EDS241: Assignment 1

Linus Ghanadan

1/23/2024

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.

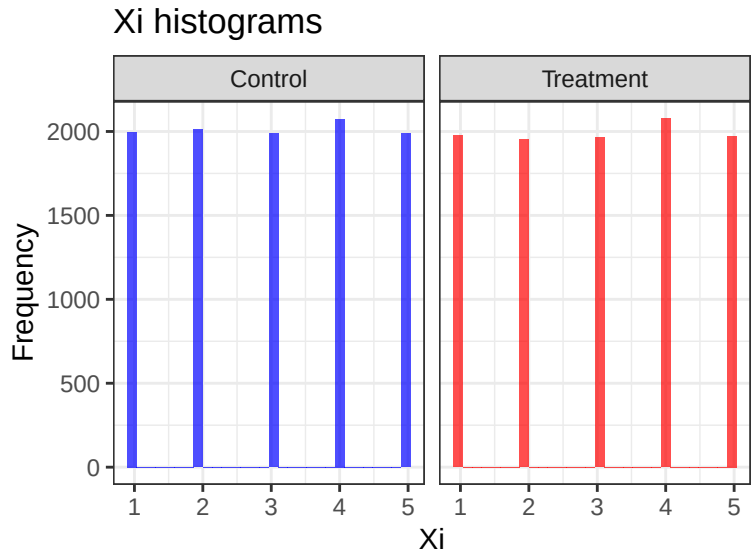
```
# Set seed
set.seed(123)

# Create binary random variable
df$Di <- sample(c(0, 1), N, replace = TRUE)
```

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?

```
# Create 'Group' column specifying whether treatment or control group
df_new <- df %>%
  mutate(Group = ifelse(Di == 1, "Treatment", "Control"))

# Plot histograms
ggplot(df_new, aes(x = Xi, fill = factor(Di))) +
  geom_histogram(alpha = 0.7, bins = 30, position = "dodge") +
  labs(title = "Xi histograms",
       x = "Xi",
       y = "Frequency") +
  scale_fill_manual(values = c("blue", "red")) +
  facet_wrap(~Group) +
  theme_bw() +
  theme(legend.position = "none")
```



The frequency of each Xi within each group is about 2000, which is what we would expect. This is expected because, in this example, Xi could take on one of five values, and we had 20000 units in total, split equally between the control and treatment groups. The difference between the control and treatment groups for Xi = 3 is somewhat surprising, as the control group has a frequency over 2000, while the treatment group has less than 2000. However, it seems reasonable to assume that this just happened by chance.

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

```
# T-test to compare mean Xi between treatment and control groups
t.test(Xi ~ Di, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  Xi by Di
## t = -0.40047, df = 19995, p-value = 0.6888
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.04712535  0.03113573
## sample estimates:
## mean in group 0 mean in group 1
##        3.004176        3.012171
```

We fail to reject the null hypothesis that the true correlation is equal to zero, and since we got an especially high p-value (0.8395), we can be very confident that treatment assignment Di is uncorrelated with pre-treatment characteristic Xi.

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).

```
# T-test to compare mean Yi_0 between treatment and control groups
t.test(Yi_0 ~ Di, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  Yi_0 by Di
## t = 0.080096, df = 19997, p-value = 0.9362
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.03246031  0.03522623
## sample estimates:
## mean in group 0 mean in group 1
##        1.503211        1.501828
```

```
# T-test to compare mean Yi_1 between treatment and control groups
t.test(Yi_1 ~ Di, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  Yi_1 by Di
## t = 0.056968, df = 19992, p-value = 0.9546
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.04632653  0.04910000
## sample estimates:
## mean in group 0 mean in group 1
##        3.001486        3.000099
```

Regarding the potential outcome variable for both the treatment (Yi_0) and control (Yi_1) groups, we fail to reject the null hypothesis that the true correlation with treatment (Di) is equal to zero. Our p-values are high (0.5668 for Yi_0 and 0.4518 for Yi_1), so we can be pretty confident that treatment assignment is uncorrelated with the potential outcome for both the treatment and control groups.

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

```
# Calculate ATE point estimate
ATE_point_estimate <- mean(df$Yi_1 - df$Yi_0)
ATE_point_estimate
```

```
## [1] 1.498273
```

```
# T-test to compare ATE between treatment and control groups
t.test(Yi_1 - Yi_0 ~ Di, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  Yi_1 - Yi_0 by Di
## t = 0.00026895, df = 19996, p-value = 0.9998
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.02753690  0.02754446
## sample estimates:
## mean in group 0 mean in group 1
##        1.498274        1.498271
```

We fail to reject the null hypothesis that the difference in ATE is not equal to zero. Our p-value is high (0.5468), so we can be pretty confident that any differences in potential outcome between the treatment and control groups are negligible.

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

```
# Run simple regression of Yi on Di
model1 <- lm(Yi_1 - Yi_0 ~ Di, data = df)
```

```r
# Display regression summary
summary(model1)
```

```
##
## Call:
## lm(formula = Yi_1 - Yi_0 ~ Di, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2872 -0.7107 -0.0126  0.7009  4.7393
##
## Coefficients:
##                 Estimate   Std. Error t value            Pr(>|t|)
## (Intercept)  1.498274457  0.009906583   151.2 <0.0000000000000002 ***
## Di          -0.000003779  0.014050831     0.0                    1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9935 on 19998 degrees of freedom
## Multiple R-squared:  3.617e-12,  Adjusted R-squared:  -5e-05
## F-statistic: 7.233e-08 on 1 and 19998 DF,  p-value: 0.9998
```

```r
# Run simple regression of Yi on Di and Xi
model2 <- lm(Yi_1 - Yi_0 ~ Di + Xi, data = df)

# Display regression summary
summary(model2)
```

```
##
## Call:
## lm(formula = Yi_1 - Yi_0 ~ Di + Xi, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7275 -0.4734 -0.0025  0.4777  3.7477
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.005632   0.012704   0.443               0.658
## Di          -0.003976   0.009952  -0.400               0.690
## Xi           0.496856   0.003525 140.945 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7037 on 19997 degrees of freedom
## Multiple R-squared:  0.4984, Adjusted R-squared:  0.4983
## F-statistic:  9933 on 2 and 19997 DF,  p-value: < 0.00000000000000022
```

In the first regression, we estimate the average ATE (across both the treatment and control group) to be 1.50, and we have >99% confidence that this average ATE is greater than 0. In addition, we find that being in the treatment group is associated with a 0.000355 unit increase in expected ATE, though the very high p-value means that we are very unsure if the change in ATE is actually greater than zero. Then, after controlling for pre-treatment characteristic Xi, we estimate the average ATE (across both the treatment and control group) to be 0.004, and our p-value is high, indicating that we have basically no idea whether ATE is greater than or less than zero. Like with the previous regression, we find that we are very uncertain

about how a change in treatment impacts ATE. Lastly, we find that a one-unit increase in the pre-treatment characteristic Xi is associated with a 0.497 unit increase in ATE, and we are >99% confident that there is a positive relationship between Xi and ATE.

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.

```
# Subset data for treatment and control groups
treatment_group <- progresa[progresa$treatment == 1, ]
control_group <- progresa[progresa$treatment == 0, ]

# T-test to compare mean 1997 household size (continuous variable) between treatment and control groups
t.test(treatment_group$hhsize97, control_group$hhsize97)
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment_group$hhsize97 and control_group$hhsize97
## t = 6.2017, df = 12393, p-value = 0.0000000005763
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1881309 0.3620133
## sample estimates:
## mean of x mean of y
##  5.782688  5.507616
```

```
# Compare proportion of units that had dirt floor in 1997 (binary variable) between treatment and contr
prop.test(x = c(sum(treatment_group$dirtfloor97, na.rm = TRUE), sum(control_group$dirtfloor97, na.rm = T
          n <- c(length(treatment_group$dirtfloor97), length(control_group$dirtfloor97)))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(treatment_group$dirtfloor97, na.rm = TRUE), sum(control_group$dirtfloor97, na.rm = TRUE)
## X-squared = 49.402, df = 1, p-value = 0.000000000002085
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.04023186 0.07167772
## sample estimates:
##    prop 1    prop 2
## 0.6394000 0.5834452
```

```
# Compare proportion of units that had electricity in 1997 (binary variable) between treatment and cont
prop.test(x = c(sum(treatment_group$electricity97, na.rm = TRUE), sum(control_group$electricity97, na.rm
```

```
          n <- c(length(treatment_group$electricity97), length(control_group$electricity97)))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(treatment_group$electricity97, na.rm = TRUE), sum(control_group$electricity97, na.rm = T
## X-squared = 73.842, df = 1, p-value < 0.00000000000000022
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.08492296 -0.05335533
## sample estimates:
##    prop 1    prop 2
## 0.5583031 0.6274422
```

```
# Compare proportion of units that had bathroom in 1997 (binary variable) between treatment and control
prop.test(x = c(sum(treatment_group$bathroom97, na.rm = TRUE), sum(control_group$bathroom97, na.rm = TRU
          n <- c(length(treatment_group$bathroom97), length(control_group$bathroom97)))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(treatment_group$bathroom97, na.rm = TRUE), sum(control_group$bathroom97, na.rm = TRUE))
## X-squared = 0.00038447, df = 1, p-value = 0.9844
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01583286  0.01641918
## sample estimates:
##    prop 1    prop 2
## 0.5027540 0.5024609
```

```
# Compare proportion of units that owned home in 1997 (binary variable) between treatment and control g
prop.test(x = c(sum(treatment_group$homeown97, na.rm = TRUE), sum(control_group$homeown97, na.rm = TRUE)
          n <- c(length(treatment_group$homeown97), length(control_group$homeown97)))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(treatment_group$homeown97, na.rm = TRUE), sum(control_group$homeown97, na.rm = TRUE)) ou
## X-squared = 25.792, df = 1, p-value = 0.0000003803
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.01891799 0.04327274
## sample estimates:
##    prop 1    prop 2
## 0.8460096 0.8149142
```

From our t-test comparing mean 1997 household size (continuous variable) between treatment and control groups, we reject our null hypothesis that there is no difference in the two means with >99% confidence. Specifically, the mean 1997 household size is 5.78 square units for the treatment group and 5.51 square units for the control group. Regarding the three of the four proportion tests that compare binary variable pre-treatment characteristics across the treatment and control group, we reject our null hypotheses that the two groups are the same, with >99% confidence in all three cases. For the dirt floor variable, we find that 64% of the treatment group and 58% of the control group had dirt floors in 1997. Secondly, for the electricity variable, we find that 56% of the treatment group and 63% of the control group had electricity in 1997. Thirdly, for the home ownership variable, we find that 85% of the treatment group and 81% of the control group were home owners in 1997. Lastly, our proportion test comparing having an exclusive

household bathroom in 1997 between treatment and control groups was the only one of our tests where we failed to reject the null hypothesis that the two groups were the same. In both groups, 50% of units had access to an exclusive bathroom. It does not necessarily matter if there are systematic differences between the treatment and control group, so long as the difference does not appear to be too extreme within the context of the specific natural experiment being conducted. In this case, even though the differences are statistically significant, they do not seem too extreme within the context of this study. In fact, the differences being statistically significant appears to largely be due to the high sample size, which also has a very positive effect on the external validity of the study. While having systematic differences certainly hurts the internal validity of the study, it doesn't disqualify it from being a useful natural experiment. If the data on these variables was collected after treatment began, it would be a mistake to the same tests. The reason why we are doing the tests is so we can understand the baseline characteristics of units in the treatment and control groups, so running these tests based on data that was collected after treatment began would give us irrelevant results. In addition, if there was a problematic difference between groups, there would not be an opportunity to reassign units to achieve a better balance in pre-treatment characteristics.

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?

```
# Run simple regression for impact of treatment on household's value of animal holdings
vani_reg <- lm(vani ~ treatment, data = progresa)

# Display regression summary
summary(vani_reg)
```

```
##
## Call:
## lm(formula = vani ~ treatment, data = progresa)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1742  -1691  -1313   -137  50495
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  1691.47      48.43  34.929 <0.0000000000000002 ***
## treatment      50.21      64.28   0.781               0.435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3702 on 13512 degrees of freedom
##   (862 observations deleted due to missingness)
## Multiple R-squared:  4.516e-05,  Adjusted R-squared:  -2.884e-05
## F-statistic: 0.6103 on 1 and 13512 DF,  p-value: 0.4347
```

On average, households that did not participate in the program had 1,691.47 dollars worth of animal holdings. Program participation is associated with a 50.21 dollar increase in value of a household's animal holdings. 50.21 dollars is an estimate of the average treatment effect on the treated (ATT), as it is using the control group as a counterfactual to compare to.

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.

```
# Run simple regression for impact of treatment on household's value of animal holdings, controlling fo
vani_reg2 <- lm(vani ~ treatment + hhsize97 + dirtfloor97 + electricity97 + bathroom97 + homeown97 + edu

# Display regression summary
```

```r
summary(vani_reg2)
```

```
##
## Call:
## lm(formula = vani ~ treatment + hhsize97 + dirtfloor97 + electricity97 +
##     bathroom97 + homeown97 + educ_sp + female_hh + age_hh + educ_hh,
##     data = progresa)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5352  -1607   -886     94  49172
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    -673.558    215.248  -3.129             0.001756 **
## treatment       145.011     63.255   2.292             0.021893 *
## hhsize97        162.169     12.686  12.784 < 0.0000000000000002 ***
## dirtfloor97   -1046.871     69.001 -15.172 < 0.0000000000000002 ***
## electricity97   349.587     68.504   5.103          0.000000339 ***
## bathroom97     -169.961     63.867  -2.661             0.007796 **
## homeown97       545.841    130.812   4.173          0.000030290 ***
## educ_sp          52.984     15.519   3.414             0.000642 ***
## female_hh      -640.608    112.314  -5.704          0.000000012 ***
## age_hh           32.039      2.483  12.905 < 0.0000000000000002 ***
## educ_hh         -32.236     14.949  -2.156             0.031064 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3604 on 13503 degrees of freedom
##   (862 observations deleted due to missingness)
## Multiple R-squared:  0.05282,    Adjusted R-squared:  0.05211
## F-statistic: 75.29 on 10 and 13503 DF,  p-value: < 0.00000000000000022
```

The impact of program participation increases nearly three-fold after we add these nine controls. Compared to an ATT of 50.21 dollars computed in the regression with no controls, we now compute an ATT of 145.01 dollars. Interpretation of 'dirtfloor97' coefficient: Having a dirt floor in 1997 is associated with $1,046.87 decrease in value of a household's animal holdings, controlling for program participation (and eight other unit characteristics).

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.

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.

```r
# Run simple regression for impact of "pseudo-treatment" on household's value of animal holdings
pseudo_treatment_reg <- lm(vani ~ pseudo_treatment, data = progresa_itt_df)

# Display regression summary
summary(pseudo_treatment_reg)
```

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

```
##
## 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
```

```
# Run simple regression for impact of treatment on household's value of animal holdings, controlling fo
pseudo_treatment_reg2 <- lm(vani ~ pseudo_treatment + hhsize97 + dirtfloor97 + electricity97 + bathroom9

# Display regression summary
summary(pseudo_treatment_reg2)
```

```
##
## Call:
## lm(formula = vani ~ pseudo_treatment + hhsize97 + dirtfloor97 +
##     electricity97 + bathroom97 + homeown97 + educ_sp + female_hh +
##     age_hh + educ_hh, data = progresa_itt_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5387  -1624   -891     93  49252
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       -541.168    205.241  -2.637              0.00838 **
## pseudo_treatment   227.120    167.901   1.353              0.17617
## hhsize97           164.748     12.418  13.267 < 0.0000000000000002 ***
## dirtfloor97      -1069.276     67.840 -15.762 < 0.0000000000000002 ***
## electricity97      336.410     66.814   5.035       0.000000483662 ***
## bathroom97        -189.776     62.667  -3.028              0.00246 **
## homeown97          570.159    127.001   4.489       0.000007197929 ***
## educ_sp             45.517     15.133   3.008              0.00264 **
## female_hh         -670.178    109.131  -6.141       0.000000000841 ***
## age_hh              31.574      2.403  13.140 < 0.0000000000000002 ***
## educ_hh            -31.391     14.557  -2.157              0.03106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3643 on 14365 degrees of freedom
## Multiple R-squared:  0.05311,    Adjusted R-squared:  0.05245
## F-statistic: 80.57 on 10 and 14365 DF,  p-value: < 0.00000000000000022
```

It is unclear whether the program had an effect on the value of animal holdings among non-participants that were eligible to be in the program. Neither of the two regressions we ran (one without controls and one with) provided us with a statistically significant result, though both did estimate that "pseudo-treatment"

had a positive effect. My inclination is that there was some minor spillovers, given that the p-value in our regression with controls was reasonably low at 0.176. It also makes sense logically to me that there would be some spillover effects because people who did get the cash transfers might share some of their cash with non-participants that live in their town, especially if the two individuals are family or friends.