EDS241: Assignment 1 Template

Luna Herschenfeld-Catalán

01/23/2024

1 Part 1

(NOTE: Uses the RCT.R code provided with lecture to generate data) DO NOT CHANGE ANYTHING BELOW UNTIL IT SAYS EXPLICITLY

1.1 BELOW YOU CAN (AND HAVE TO) CHANGE AND ADD CODE TO DO ASSIGNMENT

Part 1: Use the small program above that generates synthetic potential outcomes without treatment, Yi_0, and with treatment, Yi_1. When reporting findings, report them using statistical terminology (i.e. more than y/n.) Please do the following and answer the respective questions (briefly).

a) Create equally sized treatment and control groups by creating a binary random variable Di where the units with the *1's" are chosen randomly.

```
set.seed(123)

# divide the total sample by 2
n_half <- N/2

# make a list of equal number of 0s and 1s
Di_list <- c(rep(0, n_half),
        rep(1, n_half))

# sample from the Di_list equally
Di_sample <- sample(Di_list)

# add Di to table
df_tibble <- as_tibble(df) %>%
  mutate(Di = Di_sample)
```

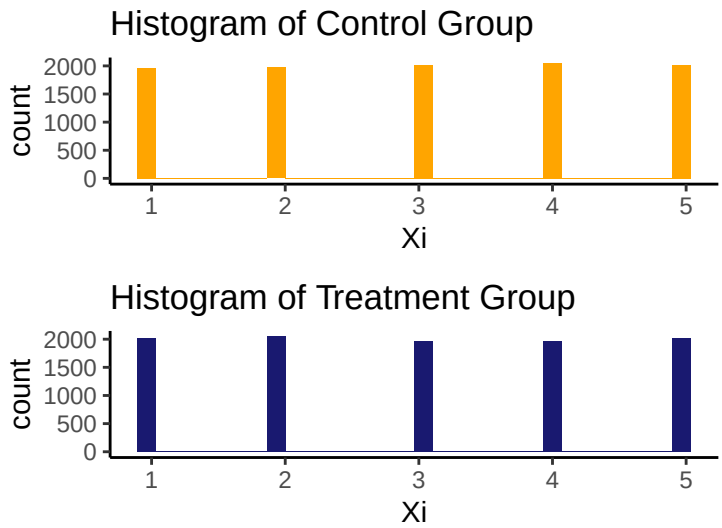
b) Make two separate histograms of Xi for the treatment and control group. What do you see and does it comply with your expectations, explain why or why not?

```
control_hist <- df_tibble %>%
  filter(Di == 0) %>%
  ggplot(aes(x = Xi)) +
  geom_histogram(fill = "orange") +
```

```
  labs(title = "Histogram of Control Group") +
  theme_classic()

treatment_hist <- df_tibble %>%
  filter(Di == 1) %>%
  ggplot(aes(x = Xi)) +
  geom_histogram(fill = "midnightblue") +
  labs(title = "Histogram of Treatment Group") +
  theme_classic()

control_hist / treatment_hist
```



c) Test whether Di is uncorrelated with the pre-treatment characteristic Xi and report your finding.

```
# test correlation between variables
cor.test(df_tibble$Di, df_tibble$Xi)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_tibble$Di and df_tibble$Xi
## t = -1.3238, df = 19998, p-value = 0.1856
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.023216616  0.004499336
## sample estimates:
##          cor
## -0.009360438
```

The correlation coefficient between Di and Xi is -0.0047, which is very small. This suggests that Di and Xi are uncorrelated.

d) Test whether Di is uncorrelated with the potential outcomes Yi_0 and Yi_1 and report your finding (only possible for this synthetic dataset where we know all potential outcomes).

```r
# test correlation between variables
print(cor.test(df_tibble$Di, df_tibble$Yi_0))
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_tibble$Di and df_tibble$Yi_0
## t = -0.43337, df = 19998, p-value = 0.6648
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01692300  0.01079512
## sample estimates:
##          cor
## -0.003064533
```

```r
print(cor.test(df_tibble$Di, df_tibble$Yi_1))
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_tibble$Di and df_tibble$Yi_1
## t = -0.66837, df = 19998, p-value = 0.5039
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.018584232  0.009133529
## sample estimates:
##          cor
## -0.004726259
```

The correlation coefficient between Di and Yi_0 is 0.00563, which is very small. This suggests that Di and Yi_0 are uncorrelated. The correlation coefficient between Di and Yi_1 is -0.00222, which is very small. This suggests that Di and Yi_1 are also uncorrelated.

e) Estimate the ATE by comparing mean outcomes for treatment and control group. Test for mean difference between the groups and report your findings.

```r
# Output the mean of the potential outcomes:
mean_y0 <- mean(df_tibble$Yi_0)
mean_y1 <- mean(df_tibble$Yi_1)

ATE = mean_y1 - mean_y0

ATE
```

```
## [1] 1.500513
```

The estimated ATE is 1.508545.

f) Estimate the ATE using a simple regression of (i) Yi on Di and (ii) Yi on Di and Xi and report your findings and include.

```r
df_tibble <- df_tibble %>%
  mutate(Yi = ifelse(Di == 0, # if Di equals 0
                     Yi_0, # then choose this value
                     Yi_1)) # if not choose this value

# the effect of treatment (Di) on the outcome (Yi)
lm_YiDi <- lm(Yi ~ Di,data = df_tibble)

# effect of treatment (Di) and value (Xi) on the outcome (Yi)
lm_YiDiXi <- lm(Yi ~ Di + Xi,data = df_tibble)

tab_model(lm_YiDi)
```

Yi			
Predictors	Estimates	CI	p
(Intercept)	1.51	1.48 – 1.54	<0.001
Di	1.49	1.45 – 1.53	<0.001
Observations	20000		
R2 / R2 adjusted	0.197 / 0.197		

```r
tab_model(lm_YiDiXi)
```

Yi			
Predictors	Estimates	CI	p
(Intercept)			

-0.76

-0.80 – -0.73

<0.001

Di

1.51

1.48 – 1.54

<0.001

Xi

0.75

0.74 – 0.76

<0.001

Observations

20000

R2 / R2 adjusted

0.603 / 0.603

The ATE for the first regression (Yi on Di), is 1.51162. This means that when there is a treatment, the outcome increases by 1.51162 units. The ATE for the second regression (Yi on Di and Xi) is 1.521709, which is similar to the first regression result. However, the standard error decreased from 0.02136 to 0.015008. This occurred because there is another variable (Xi) that was added to explain the variation, which means that the model accounts for more the variation in the data.

2 Part 2

Part 2 is based on Gertler, Martinez, and Rubio-Codina (2012) (article provided on canvas) and covers impact evaluation of the Mexican conditional cash transfer Progresa (later called Oportunidades, now Prospera). Basically, families with low-incomes received cash benefits if they complied to certain conditions, such as regular school attendance for children and regular healthcare visits. You can read more about the program in the Boxes 2.1 (p.10) & 3.1 (p.40) of the Handbook on impact evaluation: quantitative methods and practices by Khandker, B. Koolwal, and Samad (2010). The program followed a randomized phase-in design. You have data on households (hh) from 1999, when treatment hh have been receiving benefits for a year and control hh have not yet received any benefits. You can find a description of the variables at the end of the assignment. Again, briefly report what you find or respond to the questions.

a) Some variables in the dataset were collected in 1997 before treatment began. Use these variables to test whether there are systematic differences between the control and the treatment group before the cash transfer began (i.e. test for systematic differences on all 1997 variables).

- Variables: homeown97, dirtfloor97, bathroom97, electricity97, hhsize97, vani, vani1, vani2

Describe your results. Does it matter whether there are systematic differences? Why or why not? Would it be a mistake to do the same test with these variables if they were collected after treatment began and if so why? Note: If your variable is a proportion (e.g. binary variables), you should use a proportions test, otherwise you can use a t-test.

Using t.test() for the continuous variables and prop.test() for the binary variables, it was clear that there were not really any differences between the treatment and control groups in 1997. For the vani1, vani2, hhsize97 variables, the difference in means was usually very small, suggesting that the groups were similar. The proportion of homeowners in 1997 in the control group is 0.9351361, compared to 0.9410768 in the treatment group. The proportion of dirt floors in 1997 in the control group is 0.6695191, compared to 0.7112502 in the treatment group. The proportion of bathrooms that were exclusive in 1997 in the control group is 0.5765874, compared to 0.5592491 in the treatment group. The proportion of with electricity in 1997 in the control group is 0.7200068, compared to 0.6210403 in the treatment group.

It does matter if there are systematic differences because to compare the effect of treatment, we need to assume that the groups are the same in every other way / we have to be able to control for the differences so that we can isolate the true effect of the treatment on the outcomes of the groups. It would be a mistake to collect these variables after treatment began because then you wouldn't have a control and you couldn't establish what baseline was.

b) Estimate the impact of program participation on the household's value of animal holdings (vani) using a simple univariate regression. Interpret the intercept and the coefficient. Is this an estimate of a treatment effect?

The intercept means that when treatment is 0 (not-treated), the household's value of animal holdings is 1691.47 in 1997 prices. The coefficient means that when treatment is 1 (treated), the value of the household's animal holdings increases by 50.21 in 1997 prices. If vani is the 1999 measure of animal holdings value that has been translated to 1997 prices, then yes it is an estimate of treatment effect since it is comparing the value of animal holdings between those that have been in the program (receiving treatment) and those that are not in the program (control) for a year. The very high p-value for the treatment condition suggests that there is not a statistically significant effect of the treatment on value of household's animal holdings.

```r
# linear regression of treatment on vani
vaniModel <- lm(vani ~ treatment, data = progresa)

tab_model(vaniModel)
```

vani		
Predictors	Estimates	CI	p
(Intercept)	1691.47	1596.54 – 1786.39	<0.001
treatment	50.21	-75.78 – 176.20	0.435
Observations	13514		
R2 / R2 adjusted	0.000 / -0.000		

c) Now, include at least 6 independent control variables in your regression. How does the impact of program participation change? Choose one of your other control variables and interpret the coefficient.

The impact of the program participation (treatment) increases when 6 other independent control variables are added to the regression. The coefficient increased from 50.21 to 125.1, which means that given all other variables held constant, the effect of getting treatment increases the value of animal holdings by an average of 125.1 in 1997 prices.

An interesting variable is whether the homeowner is female or not. Holding all other variables constant, when homeowners are female, the value of animal holdings *decreases* by 517.75 in 1997 prices.

```r
controlModel <- lm(vani ~ treatment + female_hh + educ_sp + crop_sales + mobilehealth + nonfood + lnup_
                   data = progresa)

tab_model(controlModel)
```

vani		
Predictors	Estimates	CI	p

Estimates

CI

p

(Intercept)

699.85

-88.76 – 1488.47

0.082

treatment

125.09

-423.52 – 673.70

0.655

female hh

-517.75

-1628.12 – 592.62

0.360

educ sp

0.36

-118.75 – 119.47

0.995

crop sales

0.00

-0.00 – 0.00

0.386

mobilehealth

719.51

111.97 – 1327.05

0.020

nonfood

3.96

2.89 – 5.04

<0.001

lnup cwagepw

96.34

2.93 – 189.74

0.043

Observations

1214

R2 / R2 adjusted

0.056 / 0.050

d) The dataset also contains a variable intention_to_treat. This variable identifies eligible households in participating villages. Most of these households ended up in the treatment group receiving the cash transfer, but some did not. Test if the program has an effect on the value of animal holdings of these non-participants (spillover effects). Think of a reason why there might or might not be spillover effects.

For people that they had the intention to treat, but weren't treated, the program did not have a statistically significant effect on the value of animal holdings in 1997 prices. The difference in the impact on animal holding value for those that were treated, and those that were not BUT WERE intended to be treated is very small. The coefficient for treated was approximately 50, and the coefficient for intent to be treated was approximately 49. There is no evidence of a spillover effect, since the impact was the same.

This suggests that there was no overall effect of treatment. This might be because those that were involved in the program did not disperse the cash payout that they were given, and decided to keep it. Hypothetically, this would mean that the treatment would not spill over into the community because the program participants would not be sharing their money and the non participants would not recieve any benefits from the program.

```
# update NA values to 0 since classification of 0 are listed as NA in the data
progresa_itt_df <- progresa %>%
  mutate(treatment = replace_na(treatment, 0))

# Examine number of hh that were intended to get treatment and that ended up receiving treatment
table(treatment = progresa_itt_df$treatment,
      intention_to_treat = progresa_itt_df$intention_to_treat,
      exclude = NULL)
```

```
##          intention_to_treat
## treatment    0    1
##         0 6215  490
##         1    0 7671
```

```
# Create a new treatment variable that is:
intention <- progresa_itt_df %>%
  mutate(pseudo_treatment = ifelse(intention_to_treat == 1 & treatment == 0, #if intention_to_treat ==
                                   1,
                                   0)) # = 0 for normal control hh.

# fille treatment = 1 with NA values
intention$pseudo_treatment[intention$treatment == 1] <- NA

intentionModel <- lm(vani ~ pseudo_treatment, data = intention)

tab_model(intentionModel)
```

vani

Predictors	Estimates	CI	p
(Intercept)	1712.44	1615.53 – 1809.34	<0.001
pseudo treatment	46.83	-311.64 – 405.30	0.798
Observations	6705		
R2 / R2 adjusted	0.000 / -0.000		