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, 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).

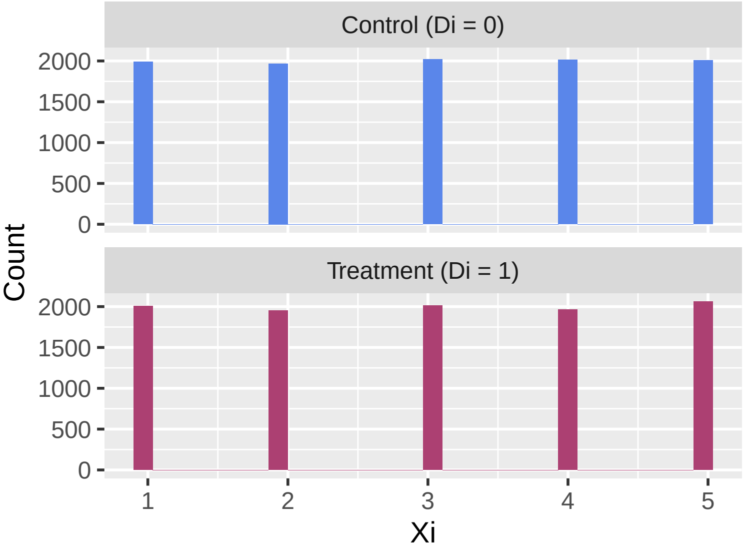
a) Create equally sized treatment and control groups by creating a binary random variable Di where the units with the "1's" are chosen randomly.

```
# set seed 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  4 2.205255 4.874748  0
## 2  4 2.583803 4.943530  0
```

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?

```
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 Di is uncorrelated with the pre-treatment characteristic Xi and report your finding.

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

```
##
##  Pearson's product-moment correlation
##
## data:  df$Di and Xi
## t = 0.14467, df = 19998, p-value = 0.885
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01283633  0.01488202
## sample estimates:
##         cor
## 0.001023043
```

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 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).

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

```
##
##  Pearson's product-moment correlation
##
## data:  df$Di and df$Yi_0
## t = 1.1689, df = 19998, p-value = 0.2425
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.005594537  0.022121950
## sample estimates:
##         cor
## 0.008265294
```

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

```
##
##  Pearson's product-moment correlation
##
## data:  df$Di and df$Yi_1
## t = 1.0117, df = 19998, p-value = 0.3117
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006705644  0.021011318
## sample estimates:
##         cor
## 0.007154211
```

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 1.512004
```

ANS: The average treatment effect (ATE) was approximately 1.5. In this hypothetical, the ATE does not have units, although we do see there's some positive affect on the control group.

f) Estimate the ATE using a simple regression of (i) Yi on Di and (ii) Yi on Di and Xi 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
##       1.506        1.534
```

```r
# 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
##     -0.7341       1.5322       0.7446
```

ANS: The ATE estimated using the linear regression of Yi on Di is approximately 1.5, which is expected as it matches the ATE value calculated in part (e).

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.

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

```
##
##  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 = progresa)
```

```
##
##  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 = progresa)
```

```
##
##  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 = progresa)
```

```
##
##  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)
# dirt floor
df_tb <- table(progresa$treatment, progresa$dirtfloor97)[, c(2,1)]
prop.test(df_tb)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  df_tb
## X-squared = 21.251, df = 1, p-value = 0.00000403
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.05088787 -0.02038203
## sample estimates:
##    prop 1    prop 2
## 0.6756152 0.7112502
```

```
# bathroom
bth_tb <- table(progresa$treatment, progresa$bathroom97)[, c(2,1)]
prop.test(bth_tb)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  bth_tb
## X-squared = 0.13311, df = 1, p-value = 0.7152
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01323423  0.01956821
## sample estimates:
##    prop 1    prop 2
## 0.5624161 0.5592491
```

```r
# electricity
elec_tb <- table(progresa$treatment, progresa$electricity97)[, c(2,1)]
prop.test(elec_tb)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  elec_tb
## X-squared = 105.98, df = 1, p-value < 0.00000000000000022
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.06616693 0.09727077
## sample estimates:
##    prop 1    prop 2
## 0.7027591 0.6210403
```

```r
# homeownership
own_tb <-table(progresa$treatment, progresa$homeown97)[, c(2,1)]
prop.test(own_tb)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  own_tb
## X-squared = 4.0869, df = 1, p-value = 0.04322
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0164618357 -0.0002182024
## sample estimates:
##    prop 1    prop 2
## 0.9327368 0.9410768
```

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?

```r
vani_model <- lm(vani ~ treatment, data = progresa)
vani_model
```

```
##
## Call:
## lm(formula = vani ~ treatment, data = progresa)
##
## Coefficients:
## (Intercept)    treatment
##     1715.86        25.82
```

ANS: Based on the results, we see that the r-squared value is zero. This indicates that applying the treatment (cash transfers) explains 0% of the variability in the value of animal holdings (vani).

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 = progresa)
```

```
##
## Call:
## lm(formula = vani ~ treatment + age_hh + educ_hh + ethnicity_hh +
##     healthcenter + min_dist + hairrigation, data = progresa)
##
## 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: Adding multiple variables slightly improves the model's predicting power, but not by a significatnt amount. The model with 6 control variables explains approximately 4.9% of the variability in vani. Despite this we can still interpret the coefficients. For example, if you were to hold all other variables constant, a one unit increase in the minimum distance between the location and an urban center (`min dist`) results in a 1.17 unit increase in `vani`.

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 = progresa$treatment,
      int_to_treat = progresa$intention_to_treat, exclude = NULL)
```

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

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

# test
lm(vani ~ pseudo_treatment, data = progresa)
```

```
##
## Call:
## lm(formula = vani ~ pseudo_treatment, data = progresa)
##
## Coefficients:
##      (Intercept)  pseudo_treatment
##          1728.59             30.68
```

ANS: This model also has an r-squared value of 0, meaning that `pseudo_treatment` has no influence on `vani`. Although it is not evident for the `vani` variable, spill over effects are still possible. As seen in the table, 7671 families qualified for the treatment and decided to participate. This is much larger that the 490 individuals who qualified and denied treatment (denied cash incentives). Since a larger number decided to participate, it's possible that the improved welfare of these select individuals will "spill over" and improve quality of life in the town all together.