Problem Set 5

Atit Wongnophadol

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(car)
library(lmtest)
library(sandwich)

# removes all objects from the current workspace (R memory)
rm(list = ls())

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

```
##    6.2   0.0   0.0   0.0   0.0   0.0     0   9.7   4.1    0.0
##    0.0   8.6   2.4   0.0   7.4   0.0     0   0.0   5.7    0.0
##    5.3   0.0   8.1   7.8   3.3   0.0     0   9.4   0.0    0.0
##    4.1   0.0   8.8   5.8   5.9   0.0     0   0.0   9.6    0.0
##    3.6   4.6   5.5   7.2   7.1   0.0     0   0.0   0.0    0.0
##    5.5   9.8   5.0   0.0   0.0   7.7     0  11.0   4.8    6.9
```

```
d1A <- d1[d1$product_b==0]
d1B <- d1[d1$product_b==1]
```

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:

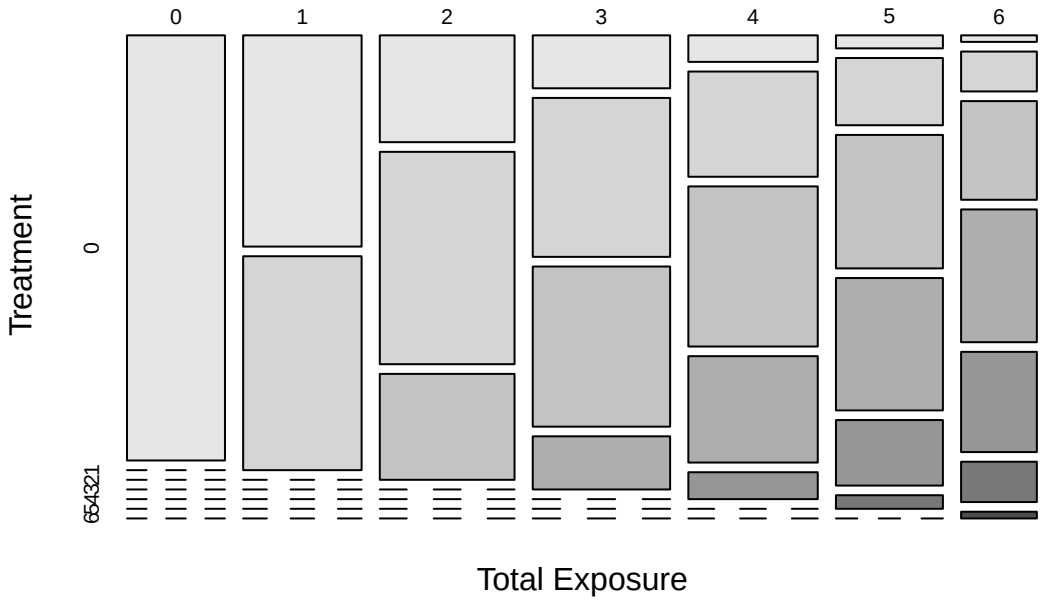
- The effect of treatment ad exposures on purchases is linear. That is, the first exposure has the same effect as the second exposure.
- There is no effect of being exposed to the odd-second ads on purchases for the product being advertised on the even second.
- Every Yahoo! user visits the Yahoo! home 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.)

```
# library to build a crosstab table
library(descr)

# crosstab table
crosstab(d1$treatment_ad_exposures_week1, d1$total_ad_exposures_week1,
         xlab = "Total Exposure", ylab = "Treatment",
         prop.c = T, prop.r = F, prop.t = F)
```




```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |          Column Percent |
## |-------------------------|
##
## ============================================================================================================
##                                    d1$total_ad_exposures_week1
## d1$treatment_ad_exposures_week1         0       1       2       3       4       5       6    Total
## ------------------------------------------------------------------------------------------------------------
## 0                                   61182   36754   21143   10683    5044    2045     729   137580
##                                     100.0%   49.7%   25.1%   12.5%    6.2%    3.1%    1.5%
## ------------------------------------------------------------------------------------------------------------
## 1                                       0   37215   42036   32073   20003   10563    4437   146327
##                                       0.0%   50.3%   50.0%   37.4%   24.8%   15.8%    9.4%
## ------------------------------------------------------------------------------------------------------------
## 2                                       0       0   20965   32314   30432   20970   10977   115658
##                                       0.0%    0.0%   24.9%   37.7%   37.7%   31.4%   23.2%
## ------------------------------------------------------------------------------------------------------------
## 3                                       0       0       0   10726   20223   20793   14771    66513
##                                       0.0%    0.0%    0.0%   12.5%   25.0%   31.1%   31.2%
## ------------------------------------------------------------------------------------------------------------
## 4                                       0       0       0       0    5115   10293   11147    26555
##                                       0.0%    0.0%    0.0%    0.0%    6.3%   15.4%   23.6%
## ------------------------------------------------------------------------------------------------------------
## 5                                       0       0       0       0       0    2131    4486     6617
##                                       0.0%    0.0%    0.0%    0.0%    0.0%    3.2%    9.5%
## ------------------------------------------------------------------------------------------------------------
## 6                                       0       0       0       0       0       0     750      750
##                                       0.0%    0.0%    0.0%    0.0%    0.0%    0.0%    1.6%
## ------------------------------------------------------------------------------------------------------------
## Total                               61182   73969   84144   85796   80817   66795   47297    5e+05
##                                      12.2%   14.8%   16.8%   17.2%   16.2%   13.4%    9.5%
## ============================================================================================================
```

ANSWER: The dataset looks fairly reasonable given that (i) the number of impression looks (i.e., treatment) are equal to or fewer than the number of exposures and (ii) the distribution of the impression looks are fairly symmetrical; for example, most people got treatments somewhere between zero and the maximum number of exposures while few others got little impresseions when exposed or all impressions whene exposed.

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?

```
# fit a linear model for total dataset
mod1a <- lm(week1 ~ treatment_ad_exposures_week1, data = d1)
mod1a.coef <- coeftest(mod1a, vcovHC(mod1a))

mod1a_placebo <- lm(week0 ~ treatment_ad_exposures_week1, data = d1)
mod1a_placebo.coef <- coeftest(mod1a_placebo, vcovHC(mod1a_placebo))

## use stargazer to print formatted tables
stargazer(mod1a, mod1a_placebo, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(mod1a.coef[3:4], mod1a_placebo.coef[3:4]),
          align = TRUE, no.space=TRUE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Total Dataset")
```

```
##
## Total Dataset
## ===============================================================
##                                    Dependent variable:
##                                ----------------------------
##                                     week1          week0
##                                      (1)            (2)
## ---------------------------------------------------------------
## treatment_ad_exposures_week1    0.299***        0.263***
##                                  (0.003)        (0.003)
## Constant                        1.615***        1.670***
##                                  (0.005)        (0.005)
## ---------------------------------------------------------------
## Observations                     500,000        500,000
## Adjusted R2                       0.018          0.014
## ===============================================================
## Note:                          *p<0.05; **p<0.01; ***p<0.001
```

```
# fit a linear model for Product A only
mod1a_A <- lm(week1 ~ treatment_ad_exposures_week1, data = d1A)
mod1a_A.coef <- coeftest(mod1a_A, vcovHC(mod1a_A))

mod1a_placebo_A <- lm(week0 ~ treatment_ad_exposures_week1, data = d1A)
mod1a_placebo_A.coef <- coeftest(mod1a_placebo_A, vcovHC(mod1a_placebo_A))

## use stargazer to print formatted tables
stargazer(mod1a_A, mod1a_placebo_A, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(mod1a_A.coef[ ,2], mod1a_placebo_A.coef[ ,2]),
          align = TRUE, no.space=TRUE,
```

```
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Product A Only")
```

```
##
## Product A Only
## ============================================================
##                                   Dependent variable:
##                               -----------------------------
##                                     week1          week0
##                                      (1)            (2)
## ------------------------------------------------------------
## treatment_ad_exposures_week1    0.297***        0.253***
##                                  (0.004)        (0.004)
## Constant                        1.513***        1.568***
##                                  (0.006)        (0.006)
## ------------------------------------------------------------
## Observations                     300,038        300,038
## Adjusted R2                       0.018          0.013
## ============================================================
## Note:                          *p<0.05; **p<0.01; ***p<0.001
```

```
# fit a linear model for Product B only
mod1a_B <- lm(week1 ~ treatment_ad_exposures_week1, data = d1B)
mod1a_B.coef <- coeftest(mod1a_B, vcovHC(mod1a_B))

mod1a_placebo_B <- lm(week0 ~ treatment_ad_exposures_week1, data = d1B)
mod1a_placebo_B.coef <- coeftest(mod1a_placebo_B, vcovHC(mod1a_placebo_B))

## use stargazer to print formatted tables
stargazer(mod1a_B, mod1a_placebo_B, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(mod1a_B.coef[ ,2], mod1a_placebo_B.coef[ ,2]),
          align = TRUE, no.space=TRUE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Product B Only")
```

```
##
## Product B Only
## ============================================================
##                                   Dependent variable:
##                               -----------------------------
##                                     week1          week0
##                                      (1)            (2)
## ------------------------------------------------------------
## treatment_ad_exposures_week1    0.245***        0.214***
##                                  (0.006)        (0.006)
## Constant                        1.870***        1.930***
##                                  (0.011)        (0.011)
## ------------------------------------------------------------
## Observations                     199,962        199,962
## Adjusted R2                       0.010          0.008
## ============================================================
## Note:                          *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: At 99.9% confidence level, the treatment ad exposure has a strong statistical significance relationshp with the purchase the week after (i.e., Week1). However the placebo test also suggests that the treatment ad exposure has a strong association (i.e., statistically significant relationship) with the prior's week purchase (i.e., Week0). That said, the placebo test fails as the treatment should not have any effect on the placebo outcome (i.e., Week0).

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.*)

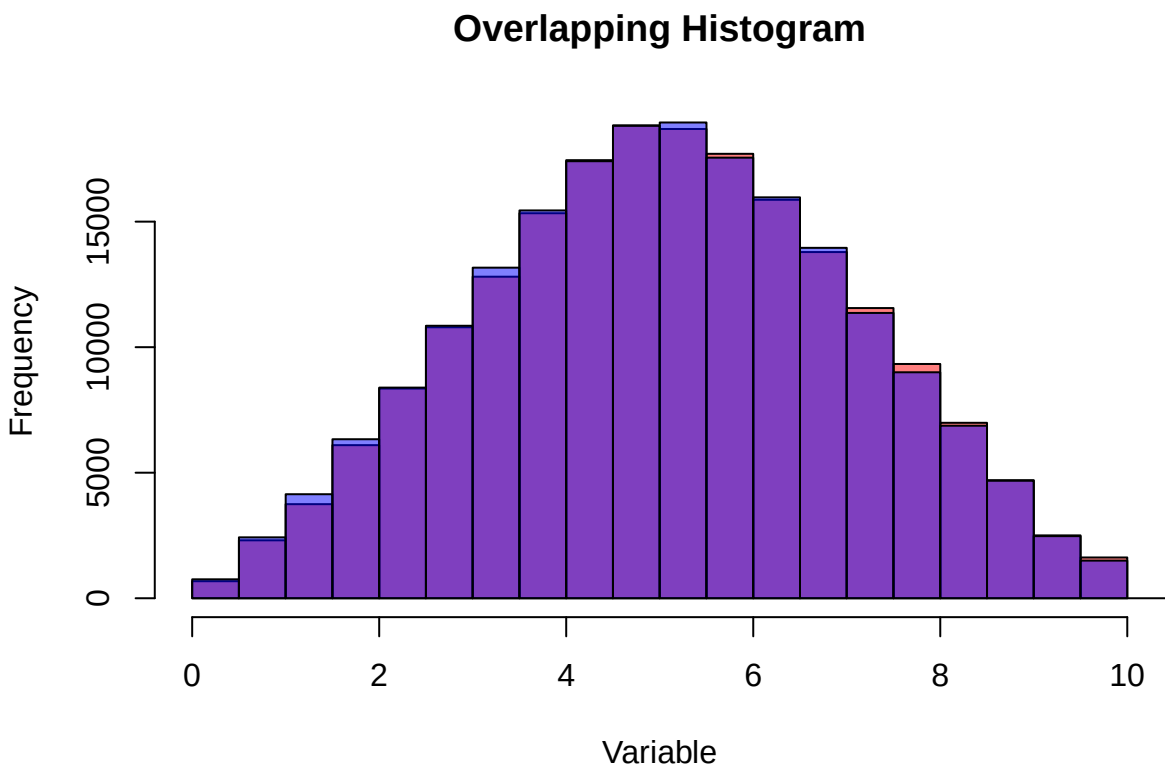
ANSWER: Randomness of the treatment variable is a concerning assumption for users who visited the site more than one day during the campaign period because they will have a higher chance to get more exposures to the treatment ad. That is, a different number of visits does not imply an equal chance of being assigned to control or treatment. A higher number of visits leads to a higher probability of receiving the treatment. For example, a subject that visits the site 6 days in a row would have an over 98% chance of getting at least one treatment (i.e., 1-(0.5)6). In other words, the probability of being assigned to treatment now becomes dependent on the number of total visits.

So to decouple this dependency, we could analyze the data by analyzing the subjects by subgroups (i.e., people who visited the site 1 day, 2 days, .., and 6 days). Then in each subgroup, the control group (i.e., got assigned with a placebo ad) would be those who did not get the treatment in any day.

To operationalize this, we could add another 6 dummy variables ('D1', 'D2', 'D3', 'D4', 'D5' and 'D6') to control for the total number of visits, where D1 represents a dummy variable when the total of visit is 1, and D2 represents a dummy variable when the total of visit is 2, and so on.

The placebo test outcome simply describes the positive relationship between the revenue the week prior the campaign and the number of treatment exposures.

```
# Compare the histograms of both week0 and week1
hist(d1[d1$week0!=0]$week0, col=rgb(1,0,0,0.5), main="Overlapping Histogram", xlab="Variable")
hist(d1[d1$week1!=0]$week1, col=rgb(0,0,1,0.5), add=T)
```



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.)

```
# create dummy variables and add them to a new dataset 'dat'
dat <- d1

dat[, D1 := as.numeric(NA)]
dat[, D1 := ifelse(total_ad_exposures_week1 == 1, 1, 0)]

dat[, D2 := as.numeric(NA)]
dat[, D2 := ifelse(total_ad_exposures_week1 == 2, 1, 0)]

dat[, D3 := as.numeric(NA)]
dat[, D3 := ifelse(total_ad_exposures_week1 == 3, 1, 0)]

dat[, D4 := as.numeric(NA)]
dat[, D4 := ifelse(total_ad_exposures_week1 == 4, 1, 0)]

dat[, D5 := as.numeric(NA)]
dat[, D5 := ifelse(total_ad_exposures_week1 == 5, 1, 0)]

dat[, D6 := as.numeric(NA)]
dat[, D6 := ifelse(total_ad_exposures_week1 == 6, 1, 0)]


# fit a linear model for total dataset
mod1d <- lm(week1 ~ treatment_ad_exposures_week1 +
              D1 + D2 + D3 + D4 + D5 + D6, data = dat)
mod1d.coef <- coeftest(mod1d, vcovHC(mod1d))
```

```r
# fit a linear model for total dataset (placebo)
mod1d_placebo <- lm(week0 ~ treatment_ad_exposures_week1 +
                      D1 + D2 + D3 + D4 + D5 + D6, data = dat)
mod1d_placebo.coef <- coeftest(mod1d_placebo, vcovHC(mod1d_placebo))


## use stargazer to print formatted tables
stargazer(mod1d, mod1d_placebo, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(mod1d.coef[ ,2], mod1d_placebo.coef[ ,2]),
          align = T, no.space=T,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Total Dataset")
```

```
##
## Total Dataset
## ============================================================
##                                   Dependent variable:
##                               ------------------------------
##                                    week1          week0
##                                     (1)            (2)
## ------------------------------------------------------------
## treatment_ad_exposures_week1    0.056***         -0.002
##                                  (0.005)        (0.005)
## D1                              0.257***        0.312***
##                                  (0.011)        (0.012)
## D2                              0.487***        0.551***
##                                  (0.012)        (0.013)
## D3                              0.693***        0.780***
##                                  (0.014)        (0.014)
## D4                              0.919***        1.003***
##                                  (0.016)        (0.017)
## D5                              1.136***        1.236***
##                                  (0.019)        (0.020)
## D6                              1.369***        1.540***
##                                  (0.024)        (0.024)
## Constant                        1.296***        1.307***
##                                  (0.008)        (0.008)
## ------------------------------------------------------------
## Observations                     500,000        500,000
## Adjusted R2                       0.028          0.026
## ============================================================
## Note:                          *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: The placebo test passes as suggested by the second model above. The regression outcome indicates that the ad treatment doesn't have impact on Week0 revenue.

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).

ANSWER: From the regression outcome of the first model in (d), the causal effect of of each ad exposure is 0.056 with the robust standard error of 0.0005, indicating a statistical significance

of the ad treatment.

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.

ANSWER: In addition to my response to question (c) about the randomness issue, the colleague's proposed specification has another flaw in the which the frequency of the site visit may suggest different shopping behavior. And therefore, a specification that decouple that dependency such as the one that I proposed (by creating dummy variables to differentiate / control for shoppers who visited the site different nubmer of times) would be more appropriate and unbiased.

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).

```
# create total purchases from week 1 to 10)

dat[, total_10weeks := as.numeric(NA)]
dat[, total_10weeks := week1 + week2 + week3 + week4 + week5 +
                           week6 + week7 + week8 + week9 + week10]


# fit a linear model for total dataset
mod1g <- lm(total_10weeks ~ treatment_ad_exposures_week1 +
              D1 + D2 + D3 + D4 + D5 + D6, data = dat)
mod1g.coef <- coeftest(mod1g, vcovHC(mod1g))


## use stargazer to print formatted tables
stargazer(mod1d, mod1g, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(mod1d.coef[ ,2], mod1g.coef[ ,2]),
          align = T, no.space=T,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Total Dataset")
```

```
##
## Total Dataset
## ===============================================================
##                                    Dependent variable:
##                                ------------------------------
##                                     week1      total_10weeks
##                                      (1)            (2)
## ---------------------------------------------------------------
## treatment_ad_exposures_week1    0.056***         0.012
##                                  (0.005)        (0.019)
## D1                              0.257***        2.600***
##                                  (0.011)        (0.048)
## D2                              0.487***        4.843***
##                                  (0.012)        (0.051)
## D3                              0.693***        7.036***
##                                  (0.014)        (0.057)
## D4                              0.919***        9.083***
##                                  (0.016)        (0.064)
```

```
## D5                                1.136***       11.196***
##                                    (0.019)        (0.074)
## D6                                1.369***       13.736***
##                                    (0.024)        (0.088)
## Constant                          1.296***       16.902***
##                                    (0.008)        (0.033)
## ------------------------------------------------------------
## Observations                       500,000        500,000
## Adjusted R2                         0.028          0.132
## ============================================================
## Note:                          *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: the causal effect of each treatment for the entire 10 weeks was 0.012, which is much lower than 0.056 on week1 alone.

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.

```
# create total purchases from week 2 to 10

dat[, total_9weeks := as.numeric(NA)]
dat[, total_9weeks := week2 + week3 + week4 + week5 +
                          week6 + week7 + week8 + week9 + week10]


# fit a linear model for total dataset
mod1h <- lm(total_9weeks ~ treatment_ad_exposures_week1 +
              D1 + D2 + D3 + D4 + D5 + D6, data = dat)
mod1h.coef <- coeftest(mod1h, vcovHC(mod1h))


## use stargazer to print formatted tables
stargazer(mod1d, mod1g, mod1h, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(mod1d.coef[ ,2], mod1g.coef[ ,2], mod1h.coef[ ,2]),
          align = T, no.space=T,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Total Dataset")
```

```
##
## Total Dataset
## ========================================================================
##                                        Dependent variable:
##                                ---------------------------------
##                                 week1   total_10weeks total_9weeks
##                                  (1)         (2)          (3)
## ------------------------------------------------------------------------
## treatment_ad_exposures_week1 0.056***     0.012       -0.044*
##                              (0.005)     (0.019)      (0.018)
## D1                           0.257***   2.600***      2.343***
##                              (0.011)     (0.048)      (0.046)
## D2                           0.487***   4.843***      4.355***
##                              (0.012)     (0.051)      (0.049)
## D3                           0.693***   7.036***      6.343***
##                              (0.014)     (0.057)      (0.054)
```

```
## D4                              0.919***   9.083***      8.164***
##                                 (0.016)     (0.064)      (0.061)
## D5                              1.136***   11.196***    10.060***
##                                 (0.019)     (0.074)      (0.071)
## D6                              1.369***   13.736***    12.367***
##                                 (0.024)     (0.088)      (0.085)
## Constant                        1.296***   16.902***    15.606***
##                                 (0.008)     (0.033)      (0.032)
## ------------------------------------------------------------------
## Observations                    500,000     500,000      500,000
## Adjusted R2                      0.028       0.132        0.116
## ==================================================================
## Note:                                  *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: Interestingly, the causal effect from week2 to week10 is negative 0.044 with a statistical significance at 95% confidence level.

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

ANSWER: What possily tells us here is that the advertisement had a practically positive effect on the shopping level during the week of the campaign. This suggests that the treatment had almost immediate effect on the purchasing behavior.

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.*)

```
# fit a linear model that includes an interactive term "product_b"
mod1j <- lm(week1 ~ treatment_ad_exposures_week1 + product_b +
              product_b*treatment_ad_exposures_week1 +
              D1 + D2 + D3 + D4 + D5 + D6, data = dat)
mod1j.coef <- coeftest(mod1j, vcovHC(mod1j))

## use stargazer to print formatted tables
stargazer(mod1j, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(mod1j.coef[ ,2]),
          align = TRUE, no.space=TRUE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Effect Comparison between Product A and B",
          dep.var.labels = c("Effect"))
```

```
##
## Effect Comparison between Product A and B
## ==================================================================
##                                          Dependent variable:
##                                      ------------------------------
##                                                 Effect
## ------------------------------------------------------------------
## treatment_ad_exposures_week1                   0.059***
##                                                (0.006)
## product_b                                      0.165***
```

```
##                                                    (0.013)
## D1                                                0.230***
##                                                    (0.012)
## D2                                                0.448***
##                                                    (0.013)
## D3                                                0.648***
##                                                    (0.015)
## D4                                                0.867***
##                                                    (0.017)
## D5                                                1.069***
##                                                    (0.020)
## D6                                                1.269***
##                                                    (0.024)
## treatment_ad_exposures_week1:product_b             -0.006
##                                                    (0.007)
## Constant                                          1.276***
##                                                    (0.008)
## ------------------------------------------------------------------------
## Observations                                       500,000
## Adjusted R2                                         0.028
## ========================================================================
## Note:                                *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: Based on the outcome of the regression above, where the interaction term between treatment ad and product B doesn't show a statistical significance, the null hypothesis cannot be rejected. The treatment ad is not more effective for product B than for product A.

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?

ANSWER: I don't have any data evidence that can be used to analyze an effect of a celebrity inclusion, so I can't say that I am confidnt in drawing such a conclusion.

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.

```
##    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
```

Questions to Answer

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?

```
# fit a regression model and calculate robust s.e.
d2a <- lm(income ~ years_education, data = d2)
d2a.coef <- coeftest(d2a, vcovHC(d2a))

## use stargazer to print formatted tables
stargazer(d2a, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(d2a.coef[ ,2]),
          align = TRUE, no.space=TRUE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Prediction of Education on Income",
          dep.var.labels = c("Annual Income"))
```

```
##
## Prediction of Education on Income
## ==========================================
##                      Dependent variable:
##                  ---------------------------
##                          Annual Income
```

```
## ------------------------------------------------
## years_education          5,750.480***
##                            (84.411)
## Constant                -23,354.640***
##                           (1,197.226)
## ------------------------------------------------
## Observations                19,567
## Adjusted R2                  0.196
## ================================================
## Note:            *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: Each additional year of education predicts $5,750.48 additional amount of annual income.

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.

ANSWER: There are two reasons that I am not convinced that the relationship is anything causal. (i) Years of education may not be the only factor that drives the outcome (i.e., Annual Income). There may be many other factors/covariates/omitted variables that actually drive the outcome (ii) Even if we believe that Annual Income has true effect on income, we still cannot be certain of its Compiler Average Causal Effect (CACE) or Local Average Treatment Effect (LATE).

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.)

```
# Create a dummy variable "high_rank"
d2[, high_rank := as.numeric(NA)]
d2[, high_rank := ifelse(draft_number <= 80, 1, 0)]

# Verify data
head(d2)
```

```
##    draft_number years_education    income high_rank
## 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
```

```
## Estimate the effect of having a high-ranked draft number on education
# Fit a linear model between high_rank and years of education
d2c <- lm(years_education ~ high_rank, data = d2)
d2c.coef <- coeftest(d2c, vcovHC(d2c))

## use stargazer to print formatted tables
stargazer(d2c, type = "text",
```

```
          omit.stat=c("LL","ser","f","rsq"),
          se = list(d2c.coef[ ,2]),
          align = TRUE, no.space=TRUE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Predicting Years of Education based on Draft Outcome",
          dep.var.labels = c("Years of Education"))
```

```
##
## Predicting Years of Education based on Draft Outcome
## ==========================================
##                    Dependent variable:
##               ---------------------------
##                    Years of Education
## ------------------------------------------
## high_rank              2.126***
##                         (0.038)
## Constant               14.434***
##                         (0.017)
## ------------------------------------------
## Observations            19,567
## Adjusted R2              0.138
## ==========================================
## Note:        *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: The effect of a high-rank draft is an additional 2.126 years of education.

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

```
## Estimate the effect of having a high-ranked draft number on income
# Fit a linear model between high_rank and income
d2d <- lm(income ~ high_rank, data = d2)
d2d.coef <- coeftest(d2d, vcovHC(d2d))

## use stargazer to print formatted tables
stargazer(d2d, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(d2d.coef[ ,2]),
          align = TRUE, no.space=TRUE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Predicting Income based on Draft Outcome",
          dep.var.labels = c("Income"))
```

```
##
## Predicting Income based on Draft Outcome
## ==========================================
##                    Dependent variable:
##               ---------------------------
##                         Income
## ------------------------------------------
## high_rank            6,637.554***
##                        (545.590)
## Constant             60,761.890***
##                        (233.387)
```

```
## ------------------------------------------
## Observations            19,567
## Adjusted R2              0.008
## ==========================================
## Note:        *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: The effect of high rank draft is an additional income of $6,637.554.

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-variables estimate

rank_effect_on_income <- summary(d2d)$coefficients[c("high_rank"), c("Estimate")]
rank_effect_on_educ_yr <- summary(d2c)$coefficients[c("high_rank"), c("Estimate")]
iv_estimate <- rank_effect_on_income / rank_effect_on_educ_yr

# print LATE (i.e., iv_estimate_)
sprintf("The instrumental-variables estimate is %.3f", iv_estimate)
```

```
## [1] "The instrumental-variables estimate is 3122.444"
```

```
# Try 2SLS
d2[ , pred_educ:= predict(d2c)]
d2e <- lm(data = d2, income ~ pred_educ)
d2e.coef <- coeftest(d2e, vcovHC(d2e))

## use stargazer to print formatted tables
stargazer(d2e, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(d2e.coef[ ,2]),
          align = TRUE, no.space=TRUE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Income based on predicted years of education (2SLS approach)",
          dep.var.labels = c("Income"))
```

```
##
## Income based on predicted years of education (2SLS approach)
## ==========================================
##                   Dependent variable:
##               ---------------------------
##                          Income
## ------------------------------------------
## pred_educ             3,122.444***
##                         (256.657)
## Constant              15,691.580***
##                        (3,810.346)
## ------------------------------------------
## Observations            19,567
## Adjusted R2              0.008
## ==========================================
## Note:        *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: The result suggests that those who had high-rank draft status are associated with a higher income of $3,122 due to an additional year of education.

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.

ANSWER: It is probably fair to assume that those with high-rank draft were subject to a relatively higher mortality rate due to them going to the war. Given this, we were left with the subjects (i) who pursued higher education to avoid being drafted, and thus earning more on average as a result of the higher education and (ii) those veterans who survived the war who then tended to average up the income given many plausible factors such as "better social connection from going to war", "more grit and tenacity gained during the war", or "better survival skills applied to career and business" (although the argument can be the opposite as well; e.g., some may have had depression, resulting in relatively low income, etc.).

If there is anything, regardless of the direction of the income in the second group, it is this second group that would likely be indicative of a violation in the "exclusion restriction" assumption in this study because the treatment (i.e., number of the draft) would have affected the outcome (i.e., income) through its effect other than the "endogenous variable" (i.e., education). As for the first group who pursued higher education, we cannot also be certain that the higher income was due to higher education, it is possible that other factors such as social status, connection and financial means enabling them to pursue a higher education are actually the driver behind their higher income in life later on. If this argument holds, then the assumption would be violated as well. In other words, a high draft rank may have nudged individuals with power and means to avoid the draft and be able to pursue further wealth without a disruption (i.e., and it just so happens that pursuing a higher degree is one of the many channels that could pursue to avoid the draft).

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.)

```
# check if there is any NA income
sum(is.na(d2$income))
```

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

```
# classifiy income of $0 as unobservable
d2[, income_obs := as.numeric(NA)]
d2[, income_obs := ifelse(income > 0, 1, 0)]

# measurement rate of income by ranking status
ate <- d2 %>%
  group_by(high_rank) %>%
  summarise(mean = mean(income_obs, na.rm=T), count = n())

# print ATE of observable income due to high-rank status
sprintf("The average treatment effect is %.4f",
        ate[1 ,c("mean")]-ate[2 ,c("mean")])
```

```
## [1] "The average treatment effect is -0.0028"
```

```r
# use regression model to test the hypothesis
d2g <- lm(income_obs~high_rank, data = d2)
d2g.coef <- coeftest(d2g, vcovHC(d2g))

## use stargazer to print formatted tables
stargazer(d2g, type = "text",
          omit.stat=c("LL","ser","f","rsq"),
          se = list(d2g.coef[ ,2]),
          align = TRUE, no.space=TRUE,
          star.cutoffs = c(0.05, 0.01, 0.001),
          digits = 4,
          title = "Measuremet of Income Observation",
          dep.var.labels = c("Measurement Rate"))
```

```
##
## Measuremet of Income Observation
## ==========================================
##                    Dependent variable:
##               ---------------------------
##                     Measurement Rate
## ------------------------------------------
## high_rank               0.0028*
##                        (0.0013)
## Constant               0.9925***
##                        (0.0007)
## ------------------------------------------
## Observations            19,567
## Adjusted R2             0.0001
## ==========================================
## Note:        *p<0.05; **p<0.01; ***p<0.001
```

ANSWER: Based on the outcome above, the measurement rate of the observable income is 0.28% higher for the high-rank status subjects than for the low-risk group. And this difference is statistically significant at 95% confidence level, which is an indication that there is a differential attrition. However, in terms of practical significance, this may not matter much; especially when we look at the large number of subjects in each group as shown below.

```r
d2 %>%
  group_by(high_rank) %>%
  summarise(mean = mean(income_obs, na.rm=T), count = n())
```

```
## # A tibble: 2 x 3
##   high_rank  mean count
##       <dbl> <dbl> <int>
## 1         0 0.993 15671
## 2         1 0.995  3896
```

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

ANSWER: The high-rank status has a relatively higher rate of observable income at 99.54% compared with 99.25% and is statistically significant at 95% confidence level. For the high-rank subjects that elected to pursue higher education, statistical evidence shows that they tend to earn more, hence higher chance that their income will be greater than 0 and observable. For the high-rank subjects that were eventually drafted, a proportion of them would be subject to

some form of government / veteran benefits that requires income report to IRS anyway. Given that the proportion of those who have to report income due to some form of benefits is likely equal (assuming randomization) among those who went to the war (regardless of high-rank or not), the more likely explanation of the result in part G is likely due to the effect of higher education among those in high-rank status who elected to do so.

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

ANSWER: As alluded to in my response to (2f), this differential attrition could be due to a group of people (among those in the high-rank draft group) who had means and power to avoid getting draft, and it is those very factors (out of education) that actually enabled and assisted them in pursuing further wealth and income later in life. It just so happens that "higher education" was one of the means they could use to absolve themselves from the war, and hence their action to pursue higher education may actually have inflated the importance of education on income (when in fact other unknown or omitted factors may have a larger role to play in driving income for this special group).

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 healp 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?
- What design and measurment improvements do you suggest?

ANSWER: The evidence in drawing the conclusion is not convincing for the following reasons. I also provide suggestions for improving the study in line.

The study concludes that a plate size leads to weight loss. The flaw in this conclusion is that plate size is only an instrument variable that leads to consumption amount which then probably leads to weight loss. Potentially, there are one or many "mediation" factors that actually have more direct link to weight loss. The study should attempt to find out what these mediators such as "ingredients of the food used in this study" might be. Further, the researcher should try to find out if different types of food (e.g., meat, vegetable) would lead to the same conclusion, and the sensitivities of those to weight loss (i.e., one study on one type of food doesn't generalize well).

Secondly, the study is a cross-sectional study, attempting to understand the treatment effect only at a point in time. In a more realistic setup, weight loss does not happen overnight, so a more proper study would be track the effect of the treatment over time. I would suggest a panel study that tracks outcomes of multiple subjects over time.

Third, the outcome from this study is only a single data point. If we are to replicate this study given a different context (e.g., location of the study, different demographic, different types of food), would we observe a similar outcome? I would suggest a replication of a similar study in more various contexts to form an evidence-based recommendation.

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.

ANSWER: The average treatment effect (ATE) is important as it tells us how outcomes would change on average were all subjects to go from untreated to treated. For each subject, the causal effect of the treatment (i.e., subject-level treatment effect) is the difference between two potential outcomes (i.e., between control/untreated and treated). ATE is the sum of the subject-level treatment effects divided by the total number of subjects.

Under an appropriate experimental design ATE is a "good" measure for difference in potential outcomes as ATE is an unbiased measure under such design, which must meet the following assumptions: (i) Random Assignment, (ii) Exclusion Restriction (i.e., only relevant causal agent is receipt of the treatment) and (iii) Non-Interference (i.e., no matter which subjects the random assignment allocates to treatment or control, a given subject's potential outcomes remain the same).