Problem Set 5

Code ▾

Hide

```
library(data.table)
library(sandwich)
library(lmtest)
library(AER)
library(ggplot2)
library(patchwork)
```

Vietnam Draft Lottery

A famous paper (http://sites.duke.edu/niou/files/2011/06/Angrist_lifetime-earningsmall.pdf) 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.
- Suppose that this data is the result of the following SQL query (this information is informative for the differential attrition question):

Hide

```
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
```

```
/bin/sh: 1: SQL: not found
```

- 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.

Hide

```
d <- fread('./draft_data.csv')
head(d)
```

draft_number	years_education	income
<int>	<int>	<dbl>
267	16	44573.90
357	13	10611.75
351	19	165467.80
205	16	71278.40
42	19	54445.09
240	11	32059.12

6 rows

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?

Hide

```
model_observational <- lm(d$income ~ d$years_education)
summary(model_observational)
```

```

Call:
lm(formula = d$income ~ d$years_education)

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 ***
d$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
```

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.

> I don’t believe that this observational result tells us anything causal because there are many factors that affect income besides years of education and wouldn’t be controlled for by randomization. For example, it costs money to go to college and it could be that young people from households that are wealthy enough to send them to college are more likely to be higher earners regardless of whether or not they go. It could be that people who attend college are more hard working in general and would have seen higher earnings regardless of whether they went or not.

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?)

Hide

```

d$high_rank = d$draft_number<81

model_education <- lm(d$years_education ~ d$high_rank)
model_education.vcovHC <- vcovHC(model_education)
me_robust_ci <- coefci(model_education, vcov.=model_education.vcovHC, level=0.95)

#summary(model_education)
summary(me_robust_ci)
```

```
     2.5 %            97.5 %      
 Min.   : 2.052   Min.   : 2.199  
 1st Qu.: 5.139   1st Qu.: 5.266  
 Median : 8.227   Median : 8.333  
 Mean   : 8.227   Mean   : 8.333  
 3rd Qu.:11.314   3rd Qu.:11.401  
 Max.   :14.401   Max.   :14.468  
```

> Individuals with high ranked draft numbers have on average 2.12 more years of education than those with low draft numbers.

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

Hide

```
model_income <- lm(d$income ~ d$high_rank)
model_income.vcovHC <- vcovHC(model_income)
mi_robust_ci <- coefci(model_income, vcov.=model_income.vcovHC, level=0.95)

#summary(model_income)
summary(mi_robust_ci)
```

```
     2.5 %           97.5 %
 Min.   : 5568   Min.   : 7707
 1st Qu.:19252   1st Qu.:21085
 Median :32936   Median :34463
 Mean   :32936   Mean   :34463
 3rd Qu.:46620   3rd Qu.:47841
 Max.   :60304   Max.   :61219
```

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.

Hide

```
model_iv <- ivreg(d$income ~ d$years_education | d$high_rank)
summary(model_iv)
```

```


Call:
ivreg(formula = d$income ~ d$years_education | d$high_rank)

Residuals:
   Min     1Q Median     3Q    Max 
-78140 -18762  -2145  16461 147217 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        15691.6     3416.4   4.593  4.4e-06 ***
d$years_education   3122.4      229.6  13.601  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27260 on 19565 degrees of freedom
Multiple R-Squared: 0.1548,	Adjusted R-squared: 0.1548 
Wald test:   185 on 1 and 19565 DF,  p-value: < 2.2e-16 
```

> The model suggest that each each additional year of education corresponds with an additional $3122 in individual earnings with a high level of significance.

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?

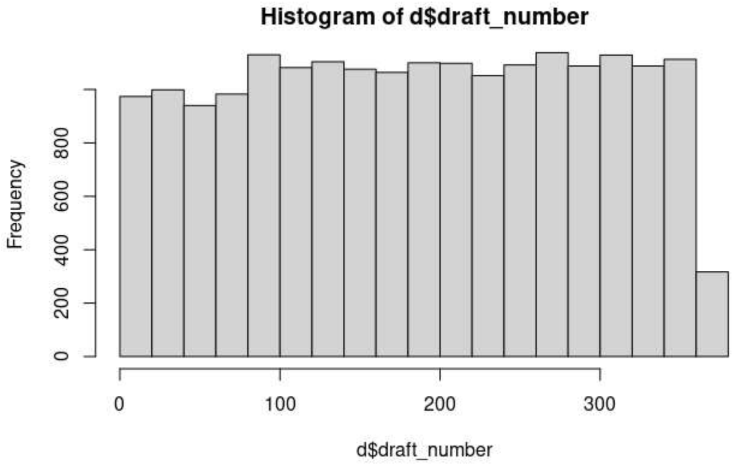
> If an individual has a high draft rank and ends up being drafted, their experiences in the military might affect their earnings. They may come back with some sort of physical disability or PTSD that reduces their earning potential. They might come back with a greater degree of personal discipline or find opportunities with companies that prioritize recruiting veterans, which could increase earning potential.

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.)**

> A brief inspection of the data in a histogram seems to indicate that earlier draft numbers are under-represented in the dataset. This can be confirmed with a t-test comparing the counts of the high and low ranked draft numbers.

Hide

```
hist(d$draft_number)
```



Hide

```
d[,.(count_group = .N), by=.(draft_number, high_rank)][,t.test(count_group~high_rank)]
```

```

    Welch Two Sample t-test

data:  count_group by high_rank
t = 6.6358, df = 123.31, p-value = 9.121e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.410934 8.160996
sample estimates:
mean in group FALSE  mean in group TRUE 
           54.98596            48.70000 
```

> The t-test confirms a significant difference in means, indicating that there is likely differential attrition.

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?

> Assuming that birthdays are scattered uniformly across the year, the differential attrition of early draft picks may indicate that individuals who are drafted end up dying in combat and are thus unable to participate in the labor market upon their return from military service. This could create bias if those who end up being drafted and dying in combat are for some reason more or less likely to have higher earnings than their surviving counterparts. This would bias the ATE toward or away from zero, respectively.

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.

> In the case of a binary RCT consisting of a treatment group and a control group, the ATE in the measure of the average difference in outcomes between these two groups. Prior to assignment, every individual in the study has two potential outcomes - one if they're assigned to treatment, and one if they're assigned to control. Given the impossibility of assigning individuals to both treatment and control simultaneously, we have no way of knowing for sure what the true treatment effect for an individual is because we will never know both of that individual's potential outcomes. The ATE is the next best thing, as it compares the outcomes of two groups that are theoretically identical except for their exposure to the treatment and lets us estimate what the *typical* treatment effect would be for whatever population the experiment is representative of. In the social sciences, we often care about how a particular treatment or policy would impact large groups of people (cities, hospital patients, app users) rather than how that treatment will impact a single individual. Because the ATE is a measure that accurately describes the effect on a population and satisfies the business case for most social experiments, it is a *good* measure.