Causal Inference 2: Targetting, Observational Methods

Lecture 16

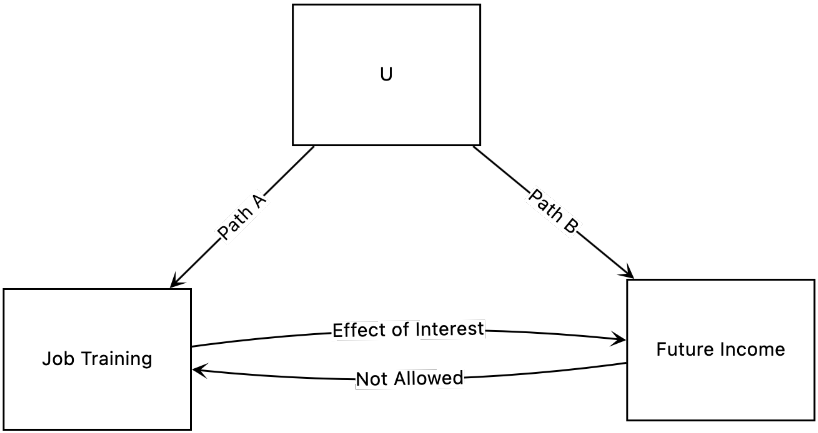
Connor Dowd

May 20th, 2021

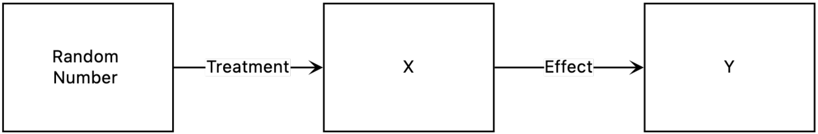
Today's Class

1. Review
   - RCTs
   - Individual TEs
2. RCTs: Targeting
3. Observational Methods: IV
   - RD
4. Observational Methods: Diff-in-Diff
   - SCM
5. HW6 Review

RCTs


U
Path A
Path B
Job Training
Effect of Interest
Not Allowed
Future Income

RCTs



Similarity

Typically the ATE is:

$$\widehat{ATE} = \bar{y}_1 - \bar{y}_0$$

But we could also define it as the average across individual treatment effects.

$$\widetilde{ATE} = \frac{1}{n}\sum_{i=1}^{n} \widehat{TE}_i$$

With a lot of rewriting of sums – we can show that when our individual treatment effects use a constant mean to predict counterfactuals:

$$\widetilde{ATE} = \widehat{ATE}$$

Proof in Data

```r
ybar0 = mean(jtpa$y[jtpa$offer == 0])
ybar1 = mean(jtpa$y[jtpa$offer == 1])
ybar1-ybar0
```

```
## [1] 1159.433
```

```r
jtpa_est = jtpa %>%
    mutate(y1 = y*offer+(1-offer)*ybar1,
           y0 = y*(1-offer)+offer*ybar0) %>%
    mutate(TE = y1-y0)
mean(jtpa_est$TE)
```

```
## [1] 1159.433
```

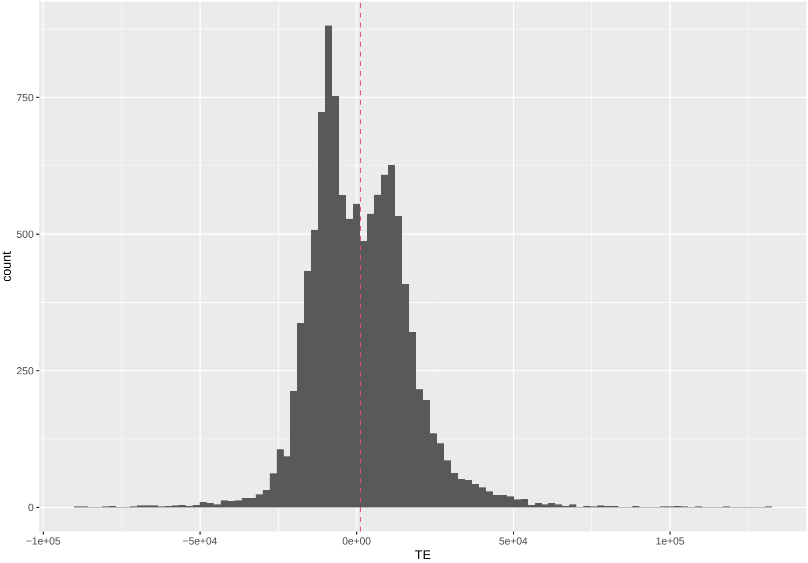
But maybe we can improve on the mean as a prediction.

ATE from predictions:

```
#Build Treat and control dfs
jtpa_cont = jtpa %>% filter(offer == 0)
jtpa_treat = jtpa %>% filter(offer == 1)
#Estimate treat and control models
mod_treat = ranger(y~.-offer,data=jtpa_treat)
mod_cont = ranger(y~.-offer,data=jtpa_cont)
#Predict counterfactuals for data from other model
jtpa_cont$confact = predict(mod_treat,data = jtpa_cont)$pre
jtpa_treat$confact  = predict(mod_cont,data = jtpa_treat)$p
#Estimate TEs
jtpa_cont$TE = jtpa_cont$confact - jtpa_cont$y
jtpa_treat$TE = jtpa_treat$y - jtpa_treat$confact
#Recombine
jtpa_est = rbind(jtpa_cont,jtpa_treat)
mean(jtpa_est$TE) #ATE
```

```
## [1] 1155.634
```

Individual TEs



Individual TEs, Do we care?

It looks like some fraction of individuals lost ~\$50k by engaging in this program. Not to mention the program cost to the government.

- What if we could target the program to people who benefit?
- In other settings, like marketing, we may wish to target groups for whom the expense of advertising is less than the gain in revenue from those individuals.

$\Longrightarrow$ Targeting. Can we use the RCT data for targeting?

Targeting

The goal of using targeting is *to identify subpopulations who benefit more*.

This will always be about *averages* in some subgroup – we can't predict outliers. Consequentially, there may be excluded individuals who benefit or suffer from exclusion.

- Do not forget about this.

Targeting

It is *also* at times very questionable, concerning, and even illegal to target *some groups*. Particularly if you are advertising financial products, or developing a government program.

- Importantly, you can wind up targeting (e.g. race) *even* if you don't observe relevant variables directly
    - Zip code can be a strong proxy.
- I'm going to drop variables for now – (and rerun prior code)

```r
jtpa = jtpa %>% select(-male,-black,-hispanic)
```

Simple Targeting

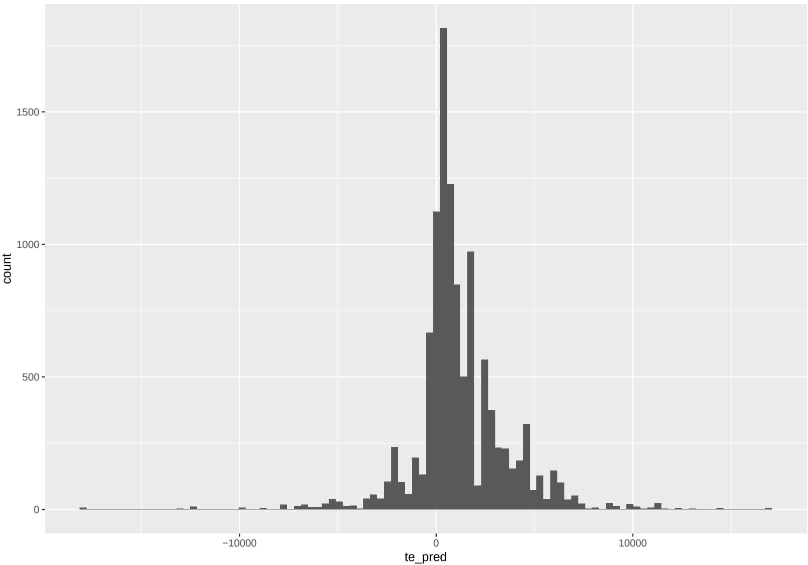
Take an observation, predict outcome under treatment and control, take the difference, compare to some threshold.

Simple Targeting

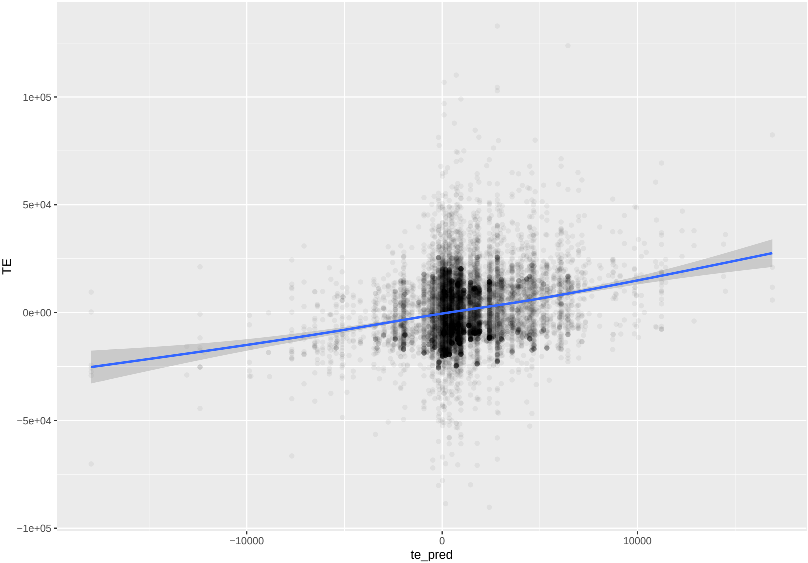
```
pred_treat = predict(mod_treat,data=jtpa_est)$predictions
pred_cont  = predict(mod_cont ,data=jtpa_est)$predictions

jtpa_est$te_pred = pred_treat-pred_cont
```

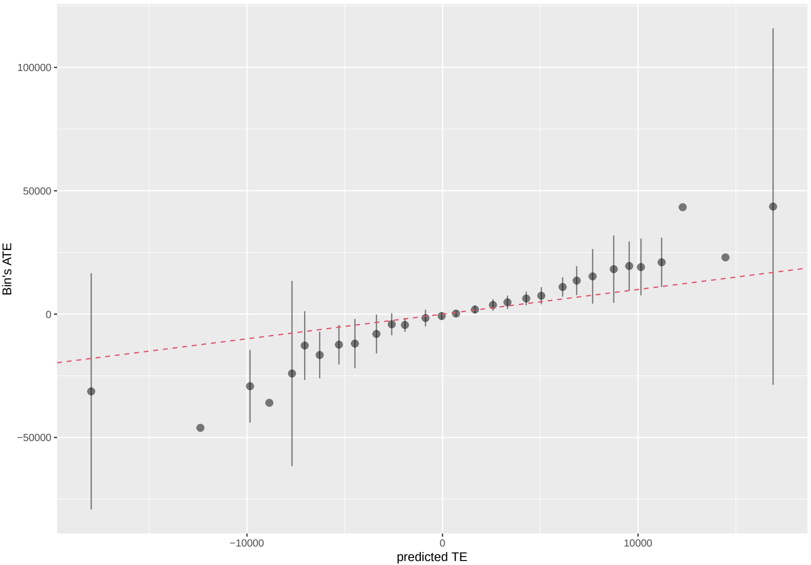
Simple Targeting



Simple Targeting – Smoothed Conditional Means



Simple Targeting – Bin Scatter



Simple Targeting – A threshold

Suppose we picked a cutoff of $0. We can look at which *predicted* treatment effects are below 0, and which *estimated* treatment effects are below 0, and think about performance.

```r
jtpa_est = jtpa_est %>% mutate(predte0 = te_pred > 0,
                               TE0 = TE > 0)
table(jtpa_est$TE0,jtpa_est$predte0,deparse.level=2)
```

```
##              jtpa_est$predte0
## jtpa_est$TE0 FALSE TRUE
##        FALSE  1288 4405
##        TRUE    824 4687
```

Simple Targeting - Quick test

```
chisq.test(jtpa_est$TE0,jtpa_est$predte0)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity corre
##
## data:  jtpa_est$TE0 and jtpa_est$predte0
## X-squared = 107.26, df = 1, p-value < 2.2e-16
```

So these aren't unrelated.

Simple Targeting – A threshold

That felt like our classification work. Could we use an ROC curve?

It will take some modification. For starters, our thresholds aren't now restricted to 0,1.

But more importantly, we aren't observing the true outcome of the classification – even that is a prediction.

- Easier to look at the mean ATE in the two groups that our classification threshold produces.
- Also important to use OOS predictions. (One step at a time)

Simple Targeting - Measuring Performance

Continuing with our Threshold of 0 for a moment. Lets calculate mean ATE for each group our predictions and threshold create.

```
jtpa_TEs = jtpa_est %>%
    group_by(offer,predte0) %>%
    summarize(y=mean(y)) %>%
    pivot_wider(names_from = predte0, values_from = y)
jtpa_TEs
```

```
## # A tibble: 2 x 3
## # Groups:   offer [2]
##   offer ‘FALSE‘ ‘TRUE‘
##   <dbl>   <dbl>  <dbl>
## 1     0  18520. 14222.
## 2     1  14503. 16592.
```

Simple Targeting - Measuring Performance

Find the actual “ATE” for each subpopulation:

```
jtpa_TEs[2,2:3]-jtpa_TEs[1,2:3]
```

```
##        FALSE     TRUE
## 1 -4017.017 2369.809
```

So for the group the proposed policy would block, the ATE was substantially negative, and for the group it would try to target, the ATE was much more positive than for the entire population.

- Substantial Improvement in our performance. *IF* it holds out of sample.

Simple Targeting

But first – we should consider other thresholds. . .

Going to plot out possible other thresholds, and the "gain" in performance from using those thresholds.

"Net Gain" is going to be "benefit-(avoided loss)". For an indiscriminate policy (all eligible), this is the pure ATE.

For a policy targeting some group and blocking some group (of equal size), this is "ATE(targeted group)-ATE(blocked group)".

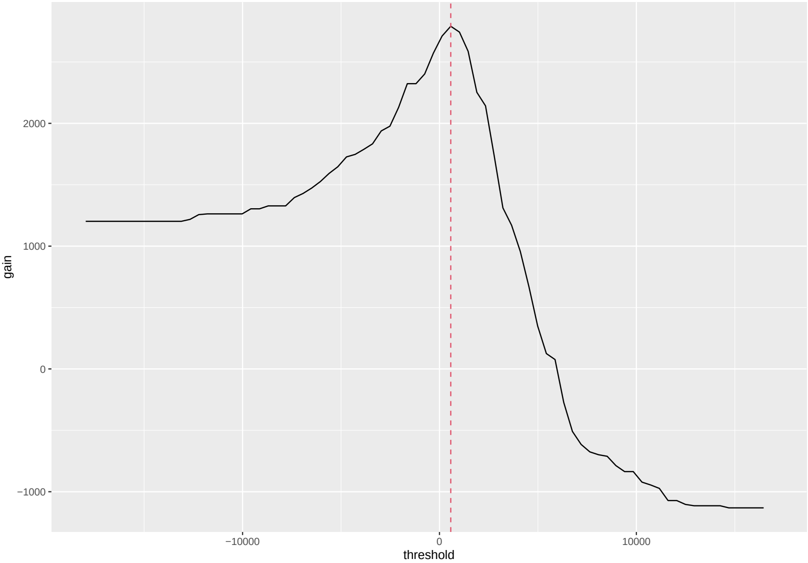
- Notice – if we block a group which would benefit – this is a loss here. (We are assuming no cost of the policy – big loss assumptions baked in)

Simple Targeting - ROC

Built a function, critical elements below

```
threshold_ATES = function(threshold,y,predte,
                          treatment,MC=0,FC=0) {
  n = length(predte)
  targets = predte > threshold
  ATE_treat   = mean(y[treatment == 1 & targets == 1])-
                mean(y[treatment == 0 & targets == 1])
  ATE_untreat = mean(y[treatment == 1 & targets == 0])-
                mean(y[treatment == 0 & targets == 0])
  ntreat = sum(targets)
  nuntreat = n-ntreat
  gain = ntreat*ATE_treat-nuntreat*ATE_untreat
  gain = gain-MC*ntreat-FC
  gain = gain/n
  c(threshold,ATE_treat,ATE_untreat,ntreat,nuntreat,gain)
}
```

Gain



```
## [1] 573.0293
```

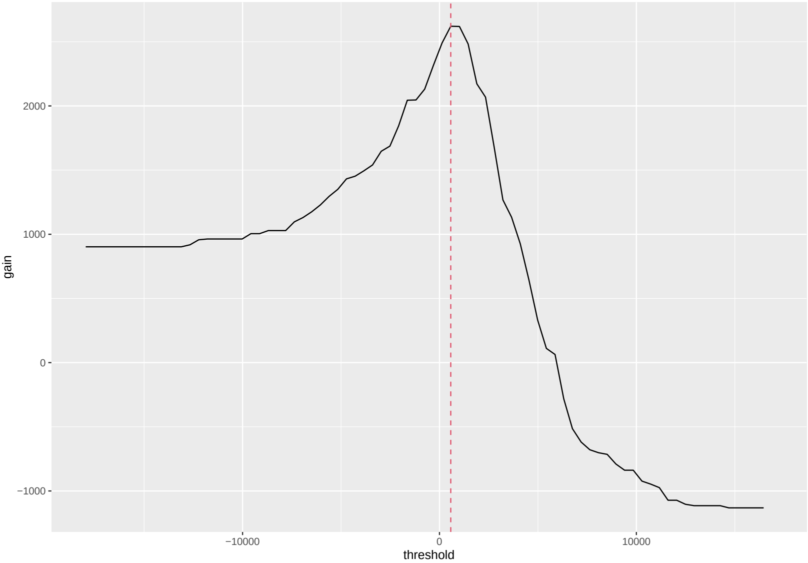
Costs

Suppose we had some cost for each person treated (call it \$300).
And a fixed cost (\$1000).

Where is the cutoff?

Changes our “gains” function.

New plot



```
## [1] 573.0293
```

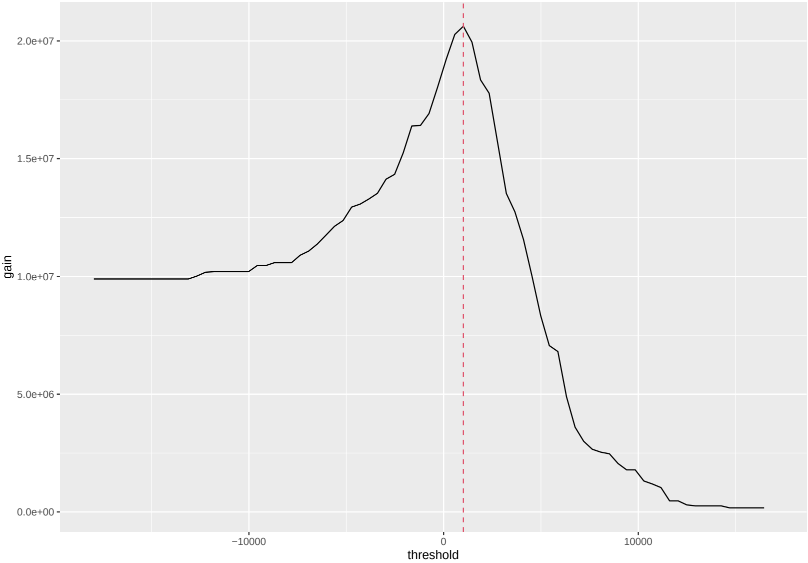
Interpretation

- First gains plot was “average benefits *to individuals* of targeting over full treatment” as a function of thresholds
- Second plot was “average benefits to individuals *above costs* of targeting over full treatment” as a function of thresholds

But maybe we aren’t currently implementing anything?

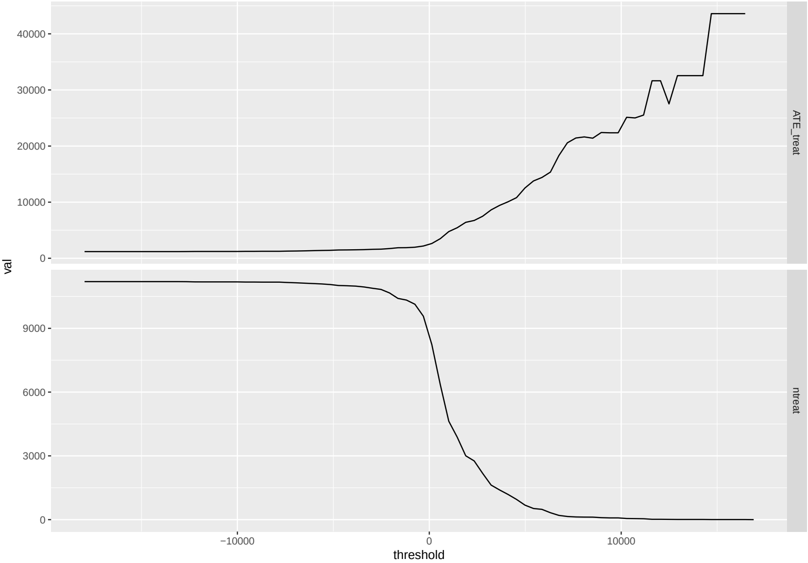
- Then it would be inappropriate to subtract off the ATE for the untreated group.

New plot



```
## [1] 1014.397
```

Plot 2



40000
30000
20000
10000
0
ATE_treat
val
9000
6000
3000
0
ntreat
−10000
0
10000
threshold

OOS Targeting measures.

USE CROSS VALIDATION.

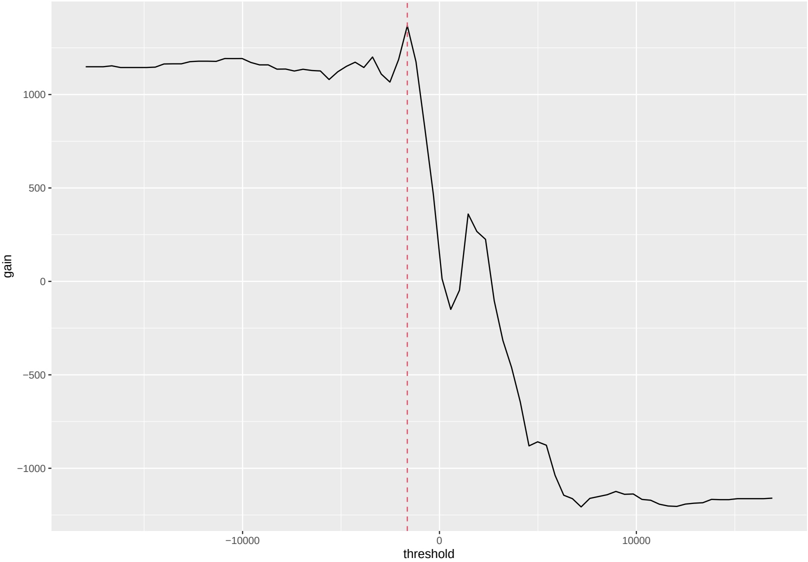
OOS Targeting

```
OOS_tepreds = function(holdout_ind,data,formula = y~.-offer
    holdout = data[holdout_ind,]
    train = data[-holdout_ind,]
    jtpa_cont = train %>% filter(offer == 0)
    jtpa_treat = train %>% filter(offer == 1)
    #Estimate treat and control models
    mod_treat = ranger(y~.-offer,data=jtpa_treat)
    mod_cont = ranger(y~.-offer,data=jtpa_cont)
    pred_treat = predict(mod_treat,data=holdout)$prediction
    pred_cont  = predict(mod_cont ,data=holdout)$prediction
    te_pred = pred_treat-pred_cont
    cbind(holdout_ind,te_pred)
}
```

Kfold CV

```r
k = 20
fold_ids = sample(rep(1:k,length.out = nrow(jtpa)))
hold_indices = lapply(1:k,function(foldk) which(fold_ids ==
oos_preds = lapply(hold_indices,OOS_tepreds,data=jtpa)
jtpa_est = jtpa
jtpa_est$oos_tepreds = NA
for (i in 1:k){
    jtpa_est$oos_tepreds[oos_preds[[i]][,1]] = oos_preds[[i
}
```

Plots



```
## [1] -1633.808
```

Targeting

- Powerful tool in many domains
  - Optimal Policy Design
  - Marketing
  - Medical treatment choice
  - Many more
- Care and Caution needed
  - Discrimination issues
  - OOS/overfitting issues
    - We may think we are finding good targeting things – and really be failing
    - Our standard tools for examining and improving predictions will help here.

RCT Wrapup

RCTs can have problems arise in several ways.

- Placebo effects: I know I recieved fancy treatment
- Spillover effects: My friend got tutoring and I benefit from that
- and others.

They can be great – but they aren't flawless.

Observational Causal Inference

Frequently we care about causal issues in contexts where we can't run an experiment.

There are still tools for trying to assess a causal effect. They will rely on *something* being **"as if"** random.

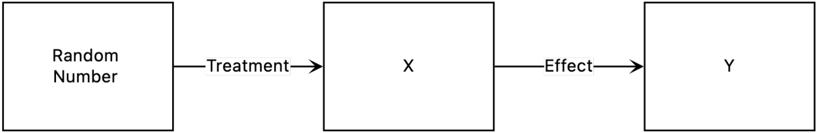
Instrumental Variables

The simplest example here is “instrumental variables”.

Recall, when we use an RCT, we assume some random number generator influences X without having any way to effect Y, and without being affected by Y.

- Thus we can use the RNG output to predict Y *only* because of the causal effect of X

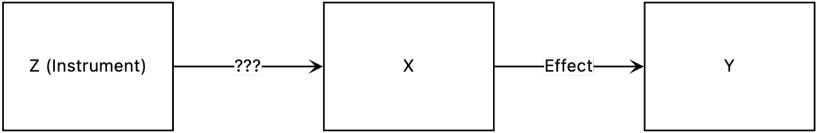
RCT DAG



Instrumental Variables

IV takes the same idea. What if I had some other variable – not a random number – which affects X, but can't affect Y or be affected by Y. Then I could predict Y as a function of that random variable – and any relationship would be only because of X.

IV DAG



IV - Example

Many Examples out there.

One classic example is testing the relationship between lead and crime (or other things like test scores)

Essentially, we can look at a city's proximity to lead foundries to predict their usage of lead pipes. Then we can try to use the proximity to lead foundries to predict crime.

"In the absence of other plausible relationships between lead foundry proximity and crime, any effect we find comes from the lead".

IV - Estimation

Simplest method (“2SLS”) runs a linear regression between Z and X, getting β_X, then another regression between Z and Y, getting β_Y.

The estimated effect of X on Y is β_Y/β_X.

IV - Problems

- Weak IV: If β_X is near 0, our estimates for the effect head towards ∞ or $-\infty$. This is a bigger problem than it may seem.
  - Z needs to be relevant to X
- Exclusion: If Z has other channels for affecting Y, then we can't interpret this as the causal effect of X on Y.
  - E.g. if lead foundries disproportionately went out of business, causing unemployed workers to form mobs that committed crimes.

IV - Other Uses

Implicitly, that 2SLS procedure of dividing one coefficient by another is how we estimated "average treatment effect on the treated" on Tuesday.

Divide the ATE by the fraction recieving treatment

Regression Discontinuity

Regression discontinuity is similar to IV.

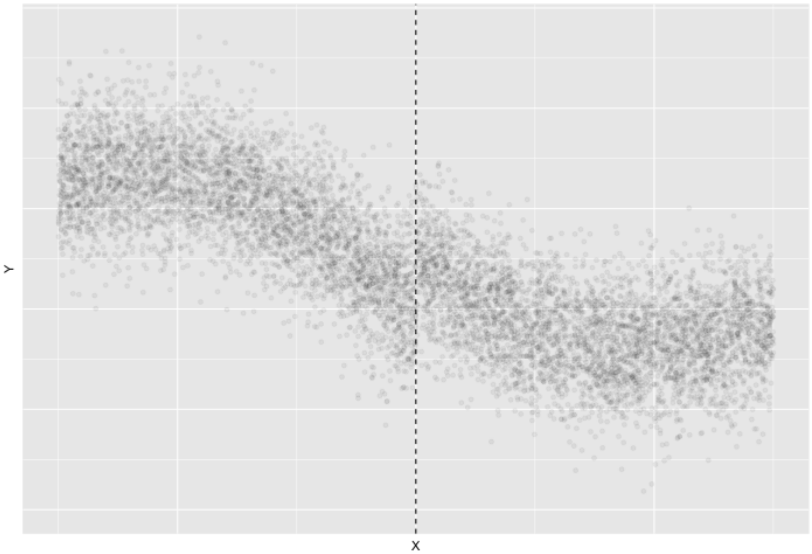
It takes advantage of some policy threshold, on one side of which individuals receive treatment, and on one side of which they don't.

RD Example

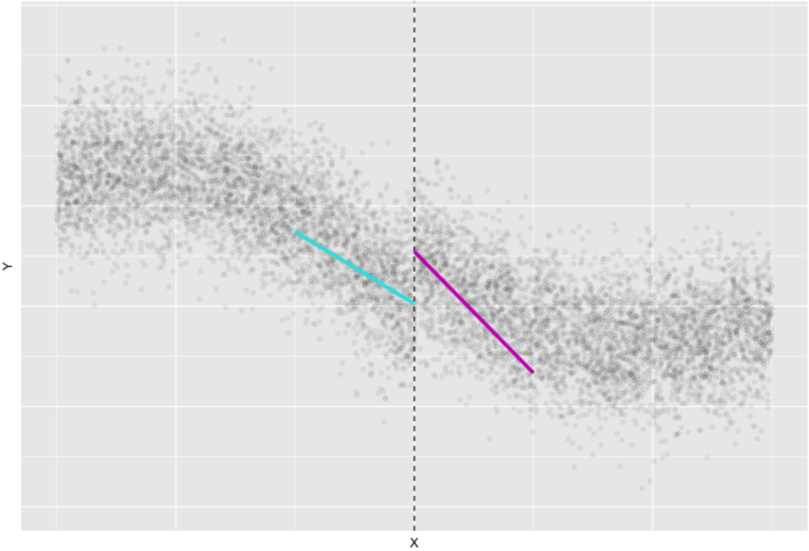
Classic examples involve looking at students.

- E.g. Some schools have a GPA cutoff for being put on academic probation. We may want to know the effect of ‘academic probation’ on
  - dropout rates
  - future GPA
  - Etc

RD



RD



RD - Estimation

Typically you use nonparametric tools (local linear regressions above), with carefully chosen bandwidths (choose with CV if you like).

That will let you estimate $E[Y|X = \mathit{threshold}, T = 1]$ and for $T = 0$ (i.e. the intercept on each side). Then we take the difference.

RD - Notes

RD gives you a Conditional Average Treatment Effect – the RD estimate of the treatment effect is the estimated *average* treatment effect *for individuals at the threshold*.

RD - Problems

RD assumes individuals are fundamentally similar on either side of the threshold.

- But if individuals are aware of the threshold, and can control their position relative to it, then they may not be fundamentally similar
  - Individuals on right side may have chosen to be there
  - And on the left side, not chosen to be on the right side
  - Selection problems

Difference-in-Differences

Suppose we observe a state implement some policy.

- E.g. California imposes a large tax (\$0.25) on cigarettes.

And we want to know the effect of the cigarette tax on consumption.

What could we do?

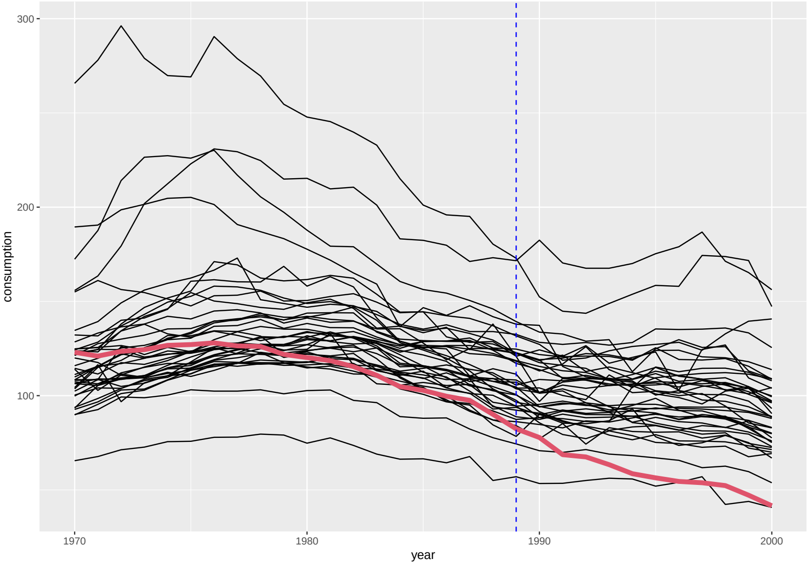
Difference-in-Differences

Recall the cigarette data. Per capita cigarette consumption in 38 states over 30 years. California policy implemented in 1989.

```
df = read_csv("https://codowd.com/bigdata/lectures/l13/ciga
df
```

```
## # A tibble: 31 x 40
##     year    AL    AR    CA    CO    CT    DE    GA    IA
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  1970  89.8  100.  123   125.  120   155   110.  108.
##  2  1971  95.4  104.  121   126.  118.  161.  116.  108.
##  3  1972 101.   104.  124.  134.  111.  156.  117   109.
##  4  1973 103.   108   124.  138.  109.  155.  120.  111.
##  5  1974 108.   110.  127.  133.  112.  151.  124.  116.
##  6  1975 112.   115.  127.  131   110.  148.  123.  120.
##  7  1976 116.   119.  128   134.  113.  153   126.  124.
##  8  1977 117.   123.  126.  132   117.  153.  128.  126.
##  9  1978 123    127.  126.  129.  118.  156.  131.  127.
## 10  1979 121    126   122   132   117   150   131   124
```

Quick Plot



Idea: State Differences

We could look at the difference between California's consumption and consumption in other states after the policy takes effect.

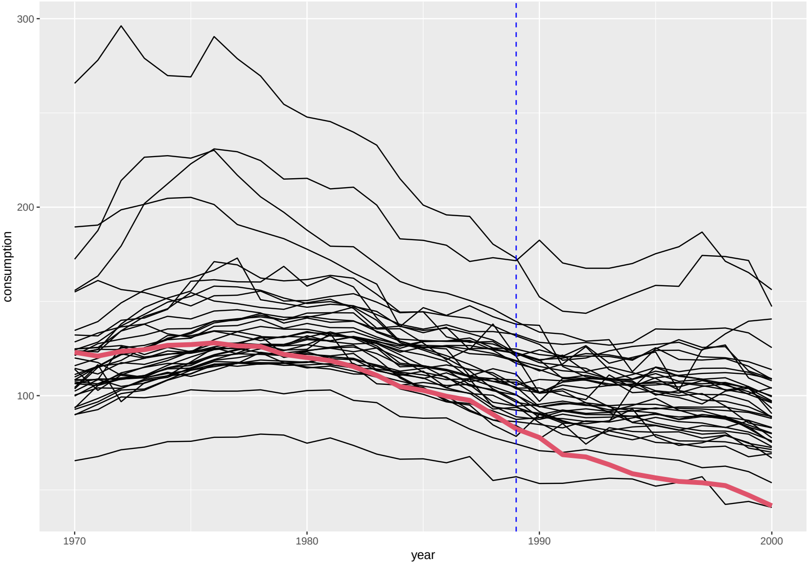
- Suppose California was randomly selected for the policy – this is how we would run an RCT analysis.
- But suppose California already consumed a different amount – we would be biased by the scale of pre-existing differences.

Idea: Time Differences

We could look at the difference between California's consumption before and after the policy takes effect.

- This would avoid concerns about differences between states.
- But if there was some kind of time trend, we would be adding a bias from the time trend.

Quick Plot

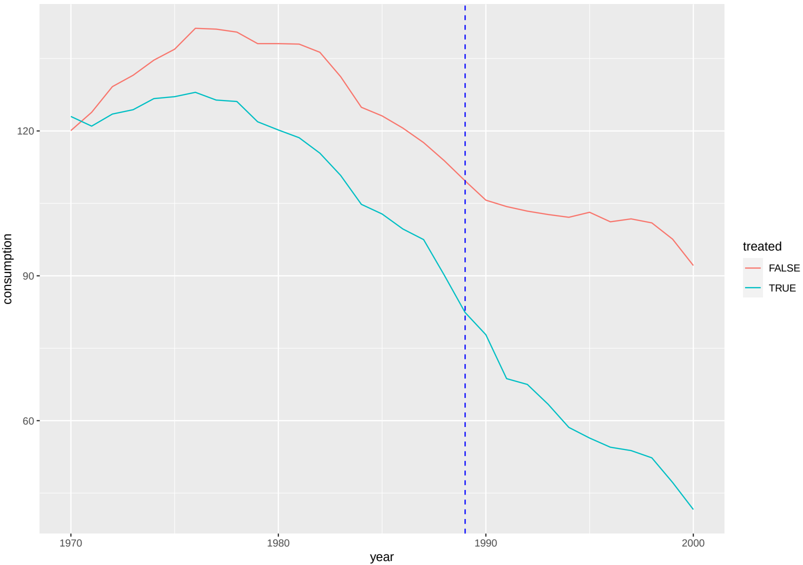


Difference-in-Difference

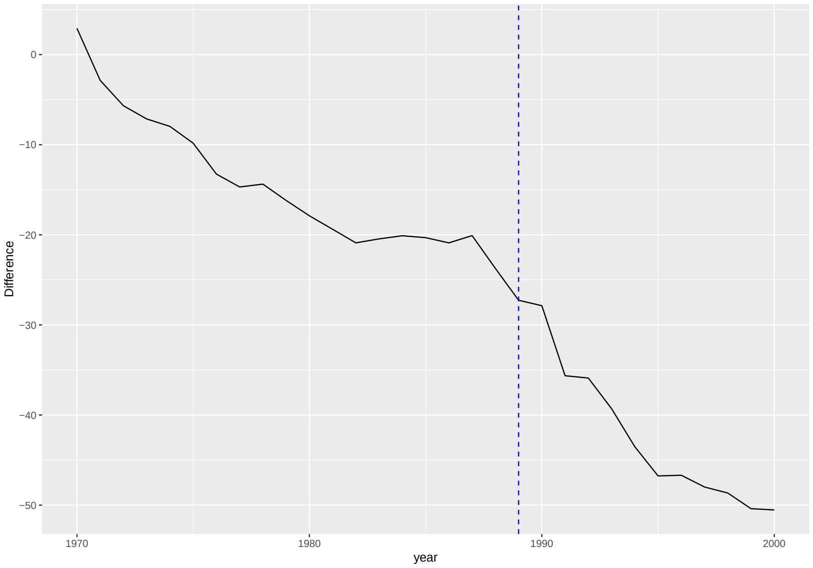
Diff-in-diff says: “why not both?”

- We can look at the difference between California and other states, and see how it changes over time.
  - Or we could look at each state’s changes over time, and see how California compares to other states.
    - These are equivalent.

Diff-in-Diff: Year means in treatment and control groups



Diff-in-Diff: Difference between year means



Diff-in-Diff

The difference looks to have gotten bigger - so there was probably an effect. The tax probably reduced consumption.

Diff-in-Diff: Notes

- Relies on the ‘parallel trends’ assumption.
  - Both treatment and control groups have time trends that are similar.
    - In that case, we can estimate the treatment effect.
- Usually relies on naive means in each of the treatment and control groups.
  - I.e. equal weighted means

Synthetic Controls

Synthetic controls builds on Diff-in-Diff by saying “can we improve those weights?”

- This should sound familiar.

We may be able to find weights for the control group that make it better at predicting the treatment group.

- This will let us relax the ‘parallel trends’ assumption somewhat
- It will also improve the statistical efficiency.

Synthetic Controls

I'm not going to go through code here. SCM is mostly beyond the scope of this class.

- But its useful for you *because the basic idea is familiar*
  - "Can we improve those weights for out of sample prediction"

Observational Causal Inference Wrapup

RCTs, IV, RD all rely on “exogenous variation” coming from something. Variation which is unrelated to the outcome of interest.

- RCTs: Varation from a random number generator
- IV: Variation from some instrument
- RD: Variation from randomness around some threshold

Diff-in-diff and SCM rely on finding the two sources of bias, and trying to eliminate them.

- Bias 1: our treated unit is fundamentally different
- Bias 2: out treated time periods are fundamentally different

These are different approaches with different strengths and weaknesses.

Causal Inference Wrapup

There is no magic bullet.

We have questions that desperately need answers.

We have data. That data *may* be able to help. But it may only mislead us.

- Critical to make decisions and implement policies which *fail gracefully*
  - We want things that work wonderfully if we're right about the world
  - But which aren't tremendously destructive if we're wrong.

HW 6 Review

If we have time I’ll review HW6 in a minute.

Wrap up

Homework 7 will be posted this afternoon, due next Wednesday. It is about causal inference.

Prediction competition 3 submissions must be in by midnight tomorrow.

Next Tuesday we will look Neural Nets, and next Thursday I'll try to review the whole of this course in 90 minutes.

See you next week!