W241 PS5 Ray Buhr

Ray Buhr

August 20, 2015

1. Vietnam Draft Lottery

```
draft_data <- read.table("C:/Users/Ray/Documents/MIDS/241/ps5_no2.csv", se
p = ",", header = T)

cl <- function(fm, cluster){
    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)
}
```

- 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?

```
ate <- lm(income ~ years_education, data = draft_data)
summary(ate)
```

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

The naive regression suggests that each additional year of education results in $5,750 more income per year.

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

The type of person who chooses to gain extra education may do so due to more gifted intelligence and thus be more likely to earn higher wages in comparison than those who choose not to obtain further education. Due to this self-selecting bias, the results of the observation may not hold true if people are forced through additional schooling.

- c.
  - 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.

```
draft_data$high_rank <- ifelse(draft_data$draft_number < 81, 1, 0)

ate2 <- lm(years_education ~ high_rank, data = draft_data)
cl(ate2, draft_data$draft_number)
```

```
## Warning: package 'sandwich' was built under R version 3.2.2
```

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

```
##
## 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
```

The effect of having a high-ranked draft number is an additional 2.12576 years of education. The standard error after clustering on draft number was 0.038188.

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

```
ate3 <- lm(income ~ high_rank, data = draft_data)
cl(ate3, draft_data$draft_number)
```

```
##
## 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
```

The effect of having a high-ranked draft number is an additional $6,637.55 in income per year. The standard error after clustering on draft number was 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?

```
ate4 <- ate3$coefficients[2] / ate2$coefficients[2]
print(ate4)
```

```
## high_rank
##  3122.444
```

The “clean” variation in both education and income that is due to high-rank draft numbers is $3,122.44 extra income per year of extra education obtained.

- 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.

Beyond having additional education, the high draft rank might have benefitted people by giving extra skills to those drafted from military training as well as undrafted who attended college. Some of the low draft rank individuals may not have obtained new skills and thus not improved their earning ability, thus bringing down the average for that cohort.

- 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.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
draft_attrition <- draft_data %>% group_by(draft_number) %>% summarise(n =
n(), high_rank = mean(high_rank))
ate5 <- lm(n ~ high_rank, data = draft_attrition)
cl(ate5, draft_attrition$draft_number)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 54.98596    0.43149 127.4326 < 2.2e-16 ***
## high_rank   -6.28596    0.94482  -6.6531 1.059e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using regression, we see that the high-rank draft number does indeed have a statistically significant effect on the observation of an individual's income for the study. The effect size was 6.28596 fewer observations for high-rank draft numbers with a standard error of 0.94482.

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

The most obvious reason why there are fewer observations available for individuals drafted are due to Vietnam War casualties of those drafted and deployed.

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

The drafted and deployed individuals who perished were unable to continue living and earning an income, so the study may be biased to survivors and they may have qualities about them that lead to higher average incomes.

2. RD question

- a.
  - Summarize the study and its conclusion. Which study did you pick? What is the outcome? What is the "treatment"? What is the discontinuity? What is the conclusion?

I picked the Manacorda, Miguel, & Vigorito (2011) study. The outcome was support for a particular political party. The treatment was a cash transfer from the government to households. The discontinuity was the program had an income eligibility cutoff. The study concluded that the discontinuity resulted in households receiving the benefit were 11 to 13 percentage points higher in favoring the pollitical party, even after the program concluded.

- b.
  - Assume the RD was not available, and someone did a simple observational study comparing individuals who happen to get treatment versus those who do not for non-random reasons. What's an alternative story for a simple association between the outcome and this non-random version of the treatment like this?

Under the observational study, we would likely find a similar association because low-income households tend to favor political parties that support social welfare since it benefits themselves directly.

- c.
  - What makes this RD evidence so convincing, relative to an observational study like in part (b)?

Unlike the observational study, we would not know whether the program itself was the reason for low-income households to favor the political party or whether it was due to other social factors like aligning with neighbors or belief in support from previous administrations. The regression discontinuity evidence makes it clear that program recipients firmly side with the political party giving them cash versus recipients who were not eligible due to being over the income limit.

- d.
  - What's an alternative story for why this RD pattern might exist even if the causal effect did not exist?

The same relationsip could be due to other social factors like aligning with neighbors or belief in support from previous administrations. The dicontinuity may have encouraged some households over the eligibility limit to stop working in order to receive the cash benefit, which if they ended up receiving the program benefit would likely be glad to have been handed free money.

3. Feedback on Class Content

I really enjoyed this class!

- I think it would have been helpful to go over the math of the cluster function a bit more since it was repeatedly used throughout the assignments, but I still only have a loose understanding of what exactly it is doing and why it is significantly different and more accurate than simply grouping by the clusters.
- Additionally, a quick primer on the notation used in Gerber and Green would be helpful prior to reading the book. Nothing fancy needed, but not everyone is used to having mathematical notation for proofs in social science backgrounds even if they have taken statistics.
- It may be useful to get help grading assignments more quickly or at least releasing answer keys or maybe even forcefully going through the answers in live sessions. It is hard to know if you are really comprehending the material when you don't know how you did on the last problem set and no one really wants to talk about it in class because they are afraid to show they were wrong.
- I think it would be super useful to do cold-calling or forced participation. There were so many awkward silences because people didn't want to speak up and be incorrect or admit they haven't done the homework. I had several breakout discussions in which no one had read and no one wanted to read or discuss the questions. That kind of ruins the whole point of paying for an online degree designed to be more like a regular classroom. One way to prevent this is to make each group speak to their breakout room discussions and choosing the person to talk at random instead of by volunteer.
- The very beginning of the asynch content, comparing apples to apples, has really stuck with me and something I make sure everyone working with me understands when we do research now. The idea is actually simple to understand, but so important and fundamental to good interpretation of results. It would be great if this concept could be introduced in other courses beyond Field Experiments, such as RDADA, because it applies to all analysis not just experiments.
- The section on blocking versus clustering as techniques to correctly segment the data was a bit confusing. I think David did a great job explaining in the lecture, but I didn't have it quite down when I went to implement myself at first. Rewatching the videos helped, but having to explain the difference to someone else helped me more. This might be useful as an in class exercise: take an experiment design and have people breakout and explain what the proper variable to block would be and what would be more appropriate to cluster.