

Causal Inference 1: RCTs

Lecture 15

Connor Dowd

May 18th, 2021

Today's Class

1. Review
 - ▶ Winsorizing
 - ▶ Bayes
 - ▶ Predictions Contest 2
2. Observational Data
3. RCTs: Average Treatment Effects
4. RCTs: Heterogenous Treatment Effects
5. RCTs: Targeting?
6. HW5 Review?

Review

Winsorizing

The standard routine does the following.

1. Pick some quantile (e.g. 1%).
2. Find that quantile – e.g. find the 99%ile $\text{abs}(\text{logerror})$
3. Set values higher than that quantile to that quantile.

Bayes Rule

$$P[\beta|X] \propto P[X|\beta]P[\beta]$$

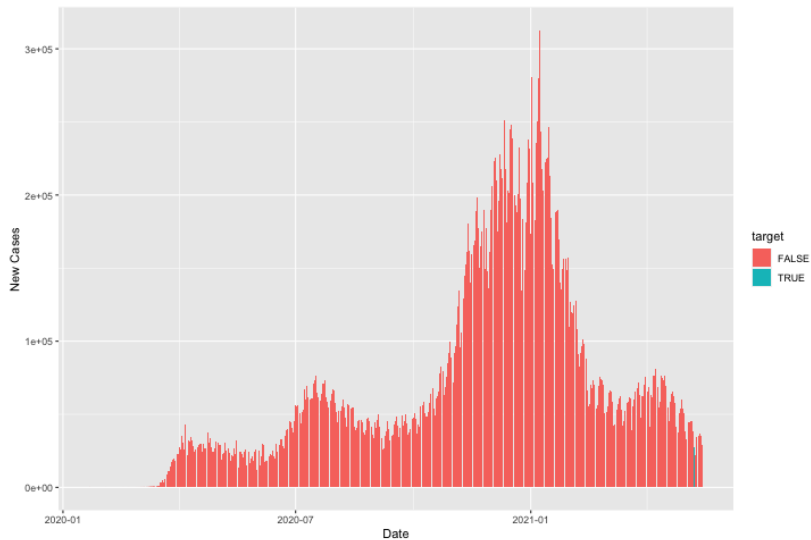
- ▶ $P[\beta|X]$ is referred to as the posterior
- ▶ $P[X|\beta]$ is the *likelihood* (we've seen before)
- ▶ $P[\beta]$ is the prior

Systematizing Uncertainty

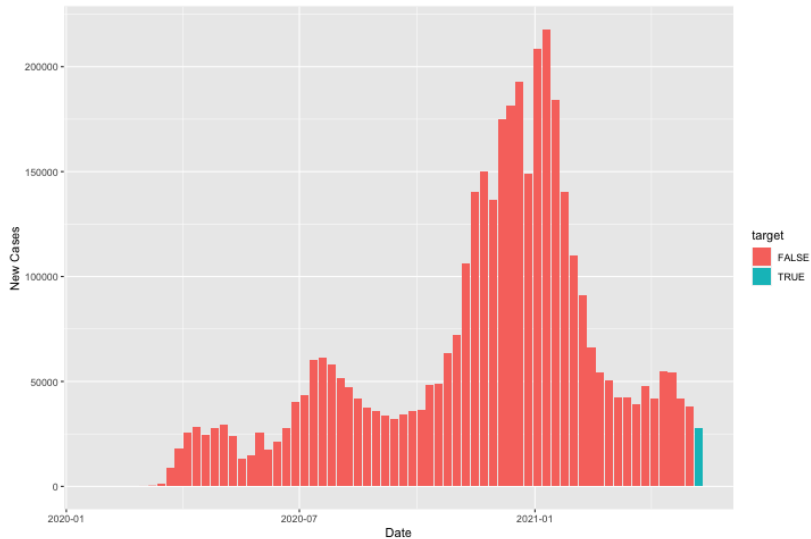
Primary Sources of Uncertainty in Modeling

1. Within the model, there is uncertainty.
 - ▶ What are parameter values
 - ▶ What value will an observation take
2. We are not certain if the model is correct
 - ▶ Should we use a different model
 - ▶ Should we average with a different model
3. We don't know if the data is correct
 - ▶ Is there a clerical error?
 - ▶ Is there measurement error?
 - ▶ E.g. Covid Case counts in the US last March – woefully undercounted.

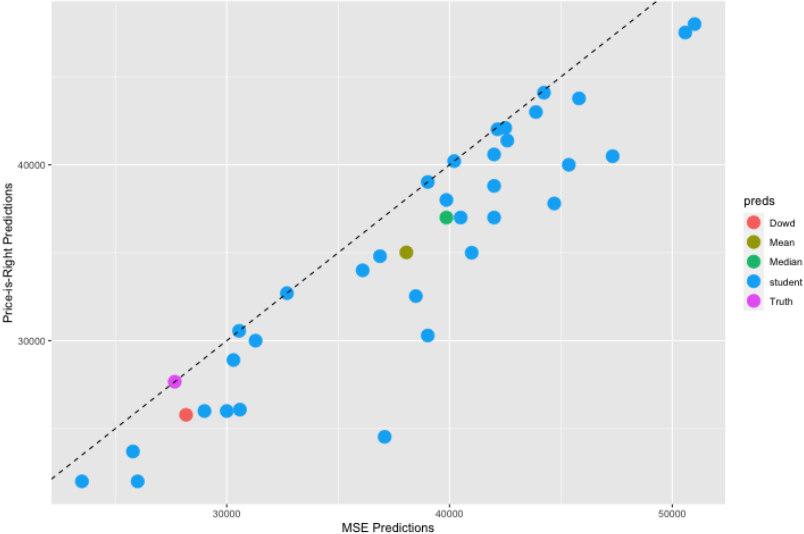
Predictions 2 - Daily Case Counts



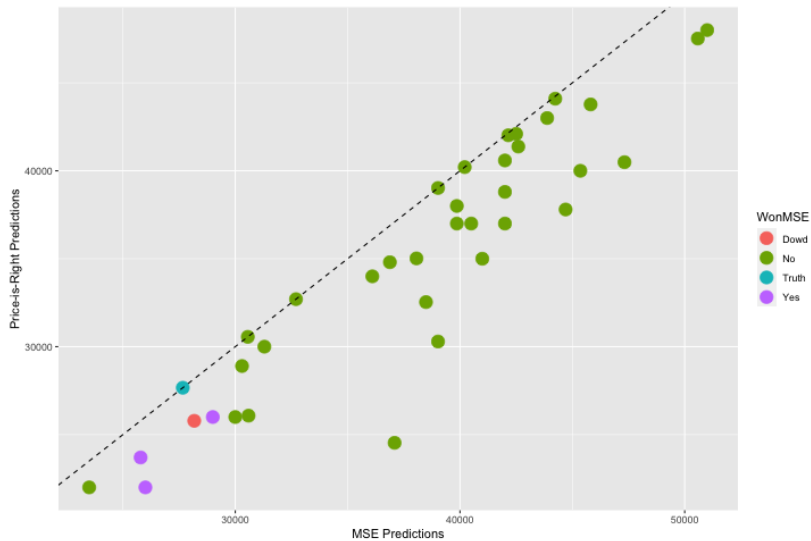
Predictions 2 - Sundays only



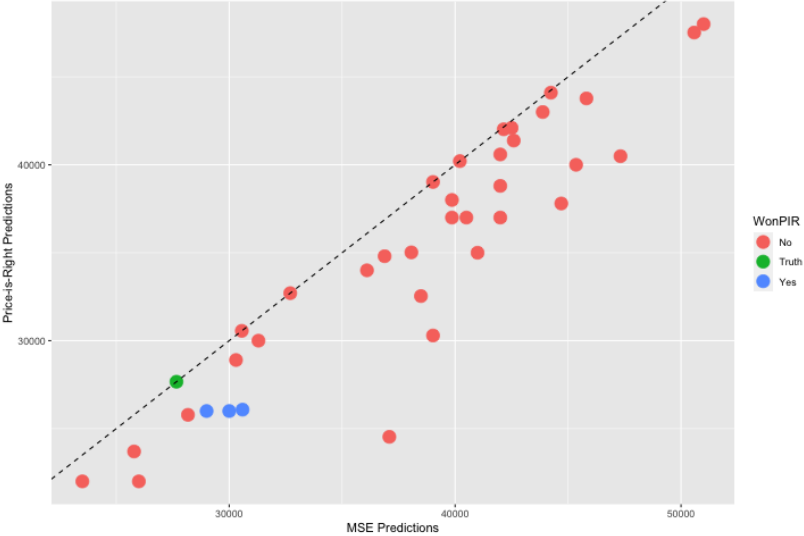
Predictions 2 - Actual Predictions



Predictions 2 - MSE Winners



Predictions 2 - Price-is-Right Winners



Predictions 2 - Wrapup

- ▶ Actual result: **27655**
- ▶ MSE Winners: Caden Kalinowski, Cagdas Okay, Matias Pietruszka
- ▶ PIR Winners: Frank Li, Caden, Amy Maldonado

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.282e+03	2.602e+03	-0.493	0.625
preda	9.537e-01	6.718e-02	14.196	1.31e-15 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2829 on 33 degrees of freedom
Multiple R-squared: 0.8593, Adjusted R-squared: 0.855
F-statistic: 201.5 on 1 and 33 DF, p-value: 1.311e-15

Predictions 3

New competition posted (this morning).

- ▶ Purely optional. Not graded. No canvas component.

US Sunday case counts *this sunday*. Two numbers:

1. Prediction
2. P[20% prediction error]

Possibly 1 more *optional* competition. Would likely run into the summer.

Observational Data

We have a question:

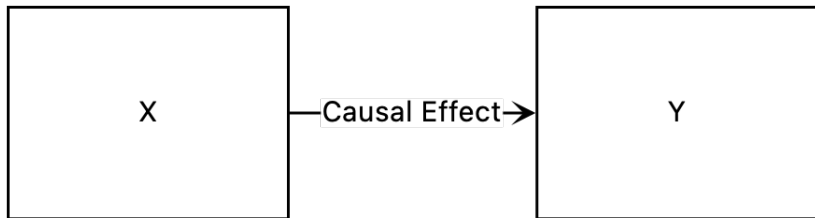
- ▶ Do job training programs for the unemployed improve worker's earnings?

We have data:

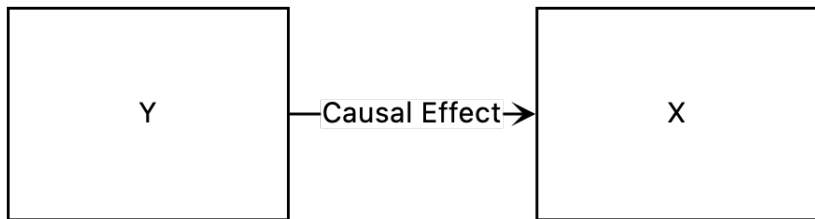
- ▶ Job training uptake, future earnings

Can we answer the question?

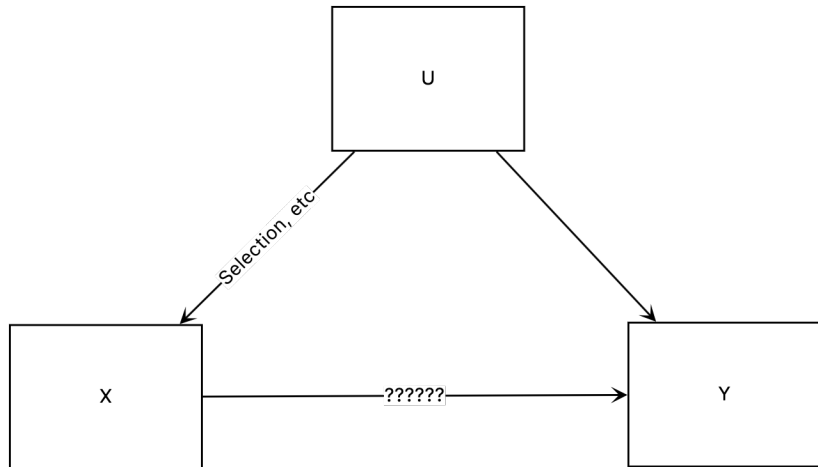
Formalizing This



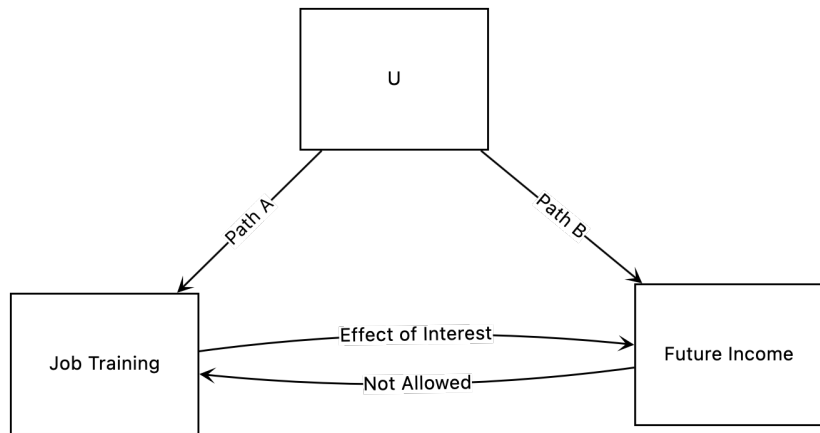
Formalizing This



Formalizing This



In Context



- ▶ Either Path A or Path B can't exist, for every other possible variable **U**.

Problem

If we have either (possibly unobserved) variables U that affect both X and Y , or Y might affect X , identifying causal effects of X on Y is *impossible* with your data.

Why? We can't disentangle the separate possible sources of changes.

E.g. Suppose extremely hard working individuals are more likely to *apply* for job training, and have higher future incomes. Or any other of a number of possible suppositions.

Assumption

In observational data, it is hard to guarantee “*No other variables have causal effects on both X and Y* ” and that “ *Y doesn't drive X* ”.

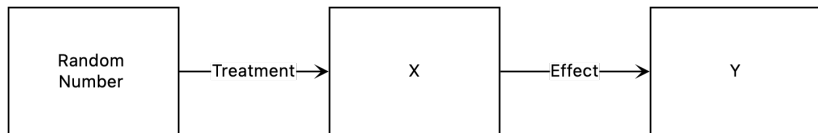
- ▶ We will see some settings, on Thursday, where we can make these assumptions more reasonable.
 - ▶ IV
 - ▶ RD
 - ▶ Diff-in-Diff
- ▶ But they will still be *assumptions*.

RCTs

For now, we turn to randomized controlled trials.

RCTs solve this problem, by introducing a variable which selects X , without having any relationship with Y – a random number.

RCTs



RCTs



RCTs

RCTs work because they force a treatment that is *unrelated* to any other possible variable, and which isn't being caused by Y .

So we can rule out the things we were worried about.

Job Training RCT

In the early 90s, a federal job training program (“JTPA”) ran an RCT where some people (~20k) were given offers immediately and some had eligibility delayed by 18 months *at random*.

There are a wide variety of outputs you could look at. We will look at the effect on income over the following 30 months on adults (~11k people).

First, some summary stats

JTPA Summary Stats

```
jtpa %>% select(-c(9:14)) %>%  
  group_by(offer) %>%  
  summarize(across(everything(), mean))
```

```
## # A tibble: 2 x 8
```

```
##   offer      y  train  male    hs black hispanic married  
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>    <dbl>  
## 1     0 15041. 0.0145 0.458 0.700 0.255    0.110    0.267  
## 2     1 16200. 0.642  0.454 0.712 0.262    0.109    0.286
```

Summary Stats 2

```
jtpa %>% select(c(2,9:14)) %>%  
  group_by(offer) %>%  
  summarize(across(everything(),mean))
```

```
## # A tibble: 2 x 7  
##   offer wkless13  AFDC  classroom OJT_JSA f2sms  age  
##   <dbl>    <dbl> <dbl>    <dbl>  <dbl> <dbl> <dbl>  
## 1     0     0.462 0.189     0.296  0.437 0.267  31.1  
## 2     1     0.465 0.186     0.304  0.431 0.277  30.9
```

Comments

There are some weird things going on.

```
summary(as.factor(jtpa$age))
```

```
##      0 23.5 27.5 32.5  40 49.5  
## 441 2638 2288 2714 2200  923
```

```
summary(as.factor(jtpa$married))
```

```
##           0 0.2111608 0.3352851           1  
##      7495           517           247      2945
```

```
summary(as.factor(jtpa$wkless13))
```

```
##           0 0.4073359 0.5332298           1  
##      5395           454           724      4631
```

Comments 2

“Treatment” here is an offer for access to the JTPA.

- ▶ Uptake (train) is not identical to offer of treatment.

```
jtpa %>% group_by(offer) %>%  
  summarize(train = mean(train))
```

```
## # A tibble: 2 x 2  
##   offer  train  
##   <dbl> <dbl>  
## 1     0 0.0145  
## 2     1 0.642
```

Comments 3

- ▶ *What treatment IS* is not uniform across genders

```
jtpa %>% group_by(male) %>%  
  select(male,OJT_JSA,classroom,train) %>%  
  summarize(across(everything(),mean))
```

```
## # A tibble: 2 x 4  
##   male OJT_JSA classroom train  
##   <dbl> <dbl> <dbl> <dbl>  
## 1     0  0.374  0.384 0.446  
## 2     1  0.504  0.203 0.419
```

- ▶ OJT: on-job-training, JSA: job-search assistance, classroom - skills training.

Observational Data – Mean Difference *on uptake*

- ▶ We could look at difference between those who use and don't use program.
 - ▶ Remember – there may be selection driving this.

```
smod = summary(lm(y~train,data=jtpa))  
signif(smod$coefficients,3)
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	14600	210	69.60	0.00e+00
## train	2790	319	8.76	2.21e-18

RCTs – Mean Difference

- ▶ Take the difference in means between group assigned treatment and group assigned control.
 - ▶ Average Treatment Effect: effect of *offer* (AKA intention-to-treat)
 - ▶ We can test for difference from 0, build CIs, etc - for this difference in means.

```
smod = summary(lm(y~offer,data=jtpa))  
signif(smod$coefficients,3)
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	15000	275	54.70	0.000000
## offer	1160	336	3.45	0.000567

RCT - complications

Basic Mean difference is unbiased on average across universes where you ran this experiment.

- ▶ But what if there was some residual variation in other variables.
 - ▶ Like, people who were AFDC recipients were more likely to receive treatment, *by chance*.
- ▶ Alternately, what if we are interested in the effect on some subpopulation? E.g. Gender differences?
- ▶ Finally, uptake in treatment group was like 60%. So the treatment effect must be larger for people *who used the program*. Can we figure that out?

RCT – Subgroup Analysis

```
smod = summary(lm(y~offer*male,data=jtpa))  
signif(smod$coefficients,3)
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	12200	367	33.20	8.84e-231
## offer	1240	449	2.77	5.62e-03
## male	6210	543	11.40	3.87e-30
## offer:male	-126	664	-0.19	8.50e-01

RCT – Basic Controls

```
smod = summary(lm(y~.,data=jtpa[,-c(3,12,11,13)]))  
signif(smod$coefficients,3)
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	12900.0	654.0	19.800	1.47e-85
## offer	1100.0	320.0	3.430	6.04e-04
## male	4510.0	324.0	13.900	1.13e-43
## hs	3730.0	345.0	10.800	4.03e-27
## black	-1350.0	362.0	-3.720	2.01e-04
## hispanic	-464.0	497.0	-0.933	3.51e-01
## married	3300.0	358.0	9.220	3.43e-20
## wkless13	-6650.0	331.0	-20.100	1.96e-88
## AFDC	-1810.0	424.0	-4.260	2.08e-05
## age	11.7	15.2	0.768	4.43e-01

► Interactions? What is the model we should choose?

RCTs - ATT

- ▶ We could try to look at the average treatment effect on the treated.

```
uptake = mean(jtpa$train[jtpa$offer == 1]) # ~64%  
smod = summary(lm(y~offer,data=jtpa))$coefficients  
smod[2,1]/uptake #Scale up Coef by uptake rate.
```

```
## [1] 1806.969
```

- ▶ Uncertainty there is not just uncertainty internal to linear model – but also around uptake. Bootstrap for CIs if you like.

RCTs - Big Data?

How is this related to this course?

1. Entirely about predictions.
 - ▶ Obscured by simplicity of analysis
2. That was *Average* treatment effects. What about individual TEs?
 - ▶ But effects are heterogenous (we saw that in gender)
 - ▶ We can estimate the TE for any given individual. (Some may be negative)
 - ▶ This gives us Conditional Average Treatment Effects (CATEs)
3. We may want to build *targeting policies* based on RCTs.
 - ▶ E.g. I run an RCT of some ad campaign.
 - 3.1 Run RCT
 - 3.2 Identify CATEs
 - 3.3 Find *optimal* policy for targeting ads.
 - 3.4 Profit

TEs are predictions?

$$\widehat{ATE} = \left[\frac{1}{n_1} \sum_{i: T=1} y_i \right] - \left[\frac{1}{n_0} \sum_{i: T=0} y_i \right] = \bar{y}_1 - \bar{y}_0$$

But suppose we define each individual's treatment effect as the sum of their observed outcome, and the *unobserved* counterfactual outcome where their treatment status was flipped:

$$\widehat{TE}_i = y_i(1) - y_i(0)$$

To do this, we would need to make a prediction about y_i in the unobserved counterfactual.

Prediction in Counterfactual

Suppose for the counterfactual, we predict the mean outcome from the other group.

$$\hat{y}_i(T_i - 1) = \bar{y}_{T_i - 1}$$

Then (for a treated observation i), we have:

$$\widehat{TE}_i = y_i - \bar{y}_0$$

And for an untreated observation, similarly:

$$\widehat{TE}_i = \bar{y}_1 - y_i$$

Similarity

We could then, form a new estimate of the ATE:

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

With a lot of rewriting of sums – we can show that:

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

Proof in Data

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

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

```
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
```

TEs as predictions: A brief diversion

How does this affect our interpretation of 'controls' and of 'subgroup analysis'?

- ▶ We can do the same basic exercise, where 'controls' or subgroup analyses affect *our counterfactual predictions* and rewrite our ATE estimate as a mean of individual treatment effects.

TEs as predictions

First off, purely for calculating ATEs, this suggests a simple heuristic:

- ▶ If we can improve our OOS predictions for either treatment or control, we can improve our ATE estimate
 - ▶ We've seen a lot of ways one could improve a prediction in this course.

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$conifact = predict(mod_treat,data = jtpa_cont)$pre
jtpa_treat$conifact = predict(mod_cont,data = jtpa_treat)$pre
#Estimate TEs
jtpa_cont$TE = jtpa_cont$conifact - jtpa_cont$y
jtpa_treat$TE = jtpa_treat$y - jtpa_treat$conifact
#Recombine
jtpa_est = rbind(jtpa_cont,jtpa_treat)
mean(jtpa_est$TE) #ATE
```

```
## [1] 1165.806
```

ATE from Predictions

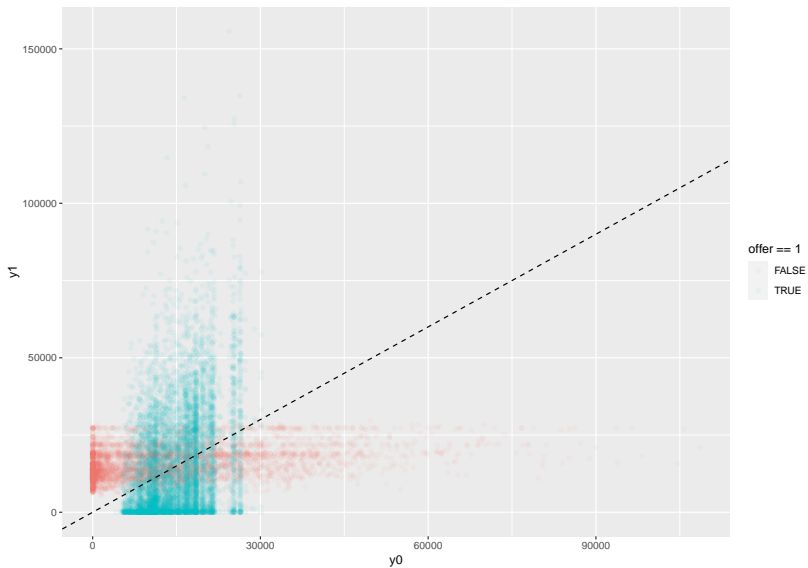
We could use *any* of the many routines we've seen, and the many more you will encounter for our *counterfactual predictions*.

- ▶ KNN
- ▶ Forests
- ▶ Logit
- ▶ LASSO
- ▶ Boosted Models
- ▶ Bagged Models
- ▶ Other Ensemble Models

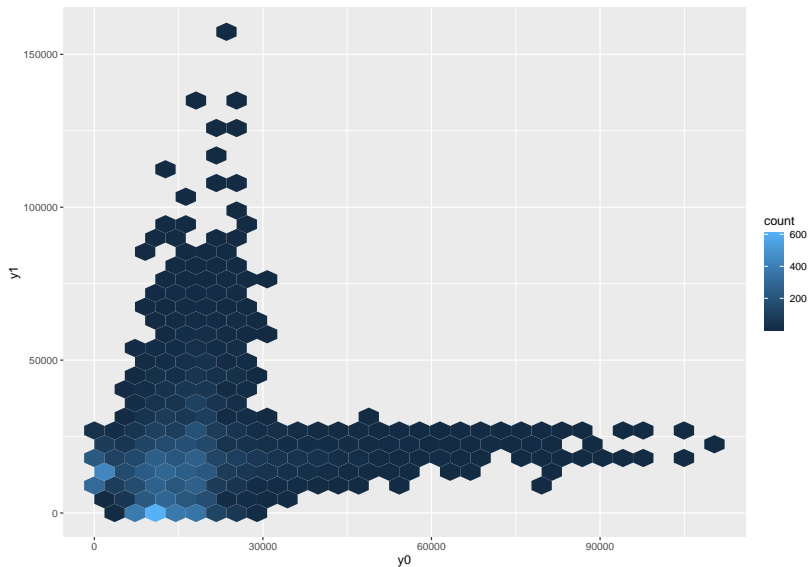
Once this is a prediction problem – we can do a lot.

TEs

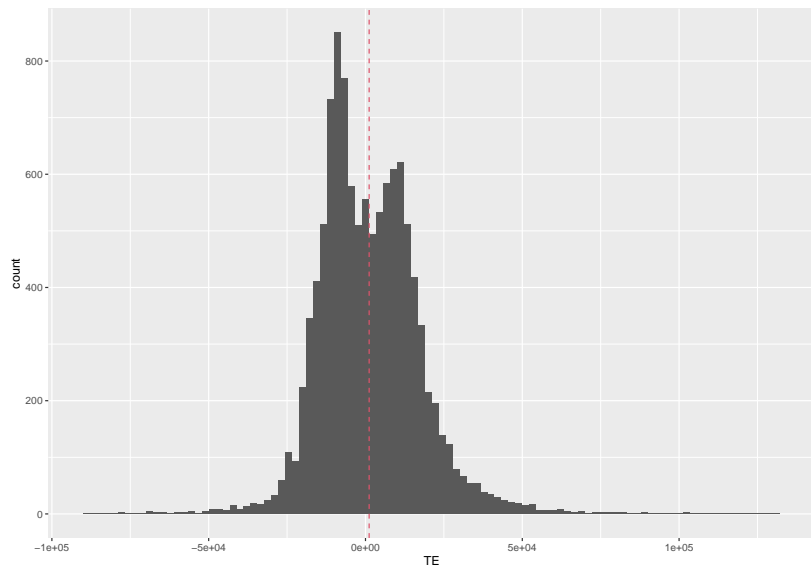
But we could also use those TE estimates from individuals.



Individual TEs



Individual TEs



There is a lot of variation around our mean of 1165.8059295 in benefits. The standard deviation is 1.588022×10^4 .

Is this variation in individual TEs sensible?

- ▶ Probably? \$50k effects either way seem large, but only a small portion, 0.0107997, experience that scale.
- ▶ Intuitively, some individuals got training that was beneficial, and some spent time doing training that they could have spent *working* at other jobs that pay. So some loss seems about right.

Individual TEs, Do we care?

It looks like some fraction of individuals lost \sim \$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.

⇒ Targeting. Can we use the RCT data for targeting?

Targeting

We estimated treatment effects for individuals in an experiment.

- ▶ This means we had to predict the unobserved counterfactual
- ▶ But it also means we witnessed the outcome for one treatment possibility.

To do targeting we need to:

- 1) Make predictions for individuals under treatment
- 2) Make predictions for individuals under control
- 3) Estimate TEs for each individual
- 4) Compare TEs to some threshold (\$0? \$1k?) to determine eligibility.

Targeting

1. Fully a counterfactual exercise.
2. We need *good* OOS predictions
 - ▶ We may want to do some CV to determine performance
3. We need something to compare to. “Opportunity Costs”

Talk more Thursday.

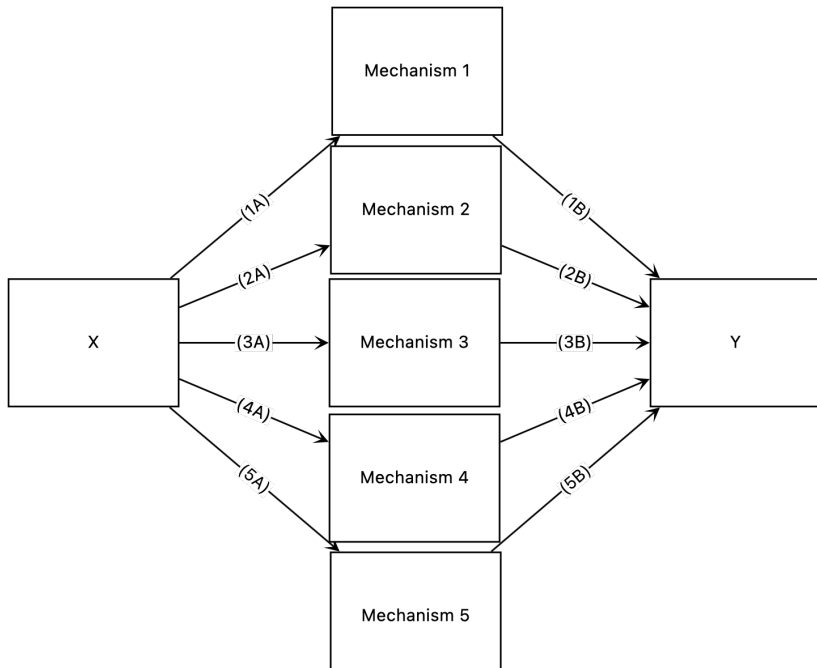
An Aside: RCTs and Mechanisms

RCTs are not good at identifying the mechanisms for effects.

This is because identifying a specific mechanism means establishing a link from X to that mechanism, from that mechanism to Y, **and** the lack of links between X and Y through other mechanisms.

You need multiple overlapping RCTs to do this kind of thing – each looking at different things.

An Aside: RCTs and Mechanisms



HW 5 Review (if time)

Wrap up

Things to do

Homework 6 is due tomorrow. New prediction competition was released yesterday – purely optional.

On Thursday we will:

1. Wrap up targeting/RCTs
2. *Briefly* encounter a few ‘observational’ causal methods:
 - a. Instrumental Variables
 - a. Useful for “Intention to Treat” vs “Average effect on Treated”
 - b. Regression Discontinuity
 - c. Diff-in-Diff
 - a. SCM

Bye!