Problem Set #4

Experiment Design: Alex & Daniel

```r
# R cell to hold functions
make_data_1a <- function() {
    ## this function will make data for the purposes of learning
    ## about how  random or non-random non-compliance will change
    ## a two-group difference estimate of those who are treated.

    require(data.table)

    d <- data.table(id = 1:6)

    d[ , y0  := c(2,4,6,8,10,12)]
    d[ , y1  := c(1,6,5,10,8,13)]
    d[ , C   := c(1,0,1,0,1,0)]

    ## return the dataset back
    return(d)
}
```

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 compliers 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.

```
data_1a = make_data_1a()
head(data_1a)
```

```
##    id y0 y1 C
## 1:  1  2  1 1
## 2:  2  4  6 0
## 3:  3  6  5 1
## 4:  4  8 10 0
## 5:  5 10  8 1
## 6:  6 12 13 0
```

```
ATE = mean(data_1a[, y1]) - mean(data_1a[, y0])
```

The ATE for this hypothetical experiment is: 0.1666667

```
CACE = mean(data_1a[C == 1, y1]) - mean(data_1a[C == 1, y0])
```

The CACE for this hypothetical experiment is: -1.3333333

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

The complier rate is $\frac{3}{6} = 0.5$. Correspondingly the estimated CACE would be $\frac{ATE}{complierrate} = \frac{0.1666666}{0.5} = \frac{1}{3}$. This would be incorrect because the actual CACE is negative.

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?

If the goal of the experiment is to measure the effect of treatment, then the set of Compliers are far more relevant to the study. However, there are no guarantees that you will be able to determine who the Compliers are. So you may have to deal with the nitty-gritty aspects of the experiment and study both Compliers and Never-takers. This is not necessairily bad and it may open you up to new ideas for experiments in general.

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?

```
turnout_t = 400/1000
turnout_c = 700/2000
turnout_delta = turnout_t - turnout_c
CACE_2a = turnout_delta/(500/1000)
```

The estimated CACE would be 0.1

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

```
CACE_2b = turnout_delta/(250/1000)
```

The estimated CACE would be 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 CACE calculated from the canvassers' exaggerated reports make their efforts seem less effective. Effectiveness in terms of CACE is looking at the compliers in the treatment and deriving the effect from the population. Whereas calculating ITT relies on everyone (compliers and non-compliers) in treatment. Hence, CACE allows us to more accurately calculate our treatment effect.

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.

```
d <- read.dta("./data/Guan_Green_CPS_2006.dta")
d3 <- data.table::data.table(d)
head(d3)
```

```
##    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.*

```
# initialize known constants
treat_contacted = 2380
treat_total = 2688
treat_voted = 2152
control_total = 1334
control_voted = 892
```

```
ITT_means = (treat_voted/treat_total) - (control_voted/control_total)
```

The ITT using the difference in the group means is: 0.1319296

```
model3 = lm(turnout ~ treat2, data=d3, na.action = na.omit)
ITT_regression = model3$coefficients['treat2']
```

```
stargazer(model3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 18, 2019 - 21:41:59

Table 1:

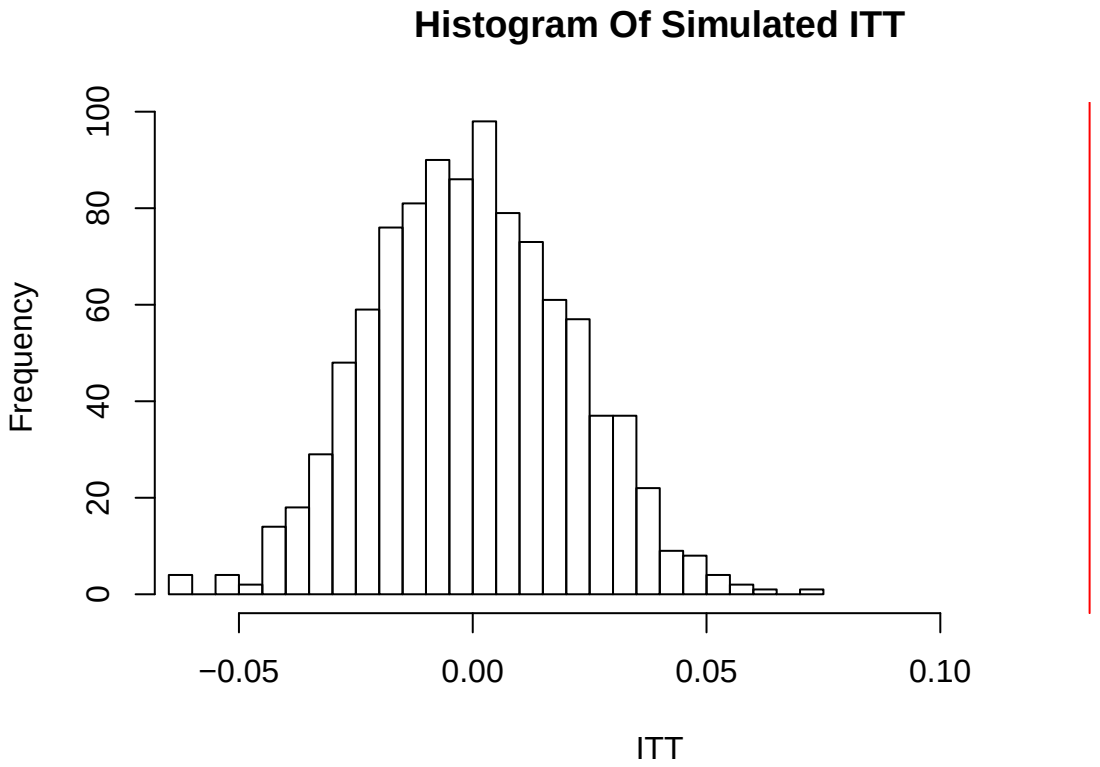
	Dependent variable:
	turnout
treat2	0.132***
	(0.014)
Constant	0.669***
	(0.012)
Observations	4,022
R^2	0.021
Adjusted R^2	0.021
Residual Std. Error	0.425 (df = 4020)
F Statistic	86.082*** (df = 1; 4020)
Note:	*p<0.1; **p<0.05; ***p<0.01

The ITT using regression is: 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.

```
d3 = na.omit(d3)
unique_clusters = unique(d3$dormid)
# specify number of simulations
n_sims = 1000
# init to NA values
res <- rep(NA , n_sims)
# loop and replicate the experiment
for(i in 1:n_sims) {
  treatment = sample(x = unique_clusters, size = length(unique_clusters)/2, replace = FALSE)
  dormids = as.numeric(d3$dormid %in% treatment)
  res[i] = mean(d3$turnout[dormids == 1]) - mean(d3$turnout[dormids == 0])
}
```

```
hist(res, breaks = 20, main = "Histogram Of Simulated ITT", xlab = "ITT", xlim = c(-0.06, 0.14))
abline(v=ITT_regression, col = "red")
```




```
mean(ITT_regression < res)
```

```
## [1] 0
```

We can see that there are no simulated instances where the replicated ITT is larger than the ITT calculated through regression. We can also see that the ITT calculated via regression is not even in the distribution. Hence, we can reject the sharp null hypothesis.

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 compliers.

```
cace_means = ITT_means/(treat_contacted/treat_total)
```

The estimated CACE using means is: 0.1490028

```
model_3c = d3[, ivreg(turnout ~ contact | treat2)]
ttest_3c = coeftest(model_3c, vcov. = cluster.vcov(model_3c, d3$dormid))
ttest_3c
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.668666   0.020241 33.0349 < 2.2e-16 ***
## contact     0.148940   0.026311  5.6607 1.613e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated CACE using the 2SLS method is: 0.1489402. The corresponding standard errors are: 0.0263114. Since we see that the CACE did not change we can interpret this to mean that the leaflet did not have any effect among compliers. However it is hard to verify

because of the experimental design. The leaflets were left regardless of whether or not the door was answered. That makes it hard to determine whether there is an independent effect just from the leaflets themselves. Some of the impact could have been absorbed by the treatment.

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.*)

```
base = 2572
treat = 486
treat_c = 2086
placebo = 470
placebo_c = 2109
base_to = round(base * .3122)
treat_to = round(treat * .3909)
treat_c_to = round(treat_c * .3274)
placebo_to = round(placebo * .2979)
placebo_c_to = round(placebo_c * .3215)
tot_obs = base + treat + treat_c + placebo_c + placebo
```

```
# re-create the table
d4 <- data.frame( "baseline" = c(rep(1,base), rep(0, tot_obs - base)),
                  "treatment" = c(rep(0, base), rep(1, treat ), rep(0, tot_obs - base - treat)),
                  "treatment_control" = c(rep(0, base), rep(0, treat ), rep(1, treat_c), rep(0, tot_obs
                  "placebo" = c(rep(0, base), rep(0, treat ), rep(0, treat_c), rep(1, placebo), rep(0,
                  "placebo_control" = c(rep(0, base), rep(0, treat ), rep(0, treat_c), rep(0, placebo),
                  "turnout" = c(rep(1, base_to), rep(0, base - base_to), rep(1, treat_to), rep(0, treat
                            rep(1, placebo_to), rep(0, placebo - placebo_to), rep(1, placebo_c_to), re
```

```
# replicate the turnout
mean(d4$turnout[d4$baseline == 1])
```

```
## [1] 0.3122084
```

```
mean(d4$turnout[d4$treatment == 1])
```

```
## [1] 0.3909465
```

```
mean(d4$turnout[d4$treatment_control == 1])
```

```
## [1] 0.3274209
```

```
mean(d4$turnout[d4$placebo == 1])
```

```
## [1] 0.2978723
```

```r
mean(d4$turnout[d4$placebo_control == 1])
```

```
## [1] 0.3214794
```

We are able to replicate the turnout using our data and it matches the values in the table.

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).

```r
treat_compliers = treat/(treat + treat_c)
placebo_compliers = placebo/(placebo + placebo_c)
```

The estimated treatment compliers are: 0.188958. The estimated placebo compliers are: 0.1822412. These two values are not statistically significantly different from each other. This makes sense because they were randomly assigned from the same population and the delivery mechanism is the same.

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

As per the table, the Never Takers in the treatment and placebo group have approximately the same rate of turnout. The comparison is informative in that we know that we have randomized well.

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

```r
placebo_ratio = (0.2979*placebo + 0.3215*placebo_c)/(placebo + placebo_c)
placebo_itt = placebo_ratio - 0.3122
placebo_cace = placebo_itt/placebo_compliers
```

The estimated CACE for the placebo group is: 0.0274313. This does support our assumption that the placebo has very little to no effect on 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.)

```r
treat_ratio = (0.3909*treat + 0.3274*treat_c)/(treat + treat_c)
treat_itt = treat_ratio - 0.3122
treat_cace = treat_itt/treat_compliers
```

The estimated CACE for the treatment group is: 0.1439412

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

```r
to_delta = 0.3909 - 0.2979
```

The delta between turnout rates is: 0.093. This means that 9.3% more turnout if the door was answered versus if the door was not answered The results are more conservative when compared to the ITT method because the ITT method does not account for the difference between compliers and non-compliers.

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?

We calculated the compliance rates in part a and they are very similar between treatment and placebo. So does adding a placebo into our experimental design make it more efficient? Based off of our calculations above we see that there is a difference when we compare treatment to placebo vs baseline. This leads us to believe that the experimental design with the placebo would be more efficient and serve as a better comparison.

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.

```
d <- read.dta("./data/Hough_WorkingPaper_2010.dta")
d5 <- data.table::data.table(d)
#d5 = na.omit(d5)
head(d5)
```

```
##    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 first thing that comes to mind is the no-anticipation assumption. The subject should not have an anticipation that running tomorrow will have an impact on tetris scores today. This is inherently hard to do because humans, in general, have anticipations as a result of the actions they perform. And because she is performing this experiment herself and not on someone else, there is no way to mask our treatment or the outputs we care about. Another one would be the no-persistance assumption. We cannot assume that our results will persist or get better as the experiment goes on. This is an interesting assumption in this case especially because the benefits of exercise are seen mainly through persistance.

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

```
ATE_b = mean(na.omit(d5$tetris[d5$run == 1])) - mean(na.omit(d5$tetris[d5$run == 0]))
```

The estimated ATE is: 13613.1

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?

```
d5[, tetris_lag := lag(d5$tetris)]
```

```
model5c = lm(tetris_lag ~ run, data=d5)
```

```
stargazer(model5c)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 18, 2019 - 21:42:00

Table 3:

	Dependent variable:
	tetris_lag
run	645.621
	(4,823.528)
Constant	18,903.830***
	(3,335.778)
Observations	23
R^2	0.001
Adjusted R^2	−0.047
Residual Std. Error	11,555.480 (df = 21)
F Statistic	0.018 (df = 1; 21)
Note:	*p<0.1; **p<0.05; ***p<0.01

We are testing whether or not the anticipation for running the next day effects the tetris scores on the current day. Hence, it is a valid test of the no anticipation assumption. We see that the results for run are not statistically significant and hence the no-anticipation assumption is verified.

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?

```
model5d = lm(tetris ~ run, data=d5)
```

```
stargazer(model5d)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 18, 2019 - 21:42:00

Table 4:

	Dependent variable:
	tetris
run	13,613.100**
	(4,855.626)
Constant	12,806.400***
	(3,708.546)
Observations	24
R^2	0.263
Adjusted R^2	0.230
Residual Std. Error	11,727.450 (df = 22)
F Statistic	7.860** (df = 1; 22)
Note:	*p<0.1; **p<0.05; ***p<0.01

From the model we see that at ATE is greater than two standard errors which leads us to

believe that the impact of running on tetris is 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?

```
model5e1 = lm(gre ~ run, data=d5)
```

```
stargazer(model5e1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 18, 2019 - 21:42:00

Table 5:

	Dependent variable:
	gre
run	−0.175
	(0.185)
Constant	0.818***
	(0.138)
Observations	25
R^2	0.038
Adjusted R^2	−0.004
Residual Std. Error	0.459 (df = 23)
F Statistic	0.898 (df = 1; 23)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
model5e2 = lm(energy ~ run, data=d5)
```

```
stargazer(model5e2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 18, 2019 - 21:42:00

Table 6:

	Dependent variable:
	energy
run	0.071
	(0.441)
Constant	3.000***
	(0.337)
Observations	24
R^2	0.001
Adjusted R^2	−0.044
Residual Std. Error	1.064 (df = 22)
F Statistic	0.026 (df = 1; 22)
Note:	*p<0.1; **p<0.05; ***p<0.01

These hypotheses are not borne out by the data. The models show that the values are statistically insignificant.

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?

The decision to do this is very unethical. Going in to the study she did not know that the tetris variable would be the most impacted as a result of her treatment. She should include the results from all the variables so as to show the details for her experiment in case someone aims to reproduce them. Additionally, including the other variables/data should not impact the results she wants to showcase using the tetris variable. The trap she has fallen for is p-hacking. Which is the misuse of data analysis to find patterns in data that can be presented as statistically significant when in fact there is no real underlying effect.

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?

```
d5[, run_lag := lag(d5$run)]
```

```
model5g = lm(tetris ~ run + run_lag, data=d5)
```

```
stargazer(model5g)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Mar 18, 2019 - 21:42:00

Table 7:

	Dependent variable:
	tetris
run	15,025.950***
	(4,985.500)
run_lag	1,688.645
	(4,947.587)
Constant	11,793.210**
	(4,753.488)
Observations	23
R^2	0.312
Adjusted R^2	0.244
Residual Std. Error	11,740.250 (df = 20)
F Statistic	4.545** (df = 2; 20)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
confint(model5g, "run", level=.95)
```

```
##       2.5 %   97.5 %
## run 4626.38 25425.52
```

```
confint(model5g, "run_lag", level=.95)
```

```
##            2.5 %   97.5 %
## run_lag -8631.84 12009.13
```

The coefficient in this regression is a little over a 1000 points higher than in part d. This delta is insignificant. We can conclude that running does not have long-lasting effects on tetris scores. This validates our no-persistence assumption.

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)?

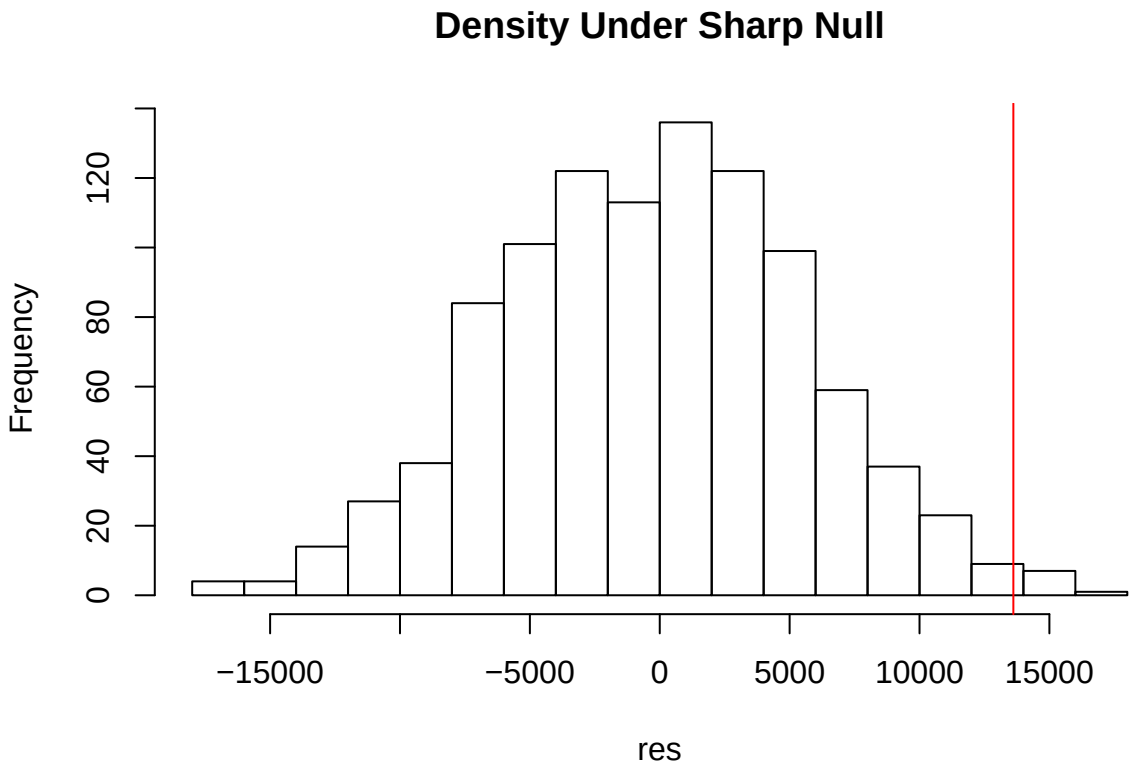
```
# specify number of simulations
n_sims = 1000
# init to NA values
```

```r
res <- rep(NA , n_sims)
# loop and replicate the experiment
for(i in 1:n_sims) {
  d5[, run_rand := sample(d5$run)]
  d5[, run_lag_rand := sample(d5$run_lag)]
  model5h = lm(tetris ~ run_rand + run_lag_rand, data=d5)
  res[i] = model5h$coefficients['run_rand']
}
```

```r
ATE = mean(res)
```

The estimated ATE is: -291.7558963

```r
hist(res, breaks=20, main="Density Under Sharp Null")
abline(v = ATE_b, col = "red")
```




```r
se = sd(res/sqrt(length(res)))
```

The standard error is: 185.0592357

The resulting confidence interval is: -661.8743676, 78.362575

The confidence interval we get for run from part h is much tighter than the interval we got from part g. This implies that there is lower variability which lends credibility to the experiment that it is less likely to obtain these results by chance.