Problem Set #4

Experiment Design: Alex & Daniel

```r
# load packages
library(foreign)
library(data.table)
library(ggplot2)
library(RColorBrewer)
library(lmtest)
library(sandwich)
library(multiwayvcov)
library(AER)
```

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.

Observation	$Y_i(d=0)$	$Y_i(d=1)$	$d_i(z=0)$	$d_i(z=1)$	Type
1	1	0	0	1	Complier
2	1	0	0	1	Complier
3	1	0	0	1	Complier
4	0	3	0	0	Never-Taker
5	0	3	0	0	Never-Taker
6	0	3	0	0	Never-Taker

$$ATE = \frac{-1 + -1 + -1 + 3 + 3 + 3}{6} = 1$$
$$CACE = \frac{-1 + -1 + -1}{3} = -1$$

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 not be the correct sign. Because the ATE is positive, the estimated CACE can only be positive, but we know the true 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?

Both are important. If compliance can be increased through some external measure, it useful to determine the CACE as an estimate of the treatment effect. If not, the set of both groups is a good estimate of the actual treatment effect when implemented over the entire population.

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?

$$ATE = \frac{400}{1000} - \frac{700}{2000} = 0.05$$

$$CACE = \frac{0.05}{\frac{500}{1000}} = 0.1$$

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

$$CACE = \frac{0.05}{\frac{250}{1000}} = 0.2$$

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' exagerated reports make their efforts seem less effective in terms of CACE. The ITT remains the same, regardless of their exageration. They differ. That's why it matters.

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.

```r
library(foreign)
library(data.table)
d <- data.table(read.dta("./data/Guan_Green_CPS_2006.dta"))
d = d[!is.na(turnout),]
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.*

```
# difference in means
(itt = mean(d[treat2 == 1, turnout]) - mean(d[treat2 == 0, turnout]))
```

```
## [1] 0.1319296
```

```
# linear regression
summary(d[, lm(turnout ~ treat2)])
```

```
##
## Call:
## lm(formula = turnout ~ treat2)
##
## 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
```

Both estimate an ITT of 0.1319.

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.

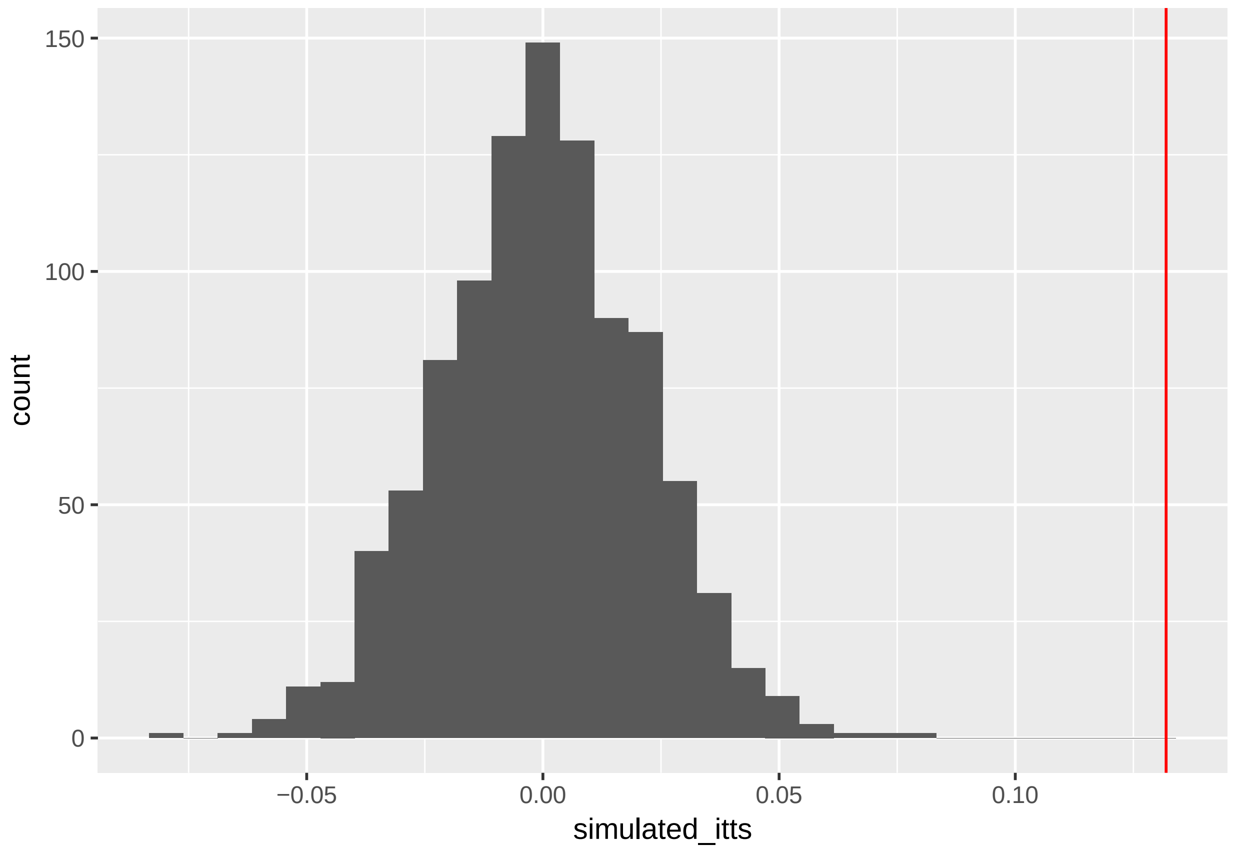
```
d$dormid = as.character(d$dormid)
dorm_ids = unique(d$dormid)

ri = function() {
    treat = sample(0:1, length(dorm_ids), replace = T)
    names(treat) = dorm_ids

    mean(d[treat[dormid] == 1, turnout]) - mean(d[treat[dormid] == 0, turnout])
}
```

```
simulated_itts = replicate(1000, ri())
mean(itt < simulated_itts)
```

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

There are no instances of a simulated ITT being larger than the observed ITT (one sided p-value = 0). We reject the sharp null hypothesis.

```
qplot(simulated_itts, bins = 30) + geom_vline(xintercept = itt, color = 'red')
```



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 contact had any causal effect among compliers.

```
# means
alpha = d[contact == 1, .N] / d[treat2 == 1, .N]
(cace = itt / alpha)
```

```
## [1] 0.1489402
```

```
# 2sls
m = d[, ivreg(turnout ~ contact | treat2)]
(co = coeftest(m, vcov. = cluster.vcov(m, d$dormid)))
```

```
##
## 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
```

Both methods predict a CACE of 0.1489. The 2-stage least squares estimates a standard error of 0.0263. The p-value is very low, so it does seem like `contact` had an effect.

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.*)

```
d = data.table(
    Z = c(rep('baseline', 2572), rep('treatment', 486+2086), rep('placebo', 470+2109)),
    D = c(rep(0, 2572), rep(1, 486), rep(0, 2086), rep(1, 470), rep(0, 2109)),
    Y = c(
        rep(1, ceiling(.3122 * 2572)), rep(0, floor((1-.3122) * 2572)),
        rep(1, ceiling(.3909 * 486)), rep(0, floor((1-.3909) * 486)),
        rep(1, ceiling(.3274 * 2086)), rep(0, floor((1-.3274) * 2086)),
        rep(1, ceiling(.2979 * 470)), rep(0, floor((1-.2979) * 470)),
        rep(1, ceiling(.3215 * 2109)), rep(0, floor((1-.3215) * 2109))
    )
)
```

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).

```
treatment = d[Z == 'treatment', ]
placebo = d[Z == 'placebo', ]
baseline = d[Z == 'baseline', ]

# estimated alpha with treatment
(alpha_treatment = mean(treatment$D))
```

```
## [1] 0.188958
```

```
# estimated alpha with placebo
(alpha_placebo = mean(placebo$D))
```

```
## [1] 0.1822412
```

```
# test of statistical difference
t.test(treatment$D, placebo$D)
```

```
##
```

```
##  Welch Two Sample t-test
## 
## data:  treatment$D and placebo$D
## 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
```

The two estimates are not significantly different. The experiment defines sucessful delivery of treatment when contact is established. What the canvasser talks about after contact, has no effect on whether the contact happens.

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
t.test(treatment[D == 0, Y], placebo[D == 0, Y])
```

```
## 
##  Welch Two Sample t-test
## 
## data:  treatment[D == 0, Y] and placebo[D == 0, Y]
## t = 0.37803, df = 4192, p-value = 0.7054
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02288769  0.03382242
## sample estimates:
## mean of x mean of y 
## 0.3274209 0.3219535
```

The data suggests the same rate of turnout for Never Takers from both groups. This comparison says that being assigned to either group has no effect on Never Takers' turnout.

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
itt = placebo[, mean(Y)] - baseline[, mean(Y)]
(cace = itt / alpha_placebo)
```

```
## [1] 0.0315203
```

```r
(co = coeftest(d[Z == 'placebo' | Z == 'baseline', ivreg(Y ~ D | Z == 'placebo')]))
```

```
## 
## t test of coefficients:
## 
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.3122084  0.0091654 34.0638   <2e-16 ***
## D           0.0315203  0.0710763  0.4435   0.6574    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated CACE is 0.0315 (0.0711). The estimate is insignificant, which is consistent with the assumption that the placebo has 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
# means
itt = mean(d[Z == 'treatment', Y]) - mean(d[Z == 'baseline', Y])
(cace = itt / alpha_treatment)
```

```
## [1] 0.1440329
```

```r
# 2SLS
m = d[Z == 'treatment' | Z == 'baseline', ivreg(Y ~ D | Z == 'treatment')]
(co = coeftest(m))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.3122084  0.0092433  33.777  < 2e-16 ***
## D           0.1440329  0.0691793   2.082  0.03739 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated CACE is 0.144 (0.0692).

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

```r
# means
(itt2 = treatment[D == 1, mean(Y)] - placebo[D == 1, mean(Y)])
```

```
## [1] 0.0909465
```

```r
# regression
m = d[(Z == 'treatment' | Z == 'placebo') & D == 1, lm(Y ~ Z)]
(co = coeftest(m))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.300000   0.021868 13.7187 < 2.2e-16 ***
## Ztreatment  0.090947   0.030670  2.9653  0.003099 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated CACE is 0.0909 (0.0307). It is much smaller compared to part d. This could suggest that there may be some placebo effect on Y. Although part c showed no significant effect, the estimate was still positive.

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?

Because the compliance rate is so low, the placebo design should be better at estimating the CACE.

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 = data.table(read.dta("./data/Hough_WorkingPaper_2010.dta"))
#d = d[!is.na(tetris),]
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?

The two assumptions (along with randomization, exclusion, and non-interference) are no-anticipation and no-persistence. The former is very plausible, especially if she assigns whether she runs on a particular day after she plays Tetris on the previous day. The latter is probably satisfied as well. Unless she is going on extremely long runs or gets injured, it is unlikely any effects on her Tetris playing persists to the next day. The fact that that she is performing the study on herself could only be problematic if she is hoping for some specific effect, e.g. if she hopes that running makes her a better Tetris player, she might try harder on days that she runs.

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

```
m = d[, lm(tetris ~ run)]
(co = coeftest(m))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12806.4     3708.5  3.4532 0.002264 **
## run          13613.1     4855.6  2.8036 0.010351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated effect is 13613.1 (4855.626).

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?

```
m = lm(d[1:.N-1, tetris] ~ d[2:.N, run])
(co = coeftest(m))
```

```
##
## t test of coefficients:
```

```
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  18903.83    3335.78  5.6670 1.264e-05 ***
## d[2:.N, run]   645.62    4823.53  0.1338    0.8948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated effect is 645.6212 (4823.528). This shows no evidence of anticipation. Whether she runs the next day has no effect on the Tetris score in the previous day.

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?

```
m = d[, lm(tetris ~ run)]
(co = coeftest(m))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12806.4     3708.5  3.4532 0.002264 **
## run          13613.1     4855.6  2.8036 0.010351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated effect is 13613.1 (4855.626). The estimate 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?

```
m = d[, lm(energy ~ run)]
(co = coeftest(m))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 3.000000   0.336618  8.9122 9.396e-09 ***
## run         0.071429   0.440736  0.1621    0.8727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated effect is 0.07142857 (0.4407364). Energy is not significantly affected by running.

```
m = d[, lm(gre ~ run)]
(co = coeftest(m))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.81818    0.13846  5.9089 5.048e-06 ***
## run         -0.17532    0.18503 -0.9475    0.3532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimated effect is -0.1753247 (0.1850315). GRE scores are not significantly affected by running.

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?

She could have gotten the Tetris results by chance. It is useful to report her other regressions, to show that she did not go fishing for results. In this case, since there are only three regressions (and the Tetris one was done first), the results would be significant even after corrections for multiple regressions.

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
d = d[, ran_yesterday := c(NA, d[1:.N-1, run])]
m = d[, lm(tetris ~ run + run*ran_yesterday)]
(co = coeftest(m))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         12150.8     6021.0  2.0181  0.05793 .
## run                 14464.1     7547.7  1.9164  0.07049 .
## ran_yesterday        1092.7     7773.1  0.1406  0.88968
## run:ran_yesterday    1038.6    10261.8  0.1012  0.92045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ate = co[2,1] # for part h
```

The estimated effect is 14464.11 (7547.72). The estimate and the standard error increased, but this could have been due to the loss of statistical power from so few days having `run == 1 & ran_yesterday == 1`. Importantly, the coefficients on `ran_yesterday` and the interaction term are low and insignificant, confirming the no anticipation effect.

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)?

```r
N = d[, .N]

ri = function() {
    run = sample(c(0,1), N, replace = T)
    m = lm(d[2:N, tetris] ~ run[1:N-1] + run[2:N])
    coeftest(m)[2,1]
}
```

```
simulated_ates = replicate(1000, ri())
mean(simulated_ates > ate)
```

```
## [1] 0.004
```

```
# confidence interval for model from part g
(ci = confint(m)[2, ])
```

```
##     2.5 %    97.5 %
## -1333.452 30261.667
```

```
# simulated confidence interval
(simulated_ci = quantile(simulated_ates, probs = c(.025, .975)))
```

```
##      2.5%     97.5%
## -11298.80  12175.54
```

```
# confidence interval range
unname(ci[2] - ci[1])
```

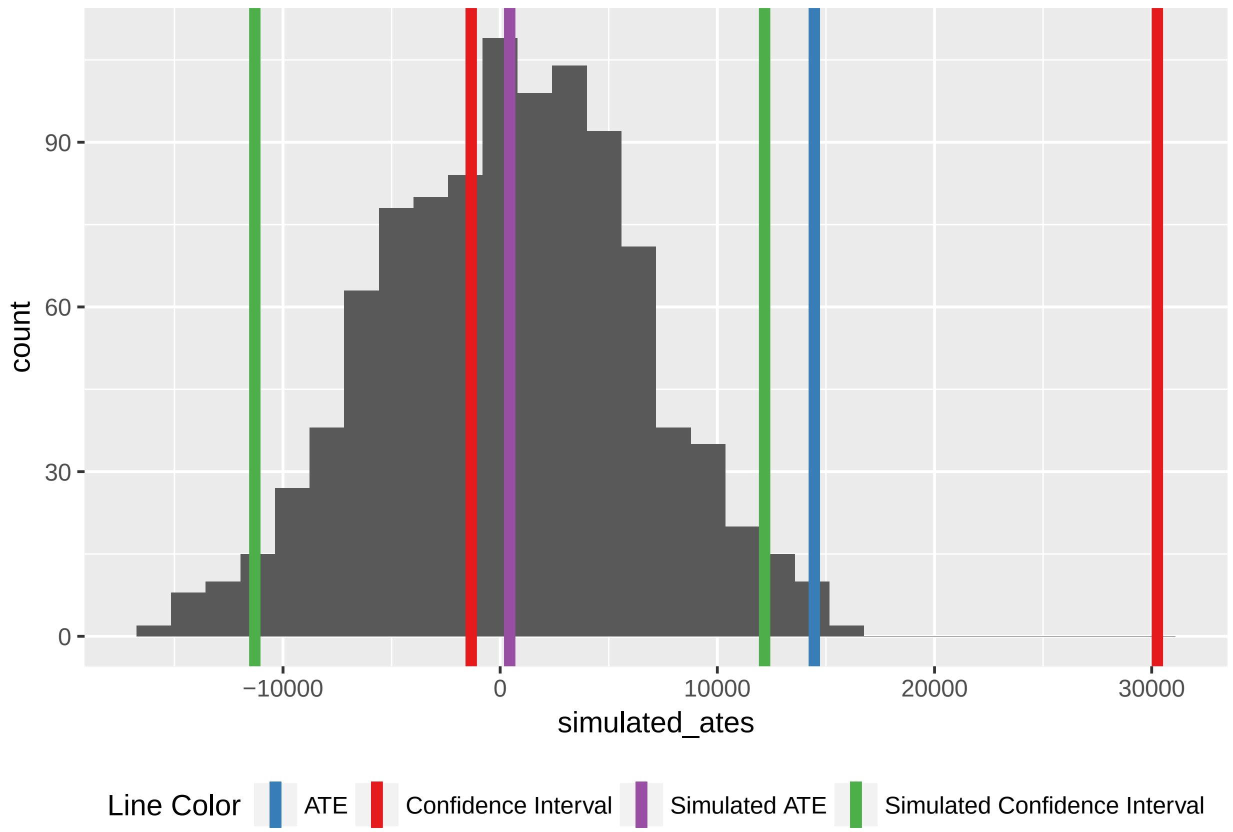
```
## [1] 31595.12
```

```
# simulated confidence interval range
unname(simulated_ci[2] - simulated_ci[1])
```

```
## [1] 23474.34
```

The simulated confidence interval is wider than the one from part g.

```
cols = brewer.pal(4, 'Set1')
names(cols) = c('Confidence Interval', 'ATE', 'Simulated Confidence Interval', 'Simulated ATE')
ggplot(mapping = aes(x = simulated_ates)) +
    geom_histogram(bins = 30) +
    geom_vline(aes(xintercept = ci[1], color = 'Confidence Interval'), size = 2) +
    geom_vline(aes(xintercept = ci[2], color = 'Confidence Interval'), size = 2) +
    geom_vline(aes(xintercept = ate, color = 'ATE'), size = 2) +
    geom_vline(aes(xintercept = simulated_ci[1], color = 'Simulated Confidence Interval'), size = 2) +
    geom_vline(aes(xintercept = simulated_ci[2], color = 'Simulated Confidence Interval'), size = 2) +
    geom_vline(aes(xintercept = mean(simulated_ates), color = 'Simulated ATE'), size = 2) +
    scale_colour_manual(name="Line Color", values=cols) +
    theme(legend.position = 'bottom')
```


count
90
60
30
0
−10000
0
10000
20000
30000
simulated_ates
Line Color
ATE
Confidence Interval
Simulated ATE
Simulated Confidence Interval