Problem Set 5

Pavan Kurapati

August 7, 2018

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)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
library(multiwayvcov)
library(lmtest)
library(AER)
library(ivpack)
d1 <- fread('./data/ps5_no1.csv')
d1
```

```
##         product_b total_ad_exposures_week1 treatment_ad_exposures_week1
##     1:          1                        4                            3
##     2:          1                        1                            1
##     3:          1                        3                            1
##     4:          0                        5                            0
##     5:          0                        1                            1
##     ---
## 499996:          1                        2                            1
```

```
## 499997:              1                              4                                0
## 499998:              0                              5                                2
## 499999:              1                              5                                1
## 500000:              0                              3                                1
##          week0 week1 week2 week3 week4 week5 week6 week7 week8 week9 week10
##       1:   5.5   6.2   0.0   0.0   0.0   0.0     0   0.0   9.7   4.1    0.0
##       2:   6.2   0.0   8.6   2.4   0.0   7.4     0   0.0   0.0   5.7    0.0
##       3:   0.0   5.3   0.0   8.1   7.8   3.3     0   0.0   9.4   0.0    0.0
##       4:   0.0   4.1   0.0   8.8   5.8   5.9     0   0.0   0.0   9.6    0.0
##       5:   7.6   3.6   4.6   5.5   7.2   7.1     0   0.0   0.0   0.0    0.0
##      ---
## 499996:   0.0   2.8   8.1   2.8   4.1   8.1     0   0.0   0.0   0.0    2.3
## 499997:   0.0   0.0   4.0   4.6   0.0   0.0     0   5.0   0.0   0.0    0.0
## 499998:   4.3   0.0   0.0   0.0   4.2   6.5     0   4.8   9.4   6.4    9.5
## 499999:   0.0   0.0   0.0   5.4   6.1   0.0    12   7.7   0.0   0.0    0.0
## 500000:   4.7   0.0   0.0   2.2   0.0   0.0     0   4.3   0.0   4.9    0.0
```

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.)

```
# Check if any total ad exposure is greater than 6
nrow(subset(d1,d1$total_ad_exposures_week1>6))
```

[1] 0

```
# Any total ad exposures is less than treatment_ad_exposures?
nrow(subset(d1,d1$total_ad_exposures_week1<d1$treatment_ad_exposures_week1))
```

[1] 0

```
# How many scenarios where the user is exposed to treatment for every login
nrow(subset(d1,d1$total_ad_exposures_week1==d1$treatment_ad_exposures_week1))
```

[1] 138084

```
# How many scenarios where the user is exposed to only control for every login
nrow(subset(d1,d1$treatment_ad_exposures_week1==0))
```

[1] 137580

There are a total of 138084 users who received treatment exclusively. There are also 137580 users who received control exlcusively. The data looks reasonable, however we need to analyze the number of times a user is exposed to a treatment and its true impact.

   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?

```
model_naive = lm(week1 ~ treatment_ad_exposures_week1, data=d1)
summary(model_naive)
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ad_exposures_week1, data = d1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.409 -2.213 -1.615  2.388  8.285
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.614684   0.005995  269.34   <2e-16 ***
## treatment_ad_exposures_week1 0.299113   0.003138   95.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.781 on 499998 degrees of freedom
## Multiple R-squared:  0.01785,    Adjusted R-squared:  0.01785
## F-statistic:  9086 on 1 and 499998 DF,  p-value: < 2.2e-16
```

```
# Run Placebo on week0
model_placebo_0 = lm(week0 ~ treatment_ad_exposures_week1, data=d1)
summary(model_placebo_0)
```

```
##
## Call:
## lm(formula = week0 ~ treatment_ad_exposures_week1, data = d1)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.248 -2.196 -1.670  2.430  8.330
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.669685   0.006027   277.0   <2e-16 ***
## treatment_ad_exposures_week1 0.263099   0.003155    83.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.796 on 499998 degrees of freedom
## Multiple R-squared:  0.01372,    Adjusted R-squared:  0.01372
## F-statistic:  6955 on 1 and 499998 DF,  p-value: < 2.2e-16
```

From the placebo test above, it implies that the treatment given in week1 has a statistically significant impact on week0 which is totally unrelated to the treatment. This implies that our naive model is not right. The placebo test has failed because we should ideally expect no statistical significance in this test.

   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.*)

There are quite a lot of users who are exposed to treatment advertisements more than once. These are repeat users who login frequently. The probability of being assigned to treatment is more for users who login frequently than those who login fewer times, so the treatment assignment is not truly random. Users who login more often might be more active buyers anyway so the treatment effect will be overstated for these users. While it is mentioned in the assumptions that effect of first exposure is same as the second exposure, the overall treatment effect may not be constant. We need to run fixed effect test on total_ad_exposures to address the impact of number of logins.

   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.)

```
# Use total_ad_exposures_week1 as factor and include in the model
# Run Placebo on week0.
model_fixed_effect = lm(week0 ~ treatment_ad_exposures_week1 +
                         as.factor(total_ad_exposures_week1), data=d1)
summary(model_fixed_effect)
```

```
##
## Call:
## lm(formula = week0 ~ treatment_ad_exposures_week1 + as.factor(total_ad_exposures_week1),
##     data = d1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.847 -2.084 -1.616  2.458  7.818
##
## Coefficients:
```

4

```
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            1.306783    0.011234 116.328   <2e-16
## treatment_ad_exposures_week1          -0.002287    0.004629  -0.494    0.621
## as.factor(total_ad_exposures_week1)1   0.311564    0.015362  20.281   <2e-16
## as.factor(total_ad_exposures_week1)2   0.551112    0.015469  35.627   <2e-16
## as.factor(total_ad_exposures_week1)3   0.779854    0.016264  47.948   <2e-16
## as.factor(total_ad_exposures_week1)4   1.003116    0.017545  57.174   <2e-16
## as.factor(total_ad_exposures_week1)5   1.236092    0.019372  63.810   <2e-16
## as.factor(total_ad_exposures_week1)6   1.539865    0.021981  70.053   <2e-16
##
## (Intercept)                          ***
## treatment_ad_exposures_week1
## as.factor(total_ad_exposures_week1)1 ***
## as.factor(total_ad_exposures_week1)2 ***
## as.factor(total_ad_exposures_week1)3 ***
## as.factor(total_ad_exposures_week1)4 ***
## as.factor(total_ad_exposures_week1)5 ***
## as.factor(total_ad_exposures_week1)6 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.779 on 499992 degrees of freedom
## Multiple R-squared:  0.02563,    Adjusted R-squared:  0.02562
## F-statistic:  1879 on 7 and 499992 DF,  p-value: < 2.2e-16
```

We now have a non-significant effect on week0, which is the correct representation of the treatment effect on the week0 outcome.

  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).

```
# Use total_ad_exposures_week1 as factor and include in the model
# Run Placebo on week0.
model_week1 = lm(week1 ~ treatment_ad_exposures_week1 +
                     as.factor(total_ad_exposures_week1), data=d1)
summary(model_week1)
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ad_exposures_week1 + as.factor(total_ad_exposures_week1),
##     data = d1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.003 -2.102 -1.552  2.447  8.110
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            1.295932   0.011185  115.87   <2e-16
## treatment_ad_exposures_week1           0.056337   0.004609   12.22   <2e-16
## as.factor(total_ad_exposures_week1)1   0.256600   0.015295   16.78   <2e-16
## as.factor(total_ad_exposures_week1)2   0.487427   0.015402   31.65   <2e-16
## as.factor(total_ad_exposures_week1)3   0.693409   0.016194   42.82   <2e-16
## as.factor(total_ad_exposures_week1)4   0.919351   0.017469   52.63   <2e-16
## as.factor(total_ad_exposures_week1)5   1.136115   0.019287   58.91   <2e-16
## as.factor(total_ad_exposures_week1)6   1.368843   0.021886   62.55   <2e-16
```

5

```
## 
## (Intercept)                             ***
## treatment_ad_exposures_week1            ***
## as.factor(total_ad_exposures_week1)1 ***
## as.factor(total_ad_exposures_week1)2 ***
## as.factor(total_ad_exposures_week1)3 ***
## as.factor(total_ad_exposures_week1)4 ***
## as.factor(total_ad_exposures_week1)5 ***
## as.factor(total_ad_exposures_week1)6 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.767 on 499992 degrees of freedom
## Multiple R-squared:  0.02783,    Adjusted R-squared:  0.02782
## F-statistic:  2045 on 7 and 499992 DF,  p-value: < 2.2e-16
```

From the above, the impact of treatment_ad_exposures is 5.6% increase in week1 sales. However, there is also a general impact on total number of ad exposures. As mentioned in the previous section, the more times the user logs in, the better their sales will be, irrespective of the treatment advertisement.

    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.

Say Sam is an active online shopper and I buy every item online, from pencils to tractor parts. Sam logs in multiple times a week to make my purchases. Tom is an infrequent online shopper, and might login once or twice a week. Tom does not mind walking into his next door retail store if needed. Naturally, Sam's probability of buying any given product is much higher than Tom's. Also, Sam has higher probability of viewing the advertisement for a given product and be enticed to purchase it. The fixed effect test tries to capture the number of times a user logs in and includes that into the model. If we exclude this, the buying of Sam is overstated as the treatment effect.

    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).

```
# Let us first add the total purchases for weeks 1 through 10
d1$total_week1_10 <- rowSums(d1[, c(5:14)])
# Now do the regression
model_1_10 <- lm(total_week1_10~treatment_ad_exposures_week1 +
                   as.factor(total_ad_exposures_week1),data=d1)
summary(model_1_10)
```

```
## 
## Call:
## lm(formula = total_week1_10 ~ treatment_ad_exposures_week1 +
##     as.factor(total_ad_exposures_week1), data = d1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.712  -7.363  -0.722   6.656  59.977
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       16.90169    0.04267 396.102   <2e-16
## treatment_ad_exposures_week1       0.01250    0.01758   0.711    0.477
```

6

```
## as.factor(total_ad_exposures_week1)1  2.59989   0.05835  44.555   <2e-16
## as.factor(total_ad_exposures_week1)2  4.84270   0.05876  82.418   <2e-16
## as.factor(total_ad_exposures_week1)3  7.03614   0.06178 113.891   <2e-16
## as.factor(total_ad_exposures_week1)4  9.08326   0.06664 136.297   <2e-16
## as.factor(total_ad_exposures_week1)5 11.19600   0.07358 152.157   <2e-16
## as.factor(total_ad_exposures_week1)6 13.73562   0.08350 164.508   <2e-16
##
## (Intercept)                            ***
## treatment_ad_exposures_week1
## as.factor(total_ad_exposures_week1)1 ***
## as.factor(total_ad_exposures_week1)2 ***
## as.factor(total_ad_exposures_week1)3 ***
## as.factor(total_ad_exposures_week1)4 ***
## as.factor(total_ad_exposures_week1)5 ***
## as.factor(total_ad_exposures_week1)6 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.55 on 499992 degrees of freedom
## Multiple R-squared:  0.1322, Adjusted R-squared:  0.1322
## F-statistic: 1.088e+04 on 7 and 499992 DF,  p-value: < 2.2e-16
```

The treatment ad exposure increases the total sales from week1-10 when by 1.2%, which is a decrease from week1 sales alone. This shows that the effect is reducing in the subsequent weeks.

    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.

```
# Let us first add the total purchases for weeks 2 through 10
d1$total_week2_10 <- rowSums(d1[, c(6:14)])
# Now do the regression
model_2_10 <- lm(total_week2_10~treatment_ad_exposures_week1 +
                 as.factor(total_ad_exposures_week1),data=d1)
summary(model_2_10)
```

```
##
## Call:
## lm(formula = total_week2_10 ~ treatment_ad_exposures_week1 +
##     as.factor(total_ad_exposures_week1), data = d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.973  -7.105  -0.705   6.394  54.266
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           15.60576    0.04087 381.800  < 2e-16
## treatment_ad_exposures_week1          -0.04384    0.01684  -2.603  0.00925
## as.factor(total_ad_exposures_week1)1   2.34329    0.05590  41.922  < 2e-16
## as.factor(total_ad_exposures_week1)2   4.35527    0.05628  77.379  < 2e-16
## as.factor(total_ad_exposures_week1)3   6.34273    0.05918 107.178  < 2e-16
## as.factor(total_ad_exposures_week1)4   8.16391    0.06384 127.884  < 2e-16
## as.factor(total_ad_exposures_week1)5  10.05988    0.07048 142.724  < 2e-16
## as.factor(total_ad_exposures_week1)6  12.36677    0.07998 154.621  < 2e-16
##
```

```
## (Intercept)                           ***
## treatment_ad_exposures_week1          **
## as.factor(total_ad_exposures_week1)1  ***
## as.factor(total_ad_exposures_week1)2  ***
## as.factor(total_ad_exposures_week1)3  ***
## as.factor(total_ad_exposures_week1)4  ***
## as.factor(total_ad_exposures_week1)5  ***
## as.factor(total_ad_exposures_week1)6  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 499992 degrees of freedom
## Multiple R-squared:  0.1156, Adjusted R-squared:  0.1156
## F-statistic:  9333 on 7 and 499992 DF,  p-value: < 2.2e-16
```

The treatment effect on week2 to week10 total sales is negative implying the fading effect of advertisement over a period of time.

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

When an advertisement is seen, its impact is short lived and can last only for a small duration of time. A user watching an advertisement in week-1 will possibly have a positive impact in purchase for that week due to short memory. The memory fades away after that, and a purchase made in subsequent weeks could only indicate purchasing by chance rather than the impact of treatment.

    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.*)

```r
# We will use product_b as indicator variable to find if product_b is more effective than product_a
model_prodB <- lm(week1~product_b+
                    treatment_ad_exposures_week1+
                    as.factor(total_ad_exposures_week1),data=d1)

summary(model_prodB)
```

```
##
## Call:
## lm(formula = week1 ~ product_b + treatment_ad_exposures_week1 +
##     as.factor(total_ad_exposures_week1), data = d1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.039 -2.195 -1.510  2.436  8.073
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.276834   0.011230  113.70   <2e-16
## product_b                            0.154743   0.008473   18.26   <2e-16
## treatment_ad_exposures_week1         0.056205   0.004607   12.20   <2e-16
## as.factor(total_ad_exposures_week1)1 0.232711   0.015346   15.16   <2e-16
## as.factor(total_ad_exposures_week1)2 0.450794   0.015527   29.03   <2e-16
## as.factor(total_ad_exposures_week1)3 0.651241   0.016352   39.83   <2e-16
## as.factor(total_ad_exposures_week1)4 0.870607   0.017666   49.28   <2e-16
```

```
## as.factor(total_ad_exposures_week1)5 1.072653    0.019592    54.75    <2e-16
## as.factor(total_ad_exposures_week1)6 1.270459    0.022532    56.38    <2e-16
##
## (Intercept)                           ***
## product_b                             ***
## treatment_ad_exposures_week1          ***
## as.factor(total_ad_exposures_week1)1 ***
## as.factor(total_ad_exposures_week1)2 ***
## as.factor(total_ad_exposures_week1)3 ***
## as.factor(total_ad_exposures_week1)4 ***
## as.factor(total_ad_exposures_week1)5 ***
## as.factor(total_ad_exposures_week1)6 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.766 on 499991 degrees of freedom
## Multiple R-squared:  0.02848,    Adjusted R-squared:  0.02847
## F-statistic:  1832 on 8 and 499991 DF,  p-value: < 2.2e-16
```

In the above regression, product_b is an indicator variable with 1 for product B and 0 for product A. Since the coefficient is positive, it implies that product B is more effective than 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?

I would not be confident after the results of 'j'. If all other parameters are same, then including celebrity endorsement seem to reduce the effectiveness of advertising at stimulating immediate purchases.

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

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?

```
##       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
##   ---
## 19563:          76              18 105795.30
## 19564:         168              14  94328.98
## 19565:          11              14  28049.30
## 19566:         113              18  52886.20
## 19567:         317              17  87578.65
```

```
model_naive = lm(income~years_education,data=d2)
summary(model_naive)
```

```
##
## Call:
## lm(formula = income ~ years_education, data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91655  -17459    -837   16346  141587
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -23354.64    1252.74  -18.64   <2e-16 ***
## years_education   5750.48      83.34   69.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26590 on 19565 degrees of freedom
## Multiple R-squared:  0.1957, Adjusted R-squared:  0.1957
```

```
## F-statistic:   4761 on 1 and 19565 DF,   p-value: < 2.2e-16
```
```
edu_income_naive <- summary(model_naive)$coeff[2,1]
```

The naive regression suggests that the income increases by 5750 USD for every additional year of education

    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.

Income is dependent on many factors. For example, ones ability to perform can have a huge contribution to how much income one can generate.Omitting this will result in large positive bias in education alone. In other words, education can have correlation to other variables which might also impact income. So we cannot control for other factors through this naive regression. We need to conduct a 2SLS and use an Instrumental Variable to control only for education.

    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.)

```
d2$high_rank <- ifelse(d2$draft_number<=80,1,0)
model_edu_rank <- lm(years_education~high_rank,data=d2)
#summary(model_edu_rank)
# Standard errors with draft number as cluster
model_edu_rank$cluster.vcov <- cluster.vcov(model_edu_rank, d2$draft_number)
cf <- coeftest(model_edu_rank, model_edu_rank$cluster.vcov)
cf
```

```
##
## 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
```
```
rank_edu_coef <- cf[2,1]
```

Having high_rank causes an increase of education by 2.12 years. Standard error is 0.038

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

```
model_income_rank <- lm(income~high_rank,data=d2)
#summary(model_income_rank)
# Standard errors with draft number as cluster
model_income_rank$cluster.vcov <- cluster.vcov(model_income_rank, d2$draft_number)
cf <- coeftest(model_income_rank, model_income_rank$cluster.vcov)
cf
```

```
##
## 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
rank_income_coef <- cf[2,1]
```

Having high_rank causes an increase of income by 6637.55 USD. Standard error is 511.90

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

```
edu_income <- rank_income_coef/rank_edu_coef
edu_income
```

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

Each year of education causes an increase in income by 3122.4437939 USD. This is a much better estimate than the overstated value of the naive model which showed 5750.4796478 USD

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

It is possible that the people who were born in 4th quarter of the year are significantly different to the first quarter of the year in terms of their maturity, abilities and seniority of the work. So, in a given year (Say for 1951 birth cohort), people born in first quarter are high ranked but they may also have greater potential to generate higher total income because of their ~1 year more experience. Especially between 1970-to-early 2000s most jobs gave higher weightage to seniority. Hence the high ranking draft lottery based on birth month violates the exclusion restriction.

It is also possible that those who are drafted and go to military end up making good income after the war by joining private firms which deal with security, irrespective of their education status.

 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 try to find the average income of high and low rank individuals
high_rank_subset = subset(d2,d2$high_rank==1)
n_high = nrow(high_rank_subset)
low_rank_subset = subset(d2,d2$high_rank==0)
n_low = nrow(low_rank_subset)

Average_high_rank_income = sum(high_rank_subset$income)/n_high
Average_low_rank_income = sum(low_rank_subset$income)/n_low

Average_high_rank_income
```

```
## [1] 67399.44
```

```
Average_low_rank_income
```

```
## [1] 60761.89
```

There is a difference in income between the two that amounts to **6637.554244 USD**, with high rank income average higher than the low rank income.

```
# How many high rank individuals have income of 0?
nrow(subset(high_rank_subset,high_rank_subset$income==0))
```

```
## [1] 18
```

```
# How many low rank individuals have income of 0?
nrow(subset(low_rank_subset,low_rank_subset$income==0))
```

```
## [1] 117
```

The presence of **18 individuals with 0 income** implies that either the people in this high rank draft are dead or the outcome was not measured. Similarly, there are **117 individuals with reported income of 0 in low ranked draft. Attrition occurs when the outcome data is missing, and this leads to biased estimates of the average treatment effect. While a 0 income is possible for those who are unemployed, it is very unlikely that anyone would report 0 income for a long duration. It is higly likely that it is a missing data for those individuals.**

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

The absence of reported outcome variable (income) could be because these are soldiers who went to war and were either dead or were permanenly disabled. Had they been alive or able to work, they would have had an income reported which would lead to an unbiased estimate of outcome. It is also possible that some of the people have changed address and the experimenters were unable to contact them.

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

Say the 18 high ranked individuals were reported (measured) but had really low income. The average treatment outcome would have been lesser in this case. Hence, our ATE is biased higher in the absence of these individuals. Similarly, the 117 low ranked individuals in the control group would have had higher income that would have increased the average income of control group.

3. Green and Gerber Practice Problems

Note, none of these require you to program anything. Instead, I'm aiming to have you think critically about these forms of designs. Have fun!

a. Field Experiments 11.10

11.10(a) average effect of assignment to the script that encouraged voting:

We follow the linear function model described in 11.4. i,e for **11-20 seconds, we divide it 2 and 21-30 seconds we divide it by 3.**

(Y_i(1-10) + Y_i(11-20)/2 + Y_i(21-30)/3 + Y_i(31-40)/4)/4

(-0.9 + (-0.9/2) + 0.8/3 + 4.5/4)/4 = 0.0104

11.10(b): DOes this table provide convincing evidence that "the longer a person listens to a recorded message that encourages voting, the more effective that medssage will be in terms of boosting voter turnout"?

No, it does not provide a convincing evidence. Here, the treatment varies in intensity, so we need to take the intentisy into account when measuring the treatment effect. After working on a simple linear function model, the effect is miuch lesser compared to what the table actually shows

b. Field Experiments 12.3

(The paper is available at `./readings/Olken.2010.pdf`.)

(a): The experimental subjects are the **49 Indonesian villages in three subdistricts** that are from different parts of rural Indonesia. The defalt mode of selecting the projects is through "Kecamatan Development Program" where villagers organize multiple meetings/gatherings to decide on a one group project and one women project. Villages that are assigned to KDP form control group. The intervention occurs for those villages where the selection of projects is through Plebiscite (direct election) method. These fom the treatment group. The treatment is the assignment of plebiscite.

(b): The sample sizes of treatment would be much smaller compared to the sample sizes of control group because the number of villages participating in plebiscite is small. It is generally common to have less sample size in treatment group for field experiments where the treatment is expensive. However, if the allocations are random, both groups exhibit same characteristics and if the sample size is large enough, central limit theorem should kick in. If the sample size is too small, I would be worried about higher standard errors.

(c): There was no reference provided as to who conducted the surveys. However, from the footnote that states time pressure, I would assume that it is local representatives either of the same village or of the same sub-district. I do not think the interviewers were blinded w.r.to the treatment assignment if they belong to the same sub-district. If the interviewers know the assingmnent, they may lead the respondents to reply in a way that satisfies their own views. So, yes I think the results of a soft parameter like "satisfaction" could be heavily biased.

(d): In page **245**, the author explains the geographic structure. "Sometimes the hamlets within a village are adjacent, but often the various hamlets in the village are separated by agricultural fields, and can be as much as **1 to 2 km** away from each other.". I think this certainly leads to spillover effects, especially between the villages that are adjacent to each other and are assigned to treatment and control groups. The villagers in treatment group might speak to the ones in control group belonging to elite, and might get influenced by the elite decisions on the projects selected. To avoid spillover effects, the treatment allocation must consider geographic boundaries that separate them far apart and avoid any potential contact that can lead to spillovers.

(e): The experiment does provide evidence in terms of statistical significance for the claims that plebiscite increases satisfaction for both general and women projects. The author does note several limitations in the experiments such as 1) Lobbying and transfers (payments of rents etc.) influencing the plebscite process and satisfaction 2) The lower sample size of the study (49 villages in 3 provinces) may not generalize it to entire Indonesia or to other countries and 3) The fading effect of satisfaction in longer term. The other limitations that do occur in such experiments is the act of coercion by village heads (elites) in voting process, attritions (not many may come and vote) etc. Although the author claims that the representation in plebiscite was high, this might change over a period of time.

(f): Democracy is very appealing in the beginning to those who are not exposed to it for certain processes. In this case, the default was a village meeting which was poorly represented. So

the first and second wave of the experiment might create the sense of excitement that creates higher initial satisfaction. However, on long run, the reality of lobbying by inter-village heads might sync in and change the perception. Also, in the first wave of surveys, hawthrone effect might cause the respondents to choose to respond more positively. This might change in subsequent surveys.

c. Field Experiments 12.5

(a): Let us assume that the treatment group assignment is done in random fashion (say every other person is given a smaller plate). Since they are attending dinner in the same room, the subjects in control group can see others in treatment group(small plate) and alter their eating habbit to match small plates. Also, the experiment states that the measurement is done through camera placed. It is very difficult to measure the intake through viewing in camera, as large plates might appear more in quantity than the actual calorie intake

(b): The first suggestion would be to conduct experiments for treatment and control groups separately so that there are no spillovers. For example, arrange a dinner for treatment group and serve small plates. Next arrange a dinner for control group and serve large plates. If such an experiment is done, the selection of people in treatment and control becomes important. One needs to pick people of similar height, weight and eating priorities which is extremely difficult to achieve. The second suggestion would be to improve the measurement. Experimenters must place people who can measure the quantity and content being served for each individual to capture the precise calorie intake per individual.

4. Skip in 2017

5. 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.

ATE takes into consideration the effect of treatment on large samples of a population. Sample size being the key, it gives us a more efficient view of causal effect. An individual's potential outcome cannot be generalized to apply to larger samples because the treatment effect may vary based on many factors. If an experiment is done where the assignment to treatments is truly randomized and other factors like clustering/blocking is taken into account, ATE gives us a better causal metric. We also learnt about other techniques such as ITT, CACE for scenarios related to non-compliance that helps us in guaging the true effect.