Problem Set #4

w241 Experiment Design

Adam Yang

```r
# load packages
library(foreign)
library(data.table)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

1. Potential Outcomes

a. Make up a hypothetical schedule of potential outcomes for three Compliers and three Never-Takers where the ATE is positive but the CACE is negative. By ATE, we mean the average treatment effect for the entire population, including both compilers and never-takers. Note that we can never compute this ATE directly in practice, because we never observe both potential outcomes for any individual, especially for never-takers. That's why this question requires you to provide a complete table of hypothetical potential outcomes for all six subjects.

Answer:

di_z0 $= d_i(z = 0):$ Subject i is assigned to control, did it receive treatment?

di_z1 $= d_i(z = 1):$ Subject i is assigned to treatment, did it receive treatment?

Yi_d0 $= Y_i(d = 0):$ Potential outcome of subject i if it is not treated.

Yi_d1 $= Y_i(d = 1):$ Potential outcome of subject i if it is treated.

- If $d_i(z = 0) = 0$, it means the subject is assigned to the control group and did not receive treatment.

- If $d_i(z = 1) = 1$, it means the subject is assigned to the treatment group and received treatment.
- If $d_i(z = 1) = 0$, it means the subject is assigned to the treatment group but did not receive treatment.

Subject	Type	di_z0	di_z1	Yi_d0	Yi_d1
1	Complier	0	1	6	5
2	Complier	0	1	7	5
3	Complier	0	1	5	4
4	Never-Taker	0	0	4	7
5	Never-Taker	0	0	4	7
6	Never-Taker	0	0	3	6

```r
# Create the data table
DT1 <- data.table(
  Type <- c("Complier","Complier","Complier","Never-Taker","Never-Taker","Never-Taker"),
  di_z0 <- c(rep(0,6)),
  di_z1 <- c(rep(1,3),rep(0,3)),
  Yi_d0 <- c(6,7,5,4,4,3),
  Yi_d1 <- c(5,5,4,7,7,6)
)

# Calculate ATE = mean(Yi_d1 - Yi_d0) > 0
paste("The ATE is", mean(Yi_d1 - Yi_d0))
```

```
## [1] "The ATE is 0.833333333333333"
```

```r
# Calculate CACE = sum[(Yi_d1 - Yi_d)*di_z1]/sum(di_z1) < 0 find ATE of only the compliers.
paste("The CACE is", sum((Yi_d1 - Yi_d0)*di_z1)/sum(di_z1))
```

```
## [1] "The CACE is -1.33333333333333"
```

b. Suppose that an experiment were conducted on your pool of subjects. In what ways would the estimated CACE be informative or misleading?

Answer: When an experiment is conducted and we wish to calculate the CACE of our findings, we would have to compare the compliers in the treatment group to the "would-be" compliers in the control group. First off, it may be difficult to figure out who the compliers are in the treatment group. Furthermore, it is even more difficult to figure out which of the people in the control group would have been compliers if they were assigned to the treatment group instead. All of these uncertainties can lead to some misleading results. One way to calculate the CACE is by finding ITT/ITT_D where ITT is the intent to treat effect and ITT_D is the compliance rate. However, the way that the compliance rate is calculated, we must make the assumption that only compliers contribute the the treatment effect while non-compliers in both groups will not have any treatment effect. However this assumption is not always true. In the example of testing blood pressure medicine, because just being placed into the treatment group, but refusing to take a pill might raise the blood pressure of the patient because they feel like their problem is not being treated. In this case, if we assume that the never-takers have no treatment effect, our results will be misleading. There are certain experiments where you can keep track of which subjects are the compliers in both the control and treatment groups. For example, if you have some kind of electric tracker of every time the subjects in both the placebo and treatment groups open up and take the pill, we can keep track of the level of compliance from each group. Even with this type of method, you might have failed compliance if the person drops decides to take two pills instead of one, or offers the pill to someone else. One other way that an estimated CACE can be misleading is if the audience of your experiment does not understand what CACE means and mistakes it for the ATE. It should be made clear that the CACE only tells you the treatment effect of those who are willing to comply to the experiment and cannot represent anyone who is unwilling to comply. However, if you managed to perform a perfect experiment and get a great estimate of the CACE, it can be very informative in telling you the treatment effect of your experiment if the treatment was successfully applied.

In terms of our specific data set of only 6 subjects, I think the sample is too small for us to get an accurate ITT_D. Because of this, the CACE that we will end up getting would not be very accurate so the results can be very misleading.

c. Which population is more relevant to study for future decision making: the set of Compliers, or the set of Compliers plus Never-Takers? Why?

Answer: I think it depends on what kind of future decision making we are talking about. The set of compilers would let us know what the treatment effect is when the treatment is successfully applied. This would be preferred in situations like developing medication where we really care about the actual effects of taking the drug versus not taking the drug. We don't care about the never-takers because they do not tell us the actual effects of the drug itself. In other cases, we may care more about the intent to treat effect such as an experiment to see the effectiveness of mailing reminders to get people to vote. In that case, whether or not the subject in the treatment group complied or not is not important. In fact, it tells us more about the effectiveness of sending those mailed reminders because if the subject can throw the mail away without opening it, or if it gets lost without reaching the subject, then it might not be a very effective way to get people to vote.

2. Turnout to Vote

Suppose that a researcher hires a group of canvassers to contact a set of 1,000 voters randomly assigned to a treatment group. When the canvassing effort concludes, the canvassers report that they successfully contacted 500 voters in the treatment group, but the truth is that they only contacted 250. When voter turnout rates are tabulated for the treatment and control groups, it turns out that 400 of the 1,000 subjects in the treatment group voted, as compared to 700 of the 2,000 subjects in the control group (none of whom were contacted).

a. If you believed that 500 subjects were actually contacted, what would your estimate of the CACE be?

```r
# CACE = ITT/ITTD
ITTD <- 0.5
ITT <- 400/1000 - 700/2000
CACE <- ITT/ITTD
paste("The CACE would be", CACE, "with 500 compliers")
```

```
## [1] "The CACE would be 0.1 with 500 compliers"
```

b. Suppose you learned that only 250 subjects were actually treated. What would your estimate of the CACE be?

```r
ITTD <- 0.25
ITT <- 400/1000 - 700/2000
CACE <- ITT/ITTD
paste("The CACE would be", CACE, "with 250 compiers")
```

```
## [1] "The CACE would be 0.2 with 250 compiers"
```

c. Do the canvassers' exaggerated reports make their efforts seem more or less effective? Define effectiveness either in terms of the ITT or CACE. Why does the definition matter?

Answer: In terms of CACE, the canvasser's exaggerated reports make their efforts seem less effective. They overestimated the number of compliers which leads to an underestimation of the CACE because we are treating some of the never-takers as compliers. In terms of ITT however, there is no difference because the ITT does not take into consideration how many subjects in the treatment group are compliers, as long as we intended to treat them all. The definition matters because ITT and CACE are very different where in CACE we are only estimating the treatment effect of the compliers while in ITT we are estimating the treatment effect on everyone we attempted to treat regardless if they complied. In the case of this experiment, I think the ITT would be a more useful estimate because it is more applicable to the real world scenario of canvassing where you will not be able to reach everyone you intend to reach.

3. Turnout in Dorms

Guan and Green report the results of a canvassing experiment conduced in Beijing on the eve of a local election. Students on the campus of Peking University were randomly assigned to treatment or control groups. Canvassers attempted to contact students in their dorm rooms and encourage them to vote. No contact with the control group was attempted. Of the 2,688 students assigned to the treatment group, 2,380 were contacted. A total of 2,152 students in the treatment group voted; of the 1,334 students assigned to the control group, 892 voted. One aspect of this experiment threatens to violate the exclusion restriction. At every dorm room they visited, even those where no one answered, canvassers left a leaflet encouraging students to vote.

```
library(foreign)
d <- read.dta("./data/Guan_Green_CPS_2006.dta")
head(d)
```

```
##   turnout contact  dormid treat2
## 1       0       0 1010101      0
## 2       0       0 1010101      0
## 3       0       0 1010101      0
## 4       0       0 1010102      0
## 5       0       0 1010102      0
## 6       0       1 1010103      1
```

a. Using the data set from the book's website, estimate the ITT. First, estimate the ITT using the difference in two-group means. Then, estimate the ITT using a linear regression on the appropriate subset of data. *Heads up: There are two NAs in the data frame. Just na.omit to remove these rows.*

```
# Use na.omit to remove the 2 NAs in the data frame.
d <- na.omit(d)

# Estimate the ITT using difference in two-group means
treatment <- mean(d[d$treat2 == 1,]$turnout)
control <- mean(d[d$treat2 == 0,]$turnout)
ITT <- treatment - control
paste("The ITT from difference in two-group means is", ITT)
```

```
## [1] "The ITT from difference in two-group means is 0.131929570928821"
```

```
# Estimate the ITT using a linear regression
model <- lm(turnout~treat2, data = d)
summary(model)
```

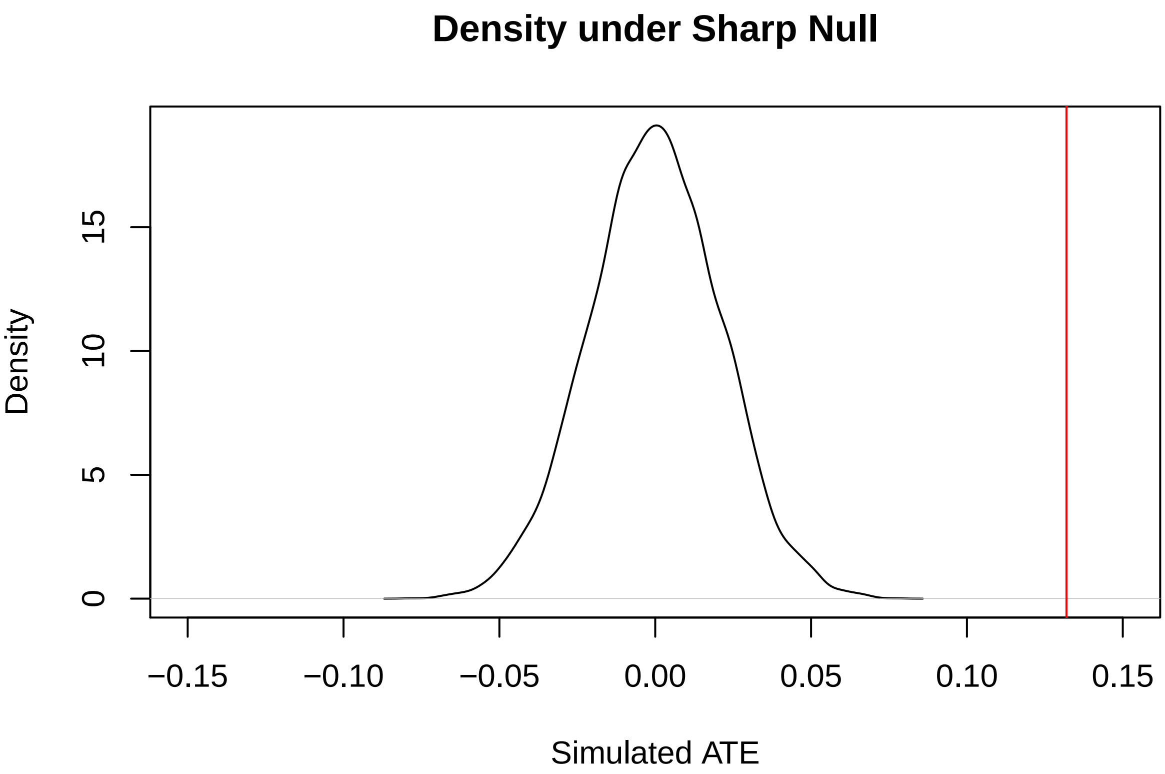
```
## 
## Call:
## lm(formula = turnout ~ treat2, data = d)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.8006  0.1994  0.1994  0.1994  0.3313 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.66867    0.01162  57.521   <2e-16 ***
## treat2       0.13193    0.01422   9.278   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 4020 degrees of freedom
## Multiple R-squared:  0.02096,    Adjusted R-squared:  0.02072
## F-statistic: 86.08 on 1 and 4020 DF,  p-value: < 2.2e-16
```

```r
paste("The ITT from a linear regression is", model$coefficients[2])
```

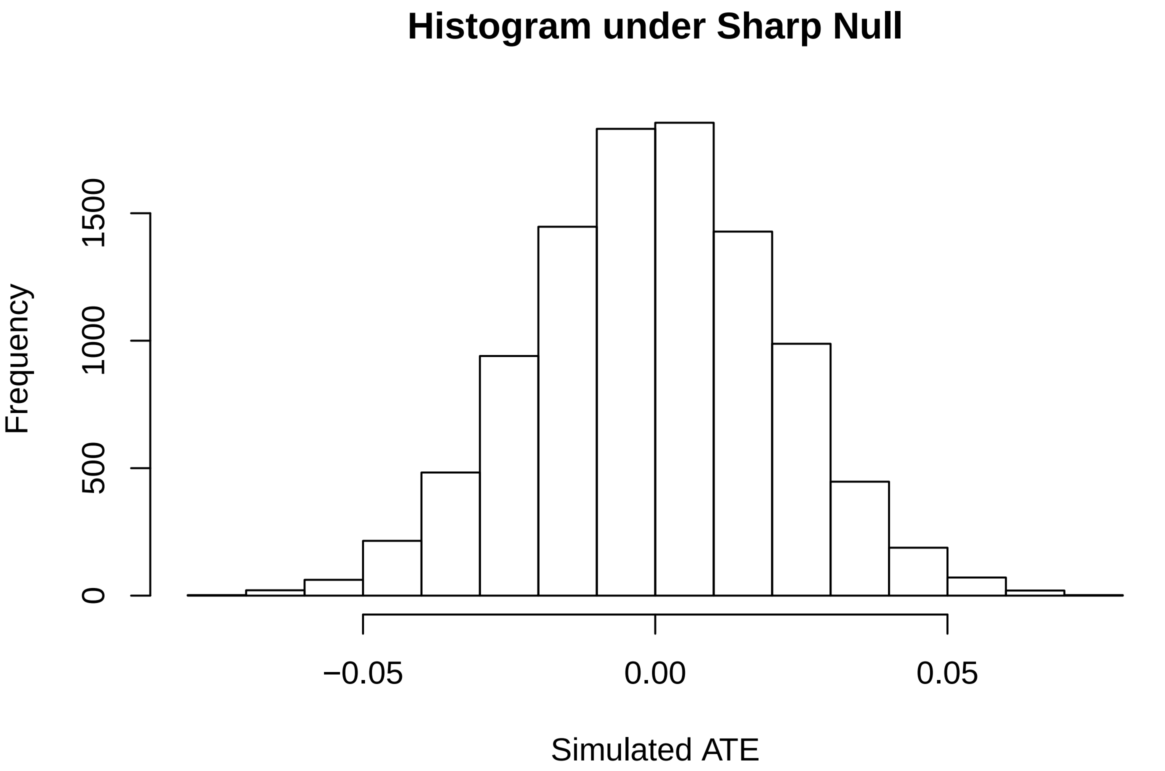
```
## [1] "The ITT from a linear regression is 0.13192957092882"
```

b. Use randomization inference to test the sharp null hypothesis that the ITT is zero for all observations, taking into account the fact that random assignment was clustered by dorm room. Interpret your results.

```r
# First process the clustered data
d <- na.omit(d)
num_clusters <- length(unique(d$dormid))
d.clustered <- aggregate(d, by = list(d$dormid), FUN = mean)

# Random Inference
simulation <- function(data) {
  # First we randomly create assignments for treatment and control group.
  # The number of 0's and 1's are not always equal but the probability of getting a 0
  # is equal to the probability of getting a 1.
  treatment <- sample(c(0,1), size = num_clusters, replace = TRUE)

  # Next we put the views data into their corresponding groups that we've randomly assigned
  treat.group <- data$turnout[treatment == 1]
  cont.group <- data$turnout[treatment == 0]

  # Now we can calculate the estimated ATE for this specific random assignment of
  # control/treatment groups.
  ate <- mean(treat.group) - mean(cont.group)

  return(ate)
}

# Now we run the simulation 10,000 times
distribution.under.sharp.null <- replicate(10000, simulation(d.clustered))

# Plot the density distribution and histogram of our simulations
plot(density(distribution.under.sharp.null),
     main = "Density under Sharp Null", xlab = "Simulated ATE", xlim = c(-0.15, 0.15))
abline(v = ITT, col = "red")
```



```r
hist(distribution.under.sharp.null,
     main = "Histogram under Sharp Null", xlab = "Simulated ATE")
abline(v = ITT, col = "red")
```



Answer: The bell curve above shows the results of our randomization inference test and the red line indicates the ITT that we calculated earlier. In this case, the p-value is essentially zero so we can reject the sharp null hypothesis that the ITT is zero for all observations.

c. Assume that the leaflet had no effect on turnout. Estimate the CACE. Do this in two ways: First, estimate the CACE using means. Second, use some form of linear model to estimate this as well. If you use a 2SLS, then report the standard errors and draw inference about whether the leaflet had any causal effect among compilers.

```
# Estimate CACE using means
treatment <- mean(d[d$treat2 == 1,]$turnout)
control <- mean(d[d$treat2 == 0,]$turnout)
ITT <- treatment - control
ITTD <- length(d[d$contact == 1,]$turnout)/length(d[d$treat2 == 1,]$turnout)
CACE <- ITT/ITTD
paste("The CACE calculated using means is", CACE)
```

```
## [1] "The CACE calculated using means is 0.14894022959121"
```

```
# Use some form of linear model to estimate CACE
# What is the effect?
model1 <- lm(turnout ~ treat2, data = d)
coeftest(model1, vcovHC(model1))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.668666   0.012897 51.8470 < 2.2e-16 ***
## treat2      0.131930   0.015025  8.7804 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# What is the compliance rate?
model2 <- lm(contact ~ treat2, data = d)
summary(model)
```

```
##
## Call:
## lm(formula = turnout ~ treat2, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8006  0.1994  0.1994  0.1994  0.3313
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.66867    0.01162  57.521   <2e-16 ***
## treat2       0.13193    0.01422   9.278   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 4020 degrees of freedom
## Multiple R-squared:  0.02096,    Adjusted R-squared:  0.02072
## F-statistic: 86.08 on 1 and 4020 DF,  p-value: < 2.2e-16
```

```
# What is the CACE?
paste("The CACE calculated using regression is", model1$coefficients[2]/model2$coefficients[2])
```

```
## [1] "The CACE calculated using regression is 0.148940229591211"
```

Did not use 2SLC for this question.

4. Why run a placebo?

Nickerson describes a voter mobilization experiment in which subjects were randomly assigned to one of three conditions: a baseline group (no contact was attempted); a treatment group (canvassers attempted to deliver an encouragement to vote); and a placebo group (canvassers attempted to deliver an encouragement to recycle). Based on the results in the table below answer the following questions

Treatment Assignment	Treated ?	N	Turnout
Baseline	No	2572	31.22%
Treatment	Yes	486	39.09%
Treatment	No	2086	32.74%
Placebo	Yes	470	29.79%
Placebo	No	2109	32.15%

First Use the information to make a table that has a full recovery of this data. That is, make a `data.frame` or a `data.table` that will have as many rows a there are observations in this data, and that would fully reproduce the table above. (*Yes, this might seem a little trivial, but this is the sort of "data thinking" that we think is important.*)

```
# Make a data.table to represent the data above
turnout1 <- round(2572*0.3122)
turnout2 <- round(486*0.3909)
turnout3 <- round(2086*0.3274)
turnout4 <- round(470*0.2979)
turnout5 <- round(2109*0.3215)
DT <- data.table(
  groups = c(rep("Baseline",2572), rep("Treatment",486+2086), rep("Placebo",470+2109)),
  baseline = c(rep(1,2572),rep(0,486+2086+470+2109)),
  treatment = c(rep(0,2572),rep(1,486+2086),rep(0,470+2109)),
  placebo = c(rep(0,2572+486+2086),rep(1,470+2109)),
  treated = c(rep(0,2572),rep(1,486),rep(0,2086),rep(1,470),rep(0,2109)),
  turnout = c(rep(1,turnout1),rep(0,2572-turnout1),rep(1,turnout2),rep(0,486-turnout2),
              rep(1,turnout3),rep(0,2086-turnout3),rep(1,turnout4),rep(0,470-turnout4),
              rep(1,turnout5),rep(0,2109-turnout5))
)
```

a. Estimate the proportion of Compliers by using the data on the Treatment group. Then compute a second estimate of the proportion of Compliers by using the data on the Placebo group. Are these sample proportions statistically significantly different from each other? Explain why you would not expect them to be different, given the experimental design. (Hint: ITT_D means "the average effect of the treatment on the dosage of the treatment." I.E., it's the contact rate α in the async).

```
# Estimate the proportion of compliers by using the data on the Treatment group
treatment_compliers <- length(DT[treatment == 1 & treated == 1]$turnout)
total_treatment <- length(DT[treatment == 1]$turnout)
ITTD.treatment <- treatment_compliers/total_treatment
paste("The proportion of compliers in the treatment group is",ITTD.treatment)
```

```
## [1] "The proportion of compliers in the treatment group is 0.18895800933126"
```

```r
# Estimate the proportion of compliers by using the data on the Placebo group
placebo_compliers <- length(DT[placebo == 1 & treated == 1]$turnout)
total_placebo <- length(DT[placebo == 1]$turnout)
ITTD.placebo <- placebo_compliers/total_placebo
paste("The proportion of compliers in the placebo group is",ITTD.placebo)
```

```
## [1] "The proportion of compliers in the placebo group is 0.182241178751454"
```

```r
# Are these sample proportions statistically significantly different from each other?
t.test(DT[treatment==1]$treated,DT[placebo==1]$treated)
```

```
##
##  Welch Two Sample t-test
##
## data:  DT[treatment == 1]$treated and DT[placebo == 1]$treated
## t = 0.61987, df = 5147.6, p-value = 0.5354
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01452611  0.02795977
## sample estimates:
## mean of x mean of y
## 0.1889580 0.1822412
```

Answer: As shown above, the proportion of compliers in the treatment group is *0.189* and the proportion of compliers in the placebo group is *0.182*, which makes a difference of *0.007*. I did a two sample t-test and it shows that the difference between the two proportions is not statistically significant. I did not expect them to be statistically significantly different because we are measuring the contact rate of each of these two groups. According to the information given, the placebo group and the treatment group are treated the same way and the only difference between the two groups is the message being delivered (encouragement to vote versus encouragement to recycle). Therefore, I do not see any obvious factor that would cause a different rate of compliance between the treatment and placebo groups.

b. Do the data suggest that Never Takers in the treatment and placebo groups have the same rate of turnout? Is this comparison informative?

```r
# Find the rate of turnout for the never-takers in the treatment and placebo groups
treatment_RoT <- mean(DT[treatment == 1 & treated == 0]$turnout)
placebo_RoT <- mean(DT[placebo == 1 & treated == 0]$turnout)

paste("The rate of turnout for the never-takers in the treatment group is", treatment_RoT)
```

```
## [1] "The rate of turnout for the never-takers in the treatment group is 0.327420901246405"
```

```r
paste("The rate of turnout for the never-takers in the placebo group is", placebo_RoT)
```

```
## [1] "The rate of turnout for the never-takers in the placebo group is 0.321479374110953"
```

```r
t.test(DT[treatment == 1 & treated == 0]$turnout,DT[placebo == 1 & treated == 0]$turnout)
```

```
##
##  Welch Two Sample t-test
##
## data:  DT[treatment == 1 & treated == 0]$turnout and DT[placebo == 1 & treated == 0]$turnout
## t = 0.41089, df = 4192, p-value = 0.6812
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02240809  0.03429115
## sample estimates:
```

```
## mean of x mean of y
## 0.3274209 0.3214794
```

Answer: As shown above, the rate of turnout for the never-takers in the treatment group is *0.327* and the rate of turnout for the never-takers in the placebo group is *0.321*, which is a difference of *0.006*. Doing a two sample t-test shows that the difference between these two numbers is not statistically significant. This comparison is informative because one of the important assumptions when calculating the CACE is that the treatment effect is driven by the compliers. In other words, there is no treatment effect on the never-takers, which is true according to what we found above.

c. Estimate the CACE of receiving the placebo. Is this estimate consistent with the substantive assumption that the placebo has no effect on turnout?

```
# I am assuming that the question is asking for the CACE of receiving the placebo in comparison to the
placebo_compliers <- length(DT[placebo == 1 & treated == 1]$turnout)
total_placebo <- length(DT[placebo == 1]$turnout)
ITTD.placebo <- placebo_compliers/total_placebo
ITT.placebo <- mean(DT[placebo == 1]$turnout) - mean(DT[baseline == 1]$turnout)
CACE.placebo <- ITT.placebo/ITTD.placebo
paste("The CACE of receiving the placebo is", CACE.placebo)
```

```
## [1] "The CACE of receiving the placebo is 0.0272649813043909"
```

Answer: The CACE of receiving the placebo is *0.027*. I think the estimate is consistent with the substantive assumption that the placebo has no effect on the turnout because of how small 0.027 is. However, since the number is not closer to zero, the placebo might have a tiny bit of an effect on the turnout.

d. Estimate the CACE of receiving the treatment using two different methods. First, use the conventional method of dividing the ITT by the ITT_{D}. (This should be a treatment vs. control comparison.)

```
# Find the CACE of receiving the treatment using ITT/ITTD
# I am assuming that the question is asking us to compare the treatment with the baseline.
treatment_compliers <- length(DT[treatment == 1 & treated == 1]$turnout)
total_treatment <- length(DT[treatment == 1]$turnout)
ITTD.treatment <- treatment_compliers/total_treatment
ITT.treatment <- mean(DT[treatment == 1]$turnout) - mean(DT[baseline == 1]$turnout)
CACE.treatment <- ITT.treatment/ITTD.treatment
paste("The CACE of receiving the treatment is", CACE.treatment)
```

```
## [1] "The CACE of receiving the treatment is 0.144032921810699"
```

Answer: The CACE of receiving the treatment is *0.144* using the ITT/ITT_D method.

e. Then, second, compare the turnout rates among the Compliers in both the treatment and placebo groups. Interpret the results.

```
# Compare the turnout rates among the compliers in both the treatment and placebo groups.
CACE <- mean(DT[treatment==1 & treated==1]$turnout) - mean(DT[placebo==1 & treated==1]$turnout)
paste("The CACE found by comparing the compliers in the treatment and placebo groups is", CACE)
```

```
## [1] "The CACE found by comparing the compliers in the treatment and placebo groups is 0.0930741616632
```

Answer: The CACE found by comparing the compliers in the treatment and placebo groups is *0.093*. It is rather different compared to the CACE that we found using the ITT/ITT_D method. That is because using the ITT/ITT_D method, we are estimating the CACE with the treatment compliance rate under the assumption that there is no treatment effect on the never-takers. In the placebo experiment, we are running the experiment in the way where we know how many of the placebo group are compliers when we figure out the CACE. The difference between these two methods may be caused when the treatment effect on the never-takers is not exactly zero. Or it can be caused if the placebo has some effect on the turnout afterall.

f. Based on what we talked about in class – that the rate of compliance determines whether one or another design is more efficient – given the compliance rate in this study, which design *should* provide a more efficient estimate of the treatment effect? If you want to review the specific paper that makes this claim, check out this link. Does it?

Answer: The rate of compliance for this study is around 18 to 19%. According to Gerber and Green, when the compliance is below 50%, the placebo design is the prefered method of calculating the CACE. This can be shown in Figure 1 in the reading provided. The figure shows that when the contact rate is small, around 0.2, the Placebo-Treatment provides a much smaller standard error, which means the estimate of the treatment effect is more efficient.

5. Tetris FTW?

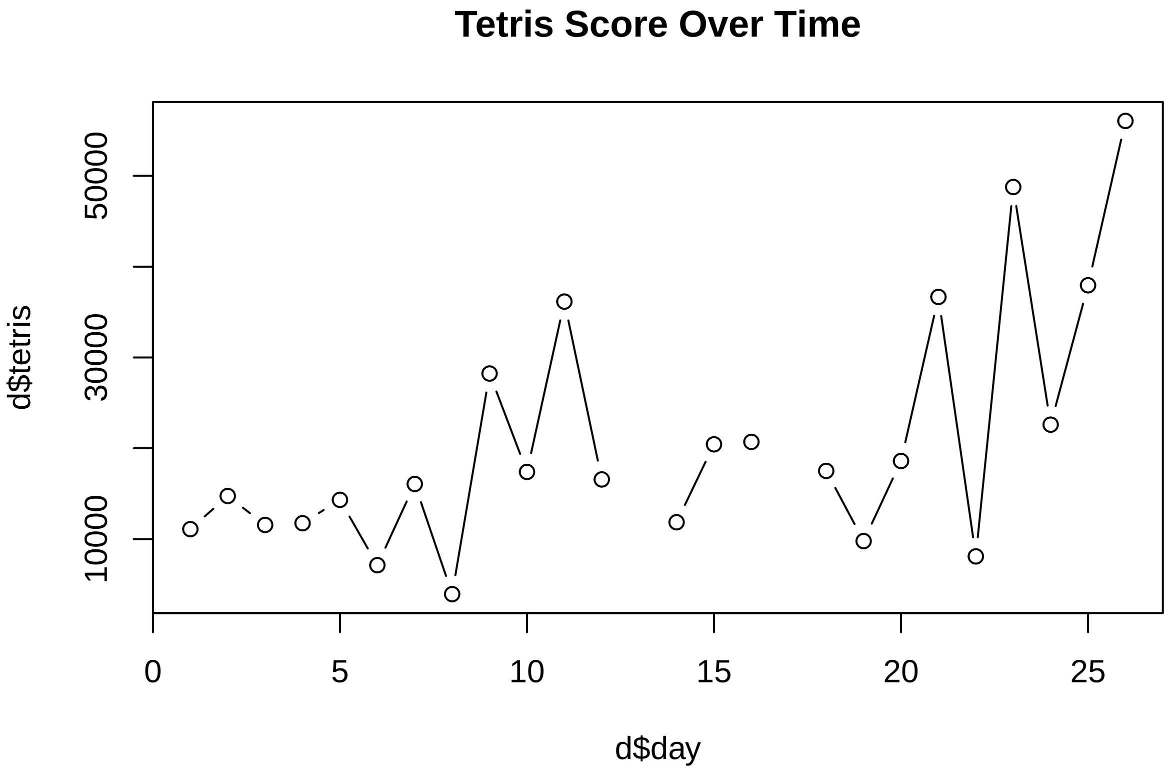
A doctoral student conducted an experiment in which she randomly varied whether she ran or walked 40 minutes each morning. In the middle of the afternoon over a period of 26 days she measured the following outcome variables: (1) her weight; (2) her score in Tetris; (3) her mood on a 0-5 scale; (4) her energy; and (5) whether she got a question right on the math GRE.

```r
d <- read.dta("./data/Hough_WorkingPaper_2010.dta")
head(d)
```

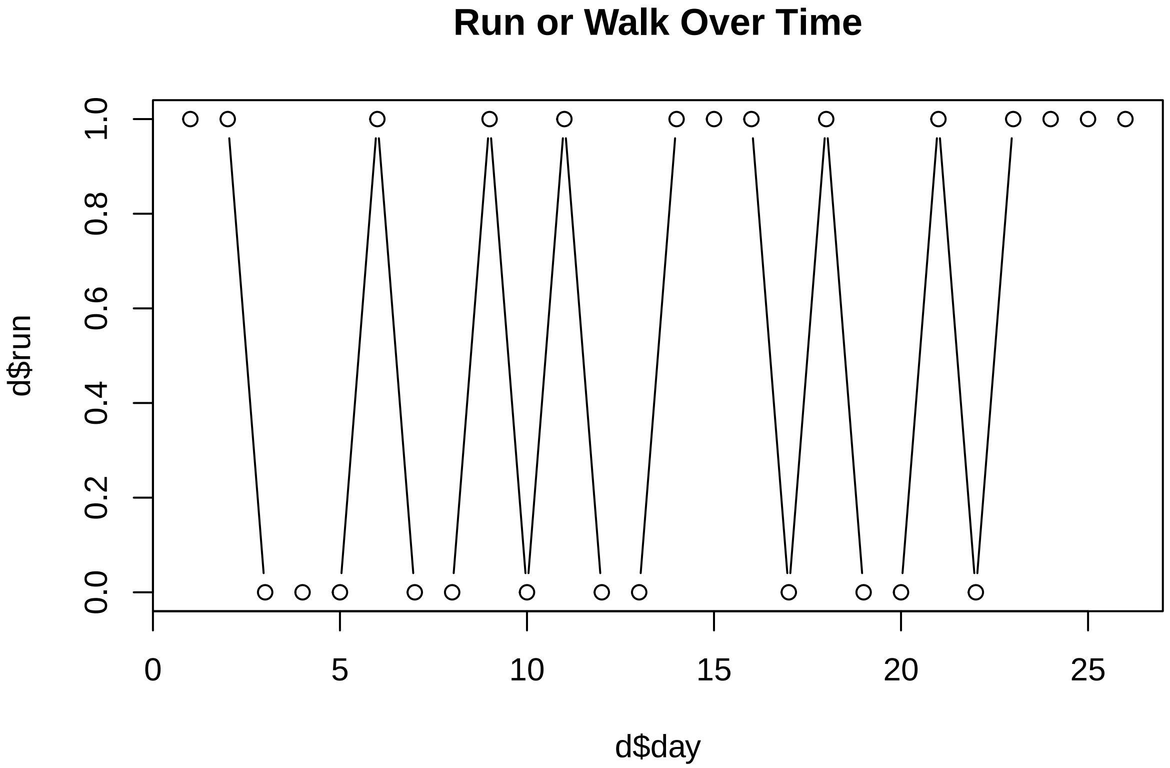
```
##   day run weight tetris mood energy appetite gre
## 1   1   1     21  11092    3      3        0   1
## 2   2   1     21  14745    3      1        2   0
## 3   3   0     20  11558    3      3        0   1
## 4   4   0     21  11747    3      1        1   1
## 5   5   0     21  14319    2      3        3   1
## 6   6   1     19   7126    3      2        0   1
```

a. Suppose you were seeking to estimate the average effect of running on her Tetris score. Explain the assumptions needed to identify this causal effect based on this within-subjects design. Are these assumptions plausible in this case? What special concerns arise due to the fact that the subject was conducting the study, undergoing the treatments, and measuring her own outcomes?

```r
plot(d$day,d$tetris, type = "b", main = "Tetris Score Over Time")
```



```
plot(d$day,d$run, type = "b", main = "Run or Walk Over Time")
```



Answer: There are 2 assumptions required to identify this causal effect based on a within-subjects design. The first assumption is the no anticipation assumption which means that the potential outcomes are unaffected by treatments that are administered in the future. The second assumption is the no persistence assumption which means that the potential outcomes in one period are unaffected by treatments administered in prior periods. It is believable that the no anticipation assumption holds in this specific experiment because I don't think one's tetris playing ability can really be affected by knowing that in the future the subject may have to run. However, the no persistance assumption is likely not plausible in this specific case. It is possible that the subject becomes more fit over time, after running or walking for 40 minutes every day and therefore have a very better mental health that can affect their tetris score as time progresses. On the other hand, it is possible that the runner becomes sore or injured over time and therefore possibly decrease their tetris score because they have the distraction of pain and soreness. Finally, the subject's tetris score can improve over time after playing every day which would mean that their score would naturally improve over time. The plot shown above shows that her tetris score has been improving over time for whatever reason which might indicate that the no persistance assumption is violated. Also in the second half of the days (day 13 onwards) she was randomly selected to run 9 out of the 12 days which might make it seem like running would result in a better tetris score when actually she was getting a better tetris score because she has been getting better at tetris.

Since the subject is conducting the study, undergoing the treatments, and measuring her own outcomes, the is always the concern that the subject would have some pre-existing assumptions of what the results of the findings are. For example, if she expects to do better in tetris on the days that she runs, she might be subconciously perform better on those days. Or maybe expecting to do better on days that she runs, she would have added pressure and then perform more poorly than she would have. The main concern is that she knows too much about the experiment and it can impact the outcomes that she measures. For other categories such as mood and energy, the score is so subjective that she may be answering the survey in a way that better fits her hypothesis. Maybe she expects to feel better on days that she runs, and therefore automatically gives herself a higher mood score for the day. Finally, one interesting note about the plot shown above is that on day 13 and 17, she was assigned to walk but she did not record her tetris score on those days. I'm not sure why she would be missing data on those days but I thought it was an interesting point.

b. Estimate the effect of running today on Tetris score. What is the ATE?

```
# Use standard method to find ATE
d2 <- na.omit(d)
run <- mean(d2[d2$run == 1,]$tetris)
walk <- mean(d2[d2$run == 0,]$tetris)
ATE <- run - walk
paste("The ATE of running on Tetris score is", ATE)
```

```
## [1] "The ATE of running on Tetris score is 13613.1"
```

```
# Use linear regression to find ATE
model <- lm(tetris ~ run, data = d)
summary(model)
```

```
##
## Call:
## lm(formula = tetris ~ run, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19294  -6707  -1154   4890  29628
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12806       3708   3.453  0.00226 **
## run            13613       4856   2.804  0.01035 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 11730 on 22 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2632, Adjusted R-squared:  0.2297
## F-statistic:  7.86 on 1 and 22 DF,  p-value: 0.01035
```

Answer: According to our findings shown above, if you run for 40 minutes in the morning and play tetris in the middle of the afternoon, you can expect your tetris score to be *13613* points higher than if you walked 40 minutes in the morning instead.

c. One way to lend credibility to with-subjects results is to verify the no-anticipation assumption. Construct a regression using the variable `run` to predict the `tetris` score *on the preceding day*. Presume that the randomization is fixed. Why is this a test of the no-anticipation assumption? Does a test for no-anticipation confirm this assumption?

```r
# First we create a new column with the staggered tetris score
d["preceding_tetris"] <- c(NA,d$tetris[1:length(d$tetris)-1])

# Create a regression using the variable run to predict the preceding tetris score.
model <- lm(preceding_tetris~run, data = d)
summary(model)
```

```
##
## Call:
## lm(formula = preceding_tetris ~ run, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15610  -7521  -2337   2424  29220
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18903.8     3335.8   5.667 1.26e-05 ***
## run            645.6     4823.5   0.134    0.895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 21 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.0008524,  Adjusted R-squared:  -0.04673
## F-statistic: 0.01792 on 1 and 21 DF,  p-value: 0.8948
```

Answer: The no anticipation assumption states that the potential outcomes are unaffected by treatments that are administered in the future. Therefore, if we try to use the `run` variable to predict the tetris score from the preceding day, we should not have a statistically significant linear regression. In the regression shown above, the p-value of the run coefficeint on the previous day's tetris score is *0.895* which is close to 90%. This means that the `run` variable does not do a good job of predicting the preceding day's tetris score which means that our no anticipation assumption is not violated.

d. Now let's use regression to put a standard error on our ATE estimate from part (b). Regress Tetris score on the the variable `run`, this time using the current rather than the future value of `run`. Is the impact on Tetris score statistically significant?

```r
# Use linear regression to find ATE
model <- lm(tetris ~ run, data = d)
summary(model)
```

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

```
## lm(formula = tetris ~ run, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19294  -6707  -1154   4890  29628
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12806       3708   3.453  0.00226 **
## run            13613       4856   2.804  0.01035 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11730 on 22 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2632, Adjusted R-squared:  0.2297
## F-statistic:  7.86 on 1 and 22 DF,  p-value: 0.01035
```

Answer: The standard error we got is *4856* and the p-value is *0.01035* which means the impact on Tetris score is statistically significant.

e. If Tetris responds to exercise, one might suppose that energy levels and GRE scores would as well. Are these hypotheses borne out by the data?

```
model.energy <- lm(energy ~ run, data = d)
summary(model.energy)
```

```
##
## Call:
## lm(formula = energy ~ run, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0714 -1.0179  0.0000  0.9286  1.9286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.00000    0.33662   8.912  9.4e-09 ***
## run          0.07143    0.44074   0.162    0.873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.064 on 22 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.001192,   Adjusted R-squared:  -0.04421
## F-statistic: 0.02627 on 1 and 22 DF,  p-value: 0.8727
```

Answer: The ATE found above is *0.07143* meaning that running will cause an average treatment effect of having 0.07 higher energy level on a scale from 1 to 5. However, the p-value is *0.873* which is very high meaning the results is not statistically significant. Therefore, the hypothesis that running for 40 minutes a day rather than walking will result in higher energy levels is not supported by the data.

```
model.gre <- lm(gre ~ run, data = d)
summary(model.gre)
```

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

```
## lm(formula = gre ~ run, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8182 -0.6429  0.1818  0.3571  0.3571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8182     0.1385   5.909 5.05e-06 ***
## run          -0.1753     0.1850  -0.948    0.353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4592 on 23 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.03757,    Adjusted R-squared:  -0.004275
## F-statistic: 0.8978 on 1 and 23 DF,  p-value: 0.3532
```

Answer: The ATE found above is *-0.1753* mean that running will cause an average treatment effect of being more likely to fail to answer the GRE question. However, the p-value in this case is *0.353* which is also not statistically significant. Therefore, the hypothesis that running for 40 minutes a day rather than walking will result in a higher likelihood of correctly answering a GRE question is not supported by the data.

f. Suppose the student decides to publish her results on Tetris, since she finds those most interesting. In the paper she writes, she chooses to be concise by ignoring the data she collected on energy levels and GRE scores, since she finds those results less interesting. How might you criticize the student's decision? What trap may she have fallen into?

Answer: First of all, she might be falling into the multiple comparisons trap where she is fishing for a statistically significant treatment effect. If she runs multiple linear regressions it is possible for her to come accross one that is statistically significant out of pure chance. She should include the other two results so that her readers will know what other linear regressions she ran before coming to this finding. She might also want to apply the Bonferoni correction to address this. Furthermore, she is also contributing to publication bias where only statistically significant or "interesting" results are being published. Maybe there is a study out there that shows that excercise produces higher energy levels with statistically significant results. Her study can counter that finding and if she does not publish it because it's not "interesting", then she is not doing her part to keep the integrity all published results.

g. After submitting her paper to a journal, the student thinks of another hypothesis. What if running has a relatively long-lasting effect on Tetris scores? Perhaps both today's running and yesterday's running will affect Tetris scores. Run a regression of today's Tetris score on both today's `run` variable and yesterday's `run` variable. How does your coefficient on running today compare with what you found in part (d)? How do you interpret this comparison?

```r
# Create new column for yesterday's run variable
d["yesterday_run"] <- lag(d$run)

# Do a linear regression of today's tetris score on both today and yesterday's run variable
model <- lm(tetris ~ run + yesterday_run, data = d)
summary(model)
```

```
##
## Call:
## lm(formula = tetris ~ run + yesterday_run, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19693.2  -7830.5    -46.2   5359.5  27539.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11793       4754   2.481  0.02211 *
## run              15026       4986   3.014  0.00686 **
## yesterday_run     1689       4948   0.341  0.73643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11740 on 20 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.3125, Adjusted R-squared:  0.2437
## F-statistic: 4.545 on 2 and 20 DF,  p-value: 0.0236
```

Answer: The new coefficient on running today is *15026* compared to *13613* found in part d. Furthermore, the p-value is now *0.00686* which means it is more statistically significant than what we found in part d. This means that keeping whether or not you ran yesterday constant, just running today would result in an average tetris score that is *15026* points higher compared to if you walked instead. This also means that running yesterday has some effect on your tetris score today. In that case, it means that the no persistence assumption is violated in this experiment.

h. (optional) Note that the observations in our regression are not necessarily independent of each other. An individual might have serially correlated outcomes, regardless of treatment. For example, I might find that my mood is better on weekends than on weekdays, or I might find that I'm terrible at playing Tetris in the few days before a paper is due, but I get better at the game once my stress level has lowered. In computing standard errors for a regression, OLS assumes that the observations are all independent of each other. If they are positively serially correlated, it's possible that OLS will underestimate the standard errors.

To check this, let's do randomization inference in the regression context. Recall that the idea of randomization inference is that under the sharp null hypothesis, we can re-randomize, recompute the ATE, and get approximately the right answer (zero) for the treatment effect. So, returning to the regression we ran in part (g), please generate 1000 new randomizations of the `run` variable, use those to replace the current and lagged values of `run` in your dataset, then run the regression again. Record the coefficient you get on the contemporaneous value of `run`, and repeat this re-randomization exercise 1000 times. Plot the distribution of beta. What are the 2.5% and 97.5% quantiles? How do they compare with the width of the 95% confidence interval you got for your main `run` coefficient in the regression in part (g)?