Problem Set #5

Alex Smith

April 11, 2015

```
# browseURL("https://drive.google.com/file/d/0B_QjOotlErJqbWtESk1zaVJtdnM/view?usp=sha
# browseURL("https://drive.google.com/file/d/0B_QjOotlErJqTTl2cXVuMENBSEk/view?usp=sha
setwd("/Users/Alex/Documents/Berkeley/1601Spring/W241/PS5")
```

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.

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.)

```
# read in the data and preview it
yahoo <- read.csv("ps5_no1.csv")
head(yahoo)
```

```
##   product_b total_ad_exposures_week1 treatment_ad_exposures_week1 week0
## 1         1                        4                            3   5.5
## 2         1                        1                            1   6.2
## 3         1                        3                            1   0.0
## 4         0                        5                            0   0.0
## 5         0                        1                            1   7.6
## 6         1                        4                            4   6.3
##   week1 week2 week3 week4 week5 week6 week7 week8 week9 week10
## 1   6.2   0.0   0.0   0.0   0.0   0.0     0   9.7   4.1    0.0
## 2   0.0   8.6   2.4   0.0   7.4   0.0     0   0.0   5.7    0.0
## 3   5.3   0.0   8.1   7.8   3.3   0.0     0   9.4   0.0    0.0
## 4   4.1   0.0   8.8   5.8   5.9   0.0     0   0.0   9.6    0.0
## 5   3.6   4.6   5.5   7.2   7.1   0.0     0   0.0   0.0    0.0
## 6   5.5   9.8   5.0   0.0   0.0   7.7     0  11.0   4.8    6.9
```

```
# create and print a frequency table that shows treat ad views by
# total ad views
crosstab_yahoo <- xtabs(~total_ad_exposures_week1 +
                          treatment_ad_exposures_week1,
                        data = yahoo)
crosstab_yahoo
```

```
##                          treatment_ad_exposures_week1
## total_ad_exposures_week1     0     1     2     3     4     5     6
##                        0 61182     0     0     0     0     0     0
##                        1 36754 37215     0     0     0     0     0
##                        2 21143 42036 20965     0     0     0     0
##                        3 10683 32073 32314 10726     0     0     0
##                        4  5044 20003 30432 20223  5115     0     0
##                        5  2045 10563 20970 20793 10293  2131     0
##                        6   729  4437 10977 14771 11147  4486   750
```

Yes, this distribution of impressions look reasonable. The zeroes in the upper right half of the table reveal that we never have more treatment ad exposures than total ad exposures. This makes intuitive sense because total ad exposures is the sum of both treatment and control ad exposures.

Also, the distribution reveals that the pseudo-randomization worked. Most people saw about half as many treatment ads as total ads. The number of seeing more or less than half of their ads as treatments decreases symmetrically.

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?

```r
# let's try out my colleague's model and see what happens
colleague_yahoo <-lm(week1 ~ treatment_ad_exposures_week1, data = yahoo)
summary(colleague_yahoo)
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ad_exposures_week1, data = yahoo)
##
## 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
```

```r
# my colleague's test shows a statistically significant effect
# we should remember that largeness of this sample size may help
# make any observed effect statistically significant

# let's run a placebo test of the ad exposures on the week's purchases
# prior to the ads
placebo_yahoo <- lm(week0 ~ treatment_ad_exposures_week1, data = yahoo)
summary(placebo_yahoo)
```

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

The placebo test fails in that is produces a statistically significant effect. The placebo test confirms our suspicions. We find both models produce coefficients that are statistically significant. Our intuition should make clear that is unreasonable to think that a person's exposure to ads would cause an uptick in sales prior to seeing the ads. Therefore, we might assume that for some reason those in the treatment group may have purchased more regardless.

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 is something spoling the data analysis. Up till now, I've been assuming that individuals have been assigned to treatment or control with equal probility. However, there is a large portion of people who were never exposed to any ads because they never visited Yahoo. This group is likely different from the group of people who do visit Yahoo. Also, the more you visit Yahoo, the more likely you are to receive treatment. This harms our idea of pure random assignment. In order to analyze the data correctly, we need to control for the total number of visits (i.e. we need to compare those who visited Yahoo with those who visited Yahoo).

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.)

```
# let's try the placebo test controlling for the total number
# of ads that an individual was exposed to
placebo_yahoo2 <- lm(week0 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
                     data = yahoo)
summary(placebo_yahoo2)
```

```
##
## Call:
## lm(formula = week0 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
##     data = yahoo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.817 -2.079 -1.589  2.455  7.823
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.345375   0.007295 184.436   <2e-16 ***
## treatment_ad_exposures_week1 -0.002245   0.004629  -0.485    0.628
## total_ad_exposures_week1      0.245348   0.003149  77.922   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.779 on 499997 degrees of freedom
## Multiple R-squared:  0.02555,    Adjusted R-squared:  0.02555
## F-statistic:  6556 on 2 and 499997 DF,  p-value: < 2.2e-16
```

When controlling for the total ads an individual sees, the placebo test passes. The number of treatment ads no longer correlates with statistical significance to the purchases pre-treatment.

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).

```r
# create a linear regression model to capture the effect of ad exposures
# on purchases when controlling for total ads seen
current_yahoo <- lm(week1 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
                    data = yahoo)
summary(current_yahoo)
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
##     data = yahoo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.003 -2.104 -1.542  2.447  8.110
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                  1.317960   0.007263  181.47   <2e-16 ***
## treatment_ad_exposures_week1 0.056340   0.004609   12.22   <2e-16 ***
## total_ad_exposures_week1     0.224478   0.003135   71.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.767 on 499997 degrees of freedom
## Multiple R-squared:  0.02782,    Adjusted R-squared:  0.02781
## F-statistic:  7153 on 2 and 499997 DF,  p-value: < 2.2e-16
```

```
# get the coefficient
current_effect <- coef(summary(current_yahoo))[2]
```

The estimated causal effect of each ad exposure on purchases during the week of the campaign is an increase of **0.0563399**.

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.

My estimation strategy is superior to my colleague's because he falsely assumes randomization. As part a of this problem reveals, there is a large portion of individuals who never visited Yahoo. In order for us to make a causal claim, we must be comparing apples-to-apples. In other words, the individuals who did not see any treatment ads must be fundamentally the same as those individuals who did see the treatment ads. Even among those who did see Yahoo ads, those individuals who saw more ads must be fundamentally the same as those who saw less ads. It is not safe to assume that those who frequent Yahoo more are the same as those who frequent Yahoo less. Perhaps those who frequent Yahoo more will purchase more anyway. In order to compare apples-to-apples, we must only compare people who have the same number of Yahoo visits. We can do this by controlling for how many total ads the individual saw.

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 a variable to store the total purchases from weeks 1 through 10
yahoo$weeks1_10 <- yahoo$week1 + yahoo$week2 + yahoo$week3 + yahoo$week4 +
  yahoo$week5 + yahoo$week6 + yahoo$week7 + yahoo$week8 + yahoo$week9 + yahoo$week10

# create a linear regression model to capture the effect of ad exposures
# on all purchases when controlling for total ads seen
total_yahoo <- lm(weeks1_10 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
                  data = yahoo)
summary(total_yahoo)
```

```
##
## Call:
## lm(formula = weeks1_10 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
##     data = yahoo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.597  -7.372  -0.731   6.654  59.782
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   17.15081    0.02771 618.949   <2e-16 ***
## treatment_ad_exposures_week1  0.01274    0.01758   0.724    0.469
## total_ad_exposures_week1      2.22834    0.01196 186.307   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.56 on 499997 degrees of freedom
## Multiple R-squared:  0.1321, Adjusted R-squared:  0.1321
## F-statistic: 3.804e+04 on 2 and 499997 DF,  p-value: < 2.2e-16
```

```
# get the coefficient
current_effect <- coef(summary(total_yahoo))[2]
```

When I look at all the purchases, I find that the effect of seeing treatment ads loses its statistical significance. The effect of each ad seen on total purchases for the 10 weeks is only **0.0127386**, which is not significantly different than zero.

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 a variable to store the total purchases from weeks 2 through 10
yahoo$weeks2_10 <- yahoo$week2 + yahoo$week3 + yahoo$week4 + yahoo$week5 +
  yahoo$week6 + yahoo$week7 + yahoo$week8 + yahoo$week9 + yahoo$week10

# create a linear regression model to capture the effect of ad exposures
# on purchases post experiment when controlling for total ads seen
post_yahoo <- lm(weeks2_10 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
                 data = yahoo)
summary(post_yahoo)
```

```
##
```

```
## Call:
## lm(formula = weeks2_10 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
##     data = yahoo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.856  -7.097  -0.697   6.382  54.079
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   15.83285    0.02654 596.493  < 2e-16 ***
## treatment_ad_exposures_week1 -0.04360    0.01684  -2.588  0.00964 **
## total_ad_exposures_week1      2.00387    0.01146 174.901  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 499997 degrees of freedom
## Multiple R-squared:  0.1154, Adjusted R-squared:  0.1154
## F-statistic: 3.261e+04 on 2 and 499997 DF,  p-value: < 2.2e-16
```

```
# get the coefficient
post_effect <- coef(summary(post_yahoo))[2]
```

In weeks 2 through 10 (after the experiment), there is a statistically significant effect of **-0.0436013** on purchases for each treatment ad seen.

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

When controlling for visits to Yahoo, we found a positive treatment effect of ads on purchases during the week of the experiment. After the experiment concluded, we found a negative treatment effect of ads on purchases in the following weeks. In layman's terms, during the week of the experiment, we saw an increase in purchases but after the experiment ended, we saw a decrease in purchases for those individuals seeing the ads. This would make perfect sense if the treatment displaced purchases from the future into the present. If people were going to make a certain number of purchases, perhaps seeing the ads let the people to remember their needs and make their purchases then but not increase their total needs for purchases.

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
# create a linear model that captures the treatment effect for product A
# controlling for the total ads seen
productA <- lm(week1 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
               data = yahoo[yahoo$product_b==0,])
summary(productA)
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
##     data = yahoo[yahoo$product_b == 0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.933 -1.978 -1.296  2.190  7.186
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.295811   0.007939  163.22   <2e-16 ***
## treatment_ad_exposures_week1 0.068154   0.006023   11.31   <2e-16 ***
## total_ad_exposures_week1     0.204690   0.004028   50.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 300035 degrees of freedom
## Multiple R-squared:  0.02623,    Adjusted R-squared:  0.02622
## F-statistic:  4041 on 2 and 300035 DF,  p-value: < 2.2e-16
```

```r
# create a linear model that captures the treatment effect for product B
# controlling for the total ads seen
productB <- lm(week1 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
               data = yahoo[yahoo$product_b==1,])
summary(productB)
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ad_exposures_week1 + total_ad_exposures_week1,
##     data = yahoo[yahoo$product_b == 1, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.021 -2.407 -1.882  2.856  8.068
##
## Coefficients:
```

```
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.444893   0.015651   92.32  < 2e-16 ***
## treatment_ad_exposures_week1 0.044232   0.007192    6.15 7.75e-10 ***
## total_ad_exposures_week1     0.218411   0.005313   41.11  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.052 on 199959 degrees of freedom
## Multiple R-squared:  0.01879,    Adjusted R-squared:  0.01878
## F-statistic:  1915 on 2 and 199959 DF,  p-value: < 2.2e-16
```

```
# get the coefficients
A_effect <- coef(summary(productA))[2]
B_effect <- coef(summary(productB))[2]

# let's also try another way (the way of the hint). let's combine everything
# into on giant model testing for the effect of the treatment ads on
# purchases while also adding a variable for the project with an interaction
# variable
product_yahoo <- lm(week1 ~ treatment_ad_exposures_week1 + product_b +
                      treatment_ad_exposures_week1 * product_b +
                      total_ad_exposures_week1, data = yahoo)
summary(product_yahoo)
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ad_exposures_week1 + product_b +
##     treatment_ad_exposures_week1 * product_b + total_ad_exposures_week1,
##     data = yahoo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.032 -2.194 -1.500  2.439  8.071
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              1.289270   0.008010 160.960
## treatment_ad_exposures_week1             0.061251   0.005637  10.867
## product_b                                0.170320   0.013186  12.917
## total_ad_exposures_week1                 0.210868   0.003230  65.277
## treatment_ad_exposures_week1:product_b -0.010100   0.006490  -1.556
##                                         Pr(>|t|)
## (Intercept)                               <2e-16 ***
## treatment_ad_exposures_week1              <2e-16 ***
```

```
## product_b                                   <2e-16 ***
## total_ad_exposures_week1                    <2e-16 ***
## treatment_ad_exposures_week1:product_b        0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.766 on 499995 degrees of freedom
## Multiple R-squared:  0.02848,    Adjusted R-squared:  0.02847
## F-statistic:  3664 on 4 and 499995 DF,  p-value: < 2.2e-16
```

The difference in effects produced by ads for product A and B is **-0.0239217**. The difference being negative means that the ads for product B were actually less effective. Similarily, if we create a large regression that captures the effects of seeing treatment ads on purchases and breaks out the product being advertised, we see a negative interaction effect of product B with the treatment.

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 am not confident at all that celebrity endorsements increase the effectiveness of stimulating immediate purchases. We did not compare ads for product A with celebrity endorsements and product A without celebrity endorsements. We only compared seeing ads for product A with not seeing ads for product A. This is not an apples-to-apples comparison where the only thing we are changing is the celebrity endorsement. In order for me to feel confident, I would need to run two simultaneous ad campaigns both for product A where the only difference between the two campaigns would be a celebrity endorsement.

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.

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.")

We have generated a fake version of this data for your download here. 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. 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?

```
# read in the data
vietnam <- read.csv("ps5_no2.csv")
head(vietnam)
```

```
##   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
```

```r
# perform a simple analysis of years of education on income
naive_vietnam <- lm(income ~ years_education, data = vietnam)
summary(naive_vietnam)
```

```
##
## Call:
## lm(formula = income ~ years_education, data = vietnam)
##
## 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
```

```r
# grab the coefficient
naive_effect <- coef(summary(naive_vietnam))[2]
```

The naive regression from an observational researcher suggests that each additional year of education produces an additional **5750.4796478** dollars of income. This effect is statistically significant.

b. Tell a concrete story, not having to do with the natural experiment, about why the observational regression in part (a) may be biased.

We cannot make a causal claim based on observational data because we don't know how the people ended up in the different groups. The people who choose to get more education might well be different in other ways than the people who choose to get less education. For example, perhaps those who get more years of education also come from wealthier families. Perhaps, the wealth of your family is truly the cause of your future income. A mere observational study, like the naive model above, would assume an effect for years of education when the real effect is the family background.

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 variable defining whether someone had a low or high draft number
vietnam$high <- ifelse(vietnam$draft_number > 80, 0, 1)

# create a linear regression to estimate the effect of having a high-ranked
# draft number on years of education obtained
high_vietnam <- lm(years_education ~ high, data = vietnam)
summary(high_vietnam)
```

```
##
## Call:
## lm(formula = years_education ~ high, data = vietnam)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5601 -1.4343 -0.4343  1.5657  5.5657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.43431    0.01691  853.40   <2e-16 ***
## high         2.12576    0.03790   56.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.117 on 19565 degrees of freedom
## Multiple R-squared:  0.1385, Adjusted R-squared:  0.1384
## F-statistic:  3145 on 1 and 19565 DF,  p-value: < 2.2e-16
```

```
# get the coefficient
high_effect <- coef(summary(high_vietnam))[2]

# in essence this randomization is clustered which means we need to
# compute the standard errors more robustly than the simple linear
# regression gives us

# we borrow the function for computing clustered standard errors from the
# third problem set
```

```r
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
cl <- function(fm, cluster){
  ## This function takes a fit model `fm` and a cluster `df$variable`
  ## and returns the cluster-correct standard errors.
  ##
  ## This is really little more than an application of the sandwich
  ## estimator inside each of the clusters, but it isn't alwasy intuitive
  ## what is happening.
  ##
  ## Adapted from Mahmood Arai & Drew Dimmery
  ##
  ## Note: - This WON'T work with missing data; different vector lenghts
  ##       - I'd strongly recommend that you read your data the first time
  ##         without converting it to a factor.
  ##       - Instead, convert it to a factor after you have read-in the
  ##         data.

  require(sandwich, quietly = TRUE)
    require(lmtest, quietly = TRUE)
    M <- length(unique(cluster))
    N <- length(cluster)
    K <- fm$rank
    dfc <- (M/(M-1))*((N-1)/(N-K))
    uj <- apply(estfun(fm),2, function(x) tapply(x, cluster, sum));
    vcovCL <- dfc*sandwich(fm, meat=crossprod(uj)/N)
    coeftest(fm, vcovCL)
}

# first we make sure to treat each draft number as a factor
vietnam$draft_number = factor(vietnam$draft_number)

# we compute the standard error using the above function
```

```r
cluster_vietnam <- cl(high_vietnam, vietnam$draft_number)

# get the standard error
high_SE <- cluster_vietnam[2,2]
```

When a draft number of 80 or less is considered a high draft number, having a high draft number increases years of education by **2.1257562**, a statistically significant effect. The standard error is **0.0381878**.

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

```r
# create a linear model to estimate the effect of a high draft number on income
high_income <- lm(income ~ high, data = vietnam)
summary(high_income)
```

```
##
## Call:
## lm(formula = income ~ high, data = vietnam)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67399 -21140  -3002  18005 151306
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60761.9      235.9  257.56   <2e-16 ***
## high          6637.6      528.7   12.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29530 on 19565 degrees of freedom
## Multiple R-squared:  0.007992,   Adjusted R-squared:  0.007941
## F-statistic: 157.6 on 1 and 19565 DF,  p-value: < 2.2e-16
```

```r
# get the coefficient
high_income_effect <- coef(summary(high_income))[2]

# use our cluster function
cluster_high_income <- cl(high_income, vietnam$draft_number)

# get the standard error
high_income_SE <- cluster_high_income[2,2]
```

The effect of having a high draft number on income is **6637.554244** with a standard error of **511.899229**.

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

```
# divide the effect of high draft number on income by the effect
# of high draft number on years of education
education_income <- high_income_effect / high_effect
```

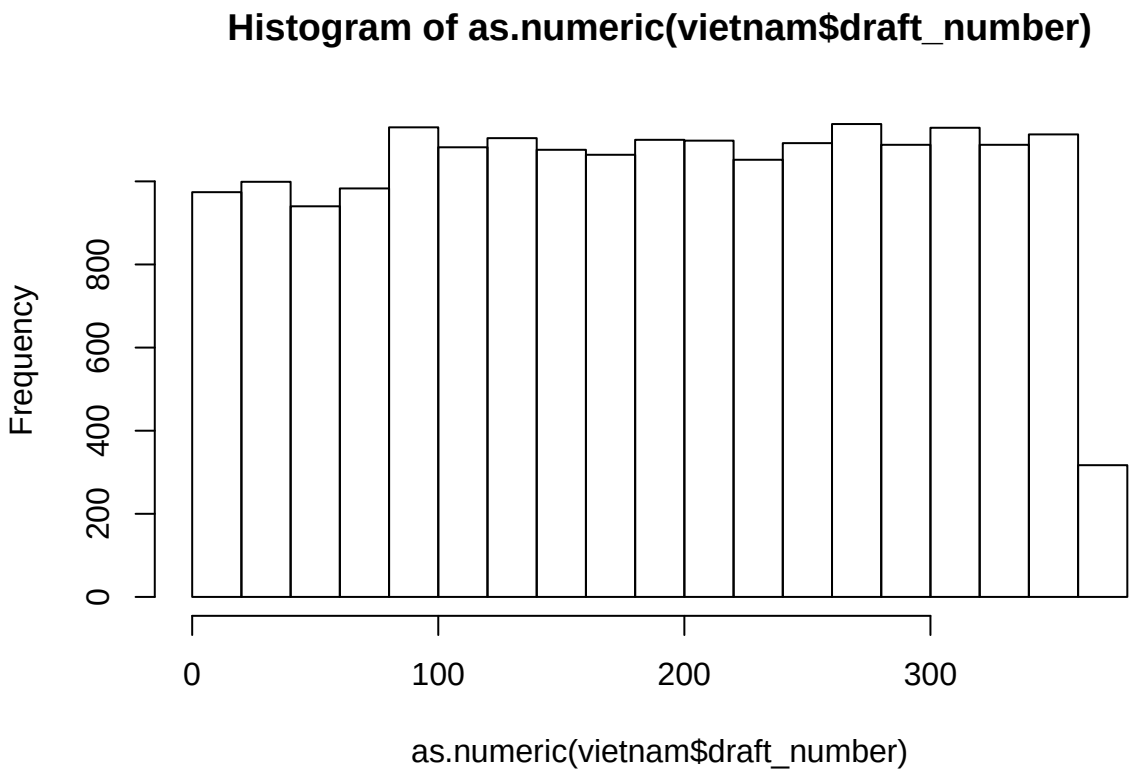
The results suggest that each additional year of education produces an increase of **3122.4437939** on income. Getting more education leads to higher income. Intuitively, we can think of this by looking at the units. When we divide the income per draft number by education per draft number, the draft numbers cancel out and we are left with income per 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.

This exclusion restriction is a pretty major assumption. People with higher ranking draft numbers were more likely to serve in the armed forces. If this was the case, perhaps serving in the armed forces is a resume booster and employers are willing to pay you more for this experience. If this is the case, then we cannot say that it was the increased education that those with higher ranking draft numbers obtained that caused them to have higher incomes.

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.

```
# birthdays, and hence draft numbers, are pretty uniformly distributed
# let's visualize the distribution of people we have records on
hist(as.numeric(vietnam$draft_number))
```




```
# run a chi squared test to test for the uniformity of the distribution
# of draft numbers
uniform <- chisq.test(as.numeric(vietnam$draft_number))
uniform
```

```
##
##  Chi-squared test for given probabilities
##
## data:  as.numeric(vietnam$draft_number)
## X-squared = 1145800, df = 19566, p-value < 2.2e-16
```

When doing a chi squared test to determine if the distribution of draft numbers is uniformly distributed in our sample of observed incomes, we find that the distribution is not uniform ($p<0.05$). In other words, we are observing the income at different rates for different draft numbers.

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

The differential attrition could well be due to death during war. Those with higher ranking draft numbers are more likely to serve in the Vietnam war, and it is reasonable to think that those who are more likely to serve in war are also more likely to die and attrit out of this study.

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

This differential attrition might bias our estimates because perhaps the people who attrited would have had lower incomes if they had not attrited. Perhaps those who are more likely to die in the Vietnam war were also more likely to be worse soldiers and perhaps, people who are worse soldiers would end up earning less income if they didn't end up dying.

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!

Field Experiments 11.10

a. Focusing only on households answered the phone, I estimate the average treatment effect of being assigned to the script that encouraged voting to be **2.2**.
b. The table does *not* provide convincing evidence that the longer someone listens to a script on voting, the more likely they are to vote because people self-selected into different treatment dosages (length of call). So, for example, perhaps people who like voting more are also more likely to be engaged by a script on voting. Because the dosage was self-selected, this part of the study would be a simple observational study.

Field Experiments 12.3

a. The experiment subjects are villagers in the 49 Indonesian villages, who are clustered by village. Each village is assigned to control or treatment. The control villages have the standard village meetings while the treatment villages have a village-wide plebiscite.
b. If in Indonesia, the plebiscite is rare but the meeting is common, I would add a little bit of hesitancy to my interpretation that plebiscites are much better. Perhaps the drastic treatment effect that we observed is not due to the plebiscite but rather to the novelty of a new governmental institution.
c. From the description on page 250 of the article, it is unclear how exactly the survey was conducted. The language of "we" implies that the researchers (or members of the research team) potentially conducted the survey. This is dangerous because it is not double-blind. The researchers could unknowingly be treating the treatment and control groups differently. This could lead to bias in the survey responses.
d. When calculating the average treatment effect, the researchers assumed that there was no spillover. In other words, they assumed that the treatment did not have an effect on anyone outside the treatment group. If the villagers did talk with one another between villages, then there is the potential for spillovers. Perhaps the control group felt jaded for not being assigned to the fancy new institution of the treatment group and reported lower satisfaction. The design that might address this issue is a higher-level regional

clustering. By clustering at a higher level, the researcher could ensure that all nearby villages (where people are more likely to talk with one another) received the same treatment and so had no spill over between them.

e. Olken oversteps his bounds. Yes, his experiment did show that people with a plebiscite had higher levels of satisfaction. However, "political legitimacy" is a tricky thing to measure and is not necessarily the same thing as being involved in government decisions or being satisfied with the political process. One could potentially be unsatisfied with the political process but think the system legitimate. This is true of most studies. The experiment proves something about responses to a survey. To draw broader conclusions from the questions about abstract notions of legitimacy is leap (though a reasonable leap). Olken mentions two major limitations in his conclusion: the small sample size (only 49 villages) and the short-run nature of his experiment (perhaps in the long run, the effects wear off). My other concerns are the novelty effect of the plebiscite (perhaps any different institution would produce positive effects on satisfaction) and the potential for spillover effects between neighboring villages. Future studies can address these concerns with a larger sample-size, a long-term experiment that lasts years, multiple variations of the treatment (e.g. plebiscite, dictators with single year terms, etc.), and higher-level clustering done at the regional level.

f. I speculate as to why the plebiscate might have a long-term negative effect on satisfaction: Perhaps, the plebiscite is actually more corruptible than the village meeting. If initially people were happy with the plebiscite, perhaps after individuals learn to game the system, villagers become even less happy with the plebiscite and become nostalgic for the days of the village meeting.

Field Experiments 12.5

a. This experiment provides convincing evidence on the effect of plate size on what people eat. It fails to show that there is any causual relationship between plate size or amount eaten with weight (though this might be a reasonable assumption). The researchers varied plate size and measured how much subjects ate. They showed that varying plate size caused changes in how much people ate. However, they did not introduce any measurements of weight. They made an assumption.

b. I would like to do a study with participants that are trapped in a room and who can only eat meals that I provide for an entire week. One half of subjects are randomly assigned to bigger plates than the other half. I measure both the food consumption and weight of the subjects. By the end of the week, I will be able to tell if varying the size of the plate actually caused an increase in weight.

4. Other Quesitons

Natural Experiments in Medicine.

Read this synopsis of an interesting study of the effects of different diabetes drugs, sent to us by a student. (I am not expecting a long response for these. Think about communicating the necessary ideas in a few sentences or less per question.)

a. What are the benefits of this study relative to a randomized controlled trial?

The benefit to this study is that it is feasible. It captures a large population and follows it for a long time. A true randomized controlled trial would have difficulty capturing such a large study or following it for as long.

b. What are the disadvantages of this study relative to a randomized controlled trial?

The disadvantage to this study is that individuals were *not* randomly assigned to one of 2 diabetes drugs (SU or TZD). Instead, the researchers assumed that the prescription patterns of the doctors acted as a random intervention. However, this is not a truly random event as patients are not randomly assigned to doctors and doctors are unlikely to assign SU or TZD at random.

c. This is a natural experiment rather than a deliberate research experiment. Therefore, practice telling a story, consistent with the reported data, about how there might be no causal difference at all between the drugs.

The researchers attempted to say that the prescription of SU or TZD was effectively random because the demographic, diagnoses, and provider quality variables were balanced between those prescribed the two drugs. Therefore, the researchers conclude that a causal effect might reasonably be estimated. However, it might be reasonably assumed that there is some hidden variables that affects treatment is not accounted for in this study. For example, perhaps sicker people are more invested in their treatments and so they research these drugs and find older observational studies that show TZD is potentially more dangerous so they all find doctors more likely to prescribe them SU and end up dying in higher rates than the healthier who randomly dispersed between SU and TZD doctors.

d. Describe the placebo test mentioned in the article. Does this test help to rule out the story you just told? Why or why not?

The researchers conduct a placebo test in an attempt to prove that randomization by prescription is real randomization. They look at populations bracketing the population in question, those slightly healthier and those slightly sicker. They find that once again the prescription patterns of physicians do not correlate with any other observable variable like

demographic, diagnoses, and provider quality. This test definitely increases confidence in the idea that perscription patterns might serve as an effective randomizer. However, it does not rule out the story I just told. It could just be that within the specific population in question, the sicker individuals seek out different perscription patterns.

e. What do you think about the prospects for such observational research in medicine? Is this kind of research a complement to, or a substitute for, deliberate field experiments?

I think the prospects of such observational research in medicine are good. While I don't think that these observational studies could substitute for real field experiments because without true randomization, we cannot make real causal claims (something pretty important in the field of medicine), I do think they can serve as a complement to real field experiments where real field experiments are too costly, too time-consuming, or ethically questionable.

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.

To understand why the average treatment effect is a good measure of the "true" causal effect of a treatment, we must consider a hypothetical world. In this hypothetical world, I can tell each person's potential outcomes for both control and treatment because I split the world into parallel universes (one for control and one for treatment) and observe the outcomes in each for each individual. From this ideal universe, I am able to tell the *actual causal effect of the treatment for each and every individual.* From the individual effects, I am able to compute an average treatment effect.

Unfortunately, the real world does not work like this, and we can only *estimate* the average treatment effect. We cannot observe an individual's potential outcomes for both treatment and control assignments because we cannot split time and space into two parallel universes. Instead, we can only observe one potential outcome, either the outcome for control or the outcome for treatment. We could get around this problem if we had two identical people (identical in every single way, even in ways we cannot observe). Because such people are not real, we do the next best thing and randomize. By randomizing, I can make it so that on aggregate I am studying two populations that are identical in both observable and non-observable features. Because people are randomized into control and treatment groups, a large enough sample would be largely the same between treatment and control. I could assume that were the samples switched, they would have identical potential outcomes. From these randomized samples, I could estimate an average treatment effect by comparing the outcomes between the groups. Because the only difference between the two groups was the treatment, this measured effect would be a good estimate of the true causal effect.