EDS241: Assignment 1 Template

Charlie Curtin

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.

```
## create equally sized treatment and control groups by creating a binary random variable where the uni
# use sample to randomly assign 0s and 1s
df$Di <- sample(c(0,1), nrow(df), replace = TRUE)

# sampling with replacement should get us nearly equal sized groups
sum(df$Di == 1)
```

```
## [1] 10025
```

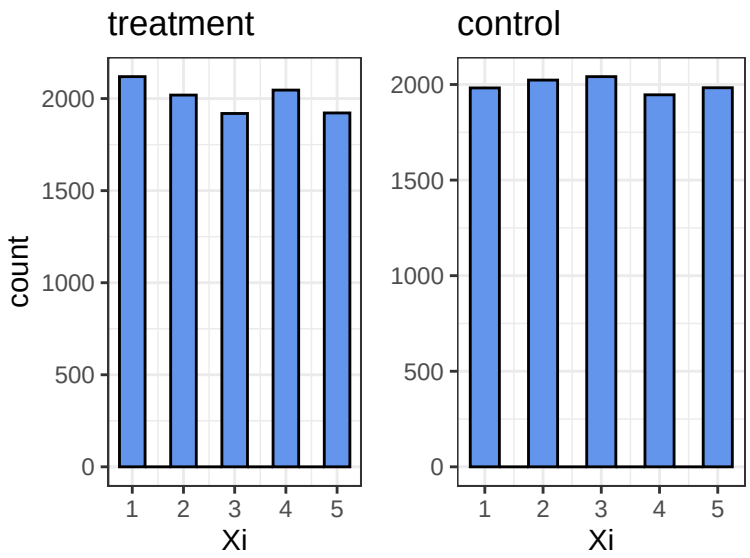
b) Make two separate histograms of Xi for the treatment and control group. What do you see and does it comply with your expectations, explain why or why not?

```
library(gridExtra)
## make two separate histograms for the treatment and control group
# treatment group histogram
treatment <- df %>%
  filter(Di == 1) %>%
  ggplot(aes(x = Xi)) +
  geom_histogram(binwidth = .5,
                 color = "black",
                 fill = "cornflowerblue") +
  labs(title = "treatment") +
  theme_bw()

# control group histogram
control <- df %>%
```

```
  filter(Di == 0) %>%
  ggplot(aes(x = Xi)) +
  geom_histogram(binwidth = .5,
                 color = "black",
                 fill = "cornflowerblue") +
  labs(title = "control") +
  theme_bw() +
  theme(axis.title.y = element_blank())

# arrange plots side-by-side
grid.arrange(treatment, control, ncol = 2)
```



- We can see roughly similarly-distributed counts of observations of different Xi values between our treatment and control groups. This makes sense because we randomly assigned units to be treated or untreated with simple random sampling

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

```
# find the correlation between our treatment assignment binary and our independent variable, Xi
print(cor(df$Di, df$Xi))
```

```
## [1] -0.01027853
```

- Our treatment assignment is uncorrelated to our pre-treatment characteristic, with an R value of -.005.

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).

```
## find the correlation between our treatment assignment binary and our potential outcomes
# treatment assignment and Yi_0 (outcome if not treated)
print(cor(df$Di, df$Yi_0))
```

```
## [1] -0.01170284
```

```
# treatment assignment and Yi_1 (outcome if treated)
print(cor(df$Di, df$Yi_1))
```

```
## [1] -0.01032088
```

- Our treatment assignment is not correlated to either of our potential outcomes.

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

```
# estimate the ATE by finding the difference in means for treatment and control groups
ATE <- mean(df$Yi_1) - mean(df$Yi_0)

print(ATE)
```

```
## [1] 1.47724
```

```
# compute a t-test for mean difference between the groups
print(t.test(df$Yi_1, df$Yi_0))
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Yi_1 and df$Yi_0
## t = 98.316, df = 36027, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.447789 1.506690
## sample estimates:
## mean of x mean of y
##  2.985408  1.508168
```

- Based on our small p-value, we can reject the null hypothesis and accept the alternative hypothesis that our true difference in means is not equal to 0. We are 95% confident that the interval [1.47, 1.53] contains the true difference in mean outcomes between the treatment and control groups.

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.

```
# append a new column "Yi" to our dataframe, which is assigned the value Yi_1 if the unit is treated (D
df <- df %>%
  mutate(Yi = ifelse(Di == 1,
                     Yi_1,
                     Yi_0))

# simple linear regression of Yi on Di
lm_a <- lm(Yi ~ Di, data = df)
```

```r
# save output to a summary table
mod_lm_a <- summary(lm_a)

# multiple linear regression of Yi on Di and Xi
lm_b <- lm(Yi ~ Di + Xi, data = df)

# save output to a summary table
mod_lm_b <- summary(lm_b)

# print model outputs
print(mod_lm_a)
```

```
##
## Call:
## lm(formula = Yi ~ Di, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0719 -1.0543 -0.0083  1.0357  5.0924
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  1.52258    0.01505  101.16 <0.0000000000000002 ***
## Di           1.44498    0.02126   67.97 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.503 on 19998 degrees of freedom
## Multiple R-squared:  0.1877, Adjusted R-squared:  0.1876
## F-statistic:  4620 on 1 and 19998 DF,  p-value: < 0.00000000000000022
```

```r
print(mod_lm_b)
```

```
##
## Call:
## lm(formula = Yi ~ Di + Xi, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1352 -0.7168 -0.0062  0.7127  4.6616
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -0.741930   0.018989  -39.07 <0.0000000000000002 ***
## Di           1.466993   0.014923   98.31 <0.0000000000000002 ***
## Xi           0.756732   0.005273  143.52 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 19997 degrees of freedom
## Multiple R-squared:  0.5998, Adjusted R-squared:  0.5998
## F-statistic: 1.499e+04 on 2 and 19997 DF,  p-value: < 0.00000000000000022
```

- Our simple linear regression tells us that the ATE is 1.53, signified by the B1 coefficient. Our multiple linear regression tells us that that the ATE is 1.52 through the B1 coefficient, but that the pre-treatment characteristic Xi influences the ATE. For each 1-unit increase in Xi, the ATE for that group is expected to increase by .75.

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.

- In a proportions test, our null hypothesis is that the proportions between our treatment and control groups is equal. With our small p-value, we can reject the null hypothesis and accept the alternative hypothesis that there is a difference in proportions between our treatment and control groups. We are 95% confident that the interval [.02, .05] contains the true difference in proportions.

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?

```
# simple regression on treatment and animal holdings
lm_vani <- lm(treatment ~ vani, data = progresa_itt_df)

# print model results
summary(lm_vani)
```

```
##
## Call:
## lm(formula = treatment ~ vani, data = progresa_itt_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5548 -0.5329  0.4640  0.4670  0.4672
##
## Coefficients:
##                 Estimate   Std. Error t value            Pr(>|t|)
## (Intercept) 0.5328041565 0.0045838906 116.234 <0.0000000000000002 ***
## vani        0.0000004588 0.0000011118   0.413                0.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4989 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
```

- The intercept means that we should expect to see households with animal value holdings have a probability of treatment of .5. The B1 coefficient means that for treated households, we expect to see the log-odds of treatment change by a miniscule amount per 1 unit increase in the value of animal holdings. The high p-value means that there is not a statistically significant relationship between treatment and animal holdings. This is not an estimate of a treatment effect because we don't have any pre-treatment animal value holdings for each household 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.

```
# multivariate regression on treatment and 6 control variables
lm_multi <- lm(treatment ~ female_hh + ethnicity_hh + ani + educ_hh + ha + foodexp, data = progresa_itt

# print model output
summary(lm_multi)
```

```
##
## Call:
## lm(formula = treatment ~ female_hh + ethnicity_hh + ani + educ_hh +
##     ha + foodexp, data = progresa_itt_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7299 -0.5211  0.3911  0.4686  0.5956
##
## Coefficients:
##                 Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)   0.44011464  0.01165629  37.758 < 0.0000000000000002 ***
## female_hh    -0.03549721  0.01429508  -2.483               0.0130 *
## ethnicity_hh  0.01993276  0.00849708   2.346               0.0190 *
## ani           0.04120556  0.00960935   4.288            0.0000181 ***
## educ_hh       0.00110214  0.00160225   0.688               0.4915
## ha           -0.00433413  0.00208943  -2.074               0.0381 *
## foodexp       0.00012885  0.00001138  11.326 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4961 on 14313 degrees of freedom
##   (56 observations deleted due to missingness)
## Multiple R-squared:  0.0116, Adjusted R-squared:  0.01119
## F-statistic:    28 on 6 and 14313 DF,  p-value: < 0.00000000000000022
```

- For "ha" (total hectares of land), the coefficient means that a 1 unit increase in total hectares of land decreases the log-odds of treatment by .004, holding every other independent variable constant.

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.

- Our high p-value indicates that we don't have enough evidence to reject the null hypothesis that there is no difference in the means between our treatment and control groups. This indicates that we don't have a spillover effect. There might not be a spillover effect because those that received the cash transfer might be using it for more pressing needs like food, not to increase the value of their animal holdings.