

EDS241: Assignment 1

Briana Barajas

01/24/2024

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, Y_{i_0} , and with treatment, Y_{i_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 D_i where the units with the “1’s” are chosen randomly.

```
# set seed to keep random column the same every time code runs
set.seed(123)

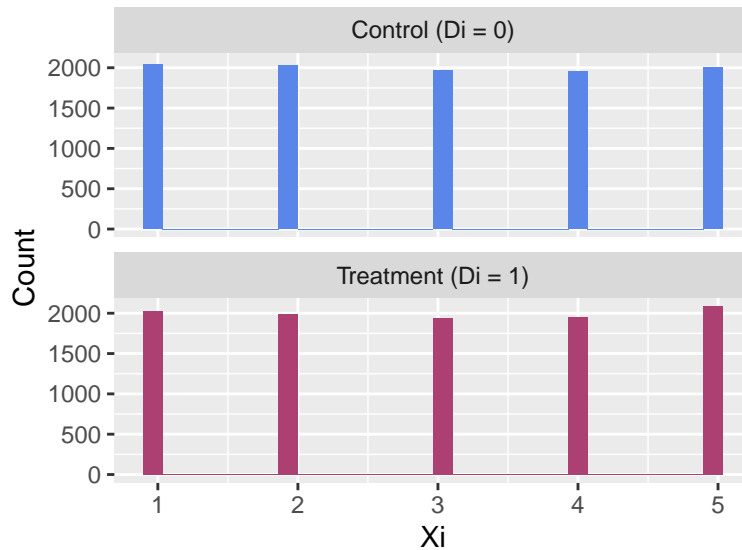
# calculate 1/2 the population
N_half <- N/2

# create Di column w/proportionate values of 0 and 1, randomly assigned
df$Di <- sample(c(rep(1, N_half), rep(0, N_half)))
head(df, 2)
```

```
##      Xi      Yi_0      Yi_1 Di
## 1  2 1.8276592 3.205607  0
## 2  3 0.2539474 1.774676  0
```

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

```
ggplot(data = df, mapping = aes(x = Xi)) +
  geom_histogram(aes(fill = Di)) +
  labs(y = 'Count') +
  facet_wrap(~Di, ncol = 1,
            labeller = labeller(Di =
              c("0" = "Control (Di = 0)",
                "1" = "Treatment (Di = 1)"))) +
  scale_fill_binned(high = 'maroon', low = 'dodgerblue') +
  guides(fill = 'none')
```



ANS: The histogram displays the expected results, there is a relatively even distribution between the treatment and control groups. Both groups also have an even number of units.

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

```
# print correlation results
cor.test(df$Di, Xi)

##
## Pearson's product-moment correlation
##
## data: df$Di and Xi
## t = 1.0722, df = 19998, p-value = 0.2837
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.006278216 0.021438571
## sample estimates:
## cor
## 0.007581634
```

ANS: The correlation value between D_i and X_i is a very small, and the p-value is large, so we fail to reject the null hypothesis that the true correlation is equal to 0. In other words, D_i is uncorrelated with the pre-treatment characteristic X_i .

- d) Test whether D_i is uncorrelated with the potential outcomes Y_{i_0} and Y_{i_1} and report your finding (only possible for this synthetic dataset where we know all potential outcomes).

```
cor.test(df$Di, df$Yi_0)

##
## Pearson's product-moment correlation
##
## data: df$Di and df$Yi_0
## t = 1.8353, df = 19998, p-value = 0.06648
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
## -0.0008825604 0.0268311534
## sample estimates:
##      cor
## 0.01297679
```

```
cor.test(df$Di, df$Yi_1)
```

```
##
## Pearson's product-moment correlation
##
## data: df$Di and df$Yi_1
## t = 1.2675, df = 19998, p-value = 0.205
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.004897135 0.022819019
## sample estimates:
##      cor
## 0.008962663
```

ANS: The correlation between D_i and Y_{i0} is slightly stronger compared to Y_{i1} , but overall there is no correlation between treatment (D_i) and Y_{i0} or Y_{i1} .

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

```
df %>% summarise(ATE = mean(Yi_1 - Yi_0))
```

```
##      ATE
## 1 1.497361
```

- f) Estimate the ATE using a simple regression of (i) Y_i on D_i and (ii) Y_i on D_i and X_i and report your findings.

```
# create a Yi column that will be used for the regressions
df <- df %>%
  mutate(Yi = (Di*Yi_1+(1-Di)*Yi_0))

# regression of Yi on Di
lm(Yi~Di, data = df)
```

```
##
## Call:
## lm(formula = Yi ~ Di, data = df)
##
## Coefficients:
## (Intercept)      Di
##      1.489      1.529
```

```
# regression of Yi on Di and Xi
lm(Yi ~ Di + Xi, data = df)
```

```
##
## Call:
## lm(formula = Yi ~ Di + Xi, data = df)
##
## Coefficients:
## (Intercept)          Di          Xi
##    -0.7354      1.5128      0.7452
```

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).** 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.

```
# creating subsets for control/treatment groups
control <- progresas %>%
  filter(treatment == 0)
treated <- progresas %>%
  filter(treatment == 1)

# continuous variables (t.test)
t.test(hhsize97 ~ treatment, data = progresas)

##
## Welch Two Sample t-test
##
## data: hhsize97 by treatment
## t = -7.7297, df = 14013, p-value = 0.00000000000001151
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
## -0.4129861 -0.2459023
## sample estimates:
## mean in group 0 mean in group 1
## 5.453244 5.782688

t.test(vani ~ treatment, data = progresas)

##
## Welch Two Sample t-test
##
## data: vani by treatment
## t = -0.41045, df = 13753, p-value = 0.6815
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
## -149.11332 97.47725
## sample estimates:
```

```
## mean in group 0 mean in group 1
##      1715.860      1741.678
```

```
t.test(vani1 ~ treatment, data = progresas)
```

```
##
## Welch Two Sample t-test
##
## data:  vani1 by treatment
## t = -2.6921, df = 14359, p-value = 0.007109
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -76.58290 -12.04909
## sample estimates:
## mean in group 0 mean in group 1
##      338.0119      382.3279
```

```
t.test(vani2 ~ treatment, data = progresas)
```

```
##
## Welch Two Sample t-test
##
## data:  vani2 by treatment
## t = 0.33301, df = 13603, p-value = 0.7391
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -90.38349 127.37943
## sample estimates:
## mean in group 0 mean in group 1
##      1377.848      1359.350
```

```
# binary variables (proportion test)
a <- table(prograsa$treatment, prograsa$dirtfloor97)[, c(2,1)]
```

- 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?

```
vani_model <- lm(vani ~ treatment, data = progresas)
summary(vani_model)
```

```
##
## Call:
## lm(formula = vani ~ treatment, data = progresas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1742   -1716   -1330    -139   50495
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
```

```
## (Intercept) 1715.86      45.71  37.541 <0.0000000000000002 ***
## treatment    25.82      62.57   0.413              0.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3743 on 14374 degrees of freedom
## Multiple R-squared:  1.184e-05, Adjusted R-squared:  -5.772e-05
## F-statistic: 0.1703 on 1 and 14374 DF,  p-value: 0.6799
```

ANS:

- 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.

```
lm(vani ~ treatment + age_hh + educ_hh + ethnicity_hh + healthcenter + min_dist + hairrigation, data = progresas)

##
## Call:
## lm(formula = vani ~ treatment + age_hh + educ_hh + ethnicity_hh +
##     healthcenter + min_dist + hairrigation, data = progresas)
##
## Coefficients:
## (Intercept)      treatment      age_hh      educ_hh  ethnicity_hh
##      1424.440         31.590       26.267         1.496       -1129.196
## healthcenter      min_dist  hairrigation
##      -791.215         1.168        651.636
```

ANS:

- 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.

```
# find individuals that qualified for AND recieved treatment (1,1)
table(treatment = progresas$treatment,
      int_to_treat = progresas$intention_to_treat, exclude = NULL)

##           int_to_treat
## treatment    0      1
##           0 6215  490
##           1    0 7671

# new treatment variable
progresas$pseudo_treatment <- ifelse(progresas$intention_to_treat == 1 & progresas$treatment == 0,
                                     1, 0)

# test
summary(lm(vani ~ pseudo_treatment, data = progresas))

##
## Call:
## lm(formula = vani ~ pseudo_treatment, data = progresas)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1759  -1729  -1332   -135   50508
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)    1728.59      31.76  54.425 <0.0000000000000002 ***
## pseudo_treatment    30.68     172.03   0.178      0.858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3743 on 14374 degrees of freedom
## Multiple R-squared:  2.212e-06, Adjusted R-squared:  -6.736e-05
## F-statistic: 0.0318 on 1 and 14374 DF,  p-value: 0.8585
```

****Hint****: Create a pseudo-treatment variable that is = 1 for individuals who were intended to get treatment but did not receive it, = 0 for the normal control group and excludes the normal treatment group.