Problem Set 5

Field Experiments

1. Online advertising natural experiment.

These are simulated data (closely, although not entirely) based on a real example, adopted from Randall Lewis' dissertation at MIT.

Problem Setup

Imagine Yahoo! sells homepage ads to advertisers that are quasi-randomly assigned by whether the user loads the Yahoo! homepage (www.yahoo.com) on an even or odd second of the day. More specifically, the setup is as follows. On any given week, Monday through Sunday, two ad campaigns are running on Yahoo!'s homepage. If a user goes to www.yahoo.com during an even second that week (e.g., Monday at 12:30:58pm), the ads for the advertiser are shown. But if the user goes to www.yahoo.com during an odd second during that week (e.g., Monday at 12:30:59), the ads for other products are shown. (If a user logs onto Yahoo! once on an even second and once on an odd second, they are shown the first of the campaigns the first time and the second of the campaigns the second time. Assignment is not persistent within users.)

This natural experiment allows us to use the users who log onto Yahoo! during odd seconds/the ad impressions from odd seconds as a randomized control group for users who log onto Yahoo! during even seconds/the ad impressions from even seconds. (We will assume throughout the problem there is no effect of viewing advertiser 2's ads, from odd seconds, on purchases for advertiser 1, the product advertised on even seconds.)

Imagine you are an advertiser who has purchased advertising from Yahoo! that is subject to this randomization on two occasions. Here is a link to (fake) data on 500,000 randomly selected users who visited Yahoo!'s homepage during each of your two advertising campaigns, one you conducted for product A in March and one you conducted for product B in August (~250,000 users for each of the two experiments). Each row in the dataset corresponds to a user exposed to one of these campaigns.

```
library(data.table)
library(stargazer)
library(dplyr)
library(gmodels)
library(descr)
library(multiwayvcov)
library(lmtest)
```

```
d1 <- fread('./data/ps5_no1.csv')
head(d1)
```

```
##    product_b total_ad_exposures_week1 treatment_ad_exposures_week1 week0
## 1:         1                        4                             3   5.5
## 2:         1                        1                             1   6.2
## 3:         1                        3                             1   0.0
## 4:         0                        5                             0   0.0
## 5:         0                        1                             1   7.6
## 6:         1                        4                             4   6.3
##    week1 week2 week3 week4 week5 week6 week7 week8 week9 week10
## 1:   6.2   0.0   0.0   0.0   0.0   0.0     0   9.7   4.1    0.0
## 2:   0.0   8.6   2.4   0.0   7.4   0.0     0   0.0   5.7    0.0
## 3:   5.3   0.0   8.1   7.8   3.3   0.0     0   9.4   0.0    0.0
```

```
## 4:   4.1   0.0   8.8   5.8   5.9   0.0     0   0.0   9.6    0.0
## 5:   3.6   4.6   5.5   7.2   7.1   0.0     0   0.0   0.0    0.0
## 6:   5.5   9.8   5.0   0.0   0.0   7.7     0  11.0   4.8    6.9
```

The variables in the dataset are described below:

- **product_b**: an indicator for whether the data is from your campaign for product A (in which case it is set to 0), sold beginning on March 1, or for product B, sold beginning on August 1 (in which case it is set to 1). That is, there are two experiments in this dataset, and this variable tells you which experiment the data belong to.
- **treatment_ad_exposures_week1**: number of ad exposures for the product being advertised during the campaign. (One can also think of this variable as "number of times each user visited Yahoo! homepage on an even second during the week of the campaign.")
- **total_ad_exposures_week1**: number of ad exposures on the Yahoo! homepage each user had during the ad campaign, which is the sum of exposures to the "treatment ads" for the product being advertised (delivered on even seconds) and exposures to the "control ads" for unrelated products (delivered on odd seconds). (One can also think of this variable as "total number of times each user visited the Yahoo! homepage during the week of the campaign.")
- **week0**: For the treatment product, the revenues from each user in the week prior to the launch of the advertising campaign.
- **week1**: For the treatment product, the revenues from each user in the week during the advertising campaign. The ad campaign ends on the last day of week 1.
- **week2-week10**: Revenue from each user for the treatment product sold in the weeks subsequent to the campaign. The ad campaign was not active during this time.

Simplifying assumptions you should make when answering this problem:

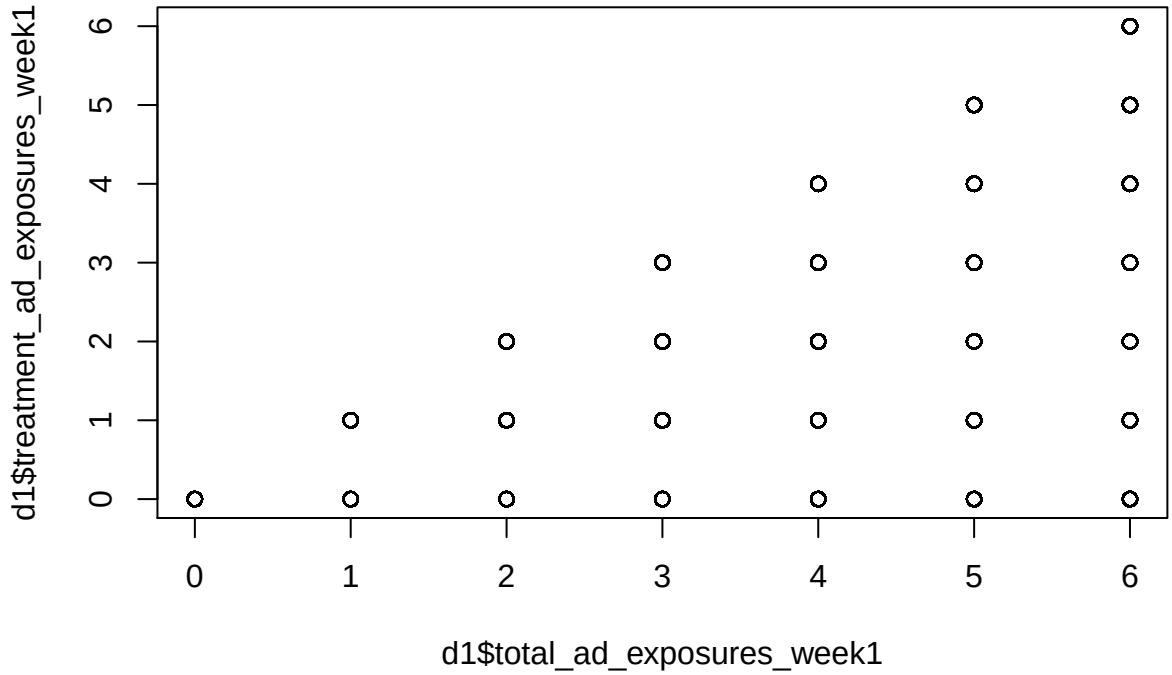
- The effect of treatment ad exposures on purchases is linear. That is, the first exposure has the same effect as the second exposure.
- There is no effect of being exposed to the odd-second ads on purchases for the product being advertised on the even second.
- Every Yahoo! user visits the Yahoo! home page at most six times a week.
- You can assume that treatment ad exposures do not cause changes in future ad exposures. That is, assume that getting a treatment ad at 9:00am doesn't cause you to be more (or less) likely to visit the Yahoo home pages on an even second that afternoon, or on subsequent days.

Questions to Answer

a. Run a crosstab of total_ad_exposures_week1 and treatment_ad_exposures_week1 to sanity check that the distribution of impressions looks as it should. Does it seem reasonable? Why does it look like this? (No computation required here, just a brief verbal response.)

	d1$treatment_ad_exposures_week1							
d1$total_ad_exposures_week1	0	1	2	3	4	5	6	Total
0	61182	0	0	0	0	0	0	61182
	116821.471	17905.157	14152.376	8138.797	3249.376	809.683	91.773	
	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.122
	0.445	0.000	0.000	0.000	0.000	0.000	0.000	
	0.122	0.000	0.000	0.000	0.000	0.000	0.000	
1	36754	37215	0	0	0	0	0	73969
	13215.670	11195.497	17110.213	9839.800	3928.494	978.906	110.954	
	0.497	0.503	0.000	0.000	0.000	0.000	0.000	0.148
	0.267	0.254	0.000	0.000	0.000	0.000	0.000	
	0.074	0.074	0.000	0.000	0.000	0.000	0.000	
2	21143	42036	20965	0	0	0	0	84144
	174.506	12310.223	115.776	11193.340	4468.888	1113.562	126.216	
	0.251	0.500	0.249	0.000	0.000	0.000	0.000	0.168
	0.154	0.287	0.181	0.000	0.000	0.000	0.000	
	0.042	0.084	0.042	0.000	0.000	0.000	0.000	
3	10683	32073	32314	10726	0	0	0	85796
	7075.933	1931.760	7832.885	41.365	4556.626	1135.424	128.694	
	0.125	0.374	0.377	0.125	0.000	0.000	0.000	0.172
	0.078	0.219	0.279	0.161	0.000	0.000	0.000	
	0.021	0.064	0.065	0.021	0.000	0.000	0.000	
4	5044	20003	30432	20223	5115	0	0	80817
	13293.701	562.797	7369.876	8345.761	157.732	1069.532	121.225	
	0.062	0.248	0.377	0.250	0.063	0.000	0.000	0.162
	0.037	0.137	0.263	0.304	0.193	0.000	0.000	
	0.010	0.040	0.061	0.040	0.010	0.000	0.000	
5	2045	10563	20970	20793	10293	2131	0	66795
	14516.852	4129.721	1971.561	15957.423	12826.563	1759.228	100.192	
	0.031	0.158	0.314	0.311	0.154	0.032	0.000	0.134
	0.015	0.072	0.181	0.313	0.388	0.322	0.000	
	0.004	0.021	0.042	0.042	0.021	0.004	0.000	
6	729	4437	10977	14771	11147	4486	750	47297
	11597.078	6389.955	0.121	11427.382	29683.865	23804.879	6499.567	
	0.015	0.094	0.232	0.312	0.236	0.095	0.016	0.095
	0.005	0.030	0.095	0.222	0.420	0.678	1.000	
	0.001	0.009	0.022	0.030	0.022	0.009	0.002	
Total	137580	146327	115658	66513	26555	6617	750	5e+05
	0.275	0.293	0.231	0.133	0.053	0.013	0.002	

```r
plot(d1$total_ad_exposures_week1, d1$treatment_ad_exposures_week1)
```



The distribution appears to look reasonable. According to our plot, the treatment ads seem to be a subset of the total ads which makes sense.

b. Your colleague proposes the code printed below to analyze this experiment: `lm(week1 ~ treatment_ad_exposures_week1, data)` You are suspicious. Run a placebo test with the prior week's purchases as the outcome and report the results. Did the placebo test "succeed" or "fail"? Why do you say so?

```
# run the experiment
model1b = lm(week1 ~ treatment_ad_exposures_week1, data=d1)
```

```
stargazer(model1b)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:40

Table 1:

	Dependent variable:
	week1
treatment_ad_exposures_week1	0.299***
	(0.003)
Constant	1.615***
	(0.006)
Observations	500,000
R^2	0.018
Adjusted R^2	0.018
Residual Std. Error	2.781 (df = 499998)
F Statistic	9,085.783*** (df = 1; 499998)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
# run the placebo
model1b_placebo = lm(week0 ~ treatment_ad_exposures_week1, data = d1)
```

```r
stargazer(model1b_placebo)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:41

Table 2:

	Dependent variable:
	week0
treatment_ad_exposures_week1	0.263***
	(0.003)
Constant	1.670***
	(0.006)
Observations	500,000
R^2	0.014
Adjusted R^2	0.014
Residual Std. Error	2.796 (df = 499998)
F Statistic	6,955.202*** (df = 1; 499998)
Note:	*p<0.1; **p<0.05; ***p<0.01

The placebo test fails as it shows the treatment to be statistically significant. This means that the test proposed by our colleague is incorrect!

c. The placebo test suggests that there is something wrong with our experiment or our data analysis. We suggest looking for a problem with the data analysis. Do you see something that might be spoiling the randomness of the treatment variable? How can you improve your analysis to get rid of this problem? Why does the placebo test turn out the way it does? What one thing needs to be done to analyze the data correctly? Please provide a brief explanation of why, not just what needs to be done. (*Note: This question, and verifying that you answered it correctly in part d below, may require some thinking. If we find many people can't figure it out, we will post another hint in a few days.*)

Our randomization could be getting spoiled because we do not account for the fact that more frequent shoppers will be exposed to more ads in general. We could improve our results by finding a way to better differentiate treatment and no-treatment subjects. This may be hinting at there being a flaw in our delivery mechanism. Since more frequent shoppers are exposed to both treatment and non-treatment ads, there is some overlap that forms between the two resulting in the placebo failing. In order to analyze the data correctly we need to regress on both treatment and total ads.

d. Implement the procedure you propose from part (c), run the placebo test for the Week 0 data again, and report the results. (This placebo test should pass; if it does not, re-evaluate your strategy before wasting time proceeding.)

```
model1d = lm(week0 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1, data = d1)
```

```
stargazer(model1d)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:41

Table 3:

	Dependent variable:
	week0
treatment_ad_exposures_week1	−0.002
	(0.005)
total_ad_exposures_week1	0.245***
	(0.003)
Constant	1.345***
	(0.007)
Observations	500,000
R^2	0.026
Adjusted R^2	0.026
Residual Std. Error	2.779 (df = 499997)
F Statistic	6,555.756*** (df = 2; 499997)
Note:	*p<0.1; **p<0.05; ***p<0.01

The placebo test passes as it shows that the treatment is not statistically significant.

e. Now estimate the causal effect of each ad exposure on purchases during the week of the campaign itself using the same technique that passed the placebo test in part (d).

```
model1e = lm(week1 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1, data = d1)
```

```
stargazer(model1e)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:41

Table 4:

	Dependent variable:
	week1
treatment_ad_exposures_week1	0.056***
	(0.005)
total_ad_exposures_week1	0.224***
	(0.003)
Constant	1.318***
	(0.007)
Observations	500,000
R^2	0.028
Adjusted R^2	0.028
Residual Std. Error	2.767 (df = 499997)
F Statistic	7,153.262*** (df = 2; 499997)
Note:	*p<0.1; **p<0.05; ***p<0.01

Each treatment ad exposure causes a 0.0563399 increase in revenue for week 1.

f. The colleague who proposed the specification in part (b) challenges your results – they make the campaign look less successful. Write a paragraph that a layperson would understand about why your estimation strategy is superior and his/hers is biased.

It's not about what the data looks like. It's about what the data actually means. If we had moved forward with the colleague's suggestion, we would be doing a disservice to our client by providing fluffed data that is false. Not to mention how that would be completely unethical given we know that is it false. Their strategy is not measuring the actual impact of treatment and is instead relying on frequent shoppers to make it seem like ad campaign is successful. We have root caused the problem, found an appropriate fix, and have the correct representation of the data and the actual impact of the ad campaign.

g. Estimate the causal effect of each treatment ad exposure on purchases during and after the campaign, up until week 10 (so, total purchases during weeks 1 through 10).

```
# compute and add total to the data table
d1$tot = rowSums(d1[,c(5,6,7,8,9,10,11,12,13,14)])
```

```
model1g = lm(tot ~ treatment_ad_exposures_week1 + total_ad_exposures_week1, data = d1)
```

```
stargazer(model1g)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:41

Table 5:

	Dependent variable:
	tot
treatment_ad_exposures_week1	0.013
	(0.018)
total_ad_exposures_week1	2.228***
	(0.012)
Constant	17.151***
	(0.028)
Observations	500,000
R^2	0.132
Adjusted R^2	0.132
Residual Std. Error	10.555 (df = 499997)
F Statistic	38,038.290*** (df = 2; 499997)
Note:	*p<0.1; **p<0.05; ***p<0.01

Each treatment ad exposure causes a 0.0127386 increase in revenue for the campaign.

h. Estimate the causal effect of each treatment ad exposure on purchases only after the campaign. That is, look at total purchases only during week 2 through week 10, inclusive.

```
d1$tot_post_camp = rowSums(d1[,c(6,7,8,9,10,11,12,13,14)])
```

```
model1h = lm(tot_post_camp ~ treatment_ad_exposures_week1 + total_ad_exposures_week1, data = d1)
```

```
stargazer(model1h)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:42

Table 6:

	Dependent variable:
	tot_post_camp
treatment_ad_exposures_week1	−0.044***
	(0.017)
total_ad_exposures_week1	2.004***
	(0.011)
Constant	15.833***
	(0.027)
Observations	500,000
R^2	0.115
Adjusted R^2	0.115
Residual Std. Error	10.111 (df = 499997)
F Statistic	32,613.680*** (df = 2; 499997)
Note:	*p<0.1; **p<0.05; ***p<0.01

Each treatment ad exposure causes a -0.0436013 decrease in revenue for the campaign.

i. Tell a story that could plausibly explain the result from part (h).

Assuming the ad campaign is focused on selling and promoting certain products it makes sense that we would see an overall negative impact after the first week of the campaign. My sister's are shopaholics. They know precisely when a a sale will be and which items they will be purchasing. As soon as the sale goes live, they will go and make the purchase, usually within the first week. After that it's on to the next sale and so on and so forth. The point being that people effected by the treatment will most likely act on it within the first week. As the ad campaign continues the appeal for buying starts going down. They may have sold out of the product(s) already so there are none left.

j. Test the hypothesis that the ads for product B are more effective, in terms of producing additional revenue in week 1 only, than are the ads for product A. (*Hint: The easiest way to do this is to throw all of the observations into one big regression and specify that regression in such a way that it tests this hypothesis.*) (*Hint 2: There are a couple defensible ways to answer this question that lead to different answers. Don't stress if you think you have an approach you can defend.*)

```
model1j = lm (week1 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1 + product_b, data=d1)
```

```
stargazer(model1j)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:42

Table 7:

	Dependent variable:
	week1
treatment_ad_exposures_week1	0.056***
	(0.005)
total_ad_exposures_week1	0.211***
	(0.003)
product_b	0.155***
	(0.008)
Constant	1.294***
	(0.007)
Observations	500,000
R^2	0.028
Adjusted R^2	0.028
Residual Std. Error	2.766 (df = 499996)
F Statistic	4,884.081*** (df = 3; 499996)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see that being product b has a 0.1545422 impact proving that ads for product b are more effective.

k. You notice that the ads for product A included celebrity endorsements. How confident would you be in concluding that celebrity endorsements increase the effectiveness of advertising at stimulating immediate purchases?

Based off of the table above we are not very confident. In the table we are looking only at week 1 and see that product b ads are more effective.

2. Vietnam Draft Lottery

A famous paper by Angrist exploits the randomized lottery for the Vietnam draft to estimate the effect of education on wages. (*Don't worry about reading this article, it is just provided to satisfy your curiosity; you can answer the question below without referring to it. In fact, it may be easier for you not to, since he has some complications to deal with that the simple data we're giving you do not.*)

Problem Setup

Angrist's idea is this: During the Vietnam era, draft numbers were determined randomly by birth date – the army would literally randomly draw birthdays out of a hat, and those whose birthdays came up sooner were higher up on the list to be drafted first. For example, all young American men born on May 2 of a given year might have draft number 1 and be the first to be called up for service, followed by November 13 who would get draft number 2 and be second, etc. The higher-ranked (closer to 1) your draft number, the likelier it was you would be drafted.

We have generated a fake version of this data for your use in this project. You can find real information (here)[https://www.sss.gov/About/History-And-Records/lotter1]. While we're defining having a high draft number as falling at 80, in reality in 1970 any number lower than 195 would have been a "high" draft number, in 1971 anything lower than 125 would have been "high".

High draft rank induced many Americans to go to college, because being a college student was an excuse to avoid the draft – so those with higher-ranked draft numbers attempted to enroll in college for fear of being drafted, whereas those with lower-ranked draft numbers felt less pressure to enroll in college just to avoid the draft (some still attended college regardless, of course). Draft numbers therefore cause a natural experiment in education, as we now have two randomly assigned groups, with one group having higher mean levels of education, those with higher draft numbers, than another, those with lower draft numbers. (In the language of econometricians, we say the draft number is "an instrument for education," or that draft number is an "instrumental variable.")

Some simplifying assumptions:

- Suppose that these data are a true random sample of IRS records and that these records measure every living American's income without error.
- Assume that the true effect of education on income is linear in the number of years of education obtained.
- Assume all the data points are from Americans born in a single year and we do not need to worry about cohort effects of any kind.

Questions to Answer

```
# read in the data
d2 = fread('./data/ps5_no2.csv')
head(d2)
```

```
##    draft_number years_education    income
## 1:          267              16  44573.90
## 2:          357              13  10611.75
## 3:          351              19 165467.80
## 4:          205              16  71278.40
## 5:           42              19  54445.09
## 6:          240              11  32059.12
```

a. Suppose that you had not run an experiment. Estimate the "effect" of each year of education on income as an observational researcher might, by just running a regression of years of education on income (in R-ish, `income ~ years_education`). What does this naive regression suggest?

```
model2a = lm(income ~ years_education, data=d2)
```

```
stargazer(model2a)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:42

Table 8:

	Dependent variable:
	income
years_education	5,750.480***
	(83.340)
Constant	−23,354.640***
	(1,252.740)
Observations	19,567
R^2	0.196
Adjusted R^2	0.196
Residual Std. Error	26,592.180 (df = 19565)
F Statistic	4,761.015*** (df = 1; 19565)
Note:	*p<0.1; **p<0.05; ***p<0.01

Our naive model suggests that every year of education results in 5750.4796478 increase in income and it is statistically significant.

b. Continue to suppose that we did not run the experiment, but that we saw the result that you noted in part (a). Tell a concrete story about why you don't believe that observational result tells you anything causal.

Tim is the son of Tom. Tom is currently the manager of a very lucrative family business and is set to retire in four years. Tim has just joined a 4-year party college and decided to major in psychology. Tim graduates, Tom retires, and Tim takes over the family business and decides to run it seriously as he has grown tired of partying. Tim now makes significanlty more than any psychologists that he graduated with. This story is meant to imply that there can be many other factors at play when comparing income to education. What field did someone study (engineers will probably earn more than artists on average)? Is there any family wealth? etc... That is exactly why the first model is naive as it ignores, or fails to account, for these.

c. Now, let's get to using the natural experiment. We will define "having a high-ranked draft number" as having a draft number of 80 or below (1-80; numbers 81-365, for the remaining 285 days of the year, can be considered "low-ranked"). Create a variable in your dataset indicating whether each person has a high-ranked draft number or not. Using regression, estimate the effect of having a high-ranked draft number, the dummy variable you've just created, on years of education obtained. Report the estimate and a correctly computed standard error. (*Hint: Pay special attention to calculating the correct standard errors here. They should match how the draft is conducted.)

```
# add variable to high draft number
d2[, high_draft := ifelse(d2$draft_number < 81, 1, 0)]
head(d2)
```

```
##    draft_number years_education    income high_draft
## 1:          267              16  44573.90          0
## 2:          357              13  10611.75          0
## 3:          351              19 165467.80          0
## 4:          205              16  71278.40          0
## 5:           42              19  54445.09          1
## 6:          240              11  32059.12          0
```

```
model2c = lm(years_education ~ high_draft, d2)
```

```
stargazer(model2c)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:42

Table 9:

	Dependent variable:
	years_education
high_draft	2.126***
	(0.038)
Constant	14.434***
	(0.017)
Observations	19,567
R^2	0.138
Adjusted R^2	0.138
Residual Std. Error	2.117 (df = 19565)
F Statistic	3,145.132*** (df = 1; 19565)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
# compute cluster se's
model2c_vcov = cluster.vcov(model2c, ~ draft_number)
coeftest(model2c, model2c_vcov)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 14.434305   0.017703 815.345 < 2.2e-16 ***
## high_draft   2.125756   0.038188  55.666 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2c_cse = sqrt(diag(model2c_vcov))
```

This model suggests that having a high draft number results in 2.1257562 (0.0381878) more years of education and it is statistically significant.

d. Using linear regression, estimate the effect of having a high-ranked draft number on income. Report the estimate and the correct standard error.

```
model2d = lm(income ~ high_draft, data=d2)
```

```
stargazer(model2d)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Apr 13, 2019 - 12:18:42

Table 10:

	Dependent variable:
	income
high_draft	6,637.554***
	(528.703)
Constant	60,761.890***
	(235.917)
Observations	19,567
R^2	0.008
Adjusted R^2	0.008
Residual Std. Error	29,532.970 (df = 19565)
F Statistic	157.613*** (df = 1; 19565)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
model2d_vcov = cluster.vcov(model2d, ~ draft_number)
coeftest(model2d, model2d_vcov)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 60761.89     244.36 248.656 < 2.2e-16 ***
## high_draft   6637.55     511.90  12.966 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2d_cse = sqrt(diag(model2d_vcov))
```

This model suggests that having a high draft number results in 6637.554244 (511.899229) higher income.

e. Divide the estimate from part (d) by the estimate in part (c) to estimate the effect of education on income. This is an instrumental-variables estimate, in which we are looking at the "clean" variation in both education and income that is due to the draft status, and computing the slope of the income-education line as "clean change in Y" divided by "clean change in X". What do the results suggest?

```
edu_effect = model2d$coefficients[2]/model2c$coefficients[2]
```

The estimated effect of education on income is: 3122.4437939

f. Natural experiments rely crucially on the "exclusion restriction" assumption that the instrument (here, having a high draft rank) cannot affect the outcome (here, income) in any other way except through its effect on the "endogenous variable" (here, education). Give one reason this assumption may be violated – that is, why having a high draft rank could affect individuals' income other than because it nudges them to attend school for longer.

It seems like a lot of folks, especially with high draft numbers, avoided joining the war. This

could lead to more open positions and more chances to move up for those that got drafted and went or wanted to go. This resulted in them moving up the chain faster and getting higher income. Another scenario could be that drafted soldiers came back with valuable experience and started their own security consulting agencies which are fairly lucrative given the niche experience required.

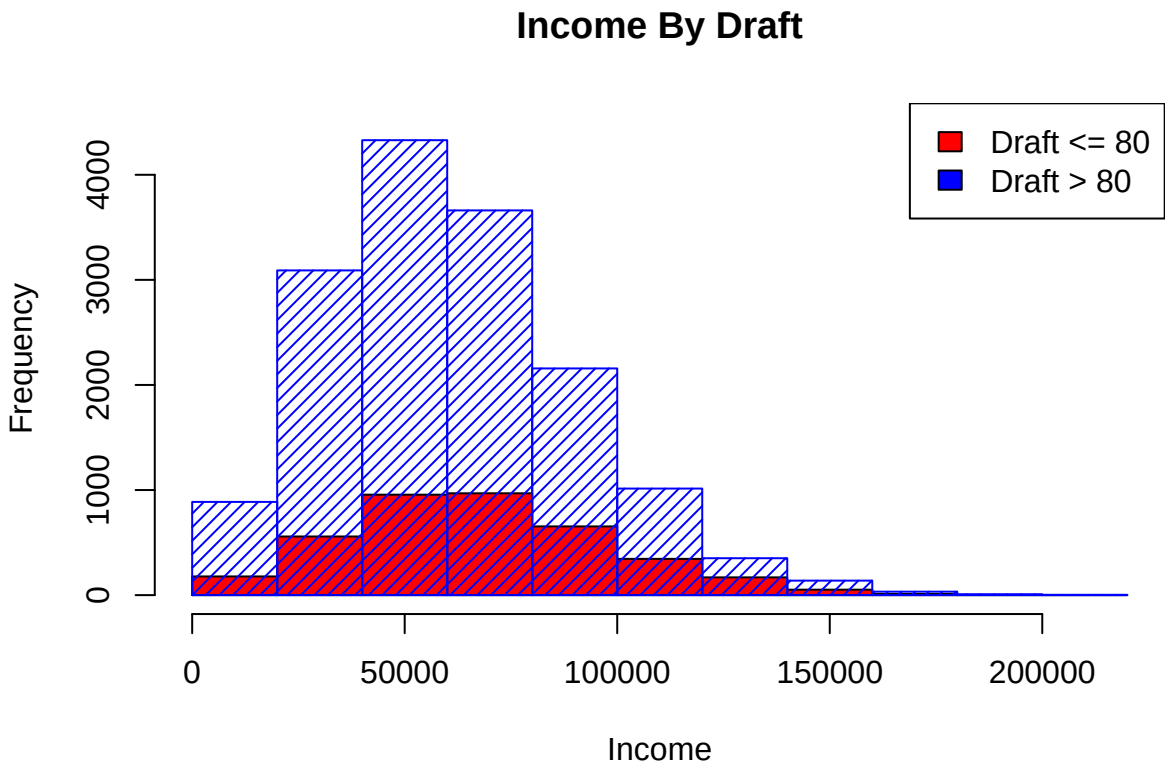
g. Conduct a test for the presence of differential attrition by treatment condition. That is, conduct a formal test of the hypothesis that the "high-ranked draft number" treatment has no effect on whether we observe a person's income. (Note, that an earning of \$0 *actually* means they didn't earn any money.)

```
d2_groups = group_by(d2, high_draft)
summarize(d2_groups, count=n())
```

```
## # A tibble: 2 x 2
##   high_draft count
##        <dbl> <int>
## 1          0 15671
## 2          1  3896
```

We see that there is almost a 5x delta between high draft and low draft.

```
# Histograms
hist(d2[which(d2$high_draft == 1) ,]$income,  ylim = c(0, 4500), col = "red", main = "Income By Draft",
hist(d2[which(d2$high_draft == 0),]$income, ylim = c(0, 4500), col= "blue", add=T, density = 20)
legend('topright',c('Draft <= 80','Draft > 80'), fill = c("red", "blue"))
```



We can see that there is a very big difference between the number of observations for our definition of high draft and not high draft. This is a little concerning but again we defined the draft threshold. We cannot conclusively say that we wee the presence of differential attrition

h. Tell a concrete story about what could be leading to the result in part (g).'

Attrition in this case seems to imply that we are unable to see income for people with high draft numbers. However, this could be attributed to a bunch of other factors. Such as, people with high draft numbers leaving the country so as to avoid being drafted. Or actually getting drafted and becoming a casualty of the war.

i. Tell a concrete story about how this differential attrition might bias our estimates.

If there is differential attrition then we cannot see all possible outcomes of the people with high draft numbers. They could have gone to war and died. The could have come back injured from the war and are unable to work due to disability. Differential attrition forces us to primarily focus only on those high draft individuals that pursued higher education. Hence, we are introducing a selection bias by not accounting for the other possible outcomes.

3. Dinner Plates

Suppose that researchers are concerned with the health consequences of what people eat and how much they weigh. Consider an experiment designed to measure the effect of a proposal to help people diet. Subjects are invited to a dinner and are randomly given regular-sized or slightly larger than regular sized plates. Hidden cameras record how much people eat, and the researchers find that those given larger plates eat substantially more food than those assigned small plates.

A statistical test shows that the apparent treatment effect is far greater than one would expect by chance. The authors conclude that a minor adjustment, reducing plate size, will help people lose weight.

- How convincing is the evidence regarding the effect of plate size of what people eat and how much they weight?

I am skeptical about the evidence concerning this experiment because I do not see the data. They check the right boxes in the sense of randomization but we know nothing about the subjects themselves. Additionally, there is no mention or discussion of the subjects weights. We have athletes that eat several times more than the average person but they are in excellent shape. The argument could be made that people eat less when given smaller plates but it is a leap to tie that to weight and weight-loss. There are several factors that come into play when talking about weight-loss. Is the subject physically active, their age, environment, and several more.

- What design and measurement improvements do you suggest?

The experimenters want to form a link between plate size for dinner and weight-loss. Weight-loss takes time so this experiment needs to be conducted on the same randomized subjects over a longer period. We additionally need the following data:

1) Weight at start of experiment

2) Age

3) Height

3) BMI

4) Are they physically active and of so how much (in hours per day)

5) Fixed sizes for regular and small plates

6) Same type of food for both types of plates

7) A way to check compliance

I believe adding all this variables and extending the experiment for a longer duration will help to form a more conclusive idea of plate size and weight-loss.

4. Think about Treatment Effects

Throughout this course we have focused on the average treatment effect. Think back to *why* we are concerned about the average treatment effect. What is the relationship between an ATE, and some individuals' potential outcomes? Make the strongest case you can for why this is *good* measure.

We are concerned about ATE because it let's us see the average difference between people in treatment vs control and it is not possible in the real world to see both potential outcomes. The ATE is not specific to any single individuals' potential outcome. It speaks more to the whole group of subjects. This is a good measure because it is a conservative way to see how the treatment is behaving in our experiment.