is the average of these differences

Propensity score weighting

- Matching is not flawless either, as exact matching can be difficult with many confounders, and approximate matching requires more data to find good matches (that will also require a bias correction)
- **Inverse Propensity score weighting (IPW)** (Rubin, 1977) is a very popular alternative
- Principle: Estimate the probability of being treated given $\boldsymbol{X}$ for each observation
- Run a regression where the dependent variable is $\boldsymbol{D_i}$ (treatment variable is on the left hand side), the covariates are $\boldsymbol{X}$
- Get $\hat{\mathbb{P}}(\boldsymbol{D_i = 1 | X_i})$ for each $\boldsymbol{i}$

Propensity score weighting

- The estimator becomes

$$\hat{\tau}_{ATE} = \sum_{i \in \{D_i = 1\}} \frac{Y_i}{\hat{\mathbb{P}}(D_i = 1 | X_i)} - \sum_{i \in \{D_i = 0\}} \frac{Y_i}{1 - \hat{\mathbb{P}}(D_i = 0 | X_i)}$$

- $\hat{\mathbb{P}}(D_i = 1 | X_i)$ is a prediction! Machine learning methods can help with that
- Prediction methods for binary variables will be needed (see lecture on Supervised learning: classification)

Randomized Control Trials (RCT)

- To be able to estimate the ***ATE*** consistently and without bias, assumptions **1** and **2** need to hold
- That can be achieved by carefully designing experiments so that:
  - Treatment is **independent** of potential outcomes and other confounders
  - The treatment doesn't create spillovers (**SUTVA**)
- In practice, it requires careful preparation and financial investments. But once everything is controlled for, the results are easy to obtain and strongly reliable. RCTs in the economic literature include:
  - Education programs: Give some kids access to some resources (internet in poor countries, financial/in-kind help) and estimate the change in education outcomes
  - Development programs: Give access to clean water, sanitary equipment, etc on poor populations and estimate the change in health outcomes
  - And many more!

Difference-in-differences methods

Difference-in-differences (DiD) estimators

- Many random experiments are "impractical, unfeasible, and maybe even unethical" (Scott Cunningham, *Causal inference: The mixtape*)
- Think about detecting the sources of a disease. Would you make patients ingest a potentially infected compound?
- And what if Assumption **2** (**independence**) is not satisfied?
- We can rely on **natural experiments**, i.e. variations in some treatment variable that affects only some individuals over time and that occur naturally
- Example: A law/policy is passed in a state/area but not another
- Estimating the difference between the treatment group and the control group after the treatment happened will be biased because groups are **fundamentally different**

Difference-in-differences (DiD): Baseline setup

- Two groups: The **treatment group** ($\boldsymbol{D = 1}$) and the **control group** ($\boldsymbol{D = 0}$)
- Two time periods: **Before** the treatment ($\boldsymbol{T = 0}$) and **after** ($\boldsymbol{T = 1}$). The treatment group is treated between the two time periods, the control group is never treated
- Assumption 1 (**SUTVA**) is satisfied
- If groups are different ex ante, then looking at the difference in average outcomes will include the effect of the treatment, but also other components that may have made groups evolve differently even without any treatment
- What we are interested in estimating here is the $\boldsymbol{ATT}$

Difference-in-differences (DiD): Crucial assumptions

Assumption 4 : Parallel trends

The following is satisfied:

$$\begin{aligned}&\mathbb{E}\left[\boldsymbol{Y_i^0}|\boldsymbol{D=1, T=1}\right] - \mathbb{E}\left[\boldsymbol{Y_i^0}|\boldsymbol{D=1, T=0}\right]\\ =&\mathbb{E}\left[\boldsymbol{Y_i^0}|\boldsymbol{D=0, T=1}\right] - \mathbb{E}\left[\boldsymbol{Y_i^0}|\boldsymbol{D=0, T=0}\right]\end{aligned}$$

- In words: The difference we observe between before and after for the control group is the **same** as for the treatment group if the treatment group had not been subject to the treatment (we do not observe the outcome variable if the treatment group is not treated)
- If Assumption 4 is not satisfied, then the estimate of the effect of the treatment is not isolated, i.e. it will include a group specific variation over time
- This assumption is impossible to verify, as it assumes something about a counterfactual
- Evidence in favor of or against it can be shown however: Check trends between groups before the treatment happens

Difference-in-differences (DiD): Parallel trends

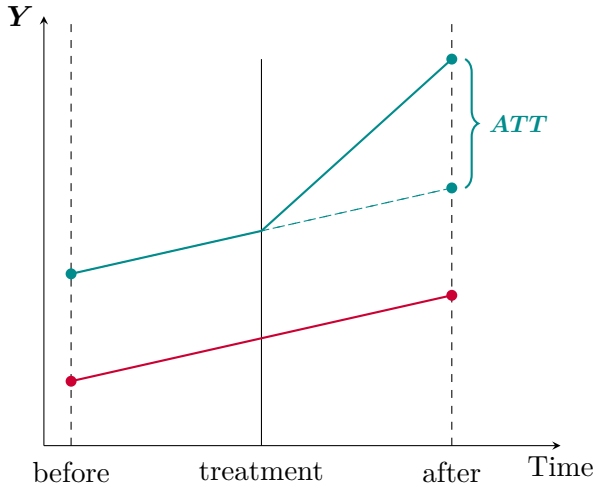


Figure 1: Parallel trends between the **treatment** and the **control group**

Difference-in-differences (DiD): Estimation

- From the graph above, one can guess how to estimate the ***ATT*** in 2 steps:
  - Take the **difference between before and after for each group**. Without treatment and under parallel trends, that difference is the same for both groups, call it a. But the treated group has ***ATT*** on top of it
  - Take the **difference of the two differences**. Since the difference for the treated group is $ATT + a$ and the difference for the control group is a, that second difference yields ***ATT***
- The first differences remove the fundamental difference between the treated and control group before treatment
- The second difference removes the time variation component a
- Being a difference in means, we can then conduct classical hypothesis testing on means to test the null hypothesis of no effect of the treatment

Difference-in-differences (DiD): Estimation decomposition

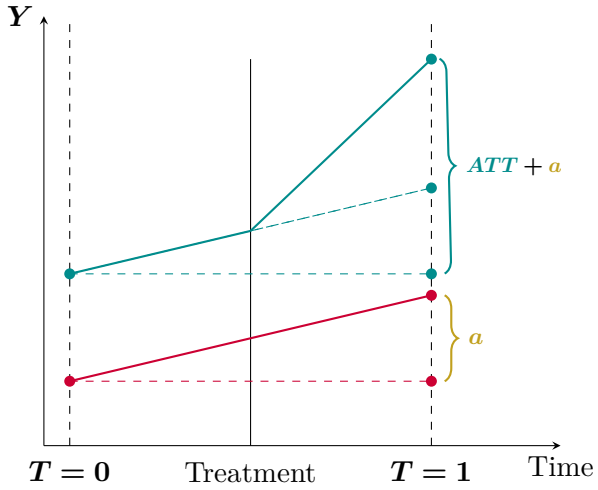


Figure 2: DiD estimation decomposition

Difference-in-differences (DiD): Regression approach

- One can estimate the same effect with a linear regression
- Advantages:
  - We can throw some $\boldsymbol{X's}$ to reduce omitted variable bias (especially by including covariates that vary over time)
  - It will improve the precision of the DiD estimates via lower residual variance
- Consider the following model:

$$\boldsymbol{Y_{i,t} = \beta_0 + \tau D_i + \lambda D_t + \delta\,(D_i \times D_t) + X'_{i,t}\beta_1 + u_{i,t}}$$

where:

- $\boldsymbol{D_i}$ is a dummy variable for treated ($\boldsymbol{D_i = 1}$) vs non treated ($\boldsymbol{D_i = 0}$)
- $\boldsymbol{D_t}$ is a dummy variable for after treatment ($\boldsymbol{D_t = 1}$) vs before ($\boldsymbol{D_t = 0}$)
- $\boldsymbol{D_i \times D_t}$ is the interaction of treatment and period: It is equal to 1 **only for treated units after the treatment happened**

Difference-in-differences (DiD): Regression decomposition

$$\boldsymbol{Y_{i,t} = \beta_0 + \tau D_i + \lambda D_t + \delta\left(D_i \times D_t\right) + X'_{i,t}\beta_1 + u_{i,t}}$$

- The $\boldsymbol{ATT}$ is thus:

$$\begin{aligned}
&\left(\mathbb{E}\left[Y_i | D_i = 1,\, D_t = 1,\, X_{i,t}\right] - \mathbb{E}\left[Y_i | D_i = 1,\, D_t = 0,\, X_{i,t}\right]\right) \\
&- \left(\mathbb{E}\left[Y_i | D_i = 0,\, D_t = 1,\, X_{i,t}\right] - \mathbb{E}\left[Y_i | D_i = 0,\, D_t = 0,\, X_{i,t}\right]\right) \\
&= \left(\beta_0 + \tau + \lambda + \delta + X'_{i,t}\beta_1 - \left(\beta_0 + \tau + X'_{i,t}\beta_1\right)\right) \\
&\quad - \left(\beta_0 + \lambda + X'_{i,t}\beta_1 - \left(\beta_0 + X'_{i,t}\beta_1\right)\right) \\
&= \delta
\end{aligned}$$

Difference-in-differences (DiD): Regression decomposition

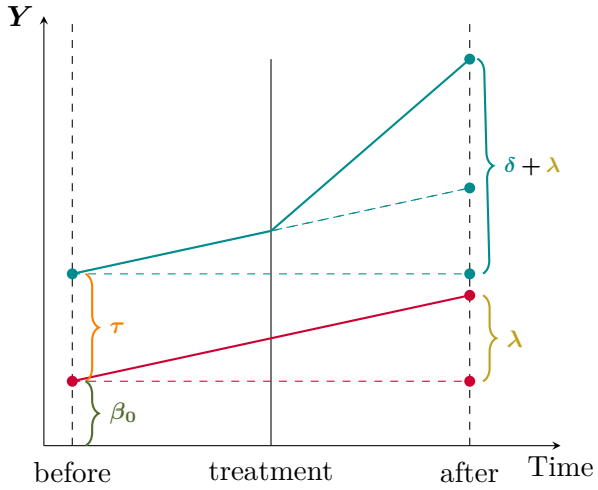


Figure 3: DiD estimation decomposition

Natural experiments in practice

- There are many natural experiments to observe and take advantage from. Units are different but one needs to make sure they have parallel paths in the absence of treatment
  - A country/state passes a law. Other states/countries around don't: health care programs, Iimmigration laws, minimum wage laws, covid-19 related practices (lockdowns, safety rules, vaccines roll outs)
- If there is another source of variation that differs across groups (so another potential confounder), one can use a **Dif-in-dif-in-difs** estimator!!
- It would come with another parallel trends assumption...

Regressions Discontinuity Designs

Regression Discontinuity Designs (RDD)

- Consider a variable $\boldsymbol{X}$ that determines whether someone will receive a treatment or not at a cutoff point $\boldsymbol{c_0}$. Such a variable is called the **running variable**
- $\boldsymbol{X}$ is obviously a **confounder** since it determines treatment
- Comparing treatment and control groups will lead to a biased estimation
- Example: a GPA cutoff of 3.5 to receive a scholarship. Those above 3.5 would do better than those below 2.5 even without the scholarship, so comparing these students is meaningless
- What about comparing those at 3.4 (hence not receiving the scholarship) vs those at 3.6 (hence receiving the scholarship)
- We are talking about **Local Average Treatment Effect (LATE)**

Regression Discontinuity Design: Continuity assumption

Assumption 5 : Continuity

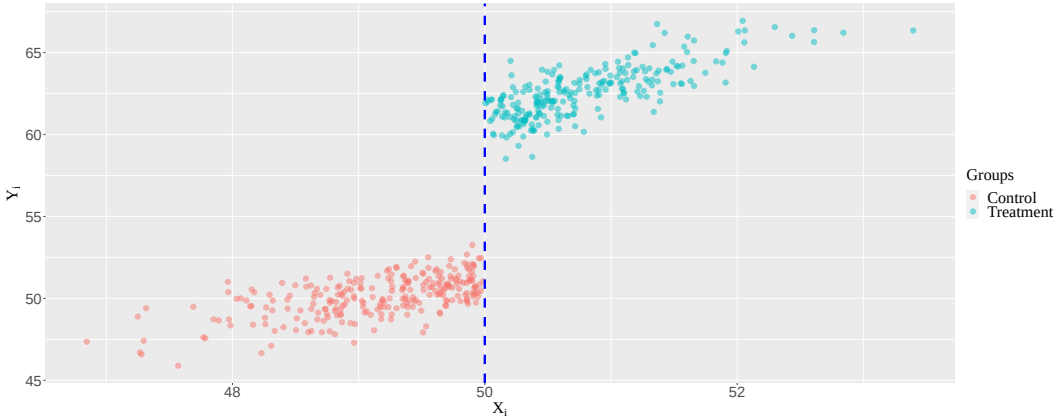
$\mathbb{E}\left[\boldsymbol{Y_i^0}|\boldsymbol{X_i} = \boldsymbol{c_0}\right]$ and $\mathbb{E}\left[\boldsymbol{Y_i^1}|\boldsymbol{X_i} = \boldsymbol{c_0}\right]$ are continuous functions of $\boldsymbol{X_i}$

- In words: At $\boldsymbol{X_i} = \boldsymbol{c_0}$, the average **potential** outcomes do not jump
- If they do jump, then it means there is something other than crossing that threshold creating an impact
- We can't directly test that assumption, but knowledge of the circumstances around the treatment can help build a case

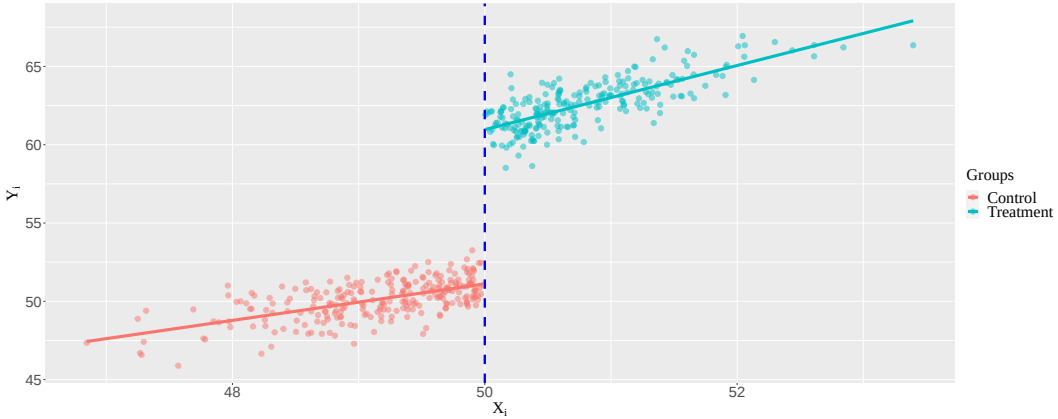
Regression Discontinuity Design: Estimation

- It is then relevant to consider observations close to the cutoff point
- Observations below the threshold, i.e. the control group, can constitute a good counterfactual for the ones above the threshold, i.e. the treatment group
- Procedure: Estimate a regression of $\boldsymbol{Y_i}$ on $\boldsymbol{X_i}$ on either side of the cutoff **separately**
- The estimate of the $\boldsymbol{LATE}$ is given by $\hat{\boldsymbol{\tau}}_{\boldsymbol{LATE}} = \hat{\boldsymbol{y}}_{\boldsymbol{i}}^{+} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}^{-}$ where $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{+}$ is the prediction at $\boldsymbol{X_i} = \boldsymbol{c_0}$ from the right regression, and $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{-}$ is the prediction at $\boldsymbol{X_i} = \boldsymbol{c_0}$ from the left regression

Regression Discontinuity Design: Illustration



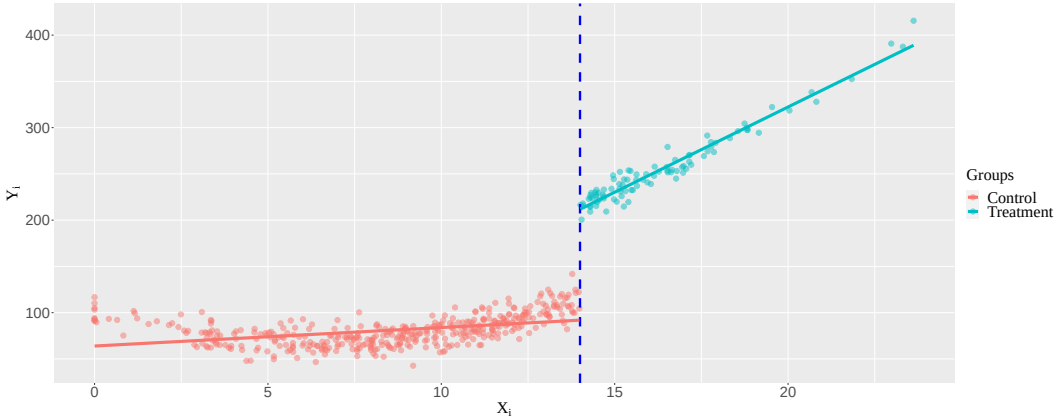
Regression Discontinuity Designs: Illustration



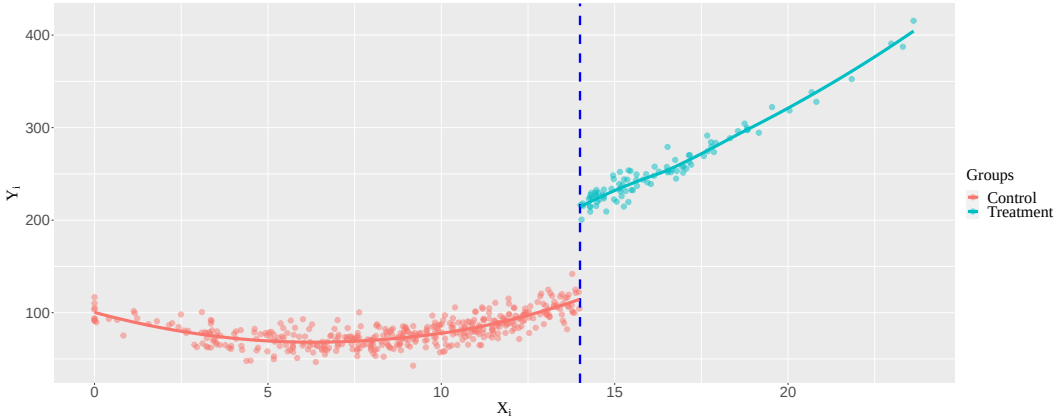
Regression Discontinuity Design: Beyond linearity

- Why use $\boldsymbol{X_i}$ linearly on either side of the cutoff point?
- We are building predictions to take a difference here, so we can allow for more flexible estimators to catch patterns beyond linearity
  - Polynomial estimators: Include $\boldsymbol{X_i^2}$, $\boldsymbol{X_i^3}$ in the regressions to add curvature
- Nonparametric estimators offer more flexibility
  - Kernel estimators
  - Splines
  - K-nearest neighbors estimators

Regression Discontinuity Design: beyond linearity



Regression Discontinuity Design: beyond linearity



Regression Discontinuity Design: Fuzzy RDD

- The RDD we considered so far is called **sharp RDD**, where the treatment goes from 0 to 1 as soon as the threshold is crossed
- Sometimes, the **probability** of being treated after crossing the threshold jumps, but not all the way to 1. So not everybody above the threshold gets treated
- It is called **fuzzy RDD**
- We still need assumptions **1** and **5** to be satisfied
- τ_{LATE} can be estimated via 2SLS. Let Z_i be a dummy variable equal to 1 if $X_i > c_0$, 0 otherwise. It is our **instrument**! Then:
  - Regress D_i on Z_i and combinations of X_i (like polynomial for instance). Get $\hat{D}_i$
  - Regress Y_i on $\hat{D}_i$ and the X_i after re-centering them around c_0

Synthetic control

Synthetic control

- So far, we have compared the average outcome between treatment and control groups, hoping (or knowing thanks to the above assumptions) that the control group can constitute a good **counterfactual** for the treatment group, i.e. that the control group somehow reflects how the outcome would have been for the treatment group units had they not been treated
- What if we don't have a plausible counterfactual, i.e. a unit to compare the treatment group to?
- Abadie and Gardeazabal (2003) studied the impact of terrorism in the Basque Country (my home!!) on economic activity
- The occurrence of terrorism acts as treatment, but there is no counterfactual Basque country for which no terrorism happened
- So they created a "**synthetic Basque country**", i.e. a Basque country without terrorism by computing a combination of other regions of Spain that resembles the Basque country before terrorism, and compared the two units

Synthetic control

- Idea: Optimally choose a set of **weights** on the control units such that the resulting **synthetic units** are close to the characteristics of the treated units in the pre-treatment period
- In the post-treatment period, the synthetic units' outcomes will be the combination of the control units' outcomes
- Then, each treated unit has its own synthetic unit. Take the difference to estimate the **ATT**
- Very powerful when only few units are treated, like a country/state/province who passes a law. Synthetic control can produce a better counterfactual for a treated unit than the control units (imagine comparing a treated country to other countries)

Synthetic control: Model and (simplified) procedure

- Let $\boldsymbol{Y_t}$ be the outcome variable at time $\boldsymbol{t = 1, ..., T}$
- Denote $\boldsymbol{i}$ for treated units, and $\boldsymbol{j}$ for control units
- The treatment happens at some time $\boldsymbol{T_0}$, so periods where $\boldsymbol{t < T_0}$ are pre-treatment periods and periods where $\boldsymbol{t > T_0}$ are post-treatment periods
- For a treated unit, we are interested in $\boldsymbol{\alpha_{i,t}}$ for $\boldsymbol{t > T_0}$
- Let $\boldsymbol{X_{i,t}}$ be the vector of $\boldsymbol{k}$ covariates that are unaffected by the treatment for unit $\boldsymbol{i}$. Let $\boldsymbol{X_0}$ be the matrix of covariates gathering the control units vectors $\boldsymbol{X_j}$
- Let $\boldsymbol{W}$ be a vector of weights of the size of the control units (one weight per control unit). Each treated unit will have a synthetic unit with a different $\boldsymbol{W}$

Synthetic control: Model and (simplified) procedure

- The vector of optimal weights for a treated unit $\boldsymbol{i}$ minimizes the distance between $\boldsymbol{i}$'s covariates and the synthetic ones $\boldsymbol{X_0 W}$, i.e. $\boldsymbol{W_i^*}$ solves:

$$\min_{\{\boldsymbol{W_i}\}} \|\boldsymbol{X_i} - \boldsymbol{X_0 W_i}\|$$

- subject to the constraints that the weights add up to one, and are positive or null
- For a treated unit $\boldsymbol{i}$, the estimated effect for $\boldsymbol{t > T_0}$ is

$$\hat{\boldsymbol{\alpha}}_{i,t} = \boldsymbol{Y}_{i,t} - \sum_j \boldsymbol{w}^*_{i,j} \boldsymbol{Y}_{j,t}$$

  where $\boldsymbol{j}$ represents units in the control group

- Repeat the same process for each treated unit

Treatment effects: Summary

- Assumption **1** (SUTVA) is needed no matter the setting
- If Assumption **2** (Conditional independence or unconfoundedness or ignorability) is satisfied, a simple OLS regression estimates τ_{ATE} or a difference in means!
- If there are confounders, subclassification, matching and inverse propensity score weighting are alternatives to alleviate the selection bias as long as we have Assumption **3** (common support)
- Natural experiments include confounders, but if Assumption **4** (parallel trends) is satisfied, we can use Dif-in-difs to estimate τ_{ATT}
- If a treatment is **determined** by a confounder, we can look at treatment effects **locally** using Regression discontinuity designs to estimate τ_{LATE}. We need Assumption **5** (continuity) to estimate that effect
- When no observations can be used as a plausible counterfactual. a **synthetic unit** can be created to estimate τ_{ATT}

Treatment effects: Takeaways

- The literature on treatment effects is vast
  - Different experimental designs
  - Different assumptions
  - Different ways to handle standard errors (clustering, bootstrapping, etc)
- Program evaluation methods are implemented by many organizations and governments. You could be working with doctors to design of experiments to test the effect of drugs or living conditions on health outcomes!
- The choice of one method over another depends on the experimental designs: What assumptions are satisfied?
- At the end of the day, we want to **predict** the effect of a policy on an outcome variable
- So there is room for machine learning algorithms!