EDS241: Assignment 1

Amanda Herbst

01/24/2024

Part 1

(NOTE: Uses the RCT.R code provided with lecture to generate data) DO NOT CHANGE ANYTHING BELOW UNTIL IT SAYS EXPLICITLY

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).

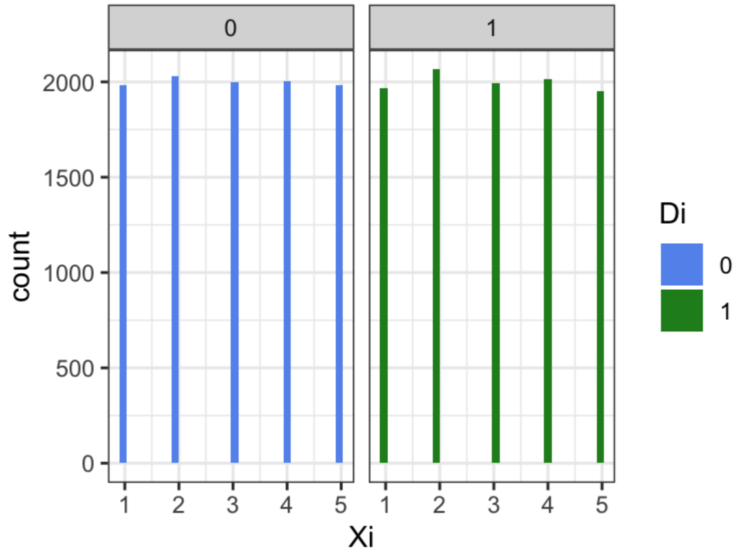
a. Create equally sized treatment and control groups by creating a binary random variable Di where the units with the *1's" are chosen randomly.

```
# by replicating the vector (0,1) to the length N, then random sampling without replacem
ent, it guarantees we will have equal 1s and 0s
Di <- sample(rep(c(0,1),length.out=N),N, replace = FALSE)

# add treatment variable to dataframe
df <- df %>%
  cbind(Di)

# # check for equal groups
# sum(dtest==1)
```

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?

The distributions of Xi for treatment and control groups are approximately the same. There are around 2000 observations of each characteristic for both treatment and control groups. This complies with my expectations because the treatment and control groups are of equal size and the characteristics were randomly sampled in a way that they would be approximately equal sizes

```
# create histogram of Xi for treatment and control groups
df %>%
  mutate(Di = as.factor(Di)) %>%
  ggplot(aes(x = Xi, fill = Di)) +
  scale_fill_manual(values = c("cornflowerblue", "forestgreen")) +
  geom_histogram() +
  # facet wrap by Di (0 = control, 1 = treatment)
  facet_wrap(~Di) +
  theme_bw()
```



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

We fail to reject the null hypothesis of the correlation test (p = 0.152) therefore Di is uncorrelated with the pre-treatment characteristics.

```
cor_Xi <- cor.test(df$Xi, df$Di)
sjPlot::tab_model(cor_Xi,
                  title = "Di Correlation with Xi",
                  dv.labels = "Xi")
```

Di Correlation with Xi

	Xi	
estimate	*CI*	*p*
-0.00	-0.02 – 0.01	0.813

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).

We fail to reject the null hypothesis for the correlation test between both Di and Yi_0 (p = 0.148) and Di and Yi_1 (p = 0.219). Therefore, Di is uncorrelated with the potential outcomes Yi_0 and Yi_1

Testing correlation between Di and Yi_0:

```
cor_Yi_0 <- cor.test(df$Yi_0, df$Di)
sjPlot::tab_model(cor_Yi_0,
                  title = "Di Correlation with Yi_0",
                  dv.labels = "")
```

Di Correlation with Yi_0

estimate	*CI*	*p*
-0.01	-0.02 – 0.00	0.129

Testing correlation between Di and Yi_1

```
cor_Yi_1 <- cor.test(df$Yi_1, df$Di)
sjPlot::tab_model(cor_Yi_1,
                  title = "Di Correlation with Yi_1",
                  dv.labels = "")
```

Di Correlation with Yi_1

estimate	*CI*	*p*
-0.00	-0.02 – 0.01	0.673

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

```
# create new "observed" Yi variable based on the assigned treatment Di
# when Di = 1, Yi_obs = Yi_1
# when Di = 0, Yi_obs = Yi_0
df <- df %>%
  mutate(Yi_obs = Di*Yi_1+(1-Di)*Yi_0)

# mean difference
mean_Yi_df <- df %>%
  group_by(Di) %>%
  summarize(mean_Yi = mean(Yi_obs))

ate_est <- mean_Yi_df$mean_Yi[2] - mean_Yi_df$mean_Yi[1]
ate_est
```

```
## [1] 1.480515
```

```
# test for mean difference = t.test
t.test(df$Yi_obs ~ df$Di)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Yi_obs by df$Di
## t = -69.878, df = 18068, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means between group 0 and group 1 is not e
qual to 0
## 95 percent confidence interval:
##  -1.522044 -1.438986
## sample estimates:
## mean in group 0 mean in group 1 
##        1.519072        2.999588
```

Estimated ATE by comparing mean outcomes is 1.48. We can reject the null hypothesis (p = 2.2e-16) so the difference in means is significant.

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 new "observed" Yi variable based on the Di
# when Di = 1, Yi_obs = Yi_1
# when Di = 0, Yi_obs = Yi_0
df <- df %>%
  mutate(Yi_obs = Di*Yi_1+(1-Di)*Yi_0)

# (i) Yi on Di
summary(lm(Yi_obs ~ Di, data = df))
```

```
##
## Call:
## lm(formula = Yi_obs ~ Di, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.9255 -1.0430 -0.0021  1.0364  5.2452 
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)    
## (Intercept)  1.51907    0.01498  101.40 <0.0000000000000002 ***
## Di           1.48052    0.02119   69.88 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.498 on 19998 degrees of freedom
## Multiple R-squared:  0.1963, Adjusted R-squared:  0.1962 
## F-statistic:  4883 on 1 and 19998 DF,  p-value: < 0.00000000000000022
```

```
# (ii) Yi on Di and Xi
summary(lm(Yi_obs ~ Di + Xi, data = df))
```

```
##
## Call:
## lm(formula = Yi_obs ~ Di + Xi, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4304 -0.7088 -0.0075  0.7069  4.4009
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -0.73006    0.01918  -38.06 <0.0000000000000002 ***
## Di           1.48404    0.01502   98.84 <0.0000000000000002 ***
## Xi           0.75049    0.00533  140.79 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.062 on 19997 degrees of freedom
## Multiple R-squared:  0.5964, Adjusted R-squared:  0.5963
## F-statistic: 1.477e+04 on 2 and 19997 DF,  p-value: < 0.00000000000000022
```

(i) The average treatment effect (ATE) is estimated to be 1.51

(ii) The (ATE) is estimated to be 1.50 holding the pre-treatment characteristics constant

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.

```
# variables collected in 1997
progresa97 <- progresa %>%
  select(dirtfloor97, bathroom97, electricity97, homeown97, hhsize97, treatment)

## For continuous variables you can use the t-test
#t.test()
# household size
t.test(progresa97$hhsize97 ~ progresa97$treatment)

## For binary variables you should use the proportions test
#prop.test()
## Dirt Floor (1 if dirt floor)
progresa97 %>%
  select(treatment, dirtfloor97) %>%
  table() %>%
  prop.test()

## Bathroom (1 if bathroom for exclusive hh use)
progresa97 %>%
  select(treatment, bathroom97) %>%
  table() %>%
  prop.test()

## Electricity (1 if electricity in dwelling)
progresa97 %>%
  select(treatment, electricity97) %>%
  table() %>%
  prop.test()

## Home ownership (1 if home in ownership)
progresa97 %>%
  select(treatment, homeown97) %>%
  table() %>%
  prop.test()
```

We can reject the null for all variables except for homeownership in 1997 (p = 0.165). Therefore there are systematic differences between the treatment and control groups in dirt floor, bathroom, electricity, and household size, but not in homeownership. Systematic differences matter because they affect our ability to create counterfactuals. It would be a mistake to do the same test after treatment began because we're looking at how circumstances impacted whether or not a household received treatment. If we looked at it after they received treatment (which impacts those circumstances), we would not have an accurate 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?

Intercept: The groups that were not treated owned an average of $1,691.47 in animals in 1997 price. coefficient: The groups that were treated owned an average of $50.21 more in animals in 1997 price than those not treated This is an estimate of ATE because there is a single treatment effect, however it may be biased because we cannot definitely say that the treatment choice is uncorrelated with the error term.

```
model1 <- lm(vani ~ treatment, data = progresa)
summary(model1)
```

```
##
## 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
```

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.

The estimated ATE of the program participation increases from approximately $50 to approximately $56 (1997 price). The average effect of having electricity in 1997 in the control group, holding years of spouse's education, if a primary school is in the community, if a health center is nearby, and if the household is using any piece of land constant, is about $600 increase in animal value (1997 price).

```
model2 <- lm(vani ~ treatment + educ_sp + primary + age_hh + healthcenter + land + elect
ricity97, data = progresa)
summary(model2)
```

```
##
## Call:
## lm(formula = vani ~ treatment + educ_sp + primary + age_hh +
##     healthcenter + land + electricity97, data = progresa)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4266  -1718   -792    129  48117
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   -345.348    231.148  -1.494              0.1352
## treatment       56.316     63.338   0.889              0.3739
## educ_sp         35.215     14.024   2.511              0.0120 *
## primary        442.072    180.426   2.450              0.0143 *
## age_hh          24.315      2.229  10.908 <0.0000000000000002 ***
## healthcenter  -970.197     82.281 -11.791 <0.0000000000000002 ***
## land          1394.163     62.958  22.144 <0.0000000000000002 ***
## electricity97  600.312     66.306   9.054 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3589 on 13506 degrees of freedom
##   (862 observations deleted due to missingness)
## Multiple R-squared:  0.06035,    Adjusted R-squared:  0.05987
## F-statistic: 123.9 on 7 and 13506 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.

There doesn't appear to be any spillover effects. The ATE of the program and the ATE of the non-participants (pseudo-treatment) are not significant (p = 0.435 and 0.798 respectively). Additionally, there is no significant difference between the ATEs. We could be observing no spillover effects because families that did receive the cash transfer are investing it into their own household and livestock and not into the surrounding community.

```r
# replace any NAs in treatment column with 0
progresa_itt_df <- progresa %>%
  mutate(treatment = replace_na(treatment, 0))

# Examine number of hh that were intended to get treatment and that ended up receiving t
reatment
table(treatment = progresa_itt_df$treatment, intention_to_treat = progresa_itt_df$intent
ion_to_treat, exclude = NULL)

# Create a new treatment variable that is:
# = 1 if intention_to_treat == 1 AND not in the actual treatment
# = 0 for normal control hh.
progresa_df <- progresa_itt_df %>%
  select(intention_to_treat, treatment, vani) %>%
  mutate(pseudo_treatment = case_when(
    intention_to_treat == 1 & treatment == 0 ~ 1,
    treatment == 0 ~ 0,
    treatment == 1 ~ NA
  ))

model3 <- lm(vani ~ pseudo_treatment, data = progresa_df)
summary(model3)
```