Recitation 12: Regression discontinuity and gradient descent

Seung-hun Lee

Columbia University
Introduction to Econometrics II Recitation

April 25th, 2022

Regression discontinuity design

RD tries to achieve 'as good as random' without IVs

- Finding an instrumental variable Z_i that satisfies relevance, exclusion, and monotonicity might be difficult
- So what about comparing the entities that barely made the treatment against those who barely missed out? Doable if there is a cutoff
  - Means-tested social program, distance, winning margin, or test score cutoffs
- RD leverages these cutoffs, with the presumption that individuals around the cutoff are otherwise equal and thus the treatment assignment is as good as random
- Key difference
  - Less emphasis on overlap condition (unlike CIA)
  - Continuity of the covariates around the cutoff (so we have 'as good as random')

RD as a potential outcome framework: Sharp RDD

- **Sharp RDD**: The assignment into the treatment is binary and determined by

$$D_i = \mathbb{I}(Z_i \geq c(W_i))$$

  - Z_i is a running variable that determines cutoffs
  - $c(W_i)$ can be a known function of W_i or just a constant
  - Everyone above (below) cutoff is treated (not treated)

- Potential outcome framework

$$\begin{aligned} y_i &= D_i y_i(1) + (1 - D_i) y_i(0) \\ &= y_i(0) + D_i(y_i(1) - y_i(0)) \\ &= y_i(0) + \mathbb{I}(Z_i \geq c(W_i)) \cdot (y_i(1) - y_i(0)) \end{aligned}$$

- Regression framework: $y_i(d) = \mu(W_i, Z_i, d) + \epsilon_i(d)$

RD as a potential outcome framework: Fuzzy RDD

- **Fuzzy RDD**: Allows jump in the probability of being treated $\Pr(D_i|W_i, Z_i)$ at $Z_i = c(W_i)$ that is not necessarily 0 to 1
  - Cases include special exceptions on ineligible and noncompliant treated individuals
  - Sharp RDD is a very special case of fuzzy RDD
- Potential outcomes: The end result is

$$\begin{aligned} y_i &= D_i y_i(1) + (1 - D_i) y_i(0) \\ &= y_i(0) + D_i(y_i(1) - y_i(0)) \\ &= y_i(0) + \Pr(D_i = 1|W_i, Z_i = c^*) \cdot (y_i(1) - y_i(0)) \end{aligned}$$

  - c^* is c^+ for intended compliers and c^- for ineligibles

Key is continuity of covariates around the cutoff

- If Z_i has a known cutoff at c, we assume that $E[y_i(1)|W_i, Z_i]$ and $E[y_i(0)|W_i, Z_i]$ are continuous around $Z_i = c$. Put if differently,

$$E[y_i(d)|W_i, Z_i = c^+] = E[y_i(d)|W_i, Z_i = c^-] \text{ for each } d \in \{0, 1\}$$

  This also implies that the distribution of $(\mu_i(W_i, Z_i, 0), \mu(W_i, Z_i, 1))$ and $(\epsilon_i(0), \epsilon_i(1))$ is continuous at $Z_i = c$
- This is to ensure that those who are treated and untreated are observationally equivalent, at least around the cutoff
- Violation implies an incorrect TE estimate or bunching of certain individuals at the cutoff (bunching)
  - The latter can be tested with McCrary test (for continuous running variables)

With sharp RD, we identify ATE for i's around the cutoff

- For a constant cutoff $c(W_i) = c$, write

$$E[y_i|W_i, Z_i] = E[y_i(0)|W_i, Z_i] + E[(y_i(1) - y_i(0))\mathbb{I}(Z_i \geq c)|W_i, Z_i]$$

- These values are different for $Z_i = c^+$ and c^-

$$E[y_i|W_i, Z_i = c^+] = E[y_i(0)|W_i, Z_i = c^+] + E[y_i(1) - y_i(0)|W_i, Z_i = c^+]$$
$$E[y_i|W_i, Z_i = c^-] = E[y_i(0)|W_i, Z_i = c^-]$$

- Subtract the two terms

$$\begin{aligned} E[y_i|W_i, Z_i = c^+] - E[y_i|W_i, Z_i = c^-] &= E[y_i(0)|W_i, Z_i = c^+] - E[y_i(0)|W_i, Z_i = c^-] \\ &\quad + E[(y_i(1) - y_i(0))|W_i, Z_i = c^+] \\ &= E[(y_i(1) - y_i(0))|W_i, Z_i = c^+] \end{aligned}$$

where $E[y_i(0)|W_i, Z_i = c^+] = E[y_i(0)|W_i, Z_i = c^-]$ by continuity assumption

With sharp RD, we identify ATE for *i*'s around the cutoff

- Again, using continuity assumptions, we get

$$E[y_i(1) - y_i(0)|W_i, Z_i = c^+] = E[y_i(1) - y_i(0)|W_i, Z_i = c]$$

- Therefore, we have

$$E[y_i(1) - y_i(0)|W_i, Z_i = c] = E[y_i|W_i, Z_i = c^+] - E[y_i|W_i, Z_i = c^-]$$

- Values of c^+ and c^-: $[c - h, c + h]$ from optimizing tradeoff between bias and variance
  - Narrow bandwidth: Increases accuracy at the cost of increased variance (vice versa)
  - Very limited external validity: What about people outside of the bandwidth?
- Like LATE, we identify ATE of local populations (think of $\mathbb{I}(Z_i \geq c)$ as possible IV)

Fuzzy RD identifies similar TE

- We have different notation for D_i

$$E[y_i|W_i, Z_i] = E[y_i(0)|W_i, Z_i] + E[(y_i(1) - y_i(0)) \Pr(D_i = 1|W_i, Z_i)|W_i, z_i]$$

- So we can do the similar approach as in sharp RDD and get

$$E[y_i|W_i = w, Z_i = c^+] = E[y_i(0)|W_i = w, Z_i = c^+] + \Pr(D_i = 1|W_i = w, Z_i = c^+)E[y_i(1) - y_i(0)|W_i = w, Z_i = c^+]$$
$$E[y_i|W_i = w, Z_i = c^-] = E[y_i(0)|W_i = w, Z_i = c^-] + \Pr(D_i = 1|W_i = w, Z_i = c^-)E[y_i(1) - y_i(0)|W_i = w, Z_i = c^-]$$

- Using the continuity assumption, the difference in the left hand side can be written as

$$\begin{aligned} E[y_i|W_i = w, Z_i = c^+] - E[y_i|W_i = w, Z_i = c^-] = [&\Pr(D_i = 1|w, c^+) - \Pr(D_i = 1|w, c^-)] \\ &\times E[y_i(1) - y_i(0)|W_i = w, Z_i = c] \end{aligned}$$

- So we get ATE at $Z_i = c$ and $W_i = w$

$$E[y_i(1) - y_i(0)|W_i = w, Z_i = c] = \frac{E[y_i|W_i = w, Z_i = c^+] - E[y_i|W_i = w, Z_i = c^-]}{\Pr(D_i = 1|W_i = w, Z_i = c^+) - \Pr(D_i = 1|W_i = w, Z_i = c^-)}$$

RD using local polynomials around cutoff

- Apply one at $[c-h, c)$ and the other at $[c, c+h]$
- Local polynomials (at least linear) $>>>>>$ local constant
- Bandwidth: Minimize the $MSE(h)$ at $Z_i = c$

$$E[(\hat{\mu}(w, c, 0) - \mu(w, c, 0))^2 + (\hat{\mu}(w, c, 1) - \mu(w, c, 1))^2]$$

which is suggested by Calonico, Cattaneo, and Titiunik (2014)

RD using parametric approaches

- Not recommended if nonparametrics is feasible since parametrics rely on extreme extrapolation assumptions
- Run this regression on $[c-h, c+h]$

$$y_i = W_i\beta + W_i \cdot 1(Z_i \geq c)\gamma + W_i \cdot (c - Z_i)\delta + W_i \cdot (c - Z_i) \cdot 1(Z_i \geq c)\mu + \epsilon_i$$

- Here, what happens is that
  - $Z_i \geq c$: $y_i = W_i(\beta + \gamma) + W_i \cdot (c - Z_i)(\delta + \mu)$
  - $Z_i < c$: $y_i = W_i(\beta) + W_i \cdot (c - Z_i)(\delta)$
- On a (y_i, Z_i) plane, γ represent jumps and we can allow slopes to differ using δ, μ

RD using parametric approaches

- Plot y_i as a function of Z_i for a given W_i and see if there is a mean shift
- Plot the propensity score $\Pr(D_i = 1|W_i, Z_i)$ and look for a shift
- Plot W_i's as function of Z_i and confirm that there is no shift
  - Check for manipulation/gaming the cutoff with McCrary test
- Unknown cutoff (or corrupted cutoff?) Check distribution of Z_i and check for largest jump (Chay, McEwan, and Urquiola (2005))

Howell, AER 2017: Discrete running variables

- Effect of R&D subsidies on opening new ventures (discrete ranking as a cutoff)

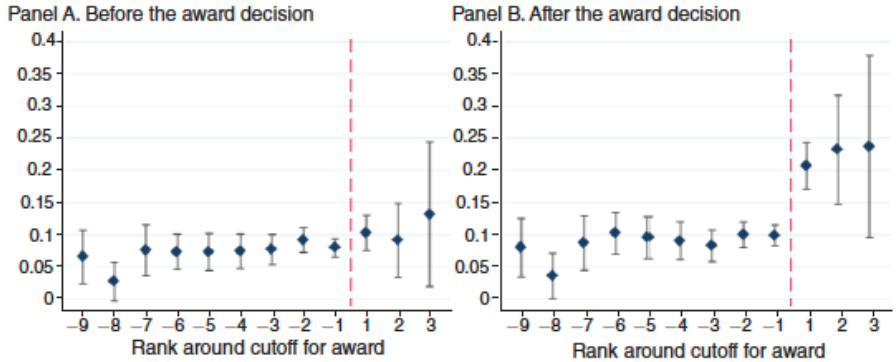
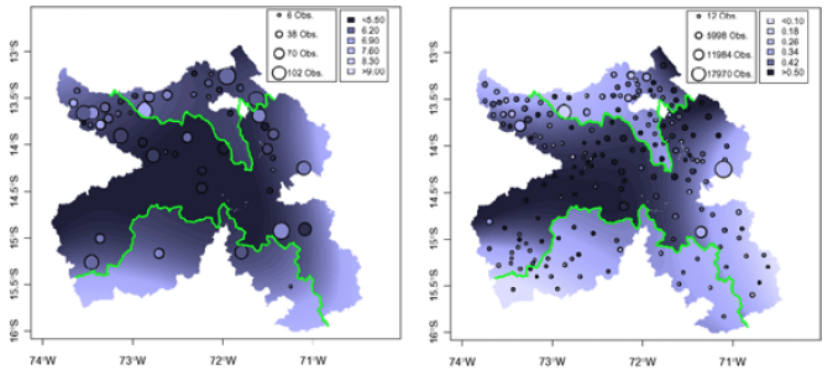


FIGURE 3. PROBABILITY OF VENTURE CAPITAL BEFORE AND AFTER GRANT BY RANK

Figure: Reception of venture capital, before and after

Dell ECMA 2010: Multivariate running variables

- Long run effect of extractive institutions on consumption and health outcomes



(a) Consumption (2001) (b) Stunting (2005)

Figure: Consumption and Stunting (Darker: less consumption and more stunting)

Machine learning: Gradient descent methods

We worry about model selection for predicting out-samples

- In determining the structure, we face the tradeoff between better fit and predictive capabilities of the model (overfitting can be bad for prediction)
- Model selection: We start with a set $\mathcal{M}$ of candidate models. Then for a model $M \in \mathcal{M}$, we define a penalty parameter $p(M)$ that increases in complexity
  - Akaike's Information Criterion: Choose M satisfying

$$\min_{M} \left(\min_{\theta} \left[-\sum_{i=1}^{n} \log l_i(\theta; M) + 2p(M) \right] \right)$$

  - Bayesian Information Criterion: Choose M satisfying

$$\min_{M} \left(\min_{\theta} \left[-\sum_{i=1}^{n} \log l_i(\theta; M) + p(M) \frac{\log n}{2} \right] \right)$$

Penalized optimization

- We minimize

$$\sum_{i=1}^{n}(Y_i - X_i\theta_0)^2 + \lambda \sum_{k=1}^{p} \theta_k^2 \quad \left(\sum_{k=1}^{p} \theta_k^2 = \|\theta\|_2^2\right) \quad \text{(Ridge)}$$

  Or

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\theta_0)^2 + \lambda \sum_{k=1}^{p} |\theta_k| \quad \text{(LASSO)}$$

- λ selected through leave-one-out, k-fold cross validation etc.

Classical gradient descent: Newton method

- We want to minimize a loss function ($l(\cdot)$ is continuously differentiable in θ)

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \theta)$$

- By Taylor expansion ($\tilde{\theta}$: Current point, $\theta^{(k)}$, θ: Next point to be reached $\theta^{(k+1)}$)

$$L(\theta) \simeq L(\tilde{\theta}) + L'(\tilde{\theta})(\theta - \tilde{\theta}) + \frac{1}{2}L''(\tilde{\theta})(\theta - \tilde{\theta})^2$$

- We solve

$$\frac{d}{d(\theta - \tilde{\theta})} = L'(\tilde{\theta}) + L''(\tilde{\theta})(\theta - \tilde{\theta}) = 0 \implies \theta^{(k+1)} = \theta^{(k)} - \frac{L'(\theta^{(k)})}{L''(\theta^{(k)})}$$

- Multiple dimensions: $\theta^{(k+1)} = \theta^{(k)} - s_k \nabla L(\theta^{(k)})$ (s_k converges to inverse Hessian matrix at minima and $\nabla L(\cdot)$ is a gradient matrix)

Stochastic gradient descent: Useful with large n and θ

- With increasing dimensions and observations, we end up doing many more iterations, which can be costly
- Stochastic gradient descent: Pick a random mini batch of m observations B_k and use the estimate of a gradient from this set of observations to optimize
- In equation,

$$\theta^{(k+1)} = \theta^{(k)} - s_k \frac{1}{m} \sum_{i \in B_k} \nabla_\theta l(y_i, \theta^{(k)})$$

  where $\frac{1}{m} \sum_{i \in B_k} \nabla_\theta l(y_i, \theta^{(k)})$ is the average gradient
- Additional noise is useful for obtaining global minima that may be far away from local ones (useful in large settings)
- The step size s_k decreases as we approach the correct global minima and it is known to give an error in $\frac{1}{\sqrt{k}}$

Minibatch gradient descent

- It splits the training data into several mini-batches and conducts the optimization on each batch iteratively
- Procedure
  - Find a minimizer in the first batch $\theta^{(1)}$ from the first batch.
  - Then use $\theta^{(1)}$ as a starting point and use the gradient descent in batch 2 to find a minimal in $\theta^{(2)}$.
  - Repeat the process until the learning process slows down and we reach a global minima.
- Uses of stochastic gradient: Backbone of artificial neural networks, geophysics, and training of linear models

Ensemble method: Combining algorithms for better predictions

- Think of ensemble as an average of pre- dictions from multiple sources, with weights determined depending on the performance of the individual algorithms
- So we solve (y : target, p_m : weight on algorithm m, $\hat{y}_m$: prediction from algorithm m)

$$\min_{p_1,..,p_q} L\left(y, \sum_{m=1}^{q} p_m \hat{y}_m\right) \leq \min_{m \in \{1,..,q\}} L(y, \hat{y}_m)$$

$$\text{s.t.} \sum_{m=1}^{q} p_m = 1, \quad p_m \geq 0$$

- Application: Remote sensing and detecting stock price manipulation

Gradient boosting: Reduce errors from weak learners with iteration

- Imagine having a simple model - regression with few covariates/decision trees with few branches
- Our goal is to minimize the prediction error $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ from the simple models in class $\mathcal{M}$, while penalizing complexity
- Start with a null predictor $\hat{y}_i^{(0)} = 0$ and residuals $r_i^{(0)} = y_i - \hat{y}_i^{(0)} = y_i$. Then:
  1. Fit sample model to data $(x_i, r_i^{(0)} = y_i)$, penalize complexity, and get new predictors $\hat{r}_i^{(1)}$
  2. Update the predictions $\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + s\hat{r}_i^{(1)}$
  3. Update the residuals $\hat{r}_i^{(1)} = \hat{r}_i^{(0)} - s\hat{r}_i^{(1)}$
  4. Iterate for each $b = 1, ..., B$

Gradient boosting: Reduce errors from weak learners with iteration

- To conduct the first step in case of parametric models with complexity penalty $p(M)$, choose M and θ_M that minimizes

$$\sum_{i=1}^{n}(r_i^{(b-1)} - r_M(x_i, \theta_M))^2 + p(M)$$

  where $\hat{r}^{(b)} = r_M(x_i, \theta_M)$
- Gradient boosting is used in any fields are the machines are trained to learn to rank (also known as machine-learned ranking) - search engines.
- Some physics fields, such as high energy physics, uses this in their data analysis.