

Problem Set #4

Experiment Design: Alex Smith

March 5, 2016

FE exercise 5.2

- 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.

```
# disable scientific notation so that r has no problems outputting to PDF
options(scipen=999)

# create the outcomes should everyone be in the control group
y0 <- c(5, 4, 7, 20, 35, 15)

# assign the first 3 individuals to be compliers and second 3 to be never takers
comp <- c(1, 1, 1, 0, 0, 0)

# create a treatment effect for compliers of -10
comp_effect <- -10

# create a treatment effect for never takers of 5
never_effect <- 40

# create the outcomes should everyone be in the treatment group
# only provide the true treatment effect to those who are compliers
y1 <- c(c(y0[comp==1]+comp_effect),c(y0[comp==0]+never_effect))

# lets think of this as a time study experiment and each person begins in control
# then receives treatment
# we can calculate the estimated treatment effect by subtracting the outcomes
# pre-treatment, y0, from the outcomes post-treatment, y1
est_ate <- mean(y1 - y0)

# let's calculate the complier average causal effect (CACE), where we only measure
# the effect for the compliers
est_cace <- mean(y1[comp==1] - y0[comp==1])
```

The actual estimated average treatment effect is **15**, which is positive. The complier average causal effect, CACE, is **-10**.

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

The estimated CACE would be informative in that it would tell use the treatment effect for those who would comply with the treatment. However, it would be misleading to assume that it was the true treatment effect because we do not know what the effect is on the never-takers. Perhaps the treatment effect on the never takers - should we actually be able to convince them to take treatment - would be different in magnitude or direction than the effect for the compliers.

- c. **In addition, please also answer this question:** Which population is more relevant to study for future decision making: the set of Compliers, or the set of Compliers plus Never-Takers? Why?

The population that is more relevant to study for future decision makers depends on the context. For example, assume that we are studying the treatment effect of some blood pressure pill. If we are the chemist designing the pill, we would be more interested in the set of compliers because we are trying to learn how the pill affects the blood pressure of those who take it. If we are a policy maker giving free pills to everyone in a certain region, we would be more interested in the set of compliers plus never-takers because we have paid for pills for both groups and so need to account for the treatment effect across both groups.

FE exercise 5.6

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?

```
# calculate the control group voter rate
turnout_control <- 700/2000

# calculate the treatment group voter rate
turnout_treat <- 400/1000

# people claimed to be contacted
treat_claim <- 500/1000

# people actually contacted
treat_actual <- 250/1000

# we can use the formula that CACE = ITT / ITT_D, where ITT is the intent
# to treat effect (difference between the treatment and control groups) and
# ITT_D is the portion of the treatment group actually treated (application rate)
claim_CACE <- (turnout_treat - turnout_control) / (treat_claim)
```

The complier average causal effect (CACE) estimated should I believe my canvassers that they contacted 500 households is **0.1**.

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

```
# CACE = ITT / ITT_D
actual_CACE <- (turnout_treat - turnout_control) / (treat_actual)
```

If I learned that the canvassers actually only reached 250 subjects, I would estimate the CACE at **0.2**.

- 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?

The canvassers' exaggerated reports make their efforts seem less effective because it shows that a fewer portion of the people supposedly contacted turned out to vote. This means that the canvassers were less effective at increasing turnout among those they contacted.

FE exercise 5.10

Guan and Green report the results of a canvassing experiment conducted 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)
peking <- read.dta("http://hdl.handle.net/10079/zkh18mj")
head(peking)
```

```
##   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.*

```
# remove the na's
peking = na.omit(peking)

# calculate the mean for the treatment and control groups
control_turnout <- mean(peking$turnout[peking$treat2==0])
treat_turnout <- mean(peking$turnout[peking$treat2==1])

# calculate the peking ITT as a difference in means
peking_itt <- treat_turnout - control_turnout

# create a linear regression model for the itt
peking_lm_model <- lm(turnout ~ treat2, data = peking)

# get the ITT from the model
peking_lm_itt <- coef(summary(peking_lm_model))[2]
```

When I compare the means of the two groups (treatment and control), I find that an intent to treat effect of **0.1319296**. Using a linear model provides me with the same estimate of **0.1319296**.

- 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.

```

# get the number of students in the study
observations = nrow(peking)

# borrow some functions from Problem Set#2 for randomization inference

# create a function that uses simple random assignment to divide observations
# into treatment and control groups
randomize <- function(random_qty, cluster_ids) {
  original_random_qty = random_qty
  if (random_qty %% 2 == 1) {random_qty = random_qty + 1}
  random_assignment <- sample(c(rep(0,random_qty/2),rep(1,random_qty/2)))
  random_assignment <- random_assignment[1:original_random_qty]

  # make sure to cluster by the cluster_ids
  random_assignment_cluster <- vector()
  for(i in 1:length(cluster_ids)) {
    if(i == 1) {random_assignment_cluster <- append(random_assignment_cluster,
                                                    random_assignment[i])}

    else {
      if(cluster_ids[i] == cluster_ids[i-1]) {random_assignment_cluster <-
        append(random_assignment_cluster,random_assignment[i-1])}
      else {
        random_assignment_cluster <- append(random_assignment_cluster,
                                            random_assignment[i])
      }
    }
  }

  return(random_assignment_cluster)
}

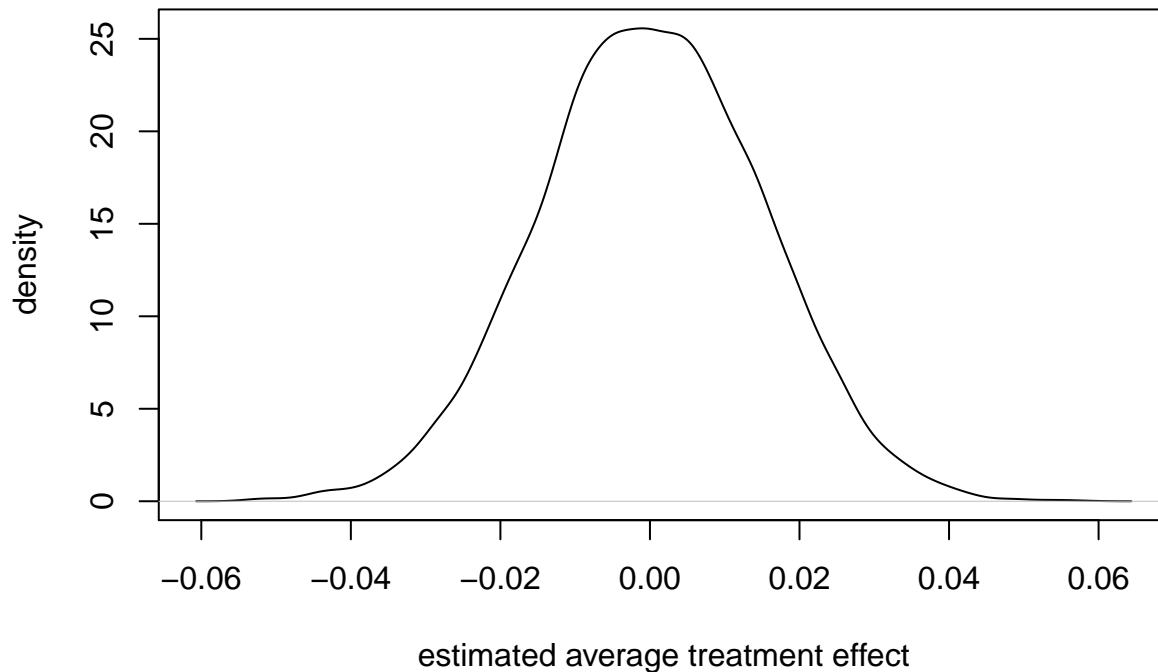
# create a function that calculates the intent to treat effect (ITT) based on two
# vectors, the first determining treatment or control assignment and the second
# the observed outcome
itt <- function(assignment_vector, outcome_vector) {
  outcomes <- assignment_vector * outcome_vector +
    (1-assignment_vector) * outcome_vector
  return(mean(outcomes[assignment_vector == 1]) -
          mean(outcomes[assignment_vector == 0]))
}

# perform 10,000 simulated random assignments and calculate estimated ITTs for each
# warning: this code runs a bit slowly
SIMULATIONS <- 10000
sharp_null_dist <- replicate(SIMULATIONS,
                             itt(randomize(observations,peking$dormid),peking[,1]))

# plot the distrubtion of ITTs
plot(density(sharp_null_dist),main = "Distribution Under the Sharp Null",
     xlab = "estimated average treatment effect", ylab = "density")

```

Distribution Under the Sharp Null



```
# calculate how many of the simulations produced an ITT at least as large as the  
# actual estimated ITT  
simulation_greater <- sum(sharp_null_dist >= peking_itt)
```

Using randomization inference, I found **0** simulated ITTs greater than the ITT that I actually calculated. The implied p-value is **0**.

- c. Assume that the leaflet had no effect on turnout. Estimate the CACE. Estimate this quantity using differences in means, and also using linear models on the appropriate subsets of data.

```
# the ITT_D is the number of contacted over the total number assigned to treatment  
peking_itt_d <- sum(peking$contact)/sum(peking$treat2)  
  
# CACE = ITT / ITT_D  
# remember that peking_itt was calculated by difference in means methods  
peking_cace <- peking_itt / peking_itt_d  
  
# pull out the complier's and control's turnouts  
contacted <- peking$turnout[peking$contact==1]  
controlled <- peking$turnout[peking$treat2==0]  
  
# calculate the ITT_D from a linear model  
peking_lm_itt_d_model <- lm(contact ~ treat2, data = peking)  
peking_lm_itt_d <- coef(summary(peking_lm_itt_d_model))[2]  
  
# CACE = ITT / ITT_D  
# remember that the peking_lm_itt is the itt calculated by the linear  
# model from part a  
peking_lm_cace <- peking_lm_itt / peking_lm_itt_d
```

Using a difference in means, I calculate the complier average causal effect (CACE) as **0.1489402**. Using a linear model, I calculate the same value, **0.1489402**.

- d. *SKIP*
- e. *SKIP*
- f. *SKIP*

FE exercise 5.11

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 as 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.*)

```
# create the data for the baseline (control) group
control_assignment <- rep(0,2572)
control_treated <- rep(0,2572)
control_pl_treated <- rep(0,2572)
control_turnout <- c(rep(1,round(.3122*2572)),rep(0,2572-round(.3122*2572)))

# create the data for the treatment group
treatment_assignment <- rep(1,486+2086)
treatment_treated <- c(rep(1,486),rep(0,2086))
treatment_pl_treated <- rep(0,486+2086)
treatment_turnout <- c(rep(1,round(.3909*486)),rep(0,486-round(.3909*486)),
                      rep(1,round(.3274*2086)),rep(0,2086-round(.3274*2086)))

# create the data for the placebo group
placebo_assignment <- rep(2,470+2109)
placebo_treated <- rep(0,470+2109)
placebo_pl_treated <- c(rep(1,470),rep(0,2109))
placebo_turnout <- c(rep(1,round(.2979*470)),rep(0,470-round(.2979*470)),
                    rep(1,round(.3215*2109)),rep(0,2109-round(.3215*2109)))

# create each column in the data table
assignment <- c(control_assignment,treatment_assignment,placebo_assignment)
treated <- c(control_treated,treatment_treated,placebo_treated)
pl_treated <- c(control_pl_treated,treatment_pl_treated,placebo_pl_treated)
turnout <- c(control_turnout,treatment_turnout,placebo_turnout)
```

```
# create the data table that produces the above summary table, name this data table
# after the experimenter
library(data.table)
nick <- data.table(assignment,treated,pl_treated,turnout)
```

- a. We are rewriting part (a) as follows: “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).

```
# calculate portion of contacted in treatment
treatment_contacted <- mean(nick$treated[nick$assignment==1])

# calculate the portion of contact in placebo
placebo_contacted <- mean(nick$pl_treated[nick$assignment==2])

# use a t-test to compute how statistically significant this difference is
complier_t <- t.test(nick$treated[nick$assignment==1],
                    nick$pl_treated[nick$assignment==2])
```

The portion of compliers in the treatment group is **0.188958**. The portion of compliers in the control group is **0.1822412**. This difference of **0.0067168** is not statistically significant. A t-test comparison of these values returns a p-value of **0.5353714**.

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

```
# create a t-test to measure the difference in turn-out between never takers in
# the treatment and placebo groups
never_t <- t.test(nick$turnout[nick$assignment==1 & nick$treated==0],
                 nick$turnout[nick$assignment==2 & nick$pl_treated==0])
```

The data do not suggest that the Never Takers of the treatment and placebo groups have the same rate of turn-out. The p-value for a t-test for the difference between these turn-outs returns a p-value of **0.681175**. This comparison is informative because it allows to calculate the turnout rate for the would-be Never Takers of the baseline control group. This would be helpful in calculating the complier average causal effect because we could use it to infer the would-be compliers in the baseline control group.

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

```
# use the method: CACE = ITT/ITT_D

# calculate the ITT for the placebo
placebo_itt <- mean(nick$turnout[nick$assignment==2]) -
  mean(nick$turnout[nick$assignment==0])

# calculate the ITT_D for the placebo
```

```

placebo_itt_d <- sum(nick$pl_treated[nick$assignment==2]==1) / sum(nick$assignment==2)

# calculate the CACE for receiving the placebo
placebo_cace <- placebo_itt / placebo_itt_d

```

The CACE of receiving the placebo is **0.027265**. Yes, this estimate is consistent with the assumption that receiving the placebo has little to no effect on turn-out.

- 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.

```

# CACE = ITT/ITT_D

# calculate the ITT for the treatment
treatment_itt <- mean(nick$turnout[nick$assignment==1]) -
  mean(nick$turnout[nick$assignment==0])

# calculate the ITT_D for the treatment
treatment_itt_d <- sum(nick$treated[nick$assignment==1]==1) /
  sum(nick$assignment==1)

# calculate the CACE
treatment_cace <- treatment_itt / treatment_itt_d

```

Using the conventional method, I calculate a CACE for the treatment of **0.1440329**.

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

```

# calculate the turnout for compliers in the treatment group
compliers_treatment <- mean(nick$turnout[nick$treated == 1])

# calculate the turnout for compliers in the placebo group
compliers_placebo <- mean(nick$turnout[nick$pl_treated == 1])

# calculate the difference in turnout among compliers
complier_difference <- compliers_treatment - compliers_placebo

```

When I compare the turnouts among the compliers in both the treatment and placebo groups, I find an effect of **0.0930742**. This is the difference in turnout rates for those who comply. With this method, I used a placebo group to accurately identify the compliers in the non-treatment group. I was able to directly calculate the turnout rate for the compliers in the non-treatment group rather than just infer it.

More Practice

Determine the direction of bias in estimating the ATE for each of the following situations when we randomize at the individual level. Do we over-estimate, or underestimate? Briefly but clearly explain your reasoning.

- a. In the advertising example of Lewis and Reiley (2014), assume some treatment-group members are friends with control-group members.

In this case, we underestimate the ATE. When members of the treatment group are friends with those in the control group, the folks who saw the advertisements might tell their friends in the control group about the ads they saw. This might increase the likelihood that their friends in the control group purchase from the marketer.

- b. Consider the police displacement example from the bulleted list in the introduction to FE 8, where we are estimating the effects of enforcement on crime.

In the police displacement study, we overestimate the ATE. When we increase enforcement in a given neighborhood, we push the crime from one neighborhood to the neighboring neighborhoods. This makes the treatment neighborhoods appear safe. However, our treatment has also affected the control group, by pushing crime onto them. This makes the control neighborhoods appear less safe.

- c. Suppose employees work harder when you experimentally give them compensation that is more generous than they expected, that people feel resentful (and therefore work less hard) when they learn that their compensation is less than others, and that some treatment-group members talk to control group members.

In this case, you are attempting to measure the impact of increasing compensation for your employees. By increasing the compensation of the treatment group, you see an improvement in their productivity. However, there is also an effect on the control group. Because they feel resentment for receiving less compensation, their productivity decreases. You overestimate the ATE because your control group (against which you are measuring the increase of the treatment) has lowered productivity due to the treatment.

- d. When Olken (2007) randomly audits local Indonesian governments for evidence of corruption, suppose control-group governments learn that treatment-group governments are being randomly audited and assume they are likely to get audited too.

When auditing local Indonesian governments, we are attempting to determine if audits lead to less corruption. The treatment is the audits. However, if the control group begins to also fear being audited, they may change their behavior. Fearing an audit, they may decrease their corruption. We attempted to measure the impact of audits on corruption compared to a lack of audits on corruption. However, the audits of the treatment may cause the control group to clean up its act. This will lead to underestimating the ATE as the control group will become more like the treatment group.

FE exercise 8.2

National surveys indicate that college roommates tend to have correlated weight. The more one roommate weighs at the end of the freshman year, the more the other freshman roommate weighs. On the other hand, researchers studying housing arrangements in which roommates are randomly paired together find no correlation between two roommates' weights at the end of their freshman year. *Explain how these two facts can be reconciled.*

The first study may be finding that students choose to live with their friends, and they choose friends who are similar to themselves. In other words, heavier students may befriend and live with other heavier students. However, when students' roommates are randomly assigned, there is no correlation between roommates' weights. This implies that your roommate's weight has no impact on your weight.

FE exercise 8.10

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.

```
library(foreign)
exercise <- read.dta("http://hdl.handle.net/10079/zcrjds4")
head(exercise)
```

##	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?

The two major assumptions of a within-subjects design are nonpersistence and nonanticipation. Nonpersistence means that the treatment effect does not outlive the treatment period. In this case, it means that the effect of her running the previous day will not impact her performance the next day. While this assumption is plausible (the effects of exercise wear off after a day), there are some concerns. For example, if she was not very healthy before she began this experiment, it could be that randomly running throughout the week will improve her general health regardless of whether or not she runs on a particular day. Improving her general health might improve her scores on the outcome variables.

The second assumption, nonanticipation, means that having treatment in the next period does not impact the results of the current period. In this case, it means that tomorrow being a running day will have no effect on her performance for today. This assumption feels more plausible than the nonpersistence assumption because if each day she decides whether she is running or walking, then on the previous day she will have no knowledge of the next day's treatment. Since she has no knowledge of the next day, she cannot anticipate it.

Special concerns arise because the subject is conducting the study, undergoing the treatments, and measuring her own outcomes. This is not a double blind (or even single blind) experiment. She knows when she has received the treatment, and so may consciously (or unconsciously) change her behavior. Or, even if she does not change her performance by realizing that she is in the treatment group for that day, she may change how she measures her results. Unconsciously, she may look to shape her results to match her expected conclusion.

- b. Estimate the effect of running on Tetris score. Use regression to test the sharp null hypothesis that running has no immediate or lagged effect on Tetris scores. (**Note** the book calls for randomization inference, but this is a bit of a tough coding problem. **HINT:** For the second part of part (b), run one regression of today's score on both today's treatment assignment and yesterday's treatment assignment. Then, calculate the p-value that both effects are zero.)

```
# estimate the effect of running on tetris score
tetris_lm_short <- lm(tetris ~ run, data = exercise)
```

```

# get the coefficient
tetris_coef <- coef(summary(tetris_lm_short))[2]

# get the p-value
run_p <- coef(summary(tetris_lm_short))[8]

# create a new vector to store yesterday's treatment
yesterday <- vector()

# iterate through the run vector and create a vector of yesterday's treatment
for(i in 1:length(exercise$run)) {
  if(i == 1) {yesterday <- append(yesterday,0)}
  else {yesterday <- append(yesterday,exercise$run[i-1])}
}

# add the vector to the data frame
exercise['yesterday'] <- yesterday

# run a regression for today's score on BOTH today's treatment
# and yesterday's treatment
tetris_lm_long <- lm(tetris ~ run + yesterday, data = exercise)

# get the p-value for yesterday's running
yesterday_p <- coef(summary(tetris_lm_long))[12]

# run an F-test ANOVA to determine if both effects adds no information compared
# to the single effect of running today
ftest <- anova(tetris_lm_short,tetris_lm_long)

# store the p-value for the ftest
f_p_value <- ftest[2,6]

```

Running causes her tetris score to improve by **13613** points. The effect of running today is statistically significant at the 0.05 level with a p-value of **0.0103512**, while the effect of running yesterday is not statistically significant at the 0.05 level with a p-value of **0.5702475**. When I compare the the long and short models, I get a p-value of **0.5702475** which means that I cannot reject the null hypothesis that there is no difference between the models. In sum, the long model (the model with both today's and yesterday's running feature) does not add much explanation for the variance.

- c. One way to lend credibility to within-subjects results is to verify the no-anticipation assumption. Use 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?

```

# create a new vector to store tomorrow's treatment
tomorrow <- vector()

# iterate through the run vector and create a vector of tomorrow's treatment
for(i in 1:length(exercise$run)) {
  if(i == length(exercise$run)) {tomorrow <- append(tomorrow,0)}
  else {tomorrow <- append(tomorrow,exercise$run[i+1])}
}

```

```

# add the vector to the data frame
exercise['tomorrow'] <- tomorrow

# create a linear model for tomorrow's treatment on tetris score
tetris_lm_tomorrow <- lm(tetris ~ tomorrow, data = exercise)

# calculate the effect of tomorrow's treatment on today's tetris score
tomorrow_coef <- coef(summary(tetris_lm_tomorrow))[2]

# calculate the p-value for tomorrow's treatment
tomorrow_p <- coef(summary(tetris_lm_tomorrow))[8]

```

To test the nonanticipation assumption, we can measure the impact of tomorrow's treatment on today's tetris results. The nonanticipation assumption is that tomorrow's treatment will have no effect on today's results. If we were to find a statistically significant effect on today's scores from tomorrow's treatment, we would suspect that the nonanticipation assumption was being violated.

We find that this assumption is not violated because while running tomorrow does produce a **-2211.5454545** effect on today's scores, this effect is not statistically significant with a p-value of **0.69554**.

- d. 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?

```

# let's create a regression model for energy levels
energy_lm <- lm(energy ~ run, data = exercise)

# calculate and store the coefficients and p-values
energy_coef <- coef(summary(energy_lm))[2]
energy_p <- coef(summary(energy_lm))[8]

# let's create a regression model for GRE scores
gre_lm <- lm(gre ~ run, data = exercise)

# calculate and store the coefficients and p-values
gre_coef <- coef(summary(gre_lm))[2]
gre_p <- coef(summary(gre_lm))[8]

```

These hypotheses are not borne out by the data. When testing if running has an impact on energy levels and GRE scores, we find p-values of **0.8727328** and **0.3532139** for energy levels and GRE scores respectively. With these p-values, we cannot reject the null hypothesis that the impact of running on energy levels and GRE problems is not different from zero. In theory, we should use a Bonferroni correction since we are performing multiple analyses from the same data set. However, in this case, the p-values are insignificant on their own. We do not need the correction to make the p-values even more insignificant.

- e. **Additional Mandatory Question:** Note that the observations in this regression are not necessarily all independent of each other. In particular, think about what happens when we lag a variable. Given this non-independence, would you expect randomization inference to give you a better answer than the regression answer you just obtained in (b)? Why? Which number(s) do you expect to be different in regression than in randomization inference? What is the direction of the bias? (This is a conceptual question, so you do not need to conduct the randomization inference to answer it. However, you are certainly welcome to try that exercise if you are curious.)

A lagged treatment variable is not independent from today's treatment variable across time. In other words, today's lagged treatment is not independent from yesterday's treatment. Or, to say it another way, my treatment today determines what my lagged treatment for tomorrow will be. In the context of this problem, whether you were assigned to run today is *not* independent of tomorrow, whether you were assigned to run the previous day. I would expect randomization inference to give me a better answer than regression because our simple linear regression models assume independence between all the treatment variables. However, with randomization inference, I can adjust my model to capture the dependence between these variables. This means that I would expect my p-value for the randomization inference model to be smaller than for the simply linear regression because I'm accounting for more of the information in my model. If the effects of my treatment today, and my lagged treatment, are both positive (or both negative), then I would expect my coefficients produced by randomization inference to be more positive (or more negative).