Experiments and Causality: Problem Set 5

Alex, Scott & Micah 12/10/2020

> I chose to answer questions 2 and 3 and skipped question 1.

```
library(data.table)

library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 3.6.2
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.6.2

## Warning: package 'zoo' was built under R version 3.6.2
```

```
library(AER)
```

```
## Warning: package 'car' was built under R version 3.6.2

## Warning: package 'carData' was built under R version 3.6.2
```

```
library(ggplot2)
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 3.6.2
```

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.

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

The variables in the dataset are described below:

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

Simplifying assumptions you should make when answering this problem:

- The effect of treatment ad exposures on purchases is linear. That is, the first exposure has the same effect as the second exposure, has the same effect as the third 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. This is a setup we need for these estimators to work, but, it might or might not hold in real life.

First things first, these variable names are frustrating. Rename them.

```
#setnames(
#  d,
#  old = c('total_ad_exposures_week1', 'treatment_ad_exposures_week1'),
#  new = c('total_ads', 'treatment_ads')
#)
```

Questions to Answer

1. Run a crosstab – which in R is `table` – of `total_ads` and `treatment_ads` to sanity check that the distribution of impressions looks as it should. After you write your code, write a few narrative sentences about whether this distribution looks reasonable. Why does it look like this? (No computation required here, just a brief verbal response.)

```
cross_tab <- 'fill this in'
```

2. A colleague of yours proposes to estimate the following model: `d[ , lm(week1 ~ tretment_ads)]` You are suspicious. Run a placebo test with `week0` purchases as the outcome and report the results. Since treatment is applied in week 1, and `week0` is purchases in week 0, *should* there be an relationship? Did the placebo test "succeed" or "fail"? Why do you say so?

```
model_colleague <- 'fill this in'
```

3. Here's the tip off: the placebo test suggests that there is something wrong with our experiment (i.e. the randomization isn't working) 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? (Hint: it should be present in the cross-tab that you wrote in the first part of this qusetion.) How can you improve your analysis to address 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.

4. Implement the procedure you propose from part 3, 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.) How can you tell this this has fixed the problem? Is it possible, even though this test now passes, that there is still some other problem?

```
model_passes_placebo <- 'fill this in'
```

5. Now estimate the causal effect of each ad exposure on purchases during Week 1. You should use the same technique that passed the placebo test in part 4. Describe how, if at all, the treatment estimate that your model produces changes from the estimate that your colleague produced.

```
model_causal <- 'fill this in'
```

6. Upon seeing these results, the colleague who proposed the specification that did not pass the placeo test challenges your results – they make the campaign look less successful! Write a short paragraph (i.e. 4-6 sentences) that argues for why your estimation strategy is better positioned to estimate a causal effect.

7. One concern raised by David Reiley is that advertisements might just shift *when* people purchase something – rather than increasing the total amount they purchase. Given the data that you have available to you, can you propose a method of evaluating this concern? Estimate the model that you propose, and describe your findings.

```
model_overall <- 'fill this in'
```

8. If you look at purchases in each week – one regression estimated for each outcome from week 1 through week 10 (that's 10 regression in a row) – what is the relationship between treatment ads and purchases in each of those weeks. This is now ranging into exploring data with models – how many have we run in this question alone!? – so consider whether a plot might help make whatever relationship exists more clear.

```
# write whatever you want to estiamte this
```

9. What might explain this pattern in your data. Stay curious when you're writing models! But, also be clear that we're fitting a **lot** of models and making up a theory/explanation after the fact.

10. We started by making the assumption that there was a linear relationship between the treatment ads and purchases. What other types of relationships might exist? After you propose at least two additional non-linear relationships, write a model that estimates these, and write a test for whether these non-linear effects you've proposed produce models that fit the data better than the linear model.

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. 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.

```
d <- fread("../data/ps5_no2.csv")
d$highrank <- ifelse(d$draft_number >80, "Low", "High")
d
```

```
##        draft_number years_education    income highrank
##     1:          267              16  44573.90      Low
##     2:          357              13  10611.75      Low
##     3:          351              19 165467.80      Low
##     4:          205              16  71278.40      Low
##     5:           42              19  54445.09     High
##    ---
## 19563:           76              18 105795.30     High
## 19564:          168              14  94328.98      Low
## 19565:           11              14  28049.30     High
## 19566:          113              18  52886.20      Low
## 19567:          317              17  87578.65      Low
```

Questions to Answer

1. 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?

```
model_observational <- lm(income~years_education,data=d)
model_observational
```

```
##
## Call:
## lm(formula = income ~ years_education, data = d)
##
## Coefficients:
##     (Intercept)  years_education
##          -23355             5750
```

> This suggests every year of education results in an added $5,570 to a subjects income.

2. Continue to suppose that we did not run the experiment, but that we saw the result that you noted in part 1. Tell a concrete story about why you don't believe that observational result tells you anything causal.

> Age and experience plays a role in income and whether or not a subject was drafted will greatly change their age and years of education when they enter the workforce.

3. Now, let's get to using the natural experiment. Define "having a high-ranked draft number" as having a draft number between 1-80. For the remaining 285 days of the year, consider them having a "low-ranked" draft number). Create a variable in your dataset indicating whether each person has a high-ranked draft number or not. Using a regression, estimate the effect of having a high-ranked draft number on years of education obtained. Report the estimate and a correctly computed standard error. (*Hint: How is the assignment to having a draft number conducted? Does random assignment happen at the individual level? Or, at some higher level?)

```
model_education <- lm(years_education~as.factor(highrank),data=d)
coeftest(model_education, .vcov = vcovHC(model_education))
```

```
##
## t test of coefficients:
##
##                         Estimate Std. Error t value  Pr(>|t|)
## (Intercept)            16.560062   0.033922 488.182 < 2.2e-16 ***
## as.factor(highrank)Low -2.125756   0.037905 -56.081 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> The individuals in the low group have 2 years less schooling on average.

4. 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 <- lm(income~as.factor(highrank),data=d)
coeftest(model_income, .vcov = vcovHC(model_income))
```

```
##
## t test of coefficients:
##
##                        Estimate Std. Error t value  Pr(>|t|)
## (Intercept)            67399.44     473.15 142.449 < 2.2e-16 ***
## as.factor(highrank)Low -6637.55     528.70 -12.554 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> The individuals in the low group get $6000 less on average which makes sense since they have less schooling as well.

5. Now, estimate the Instrumental Variables regression to estiamte the effect of education on income. To do so, use `AER::ivreg`. After you evaluate your code, write a narrative description about what you learn.

```
model_iv <-ivreg(income ~ years_education|as.factor(highrank), data = d)
model_iv
```

```
##
## Call:
## ivreg(formula = income ~ years_education | as.factor(highrank),     data = d)
##
## Coefficients:
##     (Intercept)  years_education
##           15692             3122
```

> I learned that income depends not only on education but also if you were drafted or not.

6. Just like the other experiments that we've covered in the course, natural experiments rely crucially on satisfying the "exclusion restriction".

In the case of a medical trial, we've said this means that there can't be an effect of just "being at the doctor's office" when the doctor is giving you a treatment. In the case of an instrumental variable's setup, the *instrument* (being drafted) cannot affect the outcome (income) in any other way except through its effect on the "endogenous variable" (here, education).

Give one reason this requirement might not be satisfied in this context. In what ways might having a high draft rank affect individuals' income **other** than nudging them to attend more school?

> Individuals that are drafted get experience in the war such as flying planes which gives them a higher income without needing education.

7. 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 – i.e. earning $0 does not mean that their data wasn't measured.)**

```
low = length(d[d$highrank=='Low']$highrank)
high = length(d[d$highrank=='High']$highrank)
```

```
model_differential_attrition <- prop.test(c(high,low),c(low+high,low+high),c(80/365,1-(80/365)))
model_differential_attrition
```

```
##
##  2-sample test for given proportions with continuity correction
##
## data:  c(high, low) out of c(low + high, low + high), null probabilities c(80/365, 1 - (80/365))
## X-squared = 91.85, df = 2, p-value < 2.2e-16
## alternative hypothesis: two.sided
## null values:
##    prop 1    prop 2
## 0.2191781 0.7808219
## sample estimates:
##    prop 1    prop 2
## 0.1991107 0.8008893
```

> This p value means we can reject the null that the high ranked draft number has no effect on whether we observe a person income.

8. Tell a concrete story about what could be leading to the result in part 7. How might this differential attrition create bias in the estimates of a causal effect?

> High ranked individuals get drafted and can possibly die in the war. Out of the high ranked individuals we are more often seeing the case in which they choose to go to school to avoid the draft and have a high income. We miss the case where they get drafted and pass away and cannot report their income.

3. Optional: Think about Treatment Effects

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

> We set up our experiments to find an average treatment effect so we can apply the knowledge gained to the general public. This is why randomization is crucial in the process. The average treatment effect tells you on average how a treatment will effect a subject. A person always has multiple potential outcomes and the average treatment effect attempts to tell you the difference between these for any given subject.