Class 7: Causal Inference

MFE 402

Dan Yavorsky

Last Class

- Midterm
- Stepwise and All Subset Regressions – to optimize in-sample fit
- Cross Validation – to assess out-of-sample prediction
- Bootstrapping – to estimate standard errors

Topics for Today

- Introduce Causal Inference
  - The fundamental problem of causal inference
  - Selection Bias, our formidable enemy
  - Random assignment, a simple solution

- Ways to estimate causal effects
  - Conditioning on covariates (AB tests and covariate adjustments)
  - Difference-in-Differences (Diff in Diff)
  - Regression Discontinuity Designs (RDD)

Introduction to Causal Inference

Causal Effects

Often the underling goal of econometric analysis is to measure a **causal** relationship between two variables.

For example, what is the effect of:

- class size on test scores?
- police expenditures on crime rates?
- climate change on economic activity?
- years of schooling on wages?
- institutional structure on economic growth?
- the effectiveness of rewards on behavior?
- the consequences of medical procedures on health outcomes?

Ruben's Potential Outcome Framework

Let D_i denote degree of **treatment** for unit i, and consider the binary case where $D_i \in \{0, 1\}$.

Let Y_i denote the **outcome** of interest for unit i.

Ruben proposed the following **potential outcomes** framework for causal inference:

$$Y_i = \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \end{cases}$$

Notice:

- These are potential outcomes because they "exist" irrespective of which treatment occurs
- The potential outcomes $\{Y_i(1), Y_i(0)\}$ are not changed by D_i
- D_i just switches which one you get to observe

The Fundamental Problem of Causal Inference

We would like to measure the **causal effect** of D_i on Y_i:

$$Y_i(1) - Y_i(0)$$

This is not measurable:

- we will only ever observe one of the potential outcomes for each unit i
- we will never observe both $Y_i(1)$ and $Y_i(0)$

The **fundamental problem of causal inference** is the impossibility of observing a counterfactual.

ACE and ATE

In general, there is likely to be a distribution of both $Y_i(1)$ and $Y_i(0)$ across the population, and therefore also a distribution of causal effects $Y_i(1) - Y_i(0)$. Rubin's goal was to learn about features of this distribution.

For example, the **average causal effect** (ACE) of D_i on Y_i:

$$\text{ACE} = \text{ATE} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$$

The ACE is also commonly called the **average treatment effect** (ATE).

Selection Bias

We can (almost) always measure the separate average outcomes of the treated and untreated groups, but this is not the same as the ATE:

$$\underbrace{\mathbb{E}[Y_i|D_i = 1] \text{ - } \mathbb{E}[Y_i|D_i = 0]}_{\substack{\text{Observed Difference in Means} \\ \text{Often Easy to Estimate}}} = \underbrace{\mathbb{E}[Y_i(1)|D_i = 1] \text{ - } \mathbb{E}[Y_i(0)|D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] \text{ - } \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{Selection Bias}}$$

The naive comparison of averages tells us something about potential outcomes, but usually it's not what we want to know:

- The **average treatment effect on the treated** (ATT) is the ACE for the treated group
- The **selection bias** is the difference between
  1. the average outcome of the treated group if they had not been treated and
  2. the average outcome of the untreated group

Example of Selection Effects: Earning the MFE Degree

Let Y be an individual's annual salary and D indicate if someone has earned an MFE degree.

- The average salary of MFE graduates is higher than the average salary of non-MFE graduates: $\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] > 0$
- This difference is **not** the ATE of an MFE degree on salary; it's likely too high

There is almost certainly a selection bias: MFE graduates are likely to have higher salaries even if they had not earned an MFE degree: $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] > 0$

- MFE graduates are likely to pursue jobs in finance or technology, which pay more
- MFE graduates are the "type of person" willing to endure the rigors of an MFE program, which is likely correlated with success in the labor market
- MFE graduates are willing to invest in the MFE degree, which is correlated with individuals who take their careers seriously and thus earn more
- others?

Linear Regression to Estimate Causal Effects

Denote the causal effect as $\tau_i = Y_i(1)$ - $Y_i(0)$ and the ATE as τ, we can write Y_i as:

$$\begin{aligned} Y_i &= Y_i(0) + [Y_i(1) \text{ - } Y_i(0)]D_i + (\mathbb{E}[Y_i(0)] \text{ - } \mathbb{E}[Y_i(0)]) \\ &= \underbrace{\alpha}_{\mathbb{E}[Y_i(0)]} + \underbrace{\tau_i}_{Y_i(1)\text{-}Y_i(0)} \times D_i + \underbrace{e_i}_{Y_i(0)\text{-}\mathbb{E}[Y_i(0)]} \end{aligned}$$

Evaluating the CEF with treatment status switched on and off yields:

$$\begin{aligned} \mathbb{E}[Y_i|D_i = 1] &= \alpha + \tau + \mathbb{E}[e_i|D_i = 1] \\ \mathbb{E}[Y_i|D_i = 0] &= \alpha + \mathbb{E}[e_i|D_i = 0] \end{aligned}$$

Thus the CEF is the treatment effect τ plus the selection bias where the selection bias amounts to correlation between the regression error term (e_i) and the explanatory variable (D_i)

Random Assignment

Random assignment solves the selection problem because it makes D_i **independent** of the potential outcomes.

Under random assignment:

- $\mathbb{E}[Y_i(1)|D_i = 1] = \mathbb{E}[Y_i(1)|D_i = 0] = \mathbb{E}[Y_i(1)]$
- $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)]$

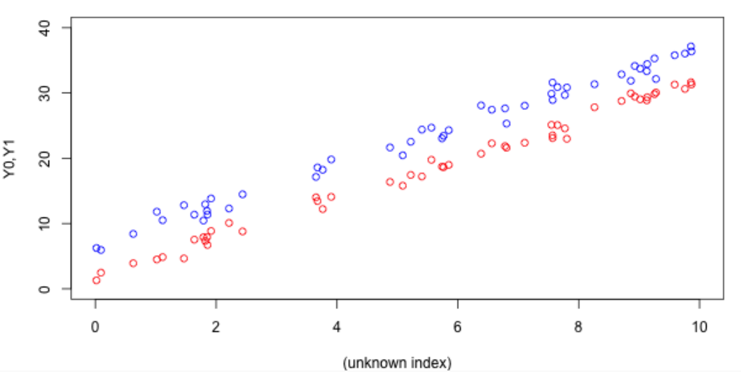
Thus:

$$\begin{aligned}\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \quad \text{because Y is indepedent of D}\\ &= \mathbb{E}[Y_i(1) - Y_i(0)]\\ &= \text{ATE}\end{aligned}$$

And from the regression model perspective, the ATE is τ, the slope coefficient on treatment D_i, which can be estimated with OLS because random assignment nullifies any concerns of endogeneity.

Graphic Example

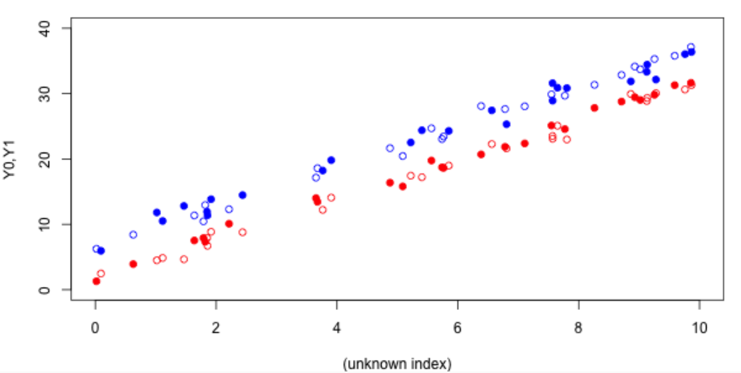
Let's graph some potential outcomes: $Y_i(0)$ in red, $Y_i(1)$ in blue.



True ATE = 5.

Graphic Example

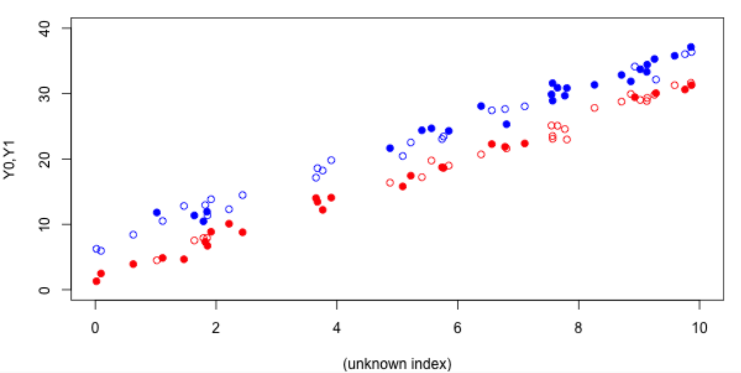
Random assignment to treatment:



$\widehat{\text{ATE}} = \hat{\mathbb{E}}[Y_i|D_i = 1]$ - $\hat{\mathbb{E}}[Y_i|D_i = 0] = 5.8$

Graphic Example

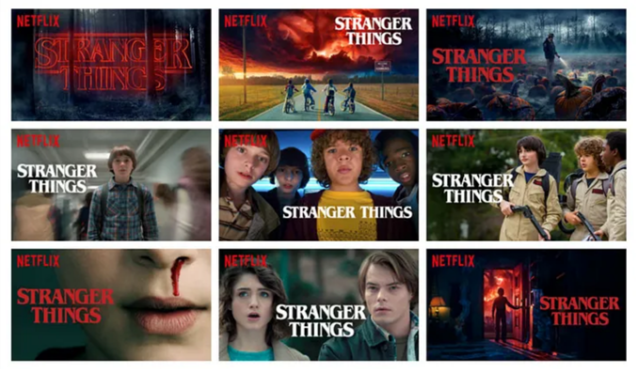
Suppose $\Pr[D_i = 1]$ increases toward the right, so that $\mathbb{E}[Y_i(0)|D_i = 1]$ is higher while $\mathbb{E}[Y_i(0)|D_i = 0]$ is lower, creating a positive selection bias:



$$\widehat{\text{ATE}} = \hat{\mathbb{E}}[Y_i|D_i = 1] - \hat{\mathbb{E}}[Y_i|D_i = 0] = 12.4$$

Example: AB Testing Thumbnail Artwork at Netflix

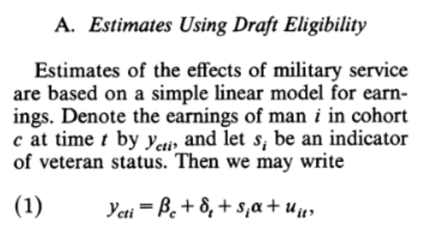
At Netflix (link), D_i is the artwork shown to a user and Y_i is the length of time spent watching.



Example: Angrist (1990) Lifetime Earnings and the Vietnam Era Draft Lottery

Angrist (1990) (link) studies the effect of military service on lifetime earnings.

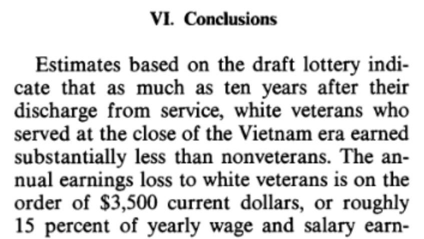
- Specifically, it measures the long-term labor market consequences (Y = earnings) of military service during the Vietnam era (D = veteran status)
- You can't just compare civilian earnings by veteran status because veterans are different from non-veterans in many ways
- Instead, Angrist uses the Vietnam draft lottery to create a randomized experiment
- He compares earnings of draft-eligible men who served in the military to those who remained civilians



A. *Estimates Using Draft Eligibility*

Estimates of the effects of military service are based on a simple linear model for earnings. Denote the earnings of man i in cohort c at time t by y_{cti}, and let s_i be an indicator of veteran status. Then we may write

$$(1) \qquad y_{cti} = \beta_c + \delta_t + s_i\alpha + u_{it},$$



VI. Conclusions

Estimates based on the draft lottery indicate that as much as ten years after their discharge from service, white veterans who served at the close of the Vietnam era earned substantially less than nonveterans. The annual earnings loss to white veterans is on the order of $3,500 current dollars, or roughly 15 percent of yearly wage and salary earn-

Always Experiment?

If random assignment is so great (and it is!) why don't we always do experiments and assign treatment via randomization?

- Too expensive
- Takes too long
- Unethical
- Technically impossible

For example:

- We cannot randomly assign people to earn an MFE degree
- We cannot randomly assign people to be exposed to climate change
- We cannot randomly assign people to undergo a medical procedure
- We cannot randomly assign people to be born in the United States
- We cannot randomly assign people to smoke cigarettes

An Often-Implicit Assumption: SUTVA

The Stable Unit Treatment Value Assumption (SUTVA) is that the potential outcomes of one unit are unaffected by the treatment status of other units (i.e., no network effects)

In other words: Y_i depends only on i's treatment status, not on the treatment status of others

We have been making this assumption implicitly in the notation up to now by writing:

$$Y_i = f(D_i) \qquad \text{instead of} \qquad Y_i = f(D_1, D_2, \ldots, D_n)$$

Be aware that you are making this assumption, even if only implicitly, especially when you are assuming a natural experiment rather than conducting your own randomized experiment.

Potential Problems

Things you should be aware of when you (assume or run your own) experiment:

1. **Network effects:** does SUTVA hold?
2. **Non-compliance:** do people assigned to treatment actually receive it?
3. **Failure of randomization:** are the treatment and control groups actually similar? True randomization takes discipline!
4. **Differential Attrition:** do people drop out of the experiment differentially by treatment status?
5. **External Validity:** does the experiment generalize to the population of interest?

The hard work in is in the design and implementation, not necessarily in the analysis!

What if we can’t experiment?

Randomization is the gold standard for causal inference because, in expectation, it balances observed **and unobserved** characteristics between treatment and control groups.

But what if we want to use observational data instead of experimental data? Can we still make causal inferences? Maybe.

The rest of this week’s slide discuss:

- Selection on Observables
- Differences in Differences (Diff in Diff)
- Regression Discontinuity Designs (RDD)

Selection on Observables
(aka Conditioning on Covariates)

Conditional Independence Assumption

The **Conditional Independence Assumption** (CIA) is the generalization that the treatment assignment is independent of the potential outcomes conditional on some set of covariates:

$$\{Y_i(0),\ Y_i(1)\} \perp D_i | X_i$$

The CIA asserts that, conditional on observed characteristics X_i, selection bias disappears

Law of Iterated Expectations

Given the CIA, conditional-on-X_i comparisons across outcomes given the treatment levels have a causal interpretation:

$$\mathbb{E}[Y_i|X_i, D_i = 1] - \mathbb{E}[Y_i|X_i, D_i = 0] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$$

We now have a separate causal effect for each value taken on by the set of conditioning variables X_i, an embarrassment of riches!

Commonly, the Law of Iterated Expectations is used to find the unconditional or overall-average causal effect:

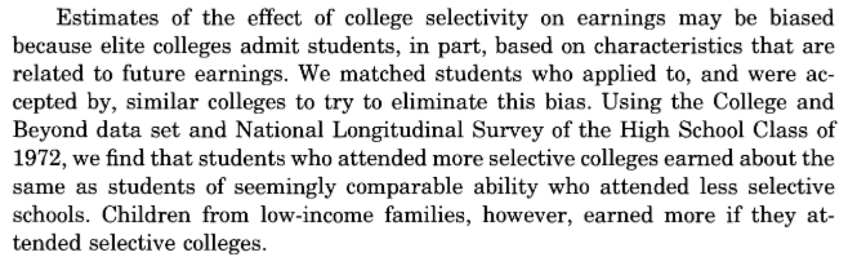
$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}\left[\mathbb{E}[Y_i(1) - Y_i(0)|X_i]\right]$$

This unconditional average effect can be computed by a averaging all of the X-specific effects, weighted by the distribution of X_i

Example: Estimating the Payoff to Attending a More Selective College

Dale & Krueger (2002) (link) studies the effect of college attended on earnings.

- Compile a dataset of 1976 high school graduates, including their SAT scores, college attended, colleges admitted to, and earnings in 1995.
- Regress log earnings (Y) on school's average SAT score (D) and other covariates (X) to estimate the effect of attending a more selective college on earnings.
- Find that attending a more selective college has a positive effect on earnings, but that this effect is small and statistically insignificant.



Estimates of the effect of college selectivity on earnings may be biased because elite colleges admit students, in part, based on characteristics that are related to future earnings. We matched students who applied to, and were accepted by, similar colleges to try to eliminate this bias. Using the College and Beyond data set and National Longitudinal Survey of the High School Class of 1972, we find that students who attended more selective colleges earned about the same as students of seemingly comparable ability who attended less selective schools. Children from low-income families, however, earned more if they attended selective colleges.

Regression Discontinuity Designs

RDD Intro via Examples

RDD applies to the scenario where treatment is assigned based on a cutoff.

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

For example, study the effect of:

- college scholarships on earnings, where scholarship are awarded to students with an SAT score above c
- polity party on economic growth, where the party in power is determined by vote share above $c = 0.5$
- a new drug on health outcomes, where the drug is prescribed to patients with a blood-test value above c
- class size on test scores, where additional classes are added if the cohort is above c
- alcohol consumption of vehicle accidents, where the U.S. legal drinking age is $c = 21$

In each of these examples, the treatment is assigned based on a cutoff c. As long as observed units cannot tailor their behavior to the cutoff, we can think of the cutoff as being randomly assigned, and thus the treatment assignment is as good as random.

Key RDD Idea Mathematically

Units just-above and just-below the cutoff are similar in all ways except for the treatment assignment.

Thus, the difference in outcomes between these units – particularly the ones close the cutoff – is a good estimate of the causal effect of the treatment.

Denote Y_i to reflect the cutoff:

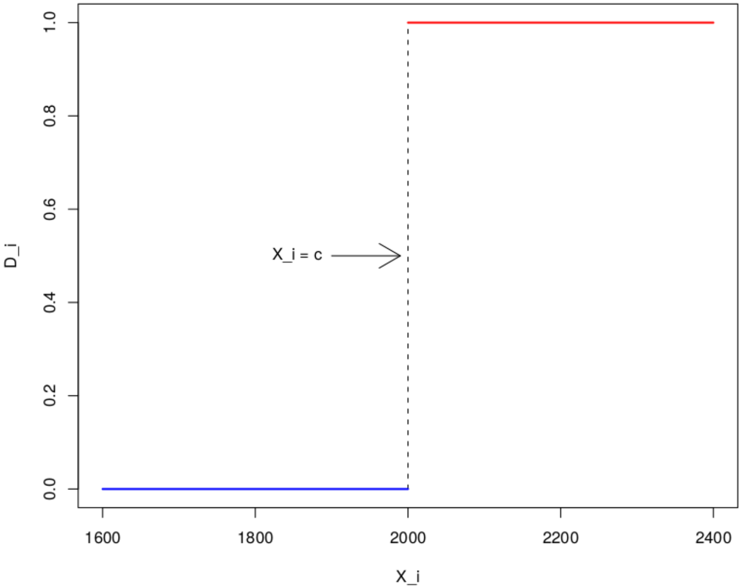
$$Y_i = Y_i(0) \times \mathbb{1}\{X_i < c\} + Y_i(1) \times \mathbb{1}\{X_i \geq c\}$$

Define $m_0(X_i) = \mathbb{E}[Y_i(0)|X_i]$ and $m_1(X_i) = \mathbb{E}[Y_i(1)|X_i]$ to be the CEFs of the potential outcomes given X_i. Then, the causal effect of the treatment is:

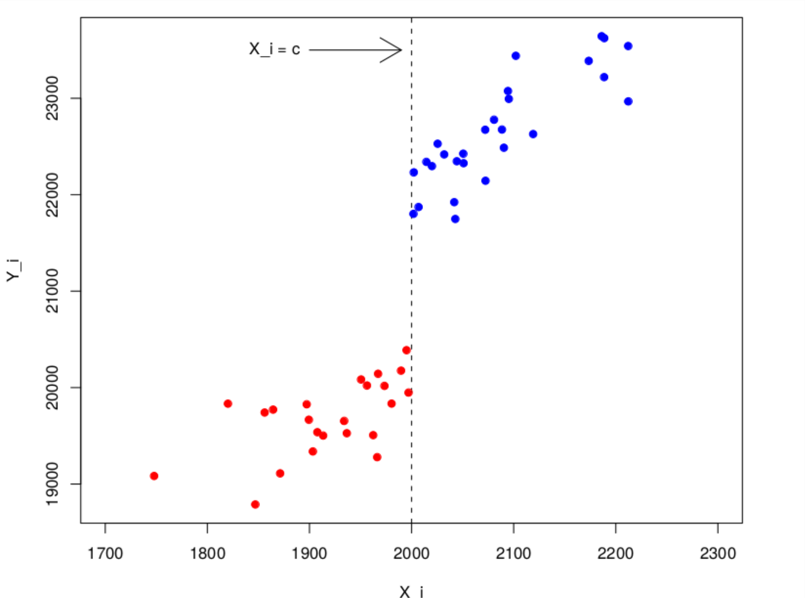
$$m(X_i) = m_0(X_i) \times \mathbb{1}\{X_i < c\} + m_1(X_i) \times \mathbb{1}\{X_i \geq c\}$$

The treatment effect at the discontinuity is $\tau(c) = m_1(c)$ - $m_0(c)$.

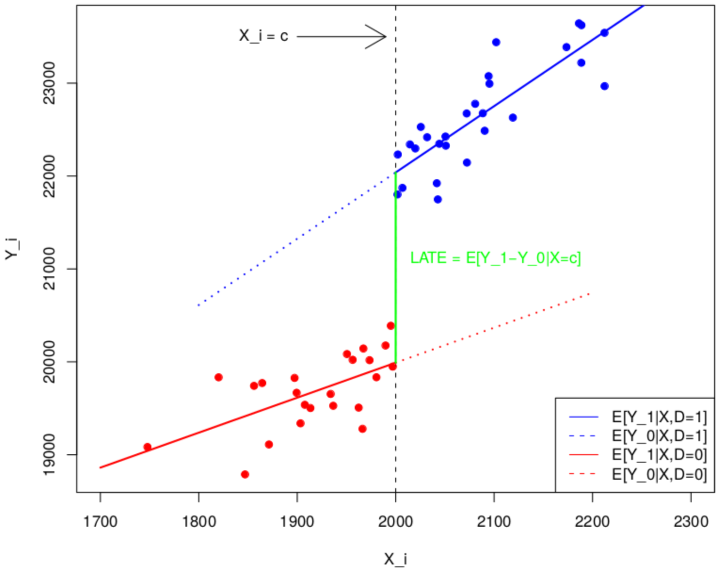
Key RDD Idea Graphically



Key RDD Idea Graphically



Key RDD Idea Graphically



Modeling a RDD

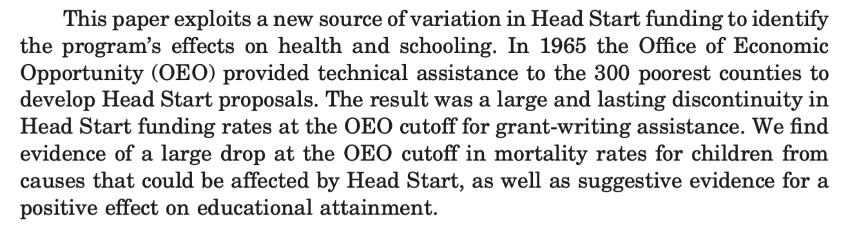
We can use the linear regression framework to model the discontinuity, centering the discontinuity at 0 aids in interpretation (ie, $\beta_2 = \tau(c) = \text{ATE}$):

$$Y_i = \beta_0 + \beta_1(X_i \text{ - } c) + \beta_2 \mathbb{1}\{X_i \geq c\} + \beta_3(X_i \text{ - } c)\mathbb{1}\{X_i \geq c\} + e_i$$

Alternatively, you can fit any model separately to each side of the cutoff, and then compare the predicted values at the cutoff. This is commonly done with a non-parametric function, such as a spline or local-linear regression.

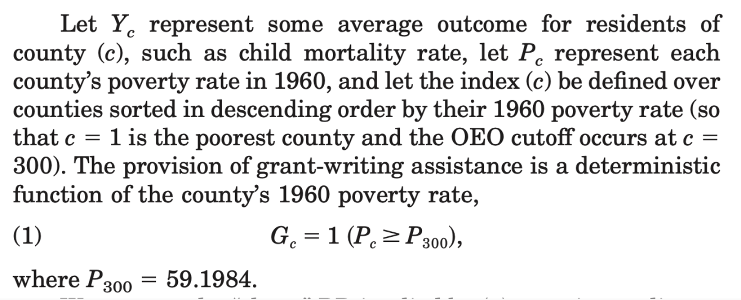
Example: Does “Head Start” Improve Children’s Life Chances?

Ludwig & Miller (2007) (link) studies children’s health and schooling outcomes related to the Head Start Program.



This paper exploits a new source of variation in Head Start funding to identify the program’s effects on health and schooling. In 1965 the Office of Economic Opportunity (OEO) provided technical assistance to the 300 poorest counties to develop Head Start proposals. The result was a large and lasting discontinuity in Head Start funding rates at the OEO cutoff for grant-writing assistance. We find evidence of a large drop at the OEO cutoff in mortality rates for children from causes that could be affected by Head Start, as well as suggestive evidence for a positive effect on educational attainment.

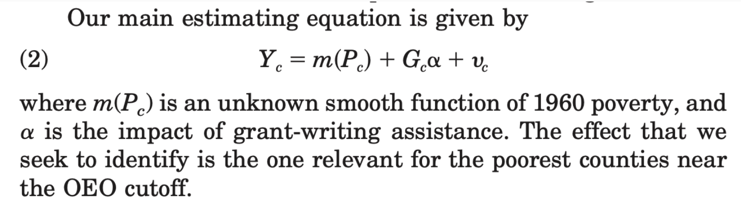
Example: Does "Head Start" Improve Children's Life Chances?



Let Y_c represent some average outcome for residents of county (c), such as child mortality rate, let P_c represent each county's poverty rate in 1960, and let the index (c) be defined over counties sorted in descending order by their 1960 poverty rate (so that $c = 1$ is the poorest county and the OEO cutoff occurs at $c = 300$). The provision of grant-writing assistance is a deterministic function of the county's 1960 poverty rate,

$$G_c = 1\,(P_c \geq P_{300}), \tag{1}$$

where $P_{300} = 59.1984$.

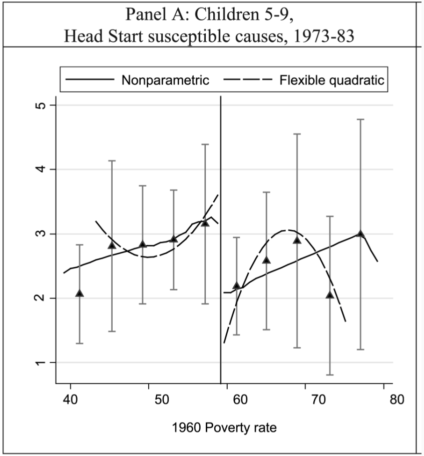


Our main estimating equation is given by

$$Y_c = m(P_c) + G_c\alpha + v_c \tag{2}$$

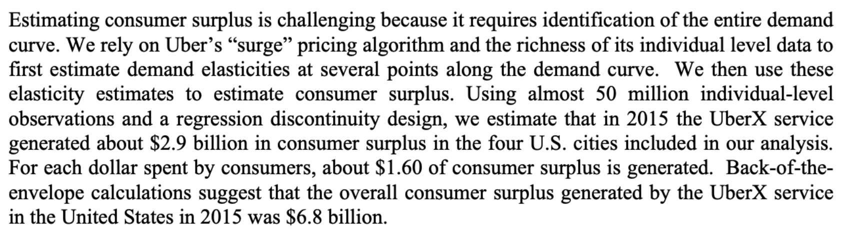
where $m(P_c)$ is an unknown smooth function of 1960 poverty, and α is the impact of grant-writing assistance. The effect that we seek to identify is the one relevant for the poorest counties near the OEO cutoff.

Example: Does “Head Start” Improve Children’s Life Chances?



Example: Using Big Data to Estimate Consumer Surplus: The Case of Uber

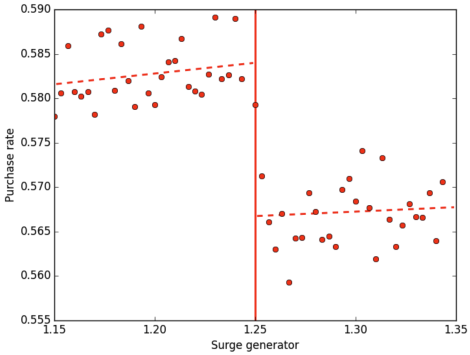
Cohen, Hahn, Levitt, & Metcalfe (2016) (link) study the consumer surplus generated by Uber.



Estimating consumer surplus is challenging because it requires identification of the entire demand curve. We rely on Uber's "surge" pricing algorithm and the richness of its individual level data to first estimate demand elasticities at several points along the demand curve. We then use these elasticity estimates to estimate consumer surplus. Using almost 50 million individual-level observations and a regression discontinuity design, we estimate that in 2015 the UberX service generated about $2.9 billion in consumer surplus in the four U.S. cities included in our analysis. For each dollar spent by consumers, about $1.60 of consumer surplus is generated. Back-of-the-envelope calculations suggest that the overall consumer surplus generated by the UberX service in the United States in 2015 was $6.8 billion.

Example: Using Big Data to Estimate Consumer Surplus: The Case of Uber

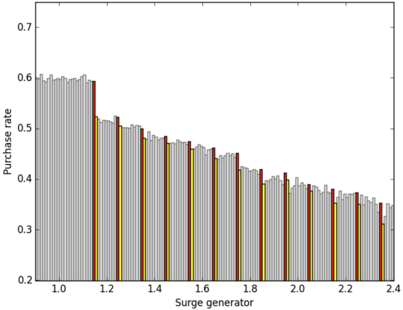
Figure 4: Example of purchase rate changes at price discontinuity



Note: This figure illustrates how purchase rates vary as a function of the surge generator over the range 1.15x to 1.35x. The vertical line when the surge generator equals 1.25 identifies the point at which the surge price changes from 1.2x to 1.3x.

Example: Using Big Data to Estimate Consumer Surplus: The Case of Uber

Figure 5: Request rate drops at pricing discontinuities



Note: This figure illustrates how purchase rates vary as a function of the surge generator when the surge generator is less than 2.4x. Red bars identify all observations within .01 units to the left of a price discontinuity. Yellow bars identify all observations within .01 units to the right of a price discontinuity. All observations not within these windows are depicted in gray.

Differences in Differences

Diff in Diff Idea

Suppose you have two virtually-identical units:

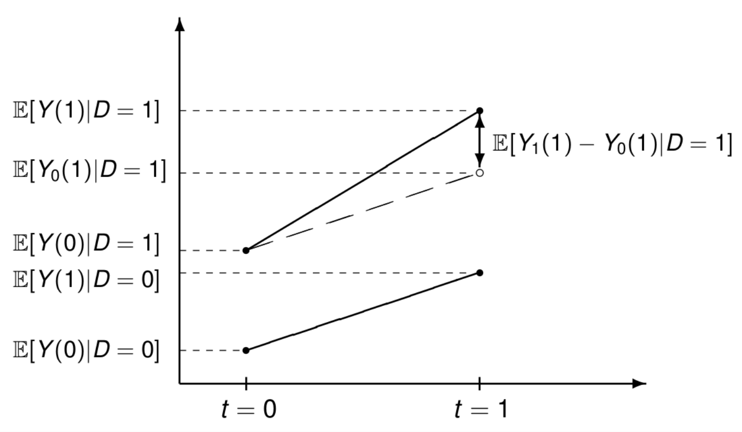
- You measure them over time
- One of them is treated with a policy intervention
- You want to know the effect of the policy intervention

The non-treated unit provides the counterfactual:

- what would have happened to the treated unit if it had not been treated?

Diff in Diff Visually

Adopt the notation $Y_d(t)$ to denote the outcome Y at time t with treatment status d.



Regression for Diff in Diff

With two time periods ($T_i = \{0, 1\}$) and two groups ($D_i = \{0, 1\}$), we can write the diff in diff as a regression:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 T_i + \beta_3 D_i T_i + e_i$$

where D is a dummy variable for treatment status and T is a dummy variable for time period.

$\beta_3 =$ ATT, the average treatment effect on the treated.

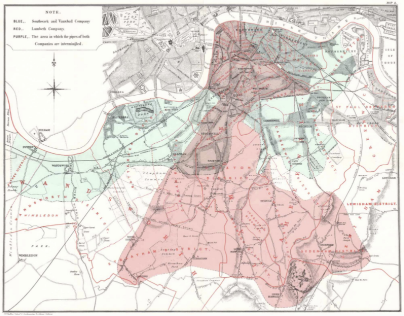
Example: Cholera in London, 1855

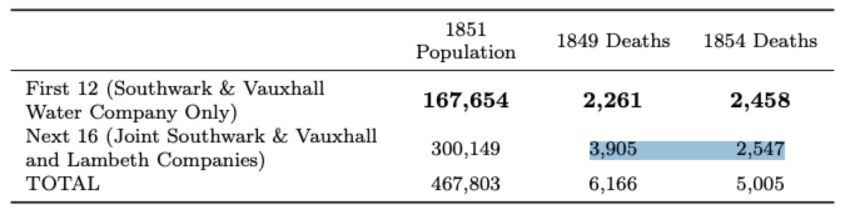
In 1949 and 1854, there were cholera outbreaks in London.

Two water companies served London:

1. Southwark & Vauxhall Company
2. Lambeth Water Company

Between the outbreaks, the Lambeth Water Company changed it's water source to be from the Thames River upstream of London, a cleaner source.

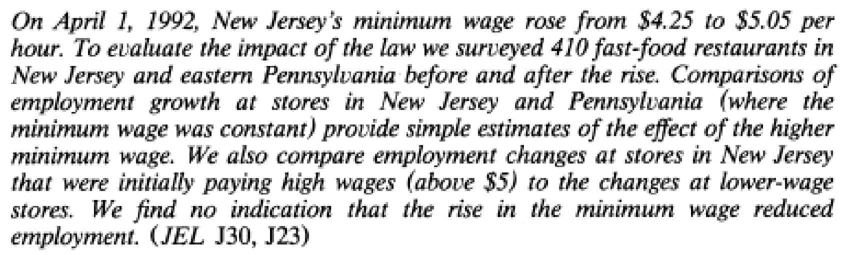

NOTE.
BLUE.. Southwark and Vauxhall Company
RED.. Lambeth Company.
PURPLE.. The area in which the pipes of both Companies are intermingled.
LEWISHAM DISTRICT



	1851 Population	1849 Deaths	1854 Deaths
First 12 (Southwark & Vauxhall Water Company Only)	**167,654**	**2,261**	**2,458**
Next 16 (Joint Southwark & Vauxhall and Lambeth Companies)	300,149	3,905	2,547
TOTAL	467,803	6,166	5,005

Example: Minimum Wage Laws and Employment

Card & Krueger (1994) study the effect of a minimum wage increase in New Jersey in 1992 on employment.



On April 1, 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above $5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment. (*JEL* J30, J23)

Example: Minimum Wage Laws and Employment

Classical economics suggests that an increase in the minimum wage will lead to a decrease in employment.

In 1992, New Jersey increased the minimum wage from $4.25 to $5.05.

Card & Krueger (1994) (link) show that Pennsylvania, which did not increase the minimum wage, had a decrease in employment relative to New Jersey.

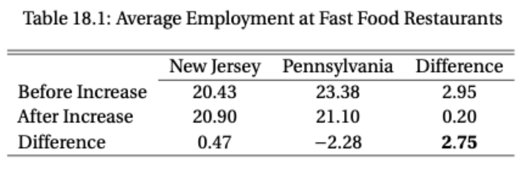
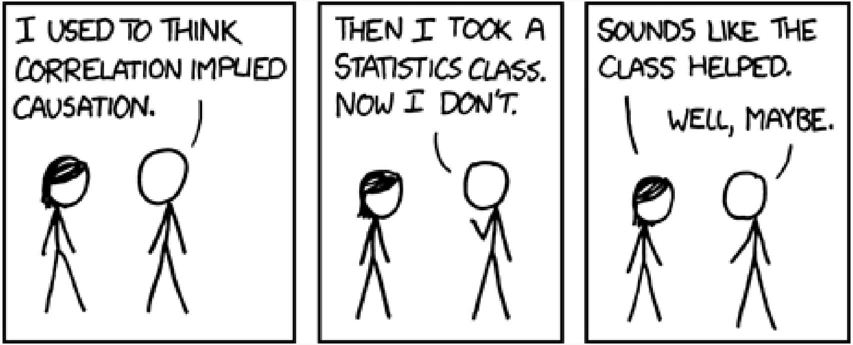


Table 18.1: Average Employment at Fast Food Restaurants

	New Jersey	Pennsylvania	Difference
Before Increase	20.43	23.38	2.95
After Increase	20.90	21.10	0.20
Difference	0.47	−2.28	**2.75**

XKCD


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Big and Growing Field

Resources:

- Angrist and Pischke, “Mostly Harmless Econometrics”
- Angrist and Pischke, “Mastering ’Metrics”
- Imbens and Rubin, “Causal Inference for Statistics, Social, and Biomedical Sciences”
- Morgan and Winship, “Counterfactuals and Causal Inference”
- Pearl, “Causality: Models, Reasoning, and Inference”
- Cunningham, “Causal Inference: The Mixtape”
- Huntington-Klein, “The Effect”

Next Time

- Define the Likelihood and Log-Likelihood functions
- Introduce the Method of Maximum Likelihood Estimation (MLE)
- Provide several one-parameter and multi-parameters examples