ECON 7130 - Microeconomics III

Spring 2016

Notes for Lecture #7

Today:

- Sharp RD
- Fuzzy RD
- How to implement RD in Stata
- Examples of thresholds/discontinuities
- Examples of RDD

Regression Discontinuity Design

- General idea: many policies make arbitrary rules resulting in treatment and control groups that are very similar, but just happen to be on either side of the policy threshold and therefore treatment is approximately random
  - e.g., Students with a 3.0 eligible for a merit-based scholarship, those with a 2.99 are not.
  - The design often arises from administrative decisions, where the incentives for units to participate in a program are partly limited for reasons of resource constraints, and clear transparent rules rather than discretion by administrators are used for the allocation of these incentives.
- There are two types of RD:
  1. Sharp Regression Discontinuity (SRD)
     - Here, there is jump from treatment to control once a threshold is crossed
     - Treatment is a deterministic function of the forcing variable
     - No "overlap" - never see observations with different treatment/control status for same value of forcing variable (this is main reason RD is different than other quasi-experimental approaches)
  2. Fuzzy Regression Discontinuity (FRD)
     - Here, there is a discontinuous change in the probability of treatment once a threshold is crossed
- Many things apply to both, so I'll start talking in general and then get to what is different for FRD
- Terms:
  - Forcing (or assignment or running) variable: the variable that determines treatment or control status (e.g. age)
  - Cutoff value/cut point: value of the forcing variable that determines if in treatment or control group (e.g. 21 yrs old)
  - Bandwidth: the amount of observations used on either side of the cut point determined as a range of the forcing variable (e.g. 6 months on either side of 21 yrs old)
- What is special about RD?
  - Unconfoundedness is satisfied by definition.
    - When $X \geq c$, T (treatment) is always 1; when $X < c$, T is always 0.
    - After conditioning on X, there is no variation left in T, so it cannot be correlated with any other factor.

    - RD is a special case of selection on observables.
  - By contrast, overlap is clearly violated in an RD since we cannot ever observe treatment and non-treatment for the same value of X.
    - A continuity assumption is required to compensate for the failure of overlap (more on this below).
  - Note a subtle but important difference from IV: the assignment variable is allowed to have a direct impact on the outcome, not just on the treatment, but it cannot have a discontinuous impact.
    - The estimated effect of the assignment variable on the outcome may reflect a true causal effect or a spurious correlation due to correlation with unobservables.

- RD is often regarded as having the highest internal validity of all quasi-experimental research methods
  - The external validity is not great since the estimate is a Local Average Treatment Effect
  - Local to the population with a values of the forcing variable near the cutoff value

- RD compared to an experiment
  - RD is often described as a "close cousin" of a randomized experiment or as a "local randomized experiment."
  - Consider an experiment in which each participant is assigned a randomly generated number, v, from a uniform distribution over the range [0,1].
    - Units with $v \geq 0.5$ assigned to treatment; units with $v < 0.5$ assigned to control
    - Because the assignment variable is random, the curves $E[Y(0)|X]$ and $E[Y(1)|X]$ are flat. And we know that they are flat.
    - The ATE can be computed as the difference in the mean value of Y on either side of the cutoff.
    - Because the functions are flat everywhere, the "optimal bandwidth" is to use all the data
  - Note that there are two main ways in which an RD differs from a randomized experiment in actuality
    - The functional form of $E[Y(0)|X]$ and $E[Y(1)|X]$ need not be flat (or linear or monotonic) and may not even be known.
    - It may be possible for units to alter their assignment to treatment by manipulating the forcing variable in a way that is not possible when it is assigned at random by the investigator.

- Identification:
  - Rules are often arbitrary so provide a natural experiment where treatment and control status is determined in a way that is uncorrelated with unobservables
  - Exploit this for quasi-random assignment to treatment/control
  - Key assumptions:
    - Discontinuity in treatment results in a discontinuity in the outcome variable
    - There is not a discontinuity in a covariate that causes the discontinuity in the outcome variable at that value of the forcing variable
    - The outcome variable is continuous in the forcing variable (other than that due to treatment) - in order to identify the average treatment effect, we need to make a smoothness assumption
    - There is a unavoidable need for extrapolation, because by design there are no units with $X_i = c$ for whom we observe $Y_i(0)$. We therefore will exploit the fact that we observe units with covariate values arbitrarily close to c.

- 3 general approaches to estimating the causal effect in an RD framework:
  1. Compare means

- In the data, we never observe $E[Y(0)|X = c]$, that is there are no units at the cutoff that don't get the treatment, but in principle it can be approximated arbitrarily well by $E[Y(0)|X = c - \varepsilon]$.
- Therefore we estimate: $E[Y|X = c + \varepsilon] - E[Y|X = c - \varepsilon]$
- This is the difference in means for those just above and below the cutoff.
- This is a nonparametric approach. A great virtue is that it does not depend on correct specification of functional forms.
- Note that I said "in principle" we can estimate means arbitrarily close to the cutoff. In practice, this depends on having lots of data within ε of the cutoff. Suppose you don't.

2. OLS with polynomials
   - The original RD design (Thistlewaite and Campbell 1960 - on scholarship and career choice) was implemented by OLS.
   - $Y = \alpha + \tau T + \beta X + \eta$
     * where T is a dummy for treatment and τ is the causal effect of interest and η is an error term.
   - This regression distinguishes the nonlinear and discontinuous jump from the smooth linear function.
   - OLS with one linear term in X is seldom used anymore because the functional form assumptions are very strong.
   - SHOW general scatter plot slides here
   - What to do?
     * Suppose the underlying functions are nonlinear and maybe unknown. In particular, suppose you want to estimate $Y = \alpha + \tau T + \beta f(X) + \eta$
     * where $f(X)$ is a smooth nonlinear function of X.
     * Perhaps the simplest way to approximate $f(X)$ is via OLS with polynomials in X. Common practice is to fit different polynomial functions on each side of the cutoff by including interactions between T and X.
     * Modeling $f(X)$ with a pth-order polynomial in this way leads to $Y = \alpha + \beta_{01}X + \beta_{02}X^2 + ... + \beta_{0p}X^p + \tau T + \beta_1 TX + \beta_2 TX^2 + ... + \beta_p TX^p + \eta$
     * Centering X at the cutoff prior to running the regression ensures that the coefficient on T is the treatment effect.
     * Common practice, for whatever reason, seems to use a 4th order polynomial, though you should be sure that your results are robust to other specifications (more on this below).
     * OLS with polynomials is a particularly simple way of allowing a flexible functional form in X. A drawback is that it provides global estimates of the regression function that use data far from the cutoff.
     * There many are other ways, but the RD setup poses a couple of problems for standard nonparametric smoothers.

3. Local linear regression
   - Instead of locally fitting a constant function (e.g.,the mean), fit linear regressions to observations within some bandwidth of the cutoff
   - A rectangular kernel seems to work best (see Imbens and Lemieux), but optimal bandwidth selection is an open question
   - A serious discussion of local linear regression is beyond the scope of this lecture. See, for example, Fan and Gijbels (1996)
   - But, really, we're just talking about running regressions on data near the cutoff.

- Graphical analysis
  - RDD lends itself especially well to graphical analysis that interpret the causal effects and test the assumptions

- Graphical inspection is an integral part of any RD analysis.
- 3 types of graphs should always be produced, where assignment variable is graphed against:
  1. The outcome
     - e.g., histogram-type estimate of the average value of the outcome by the forcing variable.
     - For some binwidth h, and for some number of bins K_0 and K_1 to the left and right of the cutoff value, respectively, construct bins $(b_k, b_{k+1}]$, for $k = 1, ..., K = K_0 + K_1$, where $b_k = c - (K_0 - k + 1)h$. Then calculate the number of observations in each bin, $N_k = \sum_{i=1}^{N} \mathbb{1}_{b_k < X_i \leq b_{k+1}}$, and the average outcome in the bin: $\bar{Y}_k = \frac{1}{N_k} \sum_{i=1}^{N} Y_i \mathbb{1}_{b_k < X_i \leq b_{k+1}}$.
     - The first plot of interest is that of the $\bar{Y}_k$, for $k = 1, K$ against the mid point of the bins, $\tilde{b}_k = (b_k + b_k + 1)/2$.
  2. Other covariates
     - e.g., Specifically, let Z_i be the M-vector of additional covariates, with m-th element Z_{im}. Then calculate $\bar{Z}_{km} = \frac{1}{N_k} \sum_{i=1}^{N} Z_{im} \mathbb{1}_{b_k < X_i \leq b_{k+1}}$.
     - The second plot of interest is that of the $\bar{Z}_{km}$, for $k = 1, K$ against the mid point of the bins, $\tilde{b}_k$, for all $m = 1, ..., M$.
  3. Density of cases
     - e.g. plot the number of observations in each bin, N_k, against the mid points $\tilde{b}_k$
- (1) should show a discontinuity at cutoff value and no other discontinuities that can't be explained
- (2) and (3) should show no discontinuity - if so, be worried about identification
- If you can't see the main result with such a simple graph, it's probably not there
- If you see a discontinuity in (2) or (3), be concerned that RDD is the right approach

- Assessing the Validity of an RD
  - It is impossible to test the continuity assumption directly, but we can test some implications of it
  - Namely, all observed predetermined characteristics should have identical distributions on either side of the cutoff, in the limit, as we approach smaller and smaller bandwidths. That is, there should be no discontinuities in the observables.
  - Again there is an analogy to an experiment: we cannot test whether unobserved characteristics are balanced, but we can test the observables. Rejection calls the randomization into question.
  - A subtle point in the RD context is that a finding a discontinuity in observable covariates indicates a violation of the continuity assumption, not a violation of unconfoundedness, which is satisfied by definition.

- The Role for Covariates in RD
  - In principle, covariates are not needed for identification in RD, but they can help reduce sampling variability in the estimator and improve precision if they are correlated with the potential outcomes.
    - This is a standard argument which also supports inclusion of covariates in analyses of randomized trials
  - Adding covariates should not affect the point estimate of the effect (very much). If it does, there is likely a problem.
  - The wider the bandwidth the more important it may be to include covariates - including additional covariates may eliminate some bias that is the result of the inclusion of these additional observations far from the cutoff point.
  - The first and most important point is that the presence of these covariates rarely changes the identification strategy. Typically, the conditional distribution of the covariates Z given X is continuous at $x = c$. If such discontinuities in other covariates are found, the justification of

the identification strategy may be questionable. If the conditional distribution of Z given X is continuous at $x = c$, then including Z in the regression

$$\min_{\alpha,\beta,\tau,\delta} \sum_{i=1}^{N} \mathbb{1}_{c-h \leq X_i \leq c+h} (Y_i - \alpha - \beta(X_i - c) - \tau W_i - \gamma(X_i - c)W_i - \delta^{'} Z_i)^2 \tag{1}$$

will have little effect on the expected value of the estimator for τ, since conditional on X being close to c, the additional covariates Z are independent of W .

- RD Pitfalls:
  - Mistaking Nonlinearity for Discontinuity
    - Consequences of using an incorrect functional form are potentially more severe for RD than for other quasi-experimental methods
    - Misspecification of the functional form may generate a bias in the treatment effect
    - The most common situation of this type is when an unaccounted for nonlinearity in the conditional mean function is mistaken for a discontinuity
    - Each of the 3 estimation methods (above) deals with this issue in a different way
      - Means - function approx linear for some small interval
      - Polynomial - more flexible functional form
      - Local linear reg - function approx linear for some small interval
  - Manipulation of assignment variable
    - If individuals have control over the assignment variable, then we should expect them to sort into (out of) treatment if treatment is desirable (undesirable)
    - Think of a means-tested income support program, or an election
    - Those just above the threshold will be a mixture of those who would have passed and those who barely failed without manipulation.
    - If individuals have precise control over the assignment variable, we would expect the density of X to be zero just below the threshold but positive just above the threshold (assuming the treatment is desirable).
    - McCrary (2008) provides a formal test for manipulation of the assignment variable in an RD. The idea is that the marginal density of X should be continuous without manipulation and hence we look for discontinuities in the density around the threshold.
    - How precise must the manipulation must be in order to threaten the RD design? See Lee and Lemieux (2010).
    - This means that when you run an RD you must know something about the mechanism generating the assignment variable and how susceptible it could be to manipulation.
  - There are discontinuities at other values of the assignment variable
    - If see this then you should be suspect of discontinuity at cut point
  - Other variables change discontinuously at the cutoff
    - If so, then can't identify what drives outcome variable
- Bandwidth Selection
  - For Local Linear Regression:
  - Bandwidth selection represents the familiar tradeoff between bias and precision
  - When the local regression function is more or less linear, there isn't much of a tradeoff so bandwidth can be larger.
  - There are two general methods for selecting bandwidth
    1. Ad hoc, or substantively derived (e.g., elections between 48-52% are "close")

2. Data driven

- Optimal bandwidth methods (Imbens and Kalyanaraman)
- Cross validation methods (Ludwig and Miller; Imbens and Lemieux)
- For Polynomial Regression:
  - Choosing the order of the polynomial is analog to the choice of bandwidth
  - Two approaches:
  - 1. Use the Akaike information criterion (AIC) for model selection: $AIC = Nln(\hat{\sigma}^2) + 2p$, where $\hat{\sigma}^2$ is the mean squared error of the regression and p is the number of model parameters (want to pick model with lowest AIC - i.e., lowest info loss)
  - 2. Select a natural set of bins (as you would for an RD graph) and add bin dummies to the model and test their joint significance. Add higher order terms to the polynomial until the bin dummies are no longer jointly significant.
  - This also turns out to be a test for the presence of discontinuities in the regression function at points other then the cutoff, which you'll want to do anyway
- In both cases:
  - In practice, you may want to focus on results for the "optimal" bandwidth, but it's important to test for lots of different bandwidths. Think of the optimal bandwidth only as a starting point.
  - If results critically depend on a particular bandwidth, they are less credible and choice of bandwidth requires a substantive justification.
  - In principle, the optimal bandwidth for testing discontinuities in covariates may not be the same as the optimal bandwidth for the treatment. Again, follow the practice of testing robustness to variations in bandwidth.

- Fuzzy Regression Discontinuity
  - Think about FRD as IV
    - First stage is a regression of treatment on polynomials of the forcing variable, a dummy for the threshold, and (optionally) interactions of these
    - The Angrist and Krueger paper on compulsory schooling was really an example of FRD - there were cutoff points for drop out and start age, but these didn't mean definitely dropped out - only affect probability drop out
  - With FRD - call "compliers" those who are affected by the treatment in the FRD (e.g. those who dropout in Angrist and Krueger's paper)
  - External Validity of results:
    - The FRD design restricts the relevant subpopulation even further to that of compliers at this value of the assignment variable.
    - Without strong assumptions justifying extrapolation to other subpopulations (e.g., homogeneity of the treatment effect), the designs never allow the researcher to estimate the overall (population) average effect of the treatment.
    - The specific average effect that is identified may well be of special interest, for example in cases where the policy question concerns changing the location of the threshold.
  - Graphical analysis with FRD:
    - In the case of FRD designs, it is also particularly useful to plot the mean values of the treatment variable W_i to make sure there is indeed a jump in the probability of treatment at the cutoff point.

- RD in Stata
  - Just run OLS with polynomials

  - Do linear OLS local to cutoff value
  - `rd.ado`
    * User written ado file
    * Creates graphs, does local linear regression
    * Choosing optimal bandwidth for local linear regression based on Imbens and Kalyanaraman (2009)

- Examples of discontinuities
  - Winner take all elections
  - IRS filing thresholds
  - SS Retirement Age
  - Legal Drinking Age
  - Age cutoffs for schooling
  - Standardized test scores/GPA for certain academic affiliations
  - Geographic boundaries like school districts

RDD: Example 1, Abdulkadiroğlu, Angrist, and Pathak, "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools", (*Econometrica*, 2014):

- Question: Do schools affect student performance?
  - What is the channel of this effect?
  - Is it peer effects or school "quality"?
- Problem: Selection into schools is nonrandom
- Solution: Use a regression discontinuity design around elite school admissions criteria
  - Within a small interval around the threshold for admittance, test scores are as good as random
  - So compare two similar groups - one who got treatment (admitted to elite school) and another who didn't
- Data:
  - Boston Public Schools
    * Standardized test scores (elementary-HS)
    * Exam school application file (includes ranking of schools (by student) and students (by school) and student test scores)
    * College Board test files
    * Student demographic info (age, race, gender, reduced lunch)
  - New York Public Schools
    * Standardized test scores (elementary-HS)
    * Exam school exam scores
    * College Board test files
    * Student demographic info (age, race, gender, reduced lunch)
- Basic Model:
  - $y_{itk} = \alpha_{tk} + \sum_j \delta_{jk} d_{ij} + (1 - Z_{ik}) f_{0k}(r_{ik}) + Z_{ik} f_{1k}(r_{ik}) + \rho_k Z_{ik} + \eta_{itk}$
  - i denotes student, t year, k school, j is year and grade of application

  - y is the outcome variable - HS/College Board test scores
  - d dummies for year and grade of application
  - α school-year effects
  - Z_{ik} are dummy variables for clearing the application cutoff at school k
  - $f(\cdot)$ are polynomial functions
  - r is the running variable - the criteria for admittance to the exam schools (basically the schools' rankings of students)
  - test scores (and the running variable) are standardized to have mean 0 and std dev = 1
- The coefficient of interest is ρ_k, the effect of school k on the outcome variable
- Identification:
  - FRD: look at students in area around school admission cutoff
  - Here, all unobservables should be the same since approx random near threshold
  - Local regressions - they look at students in a "window" around the threshold
  - Key assumption:
    * No discontinuity in student unobservables at threshold
- Results:
  - No measurable impact of elite schools on college admissions or test scores
  - No effects of racial composition
  - No significant peer effects
- Comments:
  - Discontinuities at race and baseline scores would normally be worrisome
    * I think authors justify with argument that bias would work against their results, so ok
  - Methodological contributions
    * Test external validity of estimator by looking at samples with different values of test scores that are correlated with elite school admissions criteria
    * RD estimator with deferred admissions

RDD: Example 1, Black, "Do Better Schools Matter? Parental Valuation of Elementary Education" (*QJE*, 1999):

- Idea: In the Tiebout model parents can "buy" better schools for their children by living in a neighborhood with better public schools
  - How do we measure the willingness to pay?
- Problem: Just looking in a cross section, richer parents probably live in nicer houses in areas that are better for many reasons
- Solution: Black uses the school border as a regression discontinuity
  - We could take two families who live on opposite side of the same street, but are zoned to go to different schools
  - The difference in their house price gives the willingness to pay for school quality
- Data:

  - All home sales from 1993-1995 in 3 counties in suburban Boston, MA (single-family residences only)
  - Elementary school boundaries because they are small
  - Census block data for demographic controls
- Basic Model:
  - $ln(\text{price}_{iaj}) = \alpha + X'_{iaj}\beta + Z'_j\delta + \gamma\text{test}_{aj} + \epsilon_{iaj}$
  - i denotes house, a attendance district, j school district
  - Attendance district is which school attend within a school district
  - X are house characteristics (e.g. bedrooms)
  - Z are neighborhood and school district characteristics (e.g. median income)
  - test scores are mean scores on standardized test for students in that school
- The coefficient of interest is γ, the value of good schools (proxied for by test scores)
- Identification:
  - SRD: look at houses on two sides of street that creates boundary between schools
  - Here, all unobserved neighborhood effects are same
  - Regression now: $ln(\text{price}_{iab}) = \alpha + X'_{iab}\beta + K'_b\phi + \gamma\text{test}_a + \epsilon_{iab}$, where b is boundary, and K_b are dummies for the house being on a boundary (so this an FE model, but identification is due to discontinuity at boundary)
  - Identification is within each boundary - how do home prices change in test scores are different
  - Local linear regressions - she looks at different distances from boundary
  - Key assumption:
    - No discontinuity in neighborhood amenities at boundary
- Results:
  - People pay for better schools
  - 5% increase in test scores results in 2.1% increase in home price ($3,948)

RDD: Example 2, Lee, "Randomized experiments from non-random selection in U.S. House elections", (*Journal of Econometrics*, 2008):

- Question: What is the incumbency advantage?
- Problem: Can't directly estimate looking at election rates of incumbents because incumbents are a biased sample - they won the prior election(s)!
- Solution: Compare those who narrowly one to those who narrowly lost - should be of similar "quality", but winner gets incumbency advantage
- Winning party in close elections is essentially random - use this to ID incumbency advantage
- Data: House elections, 1946-1998
- Note - he estimates the incumbent party advantage (not the advantage to a specific legislator)
- Basic Model: OLS with polynomials
- Identification:

  - "Randomly assigned" winners

- Results:
  - Incumbency advantage of 7-8% of vote share (if party won last time, this is effect)
  - Translates to 35% increase in prob winning
- One of the main points of this paper is that the running variable can be endogenous as long as it can not be perfectly chosen
- Weakness of design: What drives incumbency advantage?