Problem Set #5 - DATASCI W241

Greg Ceccarelli

December 19, 2015

Vietnam draft lottery

```
##clear global env
rm( list = ls() )

library(dplyr)
library(stargazer)
library(memisc)
library(data.table)

set.seed(123)

## set working directory
setwd(dir = "/Users/GDC/MIDS/DATASCI_W241/Assignments/Problem Set 5/")

## read in data
d <- read.csv("ps5_no2.csv")

#initialize clustered standard errors function
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?

```
##clear global env
mod <- lm(income ~ years_education, data = d)
summary(mod)
```

```
##
## Call:
## lm(formula = income ~ years_education, data = d)
##
## 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 regression suggests that each additional year of education results in an increase in income by ~$5750.

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 most believeable story, draft number aside, is that of omitted variable bias. From the previous model, the Rsquared value is ~20%... meaning that years_education is only explaining 1/5 of the variance in income and that there are other factors (currently not in the model) that are likely important. These may things like age, geography, sex, etc.

c. 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.

```
##create dummy
d <- d %>%
  dplyr::mutate(hrd = ifelse(draft_number<=80, 1, 0)
  )

#mod <- summary(lm(years_education ~ hrd, data = d))
#summary(mod)

mod.edu <- cl(lm(years_education ~ hrd,d), d$draft_number)

cat(" ATE = ", mod.edu[2,1], "conf interval = ", c(mod.edu[2,1] - 1.96 * mod.edu[2,2], "-", mod.edu[2,1
```

```
##  ATE =  2.125756 conf interval =  2.0509080248856 - 2.20060436862439
```

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

```
mod.inc <- cl(lm(income ~ hrd,d), d$draft_number)

cat(" ATE = ", mod.inc[2,1]," std. error =", mod.inc[2,2],  "conf interval = ", c(mod.inc[2,1] - 1.96 *
```

```
##  ATE =  6637.554  std. error = 511.8992 conf interval =  5634.23175523435 - 7640.87673274391
```

e. Divide the estimate from part (d) by the estimate in part (c) to estimate the effect of education on income.

```
mod.inc[2,1]/mod.edu[2,1]
```

```
## [1] 3122.444
```

The results suggest that every year of education results in a $3122 gain in income for those with high ranked draft numbers.

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.

High draft rank, for those who aren't actually excluded from the draft because they attend school, have a higher probability to actually serve in the Army which teaches a certain set of skills that, later in life, could result in higher income.

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.

```
d_test <- d %>%
  dplyr::group_by(draft_number) %>%
  dplyr::summarize(cnt = length(income))

cl(lm(cnt ~ I(ifelse(draft_number<=80, 1, 0)), d_test), d_test$draft_number)
```

```
##
## t test of coefficients:
##
##                                    Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)                        54.98596    0.43149 127.4326 < 2.2e-16
## I(ifelse(draft_number <= 80, 1, 0)) -6.28596    0.94482  -6.6531 1.059e-10
##
## (Intercept)                        ***
## I(ifelse(draft_number <= 80, 1, 0)) ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

In the above example, being drafted increases the probability of dieng since War is particularly violent. Thus, those with a high draft number could have been killed in action.

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

Differential attrition is defined as the differential loss of participants from the various comparison groups. This is a problem because the groups are different because of the people dropping (or in our case, Dieing) out rather than just the treatment. In other words, the differences due to differential attrition and the differences due to the treatments are confounded.

Regression Discontinuity

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 Mbiti & Lucas (2013) who test the impact of secondary school quality on student achievement in Kenya, using the cut-off on the primary exit exam required to get into better secondary schools. The outcome in this study is student graduation rates and achievement on the secondary school exit examination. The treatment is attending an elite secondary school in Kenya. The discontinuity, in this case, is Kenya's national standardized primary school exit exam which creates discontinuities in the probability of national school admission. In the paper, the authors ultimately conclude that there is little evidence of positive impacts on learning outcomes for students who attended the elite schools. This suggests that reputations reflect the selection of students rather than the schools ability to generate value-added test-score gains.

B. A throwback to Week 2. 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? Try to be as concrete as possible in your storytelling.

The main problem with estimating the causal effect of such an intervention is the endogeneity of assignment to treatment (e.g. school admission). Since high-performing students are more likely to be "awarded" with school admission and continue performing well at the same time, comparing the outcomes of awardees and non-recipients would lead to an upward bias of the estimates. Even if the school they were admitted into did not improve grades at all, awardees would have performed better than non-recipients, simply because school selectioin happened to students who were performing well ex ante.

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

Despite the absence of an experimental design, the fun thing about regression discontinuity is that it can exploit exogenous characteristics of the intervention to elict causal effects. If all students who received entrace into an elite secondary school above the cutoff are compared to those just below (presuming they're all very similar), you can compare the outcome of the "awardees" to the counterfactual of the control group (those below the cutoff) to estimate a local treatment effect.

D. What's an alternative story for why this RD pattern might exist even if the causal effect did not exist? This story might be hard to tell because these are good studies, but do your best.

The simplest alternative story is that the pattern might actually be caused by the school's ability to generate achievement rather than the "sorting" of students from primary school into the elite secondary schools. For example, perhaps the primary school education is extremely poor (likely not the case) and the test is a mechanism that essentially selects on IQ (or some other gauge of aptitude) and not for work effort or prior school related achievement. In this hypothetical, it could actually be the case, for example, that the elite secondary schools really do prepare students well in terms of studies + work ethic and this is ultimately reflected by the discontinuity shown by the authors.

Feedback

I loved this class. I know that's not extremely helpful feedback, but I think the level of work is just about right. I would recommend making the IRB certification material optional since not much emphasis is placed on it in the course itself and it's extremely lengthy. The problems sets were all very well constructed and I think they helped to really reinforce the concepts we learned in class.