3. Introduction to causality

Frank Edwards

9/20/2021

Introduction to Statistics

Why learn statistics?

1. Quantitative methods allow us to discover or infer patterns when we have large amounts of data

Why learn statistics?

1. Quantitative methods allow us to discover or infer patterns when we have large amounts of data
2. Statistics provide methods for testing for differences between groups of data

Why learn statistics?

1. Quantitative methods allow us to discover or infer patterns when we have large amounts of data
2. Statistics provide methods for testing for differences between groups of data
3. Always remember two things: 1) all models are wrong, but some are useful; 2) social data come from people, and are always imperfect

Causality

The key question in causal inference

- Does treatment x affect outcome y

The key question in causal inference

- Does treatment x affect outcome y
- In medicine: does a treatment affect a patient

The key question in causal inference

- Does treatment x affect outcome y
- In medicine: does a treatment affect a patient
- Typically designed by randomly assigning patients to treatment and control groups, where treatment groups are exposed to x, and control groups are not

The fundamental problem of causal inference

How much did the treatment matter?

The fundamental problem of causal inference

How much did the treatment matter?

We answer this question with counterfactuals:

What would have happened if treated units were untreated? What would have happened if untreated units were treated?

The fundamental problem of causal inference

How much did the treatment matter?

We answer this question with counterfactuals:

What would have happened if treated units were untreated? What would have happened if untreated units were treated?

For an observation i, where $Y_i(1)$ indicates treatment and $Y_i(0)$ indicates no treatment, the causal effect of the treatment is defined as

$Y_i(1) - Y_i(0)$

The fundamental problem of causal inference

How much did the treatment matter?

We answer this question with counterfactuals:

What would have happened if treated units were untreated? What would have happened if untreated units were treated?

For an observation i, where $Y_i(1)$ indicates treatment and $Y_i(0)$ indicates no treatment, the causal effect of the treatment is defined as

$Y_i(1) - Y_i(0)$

Why is this a problematic definition?

Causal questions in social science

- Does race impact hiring decisions?
    - A Black candidate applied for a job, but did not get it.
    - Would a Black candidate have been offered a job if they were white?

Causal questions in social science

- Does race impact hiring decisions?
  - A Black candidate applied for a job, but did not get it.
  - Would a Black candidate have been offered a job if they were white?
- Does the minimum wage increase unemployment?
  - Unemployment went up in a city after the minimum wage increased
  - Would unemployment have gone up were there not an increase in the minimum wage?

Causal questions in social science

- Does race impact hiring decisions?
  - A Black candidate applied for a job, but did not get it.
  - Would a Black candidate have been offered a job if they were white?
- Does the minimum wage increase unemployment?
  - Unemployment went up in a city after the minimum wage increased
  - Would unemployment have gone up were there not an increase in the minimum wage?
- Does community policing decrease crime?
  - A police department implemented community policing in certain neighborhoods, and reported crime went down
  - Would reported crime have gone down without community policing?

Experimental research

Evaluates how treatments causally effect outcomes by assigning different levels of treatment to different observations, then measuring the corresponding values of the outcome

Using an experiment to estimate the effects of a criminal record on employment

Pager, Devah. "The mark of a criminal record." American journal of sociology 108.5 (2003): 937-975.

> *With over 2 million individuals currently incarcerated, and over half a million prisoners released each year, the large and growing number of men being processed through the criminal justice system raises important questions about the consequences of this massive institutional intervention. This article focuses on the consequences of incarceration for the employment outcomes of black and white job seekers. The present study adopts an experimental audit approach—in which matched pairs of individuals applied for real entry - level jobs - to formally test the degree to which a criminal record affects subsequent employment opportunities. The findings of this study reveal an important, and much underrecognized, mechanism of stratification. A criminal record presents a major barrier to employment, with important implications for racial disparities.*

Research questions

1. Do employers use criminal histories to make hiring decisions?
2. Is racial discrimination a major barrier to employment?
3. Does the effect of a criminal record differ for white and Black applicants?

What counterfactuals are needed for each question?

1. Do employers use criminal histories to make hiring decisions?

What counterfactuals are needed for each question?

1. Do employers use criminal histories to make hiring decisions?
2. Is racial discrimination a major barrier to employment?

What counterfactuals are needed for each question?

1. Do employers use criminal histories to make hiring decisions?
2. Is racial discrimination a major barrier to employment?
3. Does the effect of a criminal record differ for white and Black applicants?

Before we begin

- grab the data from Slack

```
dat<-read.csv("./data/criminalrecord.csv")
```

Variables in the data

`jobid` Job ID number

`callback` 1 if tester received a callback, 0 if the tester did not receive a callback.

`black` 1 if the tester is black, 0 if the tester is white.

`crimrec` 1 if the tester has a criminal record, 0 if the tester does not.

`interact` 1 if tester interacted with employer during the job application, 0 if tester does not interact with employer.

`city` 1 is job is located in the city center, 0 if job is located in the suburbs.

`distance` Job's average distance to downtown.

`custserv` 1 if job is in the costumer service sector, 0 if it is not.

`manualskill` 1 if job requires manual skills, 0 if it does not.

Take a look at the data

```r
head(dat)
```

```
##   jobid callback black crimrec interact city distance custserv manualskill
## 1   108        1     0       1        1    0       15        1           0
## 2   113        0     0       0        1    0       20        0           1
## 3   101        1     0       0        0    0       15        1           0
## 4    64        1     0       0        0    1        7        1           0
## 5    33        0     0       1        0    1        5        1           0
## 6    73        0     0       1        0    1       10        0           1
```

Exploring the data: univariate crosstabs

```
dat %>% count(black)
```

```
##   black   n
## 1     0 300
## 2     1 396
```

```
dat %>% count(crimrec)
```

```
##   crimrec   n
## 1       0 349
## 2       1 347
```

Exploring the data: bivariate crosstabs

```
dat %>% count(black, crimrec)
```

```
##   black crimrec   n
## 1     0       0 150
## 2     0       1 150
## 3     1       0 199
## 4     1       1 197
```

```
dat %>% count(black, callback)
```

```
##   black callback   n
## 1     0        0 224
## 2     0        1  76
## 3     1        0 358
## 4     1        1  38
```

Using crosstabs

What was the callback rate for subjects assigned a criminal record?

```
dat %>% count(crimrec, callback)
```

```
##   crimrec callback   n
## 1       0        0 270
## 2       0        1  79
## 3       1        0 312
## 4       1        1  35
```

Using crosstabs

What was the callback rate for subjects assigned a criminal record?

```
## Divide those with a criminal record and callback
## By all those with a criminal record
dat %>%
  group_by(crimrec) %>%
  summarise(callback = sum(callback),
            n = n()) %>%
  mutate(rate = callback / n)
```

```
## # A tibble: 2 x 4
##   crimrec callback     n  rate
##     <int>    <int> <int> <dbl>
## 1       0       79   349 0.226
## 2       1       35   347 0.101
```

Recoding and conditionals

Let’s make distance categorical, with cuts at the 25th, 50th, and 75th quantile

```r
summary(dat$distance)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.00    8.00   12.00   11.96   16.00   25.00       2
```

```r
## NA???
```

Subsetting to remove missing values

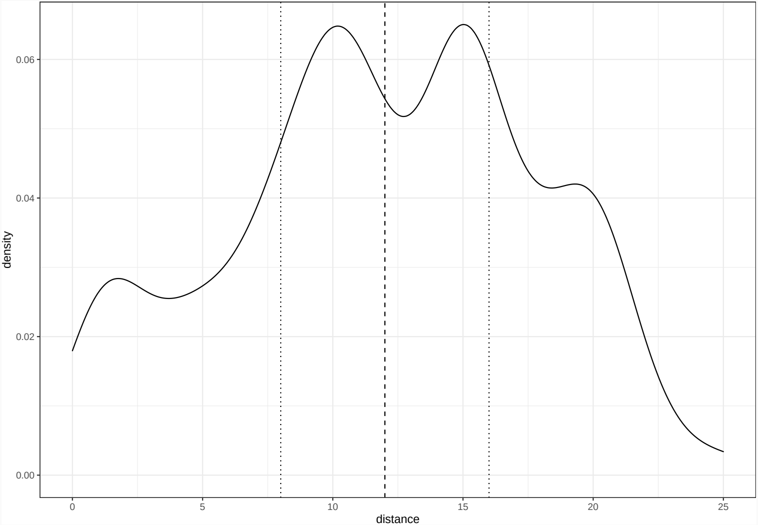
```r
## remove pesky NA values
dat_clean<-dat %>%
  filter(!(is.na(distance)))

### wait, what did you do there???!
### also works, but more agressive: dat_clean<-na.omit(dat)

summary(dat_clean$distance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   12.00   11.96   16.00   25.00
```

Visualizing quantiles: remember area under the curve?



Making a recode with one condition

Make a new variable for distance, with value T if below the median, and F if above

```
dat_clean<-dat_clean %>%
  mutate(distance_binary = distance < median(distance))
```

Making a recode with one condition: ifelse()

Make a new variable for distance, with value “near” if below the median, and “far” if above

```
dat_clean<-dat_clean %>%
  mutate(distance_binary2 = ifelse(
    distance < median(distance),
    "near",
    "far"
  ))
```

Making a recode with multiple conditions

```r
### define quartile cut points
q1<-quantile(dat_clean$distance, 0.25)
q2<-quantile(dat_clean$distance, 0.5)
q3<-quantile(dat_clean$distance, 0.75)
q1; q2; q3
```

```
## 25%
##   8
```

```
## 50%
##  12
```

```
## 75%
##  16
```

Making a recode with multiple conditions: case_when()

```
### make factor variable
dat_clean <- dat_clean %>%
  mutate(distance_quartile =
           case_when(
             distance < q1 ~ "1st",
             distance < q2 ~ "2nd",
             distance < q3 ~ "3rd",
             distance >= q3 ~ "4th"
           ))
```

Returning to Pager's experiment

The counterfactual and potential outcomes

```
c_fact<-data.frame(callback = dat$callback,
                   crimrec = dat$crimrec)

### create explicit counterfactual

c_fact <- c_fact %>%
  mutate(callback_crimT =
           ifelse(
             crimrec==1,
             callback,
             NA),
         callback_crimF =
           ifelse(crimrec==0,
                  callback,
                  NA))


head(c_fact)
```

```
##   callback crimrec callback_crimT callback_crimF
## 1        1       1              1             NA
## 2        0       0             NA              0
## 3        1       0             NA              1
## 4        1       0             NA              1
## 5        0       1              0             NA
## 6        0       1              0             NA
```

The causal effect

For observation *i*, the sample average treatment effect (SATE) is equal to:

callback_crimTRUE_i - callback_crimFALSE_i

What is the causal effect for rows 1 - 6

For observation *i*, the treatment effect is equal to:

callback_crimTRUE_i - callback_crimFALSE_i

```
head(c_fact)
```

```
##   callback crimrec callback_crimT callback_crimF
## 1        1       1              1             NA
## 2        0       0             NA              0
## 3        1       0             NA              1
## 4        1       0             NA              1
## 5        0       1              0             NA
## 6        0       1              0             NA
```

What is the causal effect for rows 1 - 6

For observation *i*, the treatment effect is equal to:

callback_crimTRUE_i - callback_crimFALSE_i

```
head(c_fact)
```

```
##   callback crimrec callback_crimT callback_crimF
## 1        1       1              1             NA
## 2        0       0             NA              0
## 3        1       0             NA              1
## 4        1       0             NA              1
## 5        0       1              0             NA
## 6        0       1              0             NA
```

The fundamental problem of causal inference is that we only observe one of these outcomes

Randomized experiments (or RCTs)

- By randomizing assignment to treatment, we can treat units as equivalent

Randomized experiments (or RCTs)

- By randomizing assignment to treatment, we can treat units as equivalent
- If units are equivalent, we can estimate the average treatment effect as a difference in means on the outcome between the treatment and control group

Randomized experiments (or RCTs)

- By randomizing assignment to treatment, we can treat units as equivalent
- If units are equivalent, we can estimate the average treatment effect as a difference in means on the outcome between the treatment and control group
- If we don't randomize, we have no assurance that the treated and control groups are equivalent, meaning we can't argue that we've observed the counterfactual

The SATE for Pager's experiment

We assume that we can estimate the counterfactual for people with criminal records (i.e. no criminal record), by using the mean value of the callback outcome for people assigned to have no criminal record.

```
### obtain the mean callback rate of those with a criminal record
### and those without
effect<-dat %>%
  group_by(crimrec) %>%
  summarise(callback = mean(callback))
### Compute the SATE
effect[2, 2] - effect[1, 2]
```

```
##     callback
## 1 -0.1254965
```

Next week

- Homework: More work with Pager's data