Problem Set 5

Anup Jha

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(foreign)
library(knitr)
library(sandwich)
library(lmtest)
library(AER)

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

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 (`table`) 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.)

```
crosstab<- table(d$total_ad_exposures_week1,d$treatment_ad_exposures_week1)
names(dimnames(crosstab)) <- c("total_ad_exposures_week1","treatment_ad_exposures_week1")
addmargins(crosstab)
```

```
##                         treatment_ad_exposures_week1
## total_ad_exposures_week1      0      1      2      3      4      5      6
##                        0    61182      0      0      0      0      0      0
##                        1    36754  37215      0      0      0      0      0
##                        2    21143  42036  20965      0      0      0      0
##                        3    10683  32073  32314  10726      0      0      0
##                        4     5044  20003  30432  20223   5115      0      0
##                        5     2045  10563  20970  20793  10293   2131      0
##                        6      729   4437  10977  14771  11147   4486    750
##                        Sum 137580 146327 115658  66513  26555   6617    750
##                         treatment_ad_exposures_week1
## total_ad_exposures_week1    Sum
##                        0    61182
##                        1    73969
##                        2    84144
##                        3    85796
##                        4    80817
##                        5    66795
##                        6    47297
##                        Sum 500000
```

Ans:Yes it looks reasonable . If the number of ad exposure increases, which means the user comes to yahoo multiple times, the distribution of number of treatment ads and non-treatment ads follow a binomial probability distribution with probability of success as 0.5. The probability is 0.5 because number of odd seconds is same as number of even seconds in a day. So for a person who visits the yahoo site 6 times the probability of receiving 0 to 6 treatment ad is:

0.015625, 0.09375, 0.234375, 0.3125, 0.234375, 0.09375, 0.015625 . We precisely see this kind of binomial distribution in our crosstab . We can look at the joint probability distribution using the prop.table and we see that for each row the probability distribution calculated from the data matches with binomial distribution of proability 0.5

```
prop.table(crosstab,1)
```

```
##                          treatment_ad_exposures_week1
## total_ad_exposures_week1          0          1          2          3
##                        0 1.00000000 0.00000000 0.00000000 0.00000000
##                        1 0.49688383 0.50311617 0.00000000 0.00000000
##                        2 0.25127163 0.49957216 0.24915621 0.00000000
##                        3 0.12451629 0.37382862 0.37663761 0.12501748
##                        4 0.06241261 0.24750981 0.37655444 0.25023201
##                        5 0.03061606 0.15814058 0.31394565 0.31129576
##                        6 0.01541324 0.09381145 0.23208660 0.31230311
##                          treatment_ad_exposures_week1
## total_ad_exposures_week1          4          5          6
##                        0 0.00000000 0.00000000 0.00000000
##                        1 0.00000000 0.00000000 0.00000000
##                        2 0.00000000 0.00000000 0.00000000
##                        3 0.00000000 0.00000000 0.00000000
##                        4 0.06329114 0.00000000 0.00000000
##                        5 0.15409836 0.03190359 0.00000000
##                        6 0.23568091 0.09484745 0.01585724
```

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?

```
lm_placebotest <- d[,lm(week0~treatment_ad_exposures_week1)]
(placebotest.Coeftest <- coeftest(lm_placebotest,vcovHC(lm_placebotest)))
```

```
##
## t test of coefficients:
##
##                              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 1.6696854  0.0054486 306.441 < 2.2e-16 ***
## treatment_ad_exposures_week1 0.2630995  0.0033545  78.431 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(lm_placebotest,se=list(sqrt(diag(vcovHC(lm_placebotest)))),type='text',header=F)
```

```
##
## ===============================================
##                                 Dependent variable:
##                             -----------------------
##                                       week0
## -----------------------------------------------
## treatment_ad_exposures_week1          0.263***
##                                        (0.003)
##
## Constant                              1.670***
##                                        (0.005)
##
```

```
## ------------------------------------------------------------
## Observations                                  500,000
## R2                                             0.014
## Adjusted R2                                    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
```

Ans:When we run the regression model for week0 as outcome and regressor as treatment_add_exposures_week1 we see that the the coefficient is 0.2631 and its statistically significant with robust standard errors having p value < 2e-16 . Which means that placebo test fails as we would not expect the week0 revenues of past to be effected by treatment in week1. We would expect the coefficient of treatment_ad_exposures_week1 to be not statistically different than 0 if the placebo test was to pass.

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.*)
Ans:What we see from the data is that when some one visits the website more he/she is more probable to get treatment also more. Also the people who would visit the website more would buy more. So the treatment_ad_exposures is not truly random and its more for frequent buyers. So the coefficient of treatment_ad_exposures also soaks up the effect of frequent buyers if we don't control for that. There is no direct variable for the frequent buyers but we can use the total_ad_exposures as a proxy for how frequent the user is in visiting the website.

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.)

```
lm_placebotest_frequent_buyer_control <- d[,lm(week0~treatment_ad_exposures_week1+total_ad_exposures_we
(placebotest.Coeftest.frequent.buyer.control <- coeftest(lm_placebotest_frequent_buyer_control,vcovHC(lm
```

```
##
## t test of coefficients:
##
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.3453746  0.0061332 219.359   <2e-16 ***
## treatment_ad_exposures_week1 -0.0022454  0.0051381  -0.437   0.6621
## total_ad_exposures_week1      0.2453481  0.0033722  72.756   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(lm_placebotest_frequent_buyer_control,se=list(sqrt(diag(vcovHC(lm_placebotest_frequent_buyer_
```

```
##
## ============================================================
##                                    Dependent variable:
##                                ---------------------------
##                                           week0
## ------------------------------------------------------------
```

```
## treatment_ad_exposures_week1            -0.002
##                                          (0.005)
##
## total_ad_exposures_week1                0.245***
##                                          (0.003)
##
## Constant                                1.345***
##                                          (0.006)
##
## ---------------------------------------------------------
## Observations                             500,000
## R2                                        0.026
## Adjusted R2                               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
```

Ans:We include the total_ad_exposures_week1 as a control variable which controls for the buyers frequency. Now we see that the treatment effect of treatment_ad_exposures_week1 is -0.00225 with p value as 0.66 . Which means that the coefficient is statistically not different than 0 and hence week0 revenues are not effected by the treatment which means now that the placebo test passes

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).

```
lm_week1 <- d[,lm(week1~treatment_ad_exposures_week1+total_ad_exposures_week1)]
(week1.Coeftest <- coeftest(lm_week1,vcovHC(lm_week1)))
```

```
##
## t test of coefficients:
##
##                               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                  1.3179604  0.0060560 217.628 < 2.2e-16 ***
## treatment_ad_exposures_week1 0.0563399  0.0051377  10.966 < 2.2e-16 ***
## total_ad_exposures_week1     0.2244777  0.0033174  67.668 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(lm_week1,se=list(sqrt(diag(vcovHC(lm_week1)))),type='text',header=F)
```

```
##
## =========================================================
##                                   Dependent variable:
##                               ----------------------------
##                                          week1
## ---------------------------------------------------------
## treatment_ad_exposures_week1           0.056***
##                                         (0.005)
##
## total_ad_exposures_week1               0.224***
##                                         (0.003)
##
## Constant                               1.318***
##                                         (0.006)
##
```

```
## ------------------------------------------------------------
## Observations                                  500,000
## R2                                             0.028
## Adjusted R2                                    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
```

Ans:We see that the treatment effect is an increase of $ 0.05634 in revenue for every extra treatment ad seen by the user. And this value is statistically significant at p value <2e-16

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.

Ans:We can argue that the amount of treatment_ad are not assigned randomly as more someone visits the website more treatment he/she gets. And some one who visits more is most probably a frequent buyer and hence gives more revenue. So if we don't control for the frequency of visit then the treatment effect of the ad would also soak up the effect of frequency of visit and hence inflating the treatment effect and making the estimation biased upwards

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).

```
d[,week_1_through_10:=week1+week2+week3+week4+week5+week6+week7+week8+week9+week10]
lm_week1_through_10 <- d[,lm(week_1_through_10~treatment_ad_exposures_week1+total_ad_exposures_week1)]
(week1_through_10.Coeftest <- coeftest(lm_week1_through_10,vcovHC(lm_week1_through_10)))
```

```
##
## t test of coefficients:
##
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  17.150813   0.024464 701.0609   <2e-16 ***
## treatment_ad_exposures_week1  0.012739   0.019022   0.6697   0.5031
## total_ad_exposures_week1      2.228344   0.012537 177.7476   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(lm_week1_through_10,se=list(sqrt(diag(vcovHC(lm_week1_through_10)))),type='text',header=F)
```

```
##
## ============================================================
##                                     Dependent variable:
##                                 ---------------------------
##                                       week_1_through_10
## ------------------------------------------------------------
## treatment_ad_exposures_week1               0.013
##                                           (0.019)
##
## total_ad_exposures_week1                  2.228***
##                                           (0.013)
##
## Constant                                 17.151***
##                                           (0.024)
##
## ------------------------------------------------------------
## Observations                              500,000
```

```
## R2                                          0.132
## Adjusted R2                                 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
```

Ans:We see that the causal effect of the treatment_ad on revenue from week1 through week10 is $0.0127 with a p value of 0.5 hence insignificant from 0 . In short the treatment has no effect on long run cumulative revenue from week1 through week10

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.

```
d[,week_2_through_10:=week2+week3+week4+week5+week6+week7+week8+week9+week10]
lm_week2_through_10 <- d[,lm(week_2_through_10~treatment_ad_exposures_week1+total_ad_exposures_week1)]
(week2_through_10.Coeftest <- coeftest(lm_week2_through_10,vcovHC(lm_week2_through_10)))
```

```
##
## t test of coefficients:
##
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  15.832852   0.023503 673.6481  < 2e-16 ***
## treatment_ad_exposures_week1 -0.043601   0.018187  -2.3973  0.01651 *
## total_ad_exposures_week1      2.003866   0.012015 166.7760  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(lm_week2_through_10,se=list(sqrt(diag(vcovHC(lm_week2_through_10)))),type='text',header=F)
```

```
##
## ============================================================
##                                    Dependent variable:
##                                ----------------------------
##                                      week_2_through_10
## ------------------------------------------------------------
## treatment_ad_exposures_week1             -0.044**
##                                          (0.018)
##
## total_ad_exposures_week1                 2.004***
##                                          (0.012)
##
## Constant                                15.833***
##                                          (0.024)
##
## ------------------------------------------------------------
## Observations                             500,000
## R2                                        0.115
## Adjusted R2                               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
```

Ans:We see that the causal effect of the treatment_ad on revenue from week2 through week10 is -$0.0436 with a p value of 0.017 hence significant at 0.05 level. In short the treatment has a negative effect on long run cumulative revenue from week2 through week10

i. Tell a story that could plausibly explain the result from part (h).
Ans:Since we saw that week1 has positive statistically significant treatment effect while treatmnent has negative effect on cumulatve revenue from week2 to week10 this means that people buy the products in the week they see the advertisement and then they refrain from buying the product again. This might be due to the fact that users generally assign a budget for their purchases and they might purcahse the product immediately after seeing the ad but once they have purchased they would not buy again in near future.

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.*)

```
lm_hte_week1 <- d[,lm(week1~treatment_ad_exposures_week1+total_ad_exposures_week1+product_b+product_b*t
(hte_week1.Coeftest <- coeftest(lm_hte_week1,vcovHC(lm_hte_week1)))
```

```
##
## t test of coefficients:
##
##                                          Estimate Std. Error  t value
## (Intercept)                             1.2892704  0.0064853 198.7975
## treatment_ad_exposures_week1            0.0612513  0.0059629  10.2721
## total_ad_exposures_week1                0.2108683  0.0033879  62.2413
## product_b                               0.1703200  0.0124985  13.6273
## treatment_ad_exposures_week1:product_b -0.0100997  0.0070911  -1.4243
##                                           Pr(>|t|)
## (Intercept)                                <2e-16 ***
## treatment_ad_exposures_week1               <2e-16 ***
## total_ad_exposures_week1                   <2e-16 ***
## product_b                                  <2e-16 ***
## treatment_ad_exposures_week1:product_b     0.1544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(lm_hte_week1,se=list(sqrt(diag(vcovHC(lm_hte_week1)))),type='text',header=F)
```

```
##
## ===============================================================
##                                         Dependent variable:
##                                     ---------------------------
##                                                week1
## ---------------------------------------------------------------
## treatment_ad_exposures_week1                  0.061***
##                                               (0.006)
##
## total_ad_exposures_week1                      0.211***
##                                               (0.003)
##
## product_b                                     0.170***
##                                               (0.012)
##
## treatment_ad_exposures_week1:product_b         -0.010
##                                               (0.007)
##
## Constant                                      1.289***
```

```
##                                                    (0.006)
##
## ---------------------------------------------------------------------
## Observations                                       500,000
## R2                                                  0.028
## Adjusted R2                                         0.028
## Residual Std. Error                          2.766 (df = 499995)
## F Statistic                               3,663.676*** (df = 4; 499995)
## =====================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

Ans:To test the hypothesis that treatment effect is more for product b than product a we create a model where we add the control for the product and also include interaction term between the treatment and product. We see that the interaction term which represents the heterogeneous treatment effect is -0.01010 at p value of 0.15 . Which makes this term statistically insignificant at 0.05 level and hence we can conclude that there is no difference between efficacy of ads between product a and product b. Though the baseline revenue for product b is higher than product a as indicated by the coefficient of variable product_b with value as 0.170 and p value <2e-16 which makes is statistically significant

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?

Ans:We would be tempted to attribute the mediation as celebrity endorsements but we cannot effectively conclude on this. To effectively conclude on whether the celebrity endorsement is the mediator we would need to experiment where we randomly manipulate the celebrity endorsements in ads and then check the outcomes.

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

```
##    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
```

```
##   draft_number   years_education     income
##  Min.   :  1.0   Min.   : 9.00   Min.   :     0
##  1st Qu.: 98.0   1st Qu.:13.00   1st Qu.: 40730
##  Median :188.0   Median :15.00   Median : 59202
##  Mean   :187.1   Mean   :14.86   Mean   : 62083
##  3rd Qu.:278.0   3rd Qu.:16.00   3rd Qu.: 80291
##  Max.   :365.0   Max.   :22.00   Max.   :215990
```

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?

```
lm_naive_model <- d[,lm(income ~ years_education)]
(lm_naive_model.Coeftest <- coeftest(lm_naive_model,vcovHC(lm_naive_model)))
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     -23354.638   1197.226 -19.507 < 2.2e-16 ***
## years_education   5750.480     84.411  68.125 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(lm_naive_model,se=list(sqrt(diag(vcovHC(lm_naive_model)))),type='text',header=F)
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                 income
## ------------------------------------------------
## years_education             5,750.480***
##                               (84.411)
##
## Constant                   -23,354.640***
```

```
##                             (1,197.226)
##
## ------------------------------------------------
## Observations                    19,567
## R2                              0.196
## Adjusted R2                     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
```

Ans: This naive regression just tells us how is income correlated with the years of education.Specifically it tells us that with increase in one year of education we can expect an increase of income by $5,750.5 in the observations

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.

Ans:As noted above the results from the naive regression just tells us how is income correlated with the years of education. It doesn't tell us any causal effect of the years of education to the income as there can be many number of confounding variables which effect the years of education and the income both. For example the ability might be both positively correlated with years of education and income. And since these confounding variables are absent from the model the years of education soaks up the effect of those variable in its estimate.

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.)

```
d[,high_rank:=ifelse(draft_number<=80,1,0)]
lm_draft_to_education_model <- d[,lm(years_education~high_rank)]
(lm_draft_to_education_model.Coeftest <- coeftest(lm_draft_to_education_model,vcovCL(lm_draft_to_educat
```

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

```
stargazer(lm_draft_to_education_model,se=list(sqrt(diag(vcovCL(lm_draft_to_education_model,d[,draft_numb
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                           years_education
## ------------------------------------------------
## high_rank                    2.126***
##                              (0.038)
##
## Constant                    14.434***
##                              (0.018)
```

```
##
## ------------------------------------------------
## Observations                    19,567
## R2                              0.138
## Adjusted R2                     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
```

Ans:We see from the model that the effect of high draft number on years of education is 2.1257562 . The clustered std error clustering on draft number is 0.0381878

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

```
lm_draft_to_income_model <- d[,lm(income~high_rank)]
(lm_draft_to_income_model.Coeftest <- coeftest(lm_draft_to_income_model,vcovCL(lm_draft_to_income_model
```

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

```
stargazer(lm_draft_to_income_model,se=list(sqrt(diag(vcovCL(lm_draft_to_income_model,d[,draft_number]))))
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                income
## ------------------------------------------------
## high_rank                   6,637.554***
##                               (511.899)
##
## Constant                    60,761.890***
##                               (244.361)
##
## ------------------------------------------------
## Observations                    19,567
## R2                               0.008
## Adjusted R2                      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
```

Ans:We see from the model that the estimate of high draft rank on income is 6637.554244 with the std errors using clustered std errors clustering on draft number is 511.899229

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?

```
(instrumental_variable_estimate <- lm_draft_to_income_model.Coeftest["high_rank","Estimate"]/lm_draft_t
```

```
## [1] 3122.444
```

Ans: We see that the instrumental variable estimate of the effects of education on income is 3122. Which implies that every one year of extra education causes an increase of $3122 in income

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.

Ans:Having a high draft number would most probably increase the chance of someone actually serving in the army. And serving in the army in war might induce other behavorial changes to a person such as dealing with PTSD which might result in lowering the income and hence affecting the outcome in question.

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.)**

```
#let us check the number of people in draft numbers
d_summary <- d[,.N,by=.(draft_number,high_rank)]
#t.test for number of obeservations per draft number
#this tests whether mean of number of observations between high draft and low draft is same
t.test(d_summary$N~d_summary$high_rank)
```

```
##
##  Welch Two Sample t-test
##
## data:  d_summary$N by d_summary$high_rank
## t = 6.6358, df = 123.31, p-value = 9.121e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.410934 8.160996
## sample estimates:
## mean in group 0 mean in group 1
##        54.98596        48.70000
```

```
#Wilcox rank test
wilcox.test(d_summary$N~d_summary$high_rank)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d_summary$N by d_summary$high_rank
## W = 16680, p-value = 2.364e-10
## alternative hypothesis: true location shift is not equal to 0
```

```
#Linear regression to estimate the difference in number of observations
lm_draft_attrition_model <- d_summary[,lm(N~high_rank)]
(lm_draft_attrition_modell.Coeftest <- coeftest(lm_draft_attrition_model,vcovCL(lm_draft_attrition_model
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 54.98596    0.43149 127.4326 < 2.2e-16 ***
```

```
## high_rank   -6.28596    0.94482  -6.6531 1.059e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
stargazer(lm_draft_attrition_model,se=list(sqrt(diag(vcovCL(lm_draft_attrition_model,d_summary[,draft_nu
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                    N
## ------------------------------------------------
## high_rank                      -6.286***
##                                 (0.945)
##
## Constant                       54.986***
##                                 (0.431)
##
## ------------------------------------------------
## Observations                      365
## R2                               0.112
## Adjusted R2                      0.110
## Residual Std. Error       7.336 (df = 363)
## F Statistic           45.861*** (df = 1; 363)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Ans:We see from the 2 group t.test on number of records in high draft number and low draft number that t.test rejects the null hypothesis of that mean number of records in two group is same with p value of 9e-10 which is statistically significant. But T tests assume about normality in distribution of obesrvations. So we can perform Wilcox rank sum test which is non parameteric test and doesn't depened on normality of distribution of observations. We also see from Wilcox Rank Sum test that the distribution of number of obeservations for high_draft and low_draft number don't coincide to same median. The Wilcox Rank sum test rejects the null hypothesis of same median with p value of 2e-10. We also see that when we regress the number of observations in each draft number to dummy variable of high_rank we see that high_rank has 6.286 fewer observations than low rank and the p value is 1.1e-10 which makes it statistically significant. So we can conclude that we have differential attrition between high rank and low rank draft number groups

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

Ans:One of the reason that we are not observing the high rank draft numbers as much as low rank may be because of casualties in the war and those who served in the army might have lost their lives in the war and hence their income cannot be captured. And people with high draft numbers are more likely to have served in war and hence have higher attrition.

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

Ans: This might bias our estimates as we are not observing few instances and the outcome income might be correlated with the attrition. As we have the data only pertaining to survivors and being a survivor might be correlated with income either negatively or positively. So when we calculate the ATE the estimate would be biased as the mean in the survivor group might not represent the mean of the random sample. One way to think about this is people who are risk takers might have lesser income in general and those would also be people if drafted in army might be victims of war. So we would have attrition for subjects whose potential outcome (income) is less . So the effect of education on income would be downwardly biased

3. Optional: 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 a *good* measure.

Ans: We are concerned about average treatment effect rather than individual treatment effect as its impossible to calculate the individual treatment effect in experimental setting as one individual cannot be both in treatment and control group. But by measuring the ATE we are trying to garner the best unbiased estimate of treatment effect for a random individual.