EDS241: Assignment 1

Heather Childers

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).

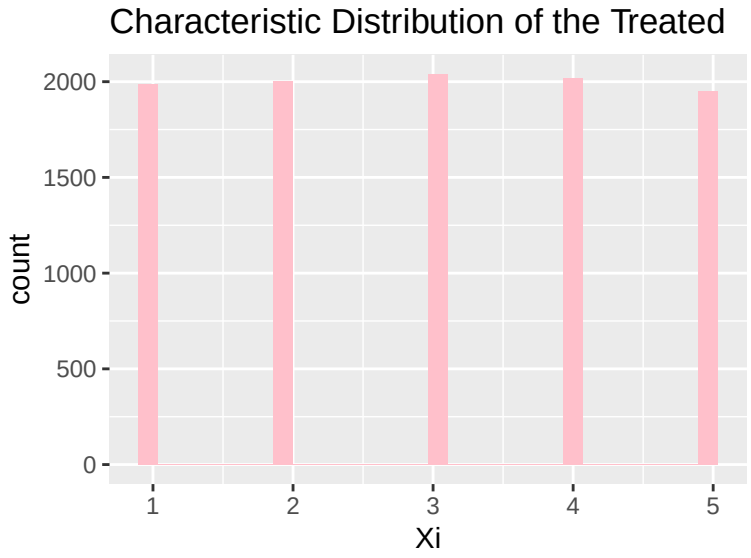
a) Create equally sized treatment and control groups by creating a binary random variable Di where the units with the *1's" are chosen randomly.

```
set.seed(456)
#Add a new column to the dataframe that randomly selects 1's or 0's
#indicating treatment or control group
df_groups <- df %>%
  mutate(groups=sample(rep(c(1, 0), length=n())))

#All treted group
case <- df_groups %>%
  filter(groups == 1)

#All non-treted group
control <- df_groups %>%
  filter(groups == 0)
```

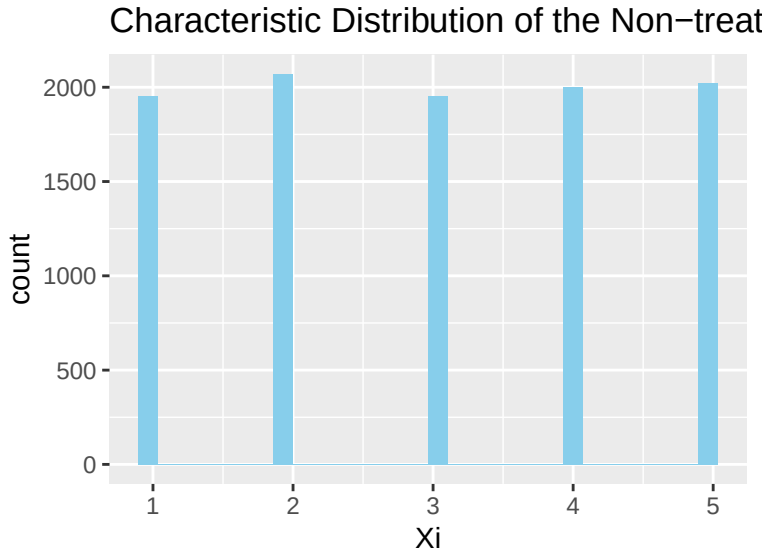
b) Make two separate histograms of Xi for the treatment and control group. What do you see and does it comply with your expectations, explain why or why not?

Yes these distributions comply with my expectations. I would assume that a truly randomly selected treatment group and control group would have roughly equal amounts of participants from each of the 5 characteristic groups. This is because the random sampling gives every participant an equal likelihood of being selected.

```
#Plot the histogram showing the counts from each characteristic in the tratment group
Treatment_group <- ggplot(case, aes(x = Xi))+
         geom_histogram(fill = "pink")+
  labs(title = "Characteristic Distribution of the Treated")
#Plot the histogram showing the counts from each characteristic in the control group
Control_group <- ggplot(control, aes(x = Xi))+
         geom_histogram(fill = "skyblue")+
  labs(title = "Characteristic Distribution of the Non-treated")

Treatment_group
```




```
Control_group
```



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

For a cor.test the null hypothesis is always that the correlation is zero. The correlation test below is showing a correlation coefficient of -0.0016 which implies almost no correlation between the pre-treatment characteristic and the group selection parameter Di. This is further justified by the high p-value showing that we will fail to reject the null hypothesis. This is generally showing that Di and Xi are uncorrelated because the correlation parameter is close to zero and we can't say with statistical significance that the correlation isn't zero.

```
cor.test(df_groups$groups, df_groups$Xi)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_groups$groups and df_groups$Xi
## t = -0.61694, df = 19998, p-value = 0.5373
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.018220725  0.009497129
## sample estimates:
##          cor
## -0.004362636
```

a) 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).

For a cor.test the null hypothesis is always that the correlation is zero.

The first correlation test below is showing a correlation coefficient of -0.006 which implies almost no correlation between the group selection parameter Di and the outcome if untreated. This is further justified by the high p-value showing that we will fail to reject the null hypothesis at a a significance level below alpha = 0.3. This is generally showing that Di and Yi_0 are uncorrelated because the correlation parameter is close to zero and we can't say with statistical significance that the correlation isn't zero.

The same analysis is true for the second correlation test analyzing the correlation between the group selection parameter Di and the outcome if treated. The correlation coefficient is -0.004 which is basically zero and has an even higher p-value showing that we again can't say the correlation isn't zero.

```
#Correlation test to see if Di and Yi_0 are correlated
cor.test(df_groups$groups, df_groups$Yi_0)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_groups$groups and df_groups$Yi_0
## t = 0.29352, df = 19998, p-value = 0.7691
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01178390  0.01593437
## sample estimates:
##         cor
## 0.002075633
```

```r
#Correlation test to see if Di and Yi_1 are correlated
cor.test(df_groups$groups, df_groups$Yi_1)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_groups$groups and df_groups$Yi_1
## t = -0.60191, df = 19998, p-value = 0.5472
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.018114454  0.009603425
## sample estimates:
##          cor
## -0.004256332
```

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

The difference in means is roughly 1.5, and the t.test verifies that estimate. The very small p-value also allows us to reject the null hypothesis that the true difference in means is equal to zero at a very high significance level (alpha < 0.01).

```r
#Calculate the mean of the treated
treated_mean <- mean(case$Yi_1)
#Calculate the mean of the untreated
control_mean <- mean(control$Yi_0)

print(treated_mean - control_mean)
```

```
## [1] 1.494608
```

```r
#Test the difference in means for the two groups
t.test(case$Yi_1, control$Yi_0)
```

```
##
##  Welch Two Sample t-test
##
## data:  case$Yi_1 and control$Yi_0
## t = 70.714, df = 17913, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.453180 1.536037
## sample estimates:
## mean of x mean of y
##  3.000764  1.506156
```

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

Based on the formula from lecture, we can use the Estimate for groups as the estimate for beta 1 which is our ATE. Base on setting our beta_0 = 1.5 and our Beta_1 = 1.5 we roughly get the means we calculated above. This showed that your expected value of untreated is ~3 and the expected value untreated is ~1.5.

This p-values for these estimates are also very high which gives us a high level of confidence around our estimates.

```
realistic_data <- df_groups %>%
  mutate(Yi = case_when(
    groups == 1 ~ Yi_1,
    groups == 0 ~ Yi_0
  ))

summary(lm(data = realistic_data, Yi ~ groups))
```

```
##
## Call:
## lm(formula = Yi ~ groups, data = realistic_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3636 -1.0458  0.0035  1.0255  5.2125
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  1.50616    0.01495  100.78 <0.0000000000000002 ***
## groups       1.49461    0.02114   70.71 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.495 on 19998 degrees of freedom
## Multiple R-squared:    0.2,  Adjusted R-squared:    0.2
## F-statistic:  5001 on 1 and 19998 DF,  p-value: < 0.00000000000000022
```

```
summary(lm(data = realistic_data, Yi ~ groups + Xi))
```

```
##
## Call:
## lm(formula = Yi ~ groups + Xi, data = realistic_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0805 -0.7159  0.0008  0.7161  4.4933
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -0.748094   0.019127  -39.11 <0.0000000000000002 ***
## groups       1.503829   0.014946  100.62 <0.0000000000000002 ***
## Xi           0.749618   0.005301  141.41 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.057 on 19997 degrees of freedom
## Multiple R-squared:    0.6,  Adjusted R-squared:    0.6
## F-statistic: 1.5e+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.

   For the household size variable there is a slight difference in the means but since the difference is less than one full person, it is generally safe to say the means are equal. As for the proportion tests, ht home ownership variable and the dirt floor variable both have high p values which shows that the difference in proportions are not different at the statistically significant level.

a) 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?

Yes, it matters. It would be a mistake to to do the same test with these variables if it showed that they were different at a statistcally significant level. This is because you don't want to run an experiment where the treatment and the control groups have vastly different characteristics. This would add in a layer of confounding variables because you wouldn't know whether the treatment would work for all individuals or if the individuals in the treatment group had some characteristic that impacted their outcome for the treatment.

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 of 1715.86 says that untreated, the average family has ~1716 animal holdings. The coefficient, also the estimate of the Average Treatment Effect, is 25.82 which means that the family's animal holdings increased by ~26 for every one unit increase in the treatment.

```
summary(lm(data = progresa_df, vani ~ treatment))
```

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

b) 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.

Each of the variables that I selected had a different impact on the participation. Increasing the animal holdings, having a female run household, and having access to a healthcenter all decreased the likelihood of participation, where as the education levels and the ethnicity of the household increased the likelihood of participation. When looking at the education level of the household (educ_hh) the coefficient shows that for every one unit increase in education, there is a 0.1% increase in likelihood that the household will participate.

```
summary(lm(data = progresa_df, treatment ~ vani + female_hh + educ_hh + healthcenter + ethnicity_hh + ed
```

```
##
## Call:
## lm(formula = treatment ~ vani + female_hh + educ_hh + healthcenter +
##     ethnicity_hh + educ_sp, data = progresa_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6791 -0.5122  0.3437  0.4829  0.5511
##
## Coefficients:
##                    Estimate    Std. Error t value             Pr(>|t|)
## (Intercept)   0.6361069495  0.0120916166  52.607 < 0.0000000000000002 ***
## vani         -0.0000004825  0.0000011253  -0.429              0.66808
## female_hh    -0.0470327146  0.0143837932  -3.270              0.00108 **
## educ_hh       0.0010016496  0.0018615611   0.538              0.59054
## healthcenter -0.1363388081  0.0108198844 -12.601 < 0.0000000000000002 ***
## ethnicity_hh  0.0204545127  0.0086410336   2.367              0.01794 *
## educ_sp       0.0014136963  0.0019918860   0.710              0.47788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.496 on 14369 degrees of freedom
## Multiple R-squared:  0.01196,    Adjusted R-squared:  0.01155
## F-statistic:    29 on 6 and 14369 DF,  p-value: < 0.00000000000000022
```

b) 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 t-test, there was a significant difference in the animal holdings between the non-participants and the treated participants. This does make some sense because when looking at the participation rates above, having more animal holdings decreased your likelihood of participating in this study. However, this

could also be due to spillover effects. Some reasons there might be spillover effects include: The money was all funneled into the community and therefore everyone recieved the benefits of the community being selected for the program, or nearby ranchers took on extra animal holdings if the groups that recieved funding grew too quickly and couldn't care for the extra animal holdings for the extent of the program. One reson there may not be spillover effects aside from the participation results stated previously is that money and property(such as the animal holdings) are valuable, and therefore wouldn't be shared among the group and there would be penalties for stealing.

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.

```
# Examine number of hh that were intended to get treatment and that ended up receiving treatment
spill <- table(treatment = progresa$treatment, intention_to_treat = progresa$intention_to_treat, exclud

#Create the new column identifying spillover individuals
progresa_df <- progresa_df %>%
  mutate(spillover = ifelse(progresa_df$intention_to_treat == 1 & progresa_df$treatment == 0, 1, 0))

#Separate the datasets into groups for the t.test
spillover<- progresa_df %>%
  filter(spillover == 1)
treated_prog <- progresa_df %>%
  filter(treatment == 1)
#Check to see if there is a difference in mean animal holdings
t.test(spillover$vani, treated_prog$vani)
```

```
##
##  Welch Two Sample t-test
##
## data:  spillover$vani and treated_prog$vani
## t = 0.09504, df = 541.03, p-value = 0.9243
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -346.0113  381.1951
## sample estimates:
## mean of x mean of y
##  1759.270  1741.678
```