EDS241: Assignment 1

Hope Hahn

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).

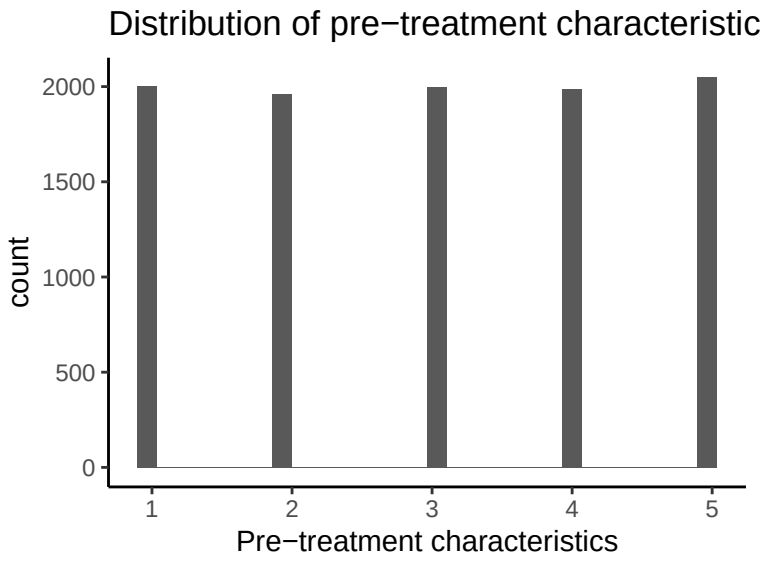
a) Create equally sized treatment and control groups by creating a binary random variable Di where the units with the *1's" are chosen randomly.

```
# add another column to dataframe where the unit 1s are chosen randomly
# 1 is treatment group
# 0 is control group
df$Di = sample(rep(c(0,1), length.out = N), N, replace = FALSE)
```

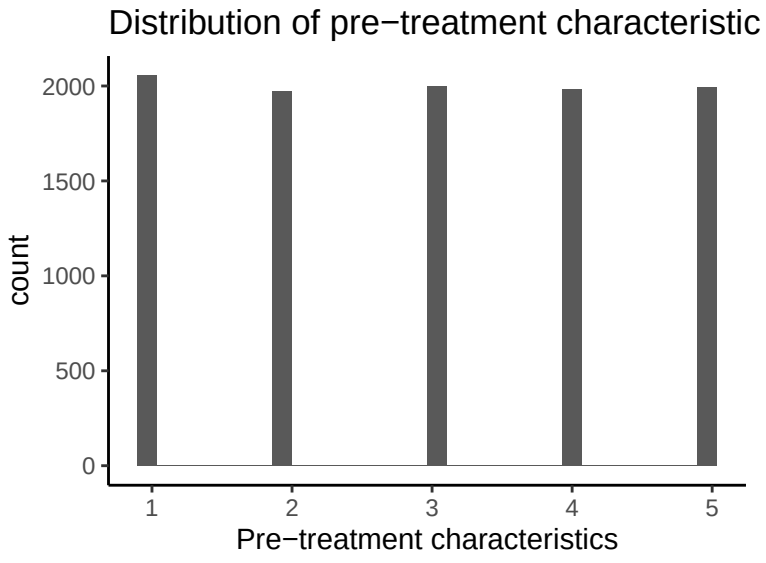
b) Make two separate histograms of Xi for the treatment and control group. What do you see and does it comply with your expectations, explain why or why not?

- **It looks like Xi is distributed somewhat equally for both the treatment and control groups. This complies with my expectations because we randomly assigned Xi as well as Di, so it makes sense that Xi is evenly distributed for both treatment/control groups.**

```
# histogram of Xi for control group
df %>%
  filter(Di == 0) %>%
  ggplot() +
  geom_histogram(aes(x = Xi)) +
  theme_classic() +
  labs(title = "Distribution of pre-treatment characteristics among control group",
       x = "Pre-treatment characteristics")
```




```
# histogram of Xi for treatment group
df %>%
  filter(Di == 1) %>%
  ggplot() +
  geom_histogram(aes(x = Xi)) +
  theme_classic() +
  labs(title = "Distribution of pre-treatment characteristics among treatment group",
       x = "Pre-treatment characteristics")
```



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

- **Di and Xi are uncorrelated.**

```r
# correlation of Di and Xi
cor(df$Di, df$Xi)
```

```
## [1] -0.008171649
```

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).

- **Di is uncorrelated with both Yi_0 and Yi_1.**

```r
# correlation of Di and control outcome
cor(df$Di, df$Yi_0)
```

```
## [1] -0.007368464
```

```r
# correlation of Di and treatment outcome
cor(df$Di, df$Yi_1)
```

```
## [1] -0.007306347
```

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

- **The mean difference between the groups is 1.5.**

```r
# find the difference between the means
mean(Yi_1) - mean(Yi_0)
```

```
## [1] 1.503986
```

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.

- **The ATE is approximately 1.51 when using a regression of Yi on Di, and the ATE is still approximately 1.51 when using a regression of Yi on Di and Xi.**

```r
# find the regression of Yi add to new column
df$Yi <- ifelse(df$Di == 1, Yi_1, Yi_0)

# Simple Regression of Yi and Di
lm1 <- lm(Yi ~ Di, df)

# regression with Yi and Di with Xi
lm2 <- lm(Yi ~ Di + Xi, df)

summary(lm1)
```

```
##
## Call:
## lm(formula = Yi ~ Di, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3517 -1.0423 -0.0143  1.0342  5.7382
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  1.50044    0.01501   99.95 <0.0000000000000002 ***
## Di           1.48227    0.02123   69.82 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.501 on 19998 degrees of freedom
## Multiple R-squared:  0.196,  Adjusted R-squared:  0.1959
## F-statistic:  4875 on 1 and 19998 DF,  p-value: < 0.00000000000000022
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = Yi ~ Di + Xi, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6937 -0.7171  0.0061  0.7174  4.2226
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -0.768239   0.018994  -40.45 <0.0000000000000002 ***
## Di           1.499746   0.014900  100.65 <0.0000000000000002 ***
## Xi           0.753362   0.005248  143.54 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.054 on 19997 degrees of freedom
## Multiple R-squared:  0.604,  Adjusted R-squared:  0.604
## F-statistic: 1.525e+04 on 2 and 19997 DF,  p-value: < 0.00000000000000022
```

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.

- **Using an significance level of alpha = 0.05, household size, value of draft animals, presence of dirtfloor, electricity in household, and homeownership in 1997 are significantly different among the treatment and control groups. It does matter that there are systematic differences in the treated and control groups. This is because the starting point of both groups are not the same, and the pre-treatment conditions/characteristics are not evenly distributed among the two groups. It would be a mistake to do the same test with these variables because the effects might be different because of starting conditions, unrelated to the treatment.**

```
## For binary variables you should use the proportions test

# dirtfloor97
dirtfloortable <- table(treatment = progresa_itt_df$treatment, progresa_itt_df$dirtfloor97, exclude = NU
dirtfloortable <- dirtfloortable[, c(2,1)]

print(prop.test(dirtfloortable))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  dirtfloortable
## 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
```

```
# bathroom97
bathroomtable <- table(treatment = progresa_itt_df$treatment, progresa_itt_df$bathroom97, exclude = NULL
bathroomtable <- bathroomtable[, c(2,1)]
```

```r
print(prop.test(bathroomtable))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  bathroomtable
## 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
# electricity97
electricitytable <- table(treatment = progresa_itt_df$treatment, progresa_itt_df$electricity97, exclude
electricitytable <- electricitytable[, c(2,1)]

print(prop.test(electricitytable))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  electricitytable
## 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
# homeown97
homeowntable <- table(treatment = progresa_itt_df$treatment, progresa_itt_df$homeown97, exclude = NULL)
homeowntable <- homeowntable[, c(2,1)]

print(prop.test(homeowntable))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  homeowntable
## 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?

- **The intercept is 1715.86 and the coefficient is 25.82. This means that when there is no treatment (no program participation), the predicted value of animal holdings is 1715.86, and when there is treatment, the predicted value of animal holdings is 1715.86 + 25.82 (1741.68). The coefficient is the estimated effect of treatment. However, I would like to note that in this context, this would not make sense because the vani data was taken before treatment occurred in 1997, but for the purposes of this assignment, if vani was collected after treatment occurred, it would be the estimate of a treatment effect.**

```
# run linear regression of vani against treatment
summary(lm(vani ~ treatment, progresa_itt_df))
```

```
##
## Call:
## lm(formula = vani ~ treatment, data = progresa_itt_df)
##
## 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
```

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.

- **When there are more independent variables, the effect of treatment (program participation) increases; instead of 25.82, the estimated treatment affect is ~243.04. In addition, the age_hh coefficient is 52.92. This means that as the head of household age increases by 1, the the predicted value of animal holdings increases by 52.92.**

```
# linear regression of vani against 6 independent variables
summary(lm(vani ~ treatment + age_hh + ani_sales + ethnicity_hh + crop_sales + educ_hh, progresa_itt_df
```

```
##
## Call:
## lm(formula = vani ~ treatment + age_hh + ani_sales + ethnicity_hh +
##     crop_sales + educ_hh, data = progresa_itt_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7386  -2638  -1232    593  45484
##
## Coefficients:
##                   Estimate    Std. Error t value        Pr(>|t|)
```

```
## (Intercept)    670.4279998   698.9370469   0.959           0.338
## treatment      243.0415726   293.9669968   0.827           0.409
## age_hh          52.9170574    11.1874594   4.730 0.0000025320922 ***
## ani_sales     3687.3728704   545.9313352   6.754 0.0000000000231 ***
## ethnicity_hh -1948.6324036   310.7174807  -6.271 0.0000000005107 ***
## crop_sales       0.0009028     0.0007713   1.170           0.242
## educ_hh         61.6428439    68.6697161   0.898           0.370
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4913 on 1116 degrees of freedom
##   (13253 observations deleted due to missingness)
## Multiple R-squared:  0.09195,    Adjusted R-squared:  0.08707
## F-statistic: 18.83 on 6 and 1116 DF,  p-value: < 0.00000000000000022
```

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.

- **Based on the pseudo-treatment variable, there appears to be spillover effects on the non-participants. This might be because those receiving cash benefits are living in close proximity to others in eligible households. This means that those in eligible households that did not receive the benefits are likely living near households that do receive benefits. It is possible that increasing value of neighbors would increase value of the area as a whole, so those in proximity also benefit. Additionally, if neighbors are increasing in value of animals, there is a chance that they are gifting/sharing animals with neighbors who are not receiving benefits, and therefore increasing the value of animals in those households as well. These situations could cause spillover effects.**

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.