EDS241: Assignment 1

Patty Park

01/25/2024

1 Part 1

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.

```
# Define variables and generate data:
  N <- 20000
  Xi <- sample(c(1,2,3,4,5),N, replace=TRUE)
  m0_Xi <- 0.5*Xi
  m1_Xi <- 1*Xi
  Di <- sample(c(0,1), N, replace = TRUE)
# Generate correlated error terms:
  ei <- mvrnorm(n=N,mu=c(0, 0),Sigma=matrix(c(1,0.75,0.75,1), ncol=2))
# Calculate potential outcomes:
  Yi_0 = m0_Xi + ei[,1]
  Yi_1 = m1_Xi + ei[,2]
# Output the mean of the potential outcomes:
  mean(Yi_0)
```

```
## [1] 1.50587
```

```
  mean(Yi_1)
```

```
## [1] 3.008077
```

```
# Create a dataframe from the vectors:
  df_1 <- data.frame(Xi, Yi_0, Yi_1, Di)
```

- **Answer**: By creating a new group (Di) for both the treatment group and the control group, we can see how it has randomly decided those that are chosen as 1's.

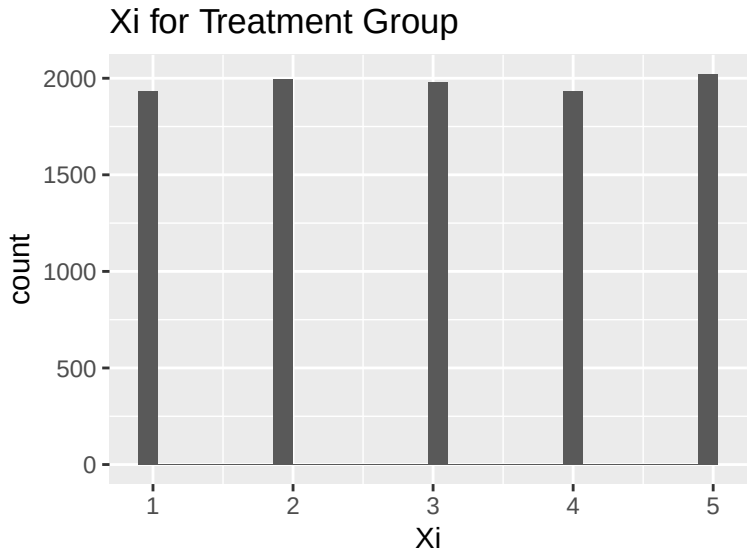
b) Make two separate histograms of Xi for the treatment and control group. What do you see and does it comply with your expectations, explain why or why not?

- **Answer**: These histogram comply to my expectations as the Xi were also randomly picked as well. It seemed as if they were picked in equal parts. And because the 0 and 1 were equally randomly picked, it was not a surprise to see both histogram to have around the same amount of count across all bars.

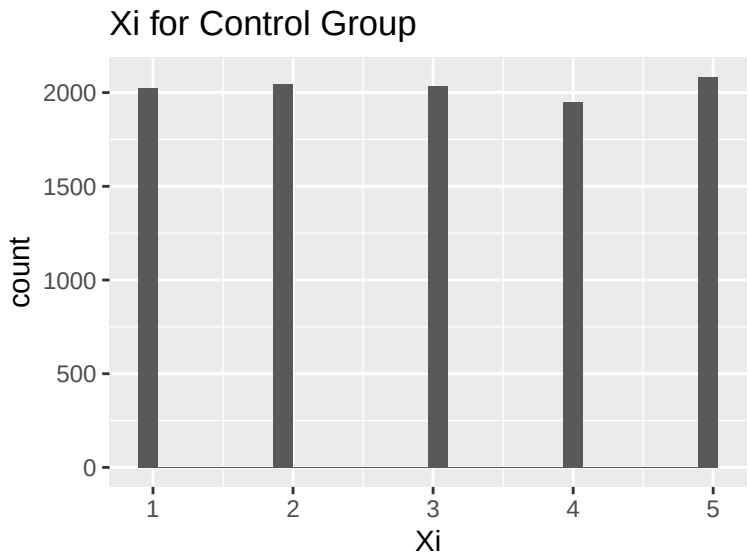
```
#seperate treatment and control for easier graph
  treatment_group <- df_1 %>%
    filter(Di == 1)

  control_group <- df_1 %>%
    filter(Di == 0)

#histogram for treatment group (1)
hist_treatment <- ggplot(data = treatment_group, aes(x = Xi)) +
  geom_histogram() +
  labs(title = "Xi for Treatment Group")
print(hist_treatment)
```




```
#histogram for control group (0)
hist_control <- ggplot(data = control_group, aes(x = Xi)) +
  geom_histogram() +
  labs(title = "Xi for Control Group")
print(hist_control)
```



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

```
#find the correlation between Di and Xi
Di_Xi_corr <- cor.test(df_1$Di, df_1$Xi)
print(Di_Xi_corr)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_1$Di and df_1$Xi
## t = 0.44362, df = 19998, p-value = 0.6573
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01072264  0.01699547
## sample estimates:
##         cor
## 0.003137021
```

-**Answer**: Looking at the coefficients, there seem to be very little correlation between Di and the pretreated Xi. This shows that the treatment has little to no effect on the characteristic Xi and has no bias being put on it.

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).

```
#Correlation between Di and Yi_0
Di_Yi_0_corr <- cor.test(df_1$Di, df_1$Yi_0)

#Correlation between Di and Yi_1
Di_Yi_1_corr <- cor.test(df_1$Di, df_1$Yi_1)

print(Di_Yi_0_corr)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_1$Di and df_1$Yi_0
## t = 1.2546, df = 19998, p-value = 0.2096
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.004988068  0.022728131
## sample estimates:
##         cor
## 0.008871735
```

```
print(Di_Yi_1_corr)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_1$Di and df_1$Yi_1
## t = 0.98702, df = 19998, p-value = 0.3236
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006880417  0.020836613
## sample estimates:
##         cor
## 0.006979438
```

-**Answer**: Same as before, because of the very low coefficient, there is little to no correlation that Di has an effect on Yi_0 (without treatment) or Yi_1 (with treatment). Because the coefficient is so low, a satifactory solution would be that Di does not introduce any bias to Yi_0 and Yi_1.

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

```
#Average mean of Y1 - Y0
ATE_mean <- mean(df_1$Yi_1) - mean(df_1$Yi_0)
print(ATE_mean)
```

```
## [1] 1.502207
```

-**Answer**: Testing for the mean difference, I get about 1.5, which is the causal effect of the treatment Di from unit i.

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.

```
#find Yi and put in new column
df_2 <- df_1 %>%
  mutate(Yi = Di*Yi_1+(1-Di)*Yi_0)

#Run regression of Yi on Di
ATE_corr_Yi_Di <- lm(Yi ~ Di, data = df_2)
```

```r
#Run regression of Yi on Di and Xi
ATE_corr_Yi_Di_Xi <- lm(Yi~ Di + Xi, data = df_2)

print(ATE_corr_Yi_Di)
```

```
##
## Call:
## lm(formula = Yi ~ Di, data = df_2)
##
## Coefficients:
## (Intercept)           Di
##       1.495        1.525
```

```r
print(ATE_corr_Yi_Di_Xi)
```

```
##
## Call:
## lm(formula = Yi ~ Di + Xi, data = df_2)
##
## Coefficients:
## (Intercept)           Di           Xi
##     -0.7472       1.5186       0.7468
```

-**Answer**: Here we see when we run a regression of Yi on Di, we get a value very close to 1.5 for our coefficient. When running a regression of Yi on Di and Xi, we also get a value very close to 1.5 for our coefficient. We can assume that Di is not affected by other variables, in this case, Xi. However, we can make the assumption that Yi, is affected by both Di and Yi as the coefficient is drastically different between the two coefficients.

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.

-**Answer**:

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?

```
#create regression
vani_regression <- lm(vani ~ treatment, data = progresa_itt_df)
#print results
print(vani_regression)
```

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

-**Answer**: Here, this tells us that at 0 units, or at no treatment, the value of the animals for each household is around 1716 dollars. With the treatment, going up one unit, it goes up about 26 dollars. This is an estimation of the treatment effect as you are looking at how much there is without the treatment and how much there is with the treatement.

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.

```
#create regression
vani_regression_6 <- lm(vani ~ treatment + age_hh + educ_hh + primary + telesec + healthcenter + female

#print results
print(vani_regression_6)
```

```
##
## Call:
## lm(formula = vani ~ treatment + age_hh + educ_hh + primary +
##     telesec + healthcenter + female_hh, data = progresa_itt_df)
##
## Coefficients:
##  (Intercept)     treatment        age_hh       educ_hh       primary
##       881.27         24.68         30.85          2.90        201.99
##      telesec  healthcenter     female_hh
##       296.31       -956.69       -884.55
```

-**Answer**: Adding in other variables change the total money amount that households have. If they are not a part of the treatment, they have about 880 dollars worth of animals. If they are part of the treatment, they have an increase of about 25 dollars. Here we can say that while other factors does impact the amount of how much their livestocks are worth, it doesn't necessary impact the treatment effects. Looking at education, we can see that there is a very small impact on the amount their livestocks are worth. For every unit increase in education, there is about a 3 unit increase in how much their livestocks are worth.

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.

NA is actually 0

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
intent_treat <- table(treatment = progresa_itt_df$treatment, intention_to_treat = progresa_itt_df$inten
#print results
print(intent_treat)
```

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

```
#create a new treatment variable (pseudo_treatment)
progresa_intent <- progresa_itt_df %>%
  mutate(pseudo_treatment = intention_to_treat == 1 & treatment == 0, 1)
  # = 1 if intention_to_treat == 1 AND not in the actual treatment
  # = 0 for normal control hh.

#run the regression
progresa_intent_reg <- lm(vani ~ pseudo_treatment, data = progresa_intent)

#print results
print(progresa_intent_reg)
```

```
##
## Call:
```

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

```
summary(progresa_intent_reg)
```

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

-**Answer**: The effect from the treatment of this program did not have any spillover effect. After doing the analysis, I receive a coefficient for my `pseudo_treatment` of 30.68, and a p-value of 0.8585. Because of this high p-value, we can assume that there is not significance and that there was not a spillover effect of those that did receive the treatment impacting those that were intending to receive the treatment but did not end up getting it. One reason why I believe there may not be a spillover effect is that families that do own livestock are isolated from others that also do. Because of this, they are not close enough to each other to help each other with buying additional animals for each other.