Canonical Research Designs VII: Regression Discontinuity I: Identification and Groundwork

Paul Goldsmith-Pinkham

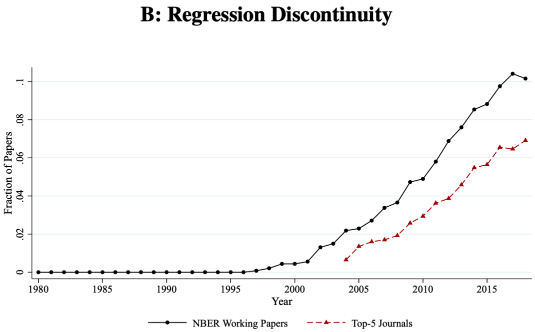
April 15, 2021

Roadmap for Today

- Today: regression discontinuity
- The goal will be to outline the simplest version of this approach, and how it works
- We will then discuss estimation in the most straightforward settings
- Next class we will touch on more complicated settings and extensions

Regression Discontinuitiy

- Regression discontinuity has exploded onto the scene for empirical designs
- A rare case of a research design with random variation that is typically caused by real world constraints (and hence much more believable)
- Also the constraint is typically of interest directly
  - The reduced form is interesting on its own, unlike some traditional IV papers
- Also allows for *graphical* presentation, a la binscatter, which creates transparency



Examples

- The intellectual history of RD begins with Thistlewaite and Campbell (1960)
- But modern empirical examples begin with three notable examples:
  - Van Der Klaauw (2002)
  - Black (1999)
  - Angrist and Lavy (1999)
- All on very different topics, but focused on discontinuous changes in some policy variables as a function of some smooth forcing variable:
  - Educational scores
  - Distance to border
  - Class size

1268 VAN DER KLAAUW

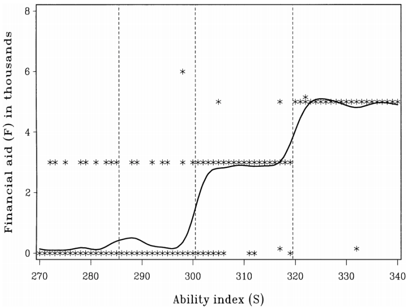


FIGURE 4

FINANCIAL AID OFFERS—NONFILERS. RAW DATA AND SPLINE SMOOTH (SOLID CURVE)

Examples

- The intellectual history of RD begins with Thistlewaite and Campbell (1960)

- But modern empirical examples begin with three notable examples:
    - Van Der Klaauw (2002)
    - Black (1999)
    - Angrist and Lavy (1999)

- All on very different topics, but focused on discontinuous changes in some policy variables as a function of some smooth forcing variable:
    - Educational scores
    - Distance to border
    - Class size

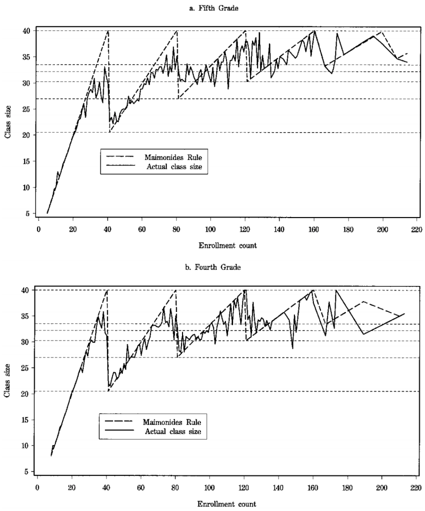


FIGURE I
Class Size in 1991 by Initial Enrollment Count, Actual Average Size and as Predicted by Maimonides' Rule

Notation for RD

- Setup notation first with traditional potential outcomes framework
  - $Y_i(0)$, $Y_i(1)$, $D_i = \{0, 1\}$, e.g. $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$
  - Running variable: Z_i (e.g. test score, distance or class size) – normalize $Z_i = 0$ as the cutoff where the treatment D_i is affected

- Key parameter to focus on is the conditional mean $\mu_Y(z) = E(Y_i | Z_i = z)$
  - Can think about more parts of distribution, but stronger requirement and will come to this later

- Need to distinguish between two cases:
  - Sharp RD: at the cutoff, $D_i = 1$ vs. $D_i = 0$
  - Fuzzy RD: at the cutoff, $E(D_i | Z_i = 0)$ changes discontinuously
    - Fuzzy RD is just IV! We can consider a scaled version of our estimate that adjusts for the compliers shifted by the design

What's the estimand? What's the goal?

- Note that since D_i discontinuously changes at $Z_i = 0$, if $E(Y_i|Z_i)$ is sufficiently smooth, we can estimate the impact of D_i on Y_i at exactly $Z_i = 0$
  - Key assumption: $E(Y_i(0)|Z_i = z)$ and $E(Y_i(1)|Z_i = z)$ are continuous in z

- Under this assumption,
  $\tau_{CATE} = E(Y_i(1) - Y_i(0)|Z_i = 0) = \lim_{z\downarrow 0} E(Y_i|Z_i = z) - \lim_{z\uparrow 0} E(Y_i|Z_i = z)$
  - Note, this is a very particular subgroup of indiviudals, right at the cutoff

- Next class, we'll discuss a design-based approach for thinking about this:
  - More in line with our intuition that those around the cutoff are effectively "randomly" assigned

- Note that this is no different than any non-parametric estimation problem that we've studied. Consider the ATE: $\tau_{ATT} = E(Y_i(1) - Y_i(0))$
  - This estimand was estimated by needing an empirical analog for an unknowable $E(Y_i(1)$ and $E(Y_i(0))$
  - With random assignment, we could estimate these.
  - The complexity of RD arrives in estimation and inference

Why is estimation harder for RD?

- We need to estimate the counterfactual means at $Z_i = 0$
  - We may not observe that point well, or at all
- If Z_i affects Y_i (e.g. the running variable affects the outcome), then we need to both account for this running variable effect *and* extrapolate
- Doing this in a flexible way asks substantially more of our data
  - If we knew the parametric relationship between Y and Z, this would be easy
- Concretely, we need to understand how to estimate $\mu(z)$ at our cutoff variable

Aside on non-parametric estimation

- What is non-parametric estimation? Model free approach to estimating features of the data

- In very simple cases, e.g. mean, variance, it is straightforward
  - $\hat{E}(Y) = n^{-1} \sum_i Y_i$

- However, you can consider non-paramterics for a wide range of estimation problems
  - The challenge becomes limitations in data
  - Let's make this concrete

Aside on non-parametric estimation

- Consider the non-parametric estimation problem that you have likely tried and solved many times: density estimation

- This is the estimation of $\hat{f}(x)$ for a random variable X_i
    - Note that in almost all cases when looking at densities, you consider scalar X variable

- Consider the case with a discrete variable X_i. In this case, estimation for $\hat{f}(x)$ is very straightforward:

$$\hat{f}(x) = n^{-1} \sum_i 1(X_i = x)$$

Aside on non-parametric estimation

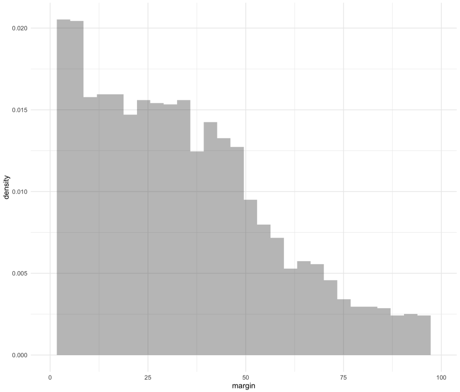
- What if X_i is continuous? The probability of $X_i = x$ is measure zero, so cannot just discretely bin

- The standard approach we learn is the histogram:

$$\hat{f}(x) = (n_k/n) \times b, \;\; n_k = \sum_i 1(X_i \in \text{k interval})$$

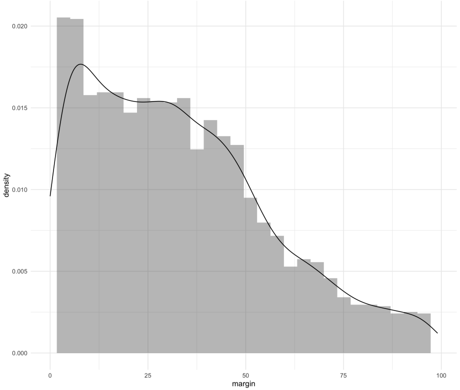
  and b is the bin-width scaled by the range of the outcome

- Example from Lee (2008) running variable



Aside on non-parametric estimation

- We can do better by using weights at each point in our dataset
- the histogram is bad because it is only “right” for certain points within the bin
  - (e.g. the approximation gets better and better as our bin size gets smaller)
- Clearly, the bandiwdth matters! What is the tradoeff?
  - Bias vs. Variance! The larger the bandiwdth, the more precisely estimated, but more bias
  - This issue comes up for RD as well



Aside on non-parametric estimation

- Formally, the density is estimated using kernel estimation as:

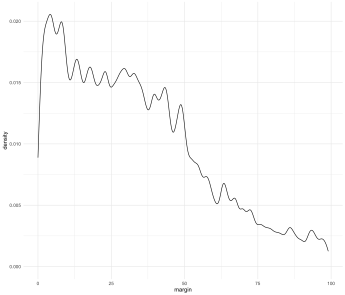
$$\hat{f}(x) = \frac{1}{Nh}\sum_i K\left(\frac{X_i - c}{h}\right),$$

  where K denotes our *kernel* weighting function

- Lots of things to know about kernels, but the key idea is that they sum to one.
  - A histogram is just a uniform kernel weighting around a given point!

- h is our choice of bandwidth / smoothing parameter. The bigger the bandwidth, the wider your window

- Next class, we will discuss data-driven approaches for this
  - However, limiting asymptotic argument requires that $h \to 0$



Aside on non-parametric estimation

- Formally, the density is estimated using kernel estimation as:

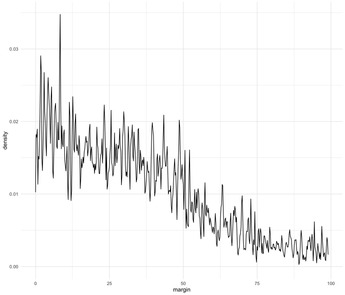
$$\hat{f}(x) = \frac{1}{Nh} \sum_i K\left(\frac{X_i - c}{h}\right),$$

  where K denotes our *kernel* weighting function

- Lots of things to know about kernels, but the key idea is that they sum to one.
  - A histogram is just a uniform kernel weighting around a given point!

- h is our choice of bandwidth / smoothing parameter. The bigger the bandwidth, the wider your window

- Next class, we will discuss data-driven approaches for this
  - However, limiting asymptotic argument requires that $h \to 0$



Challenges with non-parametric estimation

- Consider the problem of kernel estimation with two variables (or more)
- The number of datapoints necessary grows exponentilaly with the dimension of the problem
- This downside to non-parametrics becomes particularly clear once you consider non-paramteric *regression*

Non-parametric regression

- Remember what we cared about what we started was $\mu(z) = E(Y|Z_i = z)$
- Note what the expectation is:

$$\mu(z) = \int y\ f(y|z)dy$$

- If Z were discrete, this is a straightforward problem. However, once Z is continuous, we need to smoothly draw on data from nearby points
- This is the local regression approach (we will focus on the linear case)
  - This exploits the fact that the function $\mu(z)$ is locally approximable by a linear function (as we get closer and closer – the same logic of a Taylor approximation)
- Hence, consider fitting the local regression around point z with bandwidth h with uniform kernel:

$$\min_{\alpha,\beta} \sum_{i|z-h<Z_i<z} (Y_i - \alpha - \beta(Z_i - z))^2 \tag{1}$$

Aside on non-parametric estimation

- More generally, consider the following general kernel problem:

$$\hat{\mu}(z) = \min_{\alpha,\beta} \sum_{i|z-h<Z_i<z} (Y_i - \alpha - \beta(Z_i - z))^2 K_h(z - Z_i) \tag{2}$$

  where $K_h(u) = h^{-1}K(u/h)$ is our kernel weight. Two examples worth knowing:
  - Uniform: $K(u) = 0.5$ (u runs from -1 to 1)
  - Triangular: $K(u) = (1 - |u|)$ (u runs from -1 to 1)
  - Epanechnikov: $K(u) = 0.75(1 - u^2)$ (u runs from -1 to 1)

- We now have all the tools we need to do RD!
  - Recall that RD simply requires estimating μ_z at $z = 0$, using only data on the left, and only data on the right

Checklist for estimation in RD

- Choose kernel
  - Uniform is really fine for RD – if kernel matters, you likely have sensitive estimates

- Choose bandwidth
  - Can be done in a data-driven way

- Estimate on left and right:
  - $\tau_{SRD} = \lim_{z \downarrow 0} \mu(z) - \lim_{z \uparrow 0} \mu(z)$
  - And hence: $\hat{\tau}_{SRD} = \hat{\alpha}_r - \hat{\alpha}_l$ where

$$\hat{\alpha}_l, \hat{\beta}_l = \arg\min_{\alpha,\beta} \sum_{i|c-h<Z_i<c} (Y_i - \alpha - \beta(Z_i - c))^2 K_h(c - Z_i)$$

$$\hat{\alpha}_r, \hat{\beta}_r = \arg\min_{\alpha,\beta} \sum_{i|c<Z_i<c+h} (Y_i - \alpha - \beta(Z_i - c))^2 K_h(c - Z_i)$$