Causality, 2

Frank Edwards

9/17/2019

Returning to Pager's experiment

The causal effect

For observation *i* is equal to callback_crimTRUE_i - callback_crimFALSE_i

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

The counterfactual and potential outcomes

```
## # A tibble: 696 x 3
##    crimrec callback_crimTRUE callback_crimFALSE
##      <dbl>             <dbl>              <dbl>
##  1       1                 1                 NA
##  2       0                NA                  0
##  3       0                NA                  1
##  4       0                NA                  1
##  5       1                 0                 NA
##  6       1                 0                 NA
##  7       1                 0                 NA
##  8       1                 1                 NA
##  9       0                NA                  0
## 10       1                 0                 NA
## # ... with 686 more rows
```

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 don't have a strong case that we've observed the counterfactual

Obtaining a sample average treatment effect

The sample average treatment effect is defined as:

$$\text{SATE} = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

In practice, since we only observe $Y_i(1)$ OR $Y_i(0)$, we instead estimate a *difference-in-means* of the outcome between the treatment and control: $mean(Y(1)) - mean(Y(0))$. If assignment has been randomized, these values are identical.

Why we randomize

```r
data(social)
glimpse(social)
```

```
## Observations: 305,866
## Variables: 6
## $ sex         <chr> "male", "female", "male", "female", "fema
## $ yearofbirth <int> 1941, 1947, 1951, 1950, 1982, 1981, 1959,
## $ primary2004 <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1,
## $ messages    <chr> "Civic Duty", "Civic Duty", "Hawthorne",
## $ primary2006 <int> 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0,
## $ hhsize      <int> 2, 2, 3, 3, 3, 3, 3, 3, 2, 2, 1, 2, 2, 1,
```

Obtaining mean voting by treatment/control

```
control<-social %>%
  filter(messages == "Control") %>%
  summarise(primary2006 = mean(primary2006))

treatment<-social %>%
  filter(messages!="Control") %>%
  group_by(messages) %>%
  summarise(primary2006 = mean(primary2006))

treatment
```

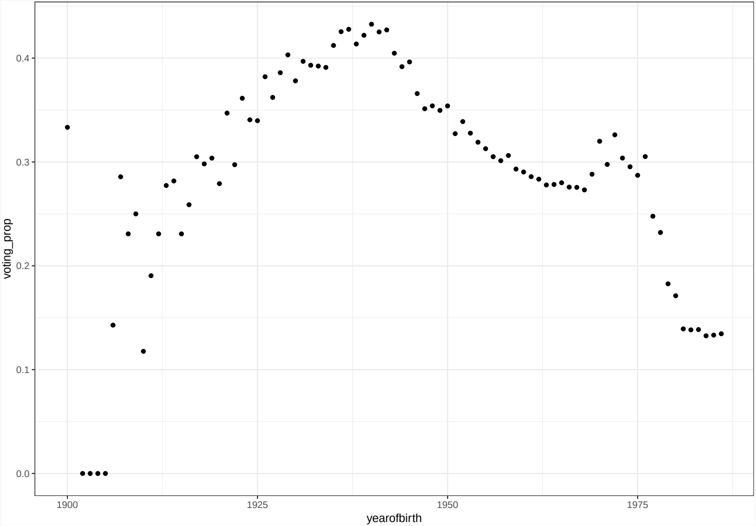
```
## # A tibble: 3 x 2
##   messages   primary2006
##   <chr>            <dbl>
## 1 Civic Duty       0.315
## 2 Hawthorne        0.322
## 3 Neighbors        0.378
```

The difference in means (causal effect)

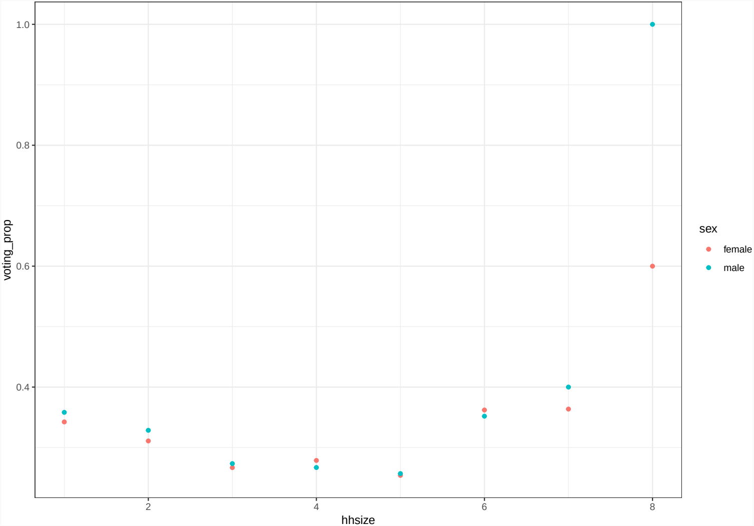
```
treatment %>%
  mutate(effect = primary2006 - control$primary2006)
```

```
## # A tibble: 3 x 3
##   messages   primary2006 effect
##   <chr>            <dbl>  <dbl>
## 1 Civic Duty       0.315 0.0179
## 2 Hawthorne        0.322 0.0257
## 3 Neighbors        0.378 0.0813
```

Why randomization matters



Why randomization matters (continued)



Randomization matters

- Because certain kinds of people are more likely to vote in primaries than others
- We note these differences between observed variables and our outcome: primary2006
- We didn't measure very much here. They could also differ across unobserved or unobservable variables!
- Randomization (given a large enough n) ensures that treatment and control groups are *identical* across all observed and unobserved/unobservable differences prior to treatment
- This condition – statistically identical treatment and control groups – is a necessary condition for causal inference. Randomization is the most straightforward way to achieve this condition.

Causal inference in observational data

Estimating the impact of a minimum wage increase

In 1992, New Jersey raised it's minimum wage from $4.25 to $5.05.

Pennsylvania did not.

```
data(minwage)
glimpse(minwage)
```

```
## Observations: 358
## Variables: 8
## $ chain      <chr> "wendys", "wendys", "burgerking", "burgerk
## $ location   <chr> "PA", "PA", "PA", "PA", "PA", "PA", "PA",
## $ wageBefore <dbl> 5.00, 5.50, 5.00, 5.00, 5.25, 5.00, 5.00,
## $ wageAfter  <dbl> 5.25, 4.75, 4.75, 5.00, 5.00, 5.00, 4.75,
## $ fullBefore <dbl> 20.0, 6.0, 50.0, 10.0, 2.0, 2.0, 2.5, 40.0
## $ fullAfter  <dbl> 0.0, 28.0, 15.0, 26.0, 3.0, 2.0, 1.0, 9.0,
## $ partBefore <dbl> 20.0, 26.0, 35.0, 17.0, 8.0, 10.0, 20.0, 3
## $ partAfter  <dbl> 36, 3, 18, 9, 12, 9, 25, 32, 39, 10, 20, 4
```

Describing the data, categoricals

```
table(minwage$chain)
```

```
##
## burgerking        kfc       roys     wendys
##        149         75         88         46
```

```
table(minwage$location)
```

```
##
## centralNJ   northNJ        PA   shoreNJ   southNJ
##        45       146        67        33        67
```

Did NJ minimum wage increase the wages paid to employees?

```r
minwage %>%
  group_by(location) %>%
  summarise(wageBefore_mn = mean(wageBefore),
            wageAfter_mn = mean(wageAfter))
```

```
## # A tibble: 5 x 3
##   location  wageBefore_mn wageAfter_mn
##   <chr>             <dbl>        <dbl>
## 1 centralNJ          4.63         5.09
## 2 northNJ            4.63         5.09
## 3 PA                 4.65         4.61
## 4 shoreNJ            4.64         5.07
## 5 southNJ            4.54         5.06
```

Another way to look at change in wages

```
minwage %>%
  group_by(location) %>%
  summarise(prop_below_before = mean(wageBefore>=5.05),
            prop_below_after = mean(wageAfter>=5.05))
```

```
## # A tibble: 5 x 3
##   location  prop_below_before prop_below_after
##   <chr>                 <dbl>            <dbl>
## 1 centralNJ            0.133            0.978
## 2 northNJ              0.0753           1
## 3 PA                   0.0597           0.0448
## 4 shoreNJ              0.121            1
## 5 southNJ              0.0746           1
```

Look at our outcome variable

```r
minwage<-minwage %>%
  mutate(prop_ft_pre = fullBefore / (fullBefore + partBefore),
         prop_ft_post = fullAfter / (fullAfter + partAfter))

minwage %>%
  group_by(location) %>%
  summarise(prop_ft_pre = mean(prop_ft_pre),
            prop_ft_post = mean(prop_ft_post))
```

```
## # A tibble: 5 x 3
##   location  prop_ft_pre prop_ft_post
##   <chr>           <dbl>        <dbl>
## 1 centralNJ       0.311        0.251
## 2 northNJ         0.321        0.375
## 3 PA              0.310        0.272
## 4 shoreNJ         0.286        0.345
## 5 southNJ         0.239        0.236
```

Assumption: PA is a no-treatment counterfactual

Estimate the causal effect

```r
control<-minwage %>%
  filter(location=="PA") %>%
  summarise(prop_ft_post = mean(prop_ft_post))

minwage %>%
  filter(location!="PA") %>%
  summarise(prop_ft_post = mean(prop_ft_post)) %>%
  mutate(effect = prop_ft_post - control$prop_ft_post)
```

```
##   prop_ft_post     effect
## 1     0.320401 0.04811886
```

Is this a valid estimate of the causal effect?

Confounding jeopardizes causal inference

- Confounding bias: a third variable is associated with both the treatment and the outcome

Confounding jeopardizes causal inference

- Confounding bias: a third variable is associated with both the treatment and the outcome
- Selection bias: a unit may choose to participate in a treatment for reasons that are correlated with the outcome

Confounding jeopardizes causal inference

- Confounding bias: a third variable is associated with both the treatment and the outcome
- Selection bias: a unit may choose to participate in a treatment for reasons that are correlated with the outcome

Correlation != Causation

- Randomize treatment!

Dealing with confounding

- Randomize treatment!
- When we can't...

Dealing with confounding

- Randomize treatment!
- When we can't...
- Statistical control: within-subgroup analysis based on confounder values

- Randomize treatment!
- When we can't...
- Statistical control: within-subgroup analysis based on confounder values

Are NJ and PA the same (at least when it comes to fast food jobs?)?

```
minwage %>%
  group_by(location=="PA") %>%
  summarise(prop_wendys = mean(chain=="wendys"),
            prop_bk = mean(chain=="burgerking"),
            prop_kfc = mean(chain=="kfc"),
            prop_roys = mean(chain=="roys"))
```

```
## # A tibble: 2 x 5
##   `location == "PA"` prop_wendys prop_bk prop_kfc prop_roys
##   <lgl>                    <dbl>   <dbl>    <dbl>     <dbl>
## 1 FALSE                    0.120   0.405    0.223     0.251
## 2 TRUE                     0.164   0.463    0.149     0.224
```

Maybe restaurant chain matters? Let's control for it!

```r
control<-minwage %>%
  filter(location=="PA") %>% group_by(chain) %>%
  summarise(prop_ft_post = mean(prop_ft_post))

minwage %>%
  filter(location!="PA") %>% group_by(chain) %>%
  summarise(prop_ft_post = mean(prop_ft_post)) %>%
  mutate(effect = prop_ft_post - control$prop_ft_post)
```

```
## # A tibble: 4 x 3
##   chain      prop_ft_post effect
##   <chr>             <dbl>  <dbl>
## 1 burgerking        0.358 0.0364
## 2 kfc               0.328 0.0918
## 3 roys              0.283 0.0697
## 4 wendys            0.260 0.0117
```

Maybe region matters: central and south vs north and shore

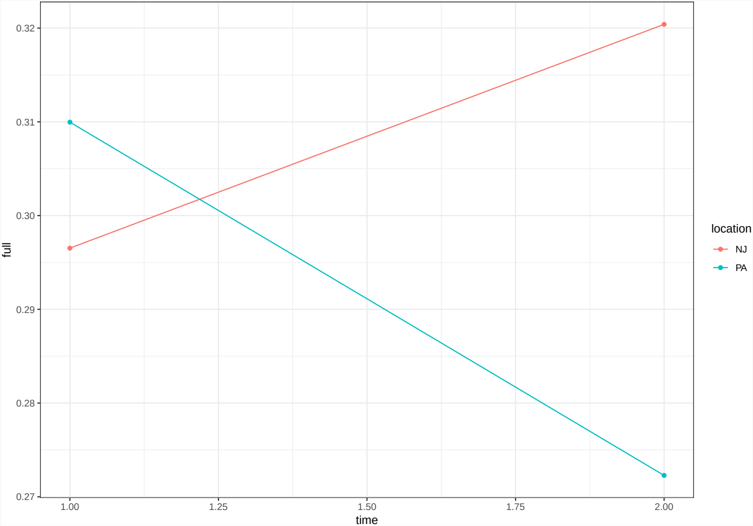
```
control<-minwage %>%
  filter(location=="PA") %>%
  summarise(prop_ft_post = mean(prop_ft_post))

minwage %>%
  filter(location!="PA") %>% group_by(location) %>%
  summarise(prop_ft_post = mean(prop_ft_post)) %>%
  mutate(effect = prop_ft_post - control$prop_ft_post)
```

```
## # A tibble: 4 x 3
##   location  prop_ft_post  effect
##   <chr>            <dbl>   <dbl>
## 1 centralNJ        0.251 -0.0210
## 2 northNJ          0.375  0.103
## 3 shoreNJ          0.345  0.0728
## 4 southNJ          0.236 -0.0366
```

Cross-sections and time series

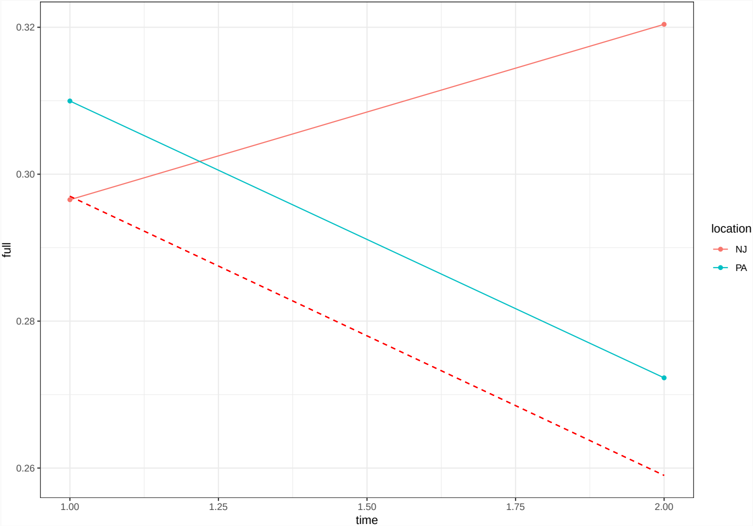
- Longitudinal data: repeated measurements of the same unit on the same variables over time
- Cross-sectional data: one measurement of many units
- Panel data (or time series cross-sectional data): repeated measurements of many units on the same variables over time
- Key advantages to panel data: variables may differ across units and within-units over time (trends).

Before and after design



- What if we treated PA as the counterfactual, and used information about it's trend in employment to estimate the effect of NJ's minimum wage increase?
- Assumption: The trend in the outcome over time would have been identical across all units if the treatment had never been imposed (parallel trends)

Difference in Differences (visual)



Estimating the causal effect

Where y_{ij} is the outcome for treatment group $i = 1$ and post-treatment time $j = 1$

$$\mathrm{DiD} = (\bar{y}_{1,1} - \bar{y}_{1,0}) - (\bar{y}_{2,1} - \bar{y}_{2,0})$$

Assuming that the counterfactual outcome for the treatment group has a parallel time trend to that observed for the control group.

Compute the DiD estimator

```
DiD<-minwage %>%
  mutate(location = ifelse(location=="PA", "PA", "NJ")) %>%
  group_by(location) %>%
  summarise(prop_ft_pre = mean(prop_ft_pre),
            prop_ft_post = mean(prop_ft_post))

control <- DiD %>% filter(location=="PA")
treatment <- DiD %>% filter(location!="PA")
(treatment$prop_ft_post - treatment$prop_ft_pre) -
  (control$prop_ft_post - control$prop_ft_pre)
```

```
## [1] 0.06155831
```

Descriptive Statistics

Summarizing a variable

Reduce a vector to a single or smaller set of values that tell us something useful

Examples we've already used: - minimum: min() - maximum: max() - median: median() - mean: mean()

Quantiles

- The median is the 0.5 quantile (50th percentile)
- Quantiles are less sensitive to outliers than are other measures (like the mean)
- Quantiles tell you the proportion of a data that falls below some cutpoint

Quantiles: example

```
quantile(minwage$wageBefore, 0.25)
```

```
##  25%
## 4.25
```

```
quantile(minwage$wageBefore, 0.75)
```

```
##    75%
## 4.9875
```

```
quantile(minwage$wageBefore, c(0.05, 0.25, 0.5, .75, 0.95))
```

```
##     5%    25%    50%    75%    95%
## 4.2500 4.2500 4.5000 4.9875 5.2500
```

Standard deviation

- The standard deviation (SD, σ) is a measure of the spread of a variable
- It provides a measure of how much each observation of a variable differs from the mean of the variable
- You can use the sd() function in R
- The variance (var() function) is the square of the standard deviation

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}^2\right)}$$

$$variance = \sigma^2$$

Calculate an SD for these variables

```r
minwage$wageBefore[1:10]
```

```
##  [1] 5.00 5.50 5.00 5.00 5.25 5.00 5.00 5.00 5.00 5.50
```

```r
minwage$fullBefore[1:10]
```

```
##  [1] 20.0  6.0 50.0 10.0  2.0  2.0  2.5 40.0  8.0 10.5
```

Homework

- Complete exercise 2.8.1
- load the data with data(STAR)
- make sure to use na.rm = TRUE for mean(), quantile() and other functions
- Recode variables to character rather than factor types using case_when() or ifelse()
- You will use group_by() and summarise() alot on this assignment
- Don’t use View(), use head() - this is a 6000 row dataset.