14.771 Development Economics: Microeconomic Issues and Policy Models
Fall 2008

14:771: Recitation Handout #5
Regression Discontinuity, Attrition/Bounds, and Education

Regression Discontinuity

Regression discontinuity is a different strategy to solve the identification problem. It supposes the existence of a discontinuous shift in the treatment variable set over a continuous set of parameters. Examples include:

- Grameen bank eligibility rule: eligible if households owns less then 0.5 hectares.
- Financial aid at NYU for college studies: step function of an index (grades in high-school, SAT scores, income of parents ...).
- Maimonides rule for class size in Israel: extra teacher added as soon as the number of pupils in class reaches multiple of 40 students.
- A firm becomes unionized if all the workers vote at more than 50% for unionization

If the rules are followed approximately, the idea is that someone just to the right of the discontinuity is very similar to one just to the left but only the one to the left will get treated. Then, a comparison of the outcome for the person to the left compared to that to the right would provide a good estimate of the treatment effect.

Formal derivation

Consider the model:

$$Y_i = \alpha_i + \beta_i X_i$$

where X_i is a dummy variable for treatment status and Y_i is an outcome of interest. There is also a variable Z, which may be related to the parameters above.

Assumptions:

$(A1)$: For a known point Z_0, the limits $X^+ = \lim_{Z \to Z_0^+} E(X_i|Z_i = Z)$ and $X^- = \lim_{Z \to Z_0^-} E(X_i|Z_i = Z)$ exists and differ

$(A2)$: $E(\alpha_i|Z_i = Z)$ is continuous at $Z=Z_0$

$(A3)$: $E(\beta_i|Z_i = Z)$ is continuous at $Z=Z_0$

$(A4)$: For Z_i in the neighborhood of Z_0, X_i is independent of β_i conditional on Z_i

In words, A1 means that treatment actually is discontinuous on each side of the cut-off. This may be obvious but is not always the case. Morduch (1999) shows that despite the Grameen official borrowing rule, people with more than 0.5 hectares of land are as likely than other people to get credit. To check A1, one should use non-parametric regression (more below on this).

A2 and A3 imply that the individuals on each side of the cut-off are not radically different. A4 means that people don't self select into treatment based on their anticipated benefit.

Assumption 4 implies that for ε sufficiently small:

$$E\left(X_i\beta_i|Z_i = Z_0 \pm \varepsilon\right) = E\left(X_i|Z_i = Z_0 \pm \varepsilon\right) * E\left(\beta_i|Z_i = Z_0 \pm \varepsilon\right)$$

Define:

$$Y^+ = \lim_{Z \to Z_0^+} E\left(Y_i|Z_i = Z\right) \text{ and } Y^- = \lim_{Z \to Z_0^-} E\left(Y_i|Z_i = Z\right)$$

Then,

$$E\left(\beta_i|Z_i = Z_0\right) = \frac{Y^+ - Y^-}{X^+ - X^-}$$

If you notice this, it looks much like another estimator we have seen before... What is it?

Proof:

$$
\begin{aligned}
E\left(\beta_i|Z_i=Z_0\right) &= \frac{Y^+-Y^-}{X^+-X^-} \\
&= \frac{\lim\limits_{Z\to Z_0^+} E\left(Y_i|Z_i=Z\right)-\lim\limits_{Z\to Z_0^-} E\left(Y_i|Z_i=Z\right)}{\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)\ -\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)} \\
&= \frac{\lim\limits_{Z\to Z_0^+} E\left(\alpha_i+X_i\beta_i|Z_i=Z\right)-\lim\limits_{Z\to Z_0^-} E\left(\alpha_i+X_i\beta_i|Z_i=Z\right)}{\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)\ -\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)} \\
&= \frac{\begin{array}{l}\lim\limits_{Z\to Z_0^+} E\left(\alpha_i|Z_i=Z\right)-\lim\limits_{Z\to Z_0^-} E\left(\alpha_i|Z_i=Z\right)+ \\ \lim\limits_{Z\to Z_0^+} E\left(X_i\beta_i|Z_i=Z\right)-\lim\limits_{Z\to Z_0^-} E\left(X_i\beta_i|Z_i=Z\right)\end{array}}{\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)\ -\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)} \\
&= \frac{\begin{array}{l}0+\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)*E\left(\beta_i|Z_i=Z\right) \\ \quad -\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)*E\left(\beta_i|Z_i=Z\right)\end{array}}{\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)\ -\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)}\text{, by A2 and A4} \\
&= \frac{\begin{array}{l}\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)\lim\limits_{Z\to Z_0^+} E\left(\beta_i|Z_i=Z\right) \\ \quad -\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)\lim\limits_{Z\to Z_0^+} E\left(\beta_i|Z_i=Z\right)\end{array}}{\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)\ -\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)} \\
&= \frac{E\left(\beta_i|Z_i=Z_0\right)\left(\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)-\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)\right)}{\lim\limits_{Z\to Z_0^+} E\left(X_i|Z_i=Z\right)\ -\lim\limits_{Z\to Z_0^-} E\left(X_i|Z_i=Z\right)}\text{, by A3} \\
&\quad QED
\end{aligned}
$$

A1 ensures that the denominator is not 0.

Given this result, what type of treatment effect is RD estimating?

A Practical Guide to RD

How do we actually compute a regression discontinuity estimate?
The usual way is to estimate a semi-parametric equation of the form:

$$Y_i = f(Z_i) + \delta 1\left(Z_i \geq Z_0\right) + u_i$$

This can be done in multiple ways. First, one could simply use a series regression. This means that the function $f(Z_i)$ is approximated by a polynomial. In theory, one is promising that as the sample size increases, the order of the polynomial will also increase.

Another method is to use a kernel regression. However, if we simply use a kernel regression for the entire domain of Z, we will "smooth out" the discontinuity. So, what is better to do is to use a one-sided kernels on each side of the discontinuity. In this context, we have:

$$\begin{aligned}
\hat{Y}^{+} &= \frac{\sum Y_i * 1\left(Z_i \geq Z_0\right) * k\left(\frac{Z_i - Z_0}{h}\right)}{\sum 1\left(Z_i \geq Z_0\right) * k\left(\frac{Z_i - Z_0}{h}\right)} \text{ and } \hat{Y}^{-} = \frac{\sum Y_i * 1\left(Z_i < Z_0\right) * k\left(\frac{Z_i - Z_0}{h}\right)}{\sum 1\left(Z_i < Z_0\right) * k\left(\frac{Z_i - Z_0}{h}\right)} \\
\hat{X}^{+} &= \frac{\sum X_i * 1\left(Z_i \geq Z_0\right) * k\left(\frac{Z_i - Z_0}{h}\right)}{\sum 1\left(Z_i \geq Z_0\right) * k\left(\frac{Z_i - Z_0}{h}\right)} \text{ and } \hat{X}^{-} = \frac{\sum X_i * 1\left(Z_i < Z_0\right) * k\left(\frac{Z_i - Z_0}{h}\right)}{\sum 1\left(Z_i < Z_0\right) * k\left(\frac{Z_i - Z_0}{h}\right)}
\end{aligned}$$

where k(.) is the kernel function, it integrates to 1 and is non-negative for all values of its domain. h is called the bandwidth. The larger the bandwith, the more observations are used to compute the estimate at each point. This gives more precision but will be "too" smooth. The problem with this method is that around the discontinuity (the point where we care the most about the estimate), we have fewer and fewer points available for our estimate. This is why this method is not so popular.

We can also use the Fan regression (local polynomal regression) on each side. This suffers in part from the same problem as the one above, but is much better.

So, how to implement this in STATA? It's actually quite easy. Suppose we have a cutoff at $X = c$. Then let $D_i = 1\left(X_i > c\right)$. Then run the regression

$$y = f\left(X_i\right) + D_i$$

where $f\left(X_i\right)$ is either a series of polynomials (you can allow for different movements on either side of the discontinuity by also including a set of polynomials interacted with D_i) or nonparametrically modeled. The first method can be done in a single OLS regression - just regress y on polynomials of X, D_i, and possibly interactions of the X polynomials and D_i. The impact is just the coefficient on D_i. For fuzzy RD, we do the same, but now D_i is the exluded instrument for actual treatment status, and the polynomials are included exogenous variables.

Finally, remember that it is always, always, always important to check that there is no evidence of sorting at the disconinuity, as evidenced by jumps in covariates that don't determine the cutoff (this is the Urquiola and Verhoogen check).

Dealing with attrition

As we have seen in class, attrition, in particular for randomized experiments, can be a significant problem. When does attrition leads to bias? Is it sufficient that the attrition is balanced between treatment and control?

Heckman selection model

The traditional strategy for dealing with attrition was to use the same tools employed in selection problems. Heckman's selection model is probably the best known work in this regard. Heckman suggests that it is possible to model the selection and assumes normal errors to derive a *sample correction* term which should compensate for the effect of attrition. While it is possible to identify this parameter without any additional variable in the selection equation due to non-linearities, it is now recommended that one has at least one regressors that only affect selection and not outcome. In the case of attrition, we need a variable that modifies the probability that an individual stays in the sample but does not modify the effect of the program. Those are not easy to find...
A newer method consists in bounding the effects of the program using assumptions about those who have left the sample. We will look at two different methods.

Manski-Horowitz (2000) bounds

This strategy is fairly simple. It consists in assuming that all the ones who have left the sample have "extreme" outcomes. Denote the best possible outcome of the sample as $\bar{Y}$ and the worst possible outcome as $\underline{Y}$. Denote T as the treatment group. Denote S as an indicator for whether you are still in the sample when the data is collected (non-attrition). Then, one can construct the upper and lower bound of the treatment effect as

$$\begin{aligned}
\bar{\theta}_M &= P(S=1|T=1) * E(Y|T=1) + (1 - P(S=1|T=1))\bar{Y} - \\
&\quad [P(S=1|T=0) * E(Y|T=0) + (1 - P(S=1|T=0))\underline{Y}] \\
\underline{\theta}_M &= P(S=1|T=1) * E(Y|T=1) + (1 - P(S=1|T=1))\underline{Y} - \\
&\quad \left[P(S=1|T=0) * E(Y|T=0) + (1 - P(S=1|T=0))\bar{Y}\right]
\end{aligned}$$

Thus, the upper bound is given by assuming that all attriters in the treatment group had the highest outcome and the attriters in the control group had the worst. The lower bound assumes the opposite.

Lee (2005) bounds

As you can imagine the bounds described above can be quite large. Lee (2005) proposes, under two simple assumptions, a process that leads to tighter bounds. The two assumptions are that treatment is randomly assigned and that it affects attrition in only one direction. Thus, either being assigned to the treatment makes you less likely to leave or more likely to leave but it cannot have the two impacts on different individuals. Can someone think of what this assumption relates to?
Then, one can find an estimate of the average treatment effect for the "never attriters". The bounds are given by

$$\begin{aligned}
\underline{\theta}_L &= E\left(Y|T=1, S=1, Y \leq \gamma_{(1-p_0)}\right) - E(Y|T=0, S=1) \\
\bar{\theta}_L &= E\left(Y|T=1, S=1, Y \geq \gamma_{(1-p_0)}\right) - E(Y|T=0, S=1) \ \textit{where} \\
\gamma_{(1-p_0)} &= G^{-1}_{S=1,T=1}(p_0) \\
p_0 &= \frac{P(S=1|T=1) - P(S=1|T=0)}{P(S=1|T=0)}
\end{aligned}$$

To illustrate, suppose that 50 percent of the treatment group has not attrited, but that only 40 percent of the control group remain. We trim observations from the group that is more frequently observed. Thus, in this case, we trim observations from the treatment group. The trimming fraction is given by p0 $=\frac{0.5-0.4}{0.5}$ =0.2. The procedure to compute the upper bound for the treatment effect amounts to the following:
1. Compute the mean outcome for the control group
2. Drop the lowest 20 percent of outcomes from the treatment group and calculate the mean for the remaining members of the treatment group
3. Calculate the difference between the trimmed treatment group mean and the control group mean.
This is the estimate of the lower bound. The upper bound is computed in a similar way but one trims observations in the treatment group where the values of the outcome are above the 80th percentile for the treatment group.
Lee (2005) shows that it is possible to tighten the bounds by using a covariate that predicts attrition. Let us imagine this variable, denote it by Z, is binary. One computes the bounds for all observations with Z=1 and separately for all those with Z=0. Then, the overall bounds are given by

$$\begin{aligned}\bar{\theta}_L &= P(Z=1|T=1) * \bar{\theta}_L^{Z=1} + (1 - P(Z=1|T=1)) * \bar{\theta}_L^{Z=0} \\ \bar{\theta}_L &= P(Z=1|T=1) * \underline{\theta}_L^{Z=1} + (1 - P(Z=1|T=1)) * \underline{\theta}_L^{Z=0}\end{aligned}$$

It is possible to show that these bounds are tighter than the simple ones. Lee also provides analytical standard errors for these bounds and thus we can also compute fairly easily the confidence intervals around those bounds.

Education financing through vouchers

As we have seen in class, the quality of public school education in developing is very low. Low teacher attendance and poor performance of teachers once in the classroom appears to be very problematic. What are other solutions that governments could take to increase human capital accumulation?

- Change opportunity costs: ban child labor, mandatory schooling, school meals, etc
- Increase returns by improving school quality: teacher incentives, class size, textbooks, (flipcharts?)
- Improve income levels: unconditional transfers
- Do nothing: faster growth will lead to higher enrollments as returns to education increase

A solution that has been proposed and implemented in many countries (including some states in the US) is to use vouchers. Vouchers are simply a system where children can select which school to attend (rather than, for example, being forced to attend the local public school). Why may we think this would be better?

- Encourage competition across schools: increase quality
- Allow individuals from poorer environment to attend better schools
- May encourage learning (particularly if the voucher is conditional)

This policy has raised a lot of interest among economists and I will here present one paper which is looking at this situation in the context of Columbia, and another one from Chile.

Angrist, Bettinger, Bloom, King and Kremer (AER 2002)

This policy was motivated by low secondary school enrollment in Colombia (55% of eligible kids in poorest quintile). It is a very large school voucher program: over 125,000 students participate. In many areas, vouchers are allocated by lottery. Lottery winners receive $190 which corresponds to average tuition of low-to-middle cost private schools in Columbia's largest cities. Nevertheless, the average fees for the private schools attended by the voucher applicants are much larger $340.
Before a child can apply to the lottery, she must have already been admitted in a private school. What does this imply for the estimate we get from this program? Does it generate selection bias?
Most elite private schools opted out of program and the overall characteristics of these schools are different while the participating private schools are very similar to existing public schools.
The empirical strategy is simple given the randomized nature of the program. The regression equation is given by

$$y_{ic} = X_i'\beta_0 + \alpha_0 Z_i + \delta_c + \varepsilon_{ic}$$

where y is an outcome for a child i, in cohort c, X is a vector of control variables and Z indicates whether a child wins the lottery. Finally, fixed effects for each applying cohort are included.
The results suggest that there is no difference in overall enrollment rates. However, lottery winners are 6-7 percentage points more likely to be in private school. Lottery winners were more likely to complete more schooling and were less likely to repeat grades.
In addition, a sample of 283 students from Bogota were tested. The results indicate that lottery winners scored 0.2 standard deviations more than lottery losers. Effect of girls is more precise. Lottery winners worked 1.2 fewer hours per week indicating that maybe the effect on test scores is due to increased effort by the student.

Let's think a little bit more about this experiment. Here are some questions from a previous final exam

1. Suppose you want to use this voucher experiment to evaluate the effect of attending private school on grade repetition and learning. Set up the reduced form, the first stage, and the Wald (IV) estimate.
2. Not everyone goes to private school even if they get the voucher. On the other hand, many people go to private school even without the voucher. If the treatment effects are different for different peoples, for which population is the Wald estimator identifying the effects?

3. Discuss whether or not the Wald estimate is a consistent estimator of the private school effect on this population. For which of the following is the estimator likely to be worse: for grade repetition or for results on a standardized test?

4. Propose a second experiment to address the difficulties you mentioned in the previous question. The experiment should allow you to estimate the pure effect of private school on the winners (for example, you could combine the voucher experiment, as it was set up, with another experiment).

5. Returning to the reduced form impact of the voucher program: some have argued that comparing winners and losers of a small randomized voucher program is not a good way to evaluate the impact of a comprehensive voucher program, offered to all students in a school system. What is the basis for this argument?

6. What would be the right research design to estimate the overall effect of a voucher system?

Later life outcomes? The authors link the lottery data to the university entrance exams. Can we simply compare the scores of lottery winners and losers? What is the problem? The authors use something akin to Lee bounds.

Hsieh and Urquiola (JPubE, 2006)

A problem with many randomized experiments is that they are often small, so we get partial equilibrium effects (hopefully you saw this in discussing question 5 above). However, when thinking about large scale policy changes - like "should we implement a voucher system in the US?", we really care about the general equilibrium impact, and there are often reasons to think that partial and general equilibrium effects are not the same. *In the voucher context, how might the general and partial equilibrium effects differ?*

The issue with evaluating GE effects is that we need a treatment delivered at a sufficiently large economic level - which often means running cross country or cross state regressions, so we sometimes have to compromise on the quality of the "experiment". In these cases, rigorous arguments and careful thinking can really improve a paper. Given these caveats, let's consider Hsieh and Urquiola (2006)

Background

In 1981, Chile introduced nationwide school choice by providing vouchers to any student wishing to attend private school. Voucher private schools could not change tuition, and got the same per-student funding from the government as public schools. However, they could receive outside funding and they had lots of latitude regarding student admission (public schools could only turn students away if oversubscribed). Also, before the reform, publich schools were not explicitly funded based on enrollment, but after they were funded on a per-student basis, just like the voucher private schools. Finally, it bears mentioning that there is also a group of non-voucher private schools that charge tuition. These are generally elite schools that cater to very high SES families. After the passage of the law, voucher

private school enrollment shot up markedly - this was accompanied by a symmetric decline in public school enrolment (see figure 1).

Sorting vs. Productivity

Ideally, if we see an impact of the program, we would like to be able to tell why the program worked. Unfortunately, there could be several things that could be going on with a voucher program:

1. Competition forces existing schools to improve - all go up in productivity and test scores increase
2. Private schools are just better, so students attending them do better - increased private enrollment leads to increased test scores
3. There is sorting - the best (or perhaps the worst) students select into the private schools
4. On top of sorting, if there are peer effects, test scores may improve in schools that cream skim even if these schools are no more productive than the schools left with the worst students

Is there any way we can ever separate the productivity and sorting effects?

Identification

The whole country was treated, so what to do? The authors note that different communes (Chile has 300 of them, with an average population of 39,000, so you can think of these as being like US counties) saw different private school enrollment gains, and that these gains were correlated with observable characteristics. *Do you think this is a reasonable geographical area to use to examine GE effects?* Specifically, private enrollment grew more in urban communes, highly populated communes, and communes with more educations inequality (measured by the inter-quartile range in years of schooling among working adults). So, the OLS specification is to regress changes in education outcomes on changes in the private enrollment rate.[1] The IV instruments change in private enrollment with the characteristics mentioned above.

This idea will probably make you wary, and rightly so. *So, let's take a moment to think - in what direction would you expect these estimates to be biased?*

[1] The authors also attempt to control for trends by controlling for the 1970-1982 change in average years of schooling, the 1980-1982 change in private enrollment, and the 1978-1982 change in the proportion of schools that are private in each commune - these are "pre-existing" trends controls. They then attempt to control for concurrent trends by controlling for 1982-1988 changes in population, labor income, and average years of schooling among adults.

Results

So, let's have a look at the results. What do tables 3 (the OLS results) and 4 (the IV results) tell us about the impact of the voucher program? Does your prior on how you thought the reusults would be biased change how you percieve these results?

To support the results, the authors also present some cross country comparisons (see figure 4). They use an international test (the IEA in 1970 and the TIMSS in 1999). They normalize the scores and take the deviations from the average for 13 countries in both years, and then graph the results. *What do they tell us?*

Finally, they try and look at sorting (remember, it is difficult to separate this from productivity changes, but the authors argue that the aggregate results seem to indicate that there was little impact on productivity). They regress private enrollment on average economic characteristics of public school children over average characteristics of all school children. *What would you expect if sorting is going on? Have a look at table 5 - what do the results tell you? Suppose people evaluate the success of vouchers by comparing the outcomes of private and public school children in the same area - how would they be biased in this case?*