Experiments and Causality: Problem Set 5

Alex, Scott & Micah

12/10/2020

```
library(data.table)

library(sandwich)
library(lmtest)

library(AER)

library(ggplot2)
library(patchwork)
```

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.
- Suppose that this data is the result of the following SQL query (this information is informative for the differential attrition question):

```
SELECT
  ssearning AS earnings
    years_of_schooling AS years_education
    ein AS id
FROM irs_income_1980
JOIN draft_status
ON ssearning.id = draft_status.id
DROP id
```

- 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('./draft_data.csv')

head(d)
```

```
##    draft_number years_education    income
## 1:          267              16  44573.90
## 2:          357              13  10611.75
## 3:          351              19 165467.80
## 4:          205              16  71278.40
## 5:           42              19  54445.09
## 6:          240              11  32059.12
```

Questions to Answer

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 <- 'fill this in'
```

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.

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 called `high_draft` that indicates 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 <- 'fill this in'
```

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 <- 'fill this in'
```

5. Now, estimate the Instrumental Variables regression to estimate 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 <- 'fill this in'
```

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?

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.)**

```
model_differential_attrition <- 'fill this in'
```

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?

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.