Gov 2002: 11. Regression Discontinuity Designs

Matthew Blackwell

November 12, 2015

1. Sharp Regression Discontinuity Designs

2. Estimation in the SRD

3. Fuzzy Regression Discontinuity Designs

4. Bandwidth selection

Introduction

- Causal for us so far: selection of observables, instrumental variables for when this doesn't hold
- Basic idea behind both: find some plausibly exogeneous variation in the treatment assignment
- Selection on observables: treatment as-if random conditional on X_i
- IV: instrument provides exogeneous variation
- Regression Discontinuity: exogeneous variation from a discontinuity in treatment assignment

Plan of attack

1. Sharp Regression Discontinuity Designs

2. Estimation in the SRD

3. Fuzzy Regression Discontinuity Designs

4. Bandwidth selection

1/ Sharp Regression Discontinuity Designs

Setup

- The basic idea behind RDDs:
  - X_i is a forcing variable.
  - Treatment assignment is determined by a cutoff in X_i.
- X_i can be related to the potential outcomes, but we assume that relationship is smooth,
- $\rightsquigarrow$ changes in the outcome around the threshold can be interpretted as a causal effect.
- The classic example of this is in the educational context:
  - Scholarships allocated based on a test score threshold (Thistlethwaite and Campbell, 1960)
  - Class size on test scores using total student thresholds to create new classes (Angrist and Lavy, 1999)

Notation

- Treatment: $D_i = 1$ or $D_i = 0$
- Potential outcomes, $Y_i(1)$ and $Y_i(0)$
- Observed outcomes:

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$$

- Forcing variable: $X_i \in \mathbb{R}$
- Covariates: an M-length vector $Z_i = (Z_{1i}, \dots, Z_{Mi})$

Design

- Sharp RD: treatment assignment is a deterministic function of the forcing variable and the threshold:

Assumption SRD

$$D_i = 1\{X_i \geq c\} \qquad \forall i$$

- When test scores are above 1500 → offered scholarship
- When test scores are below 1500 → not offered scholarship
- Key assumption: no compliance problems (deterministic)
- At the threshold, c, we only see treated units and below the threshold $c - \varepsilon$, we only see control values:

$$\mathbb{P}(D_i = 1 | X_i = c) = 1$$
$$\mathbb{P}(D_i = 1 | X_i = c - \varepsilon) = 0$$

Threshold

- Intuitively, we are interested in the discontinuity in the outcome at the discontinuity in the treatment assignment.
- We want to investigate the behavior of the outcome around the threshold:

$$\lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x]$$

- Under certain assumptions, this quantity identifies the ATE at the threshold:

$$\tau_{SRD} = E[Y_i(1) - Y_i(0)|X_i = c]$$

Plotting the RDD (Imbens and Lemieux, 2008)

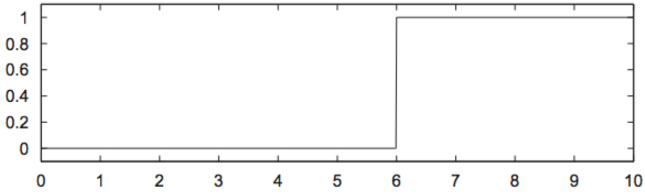


Fig. 1. Assignment probabilities (SRD).

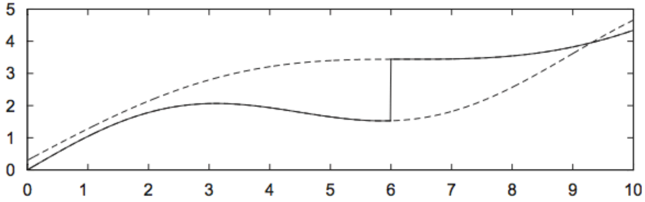


Fig. 2. Potential and observed outcome regression functions.

Comparison to traditional setup

- Note that ignorability here hold by design, because condition on the forcing variable, the treatment is deterministic.

$$Y_i(1), Y_i(0) \perp\!\!\!\perp D_i | X_i$$

- Remember the positivity/overlap assumption:

$$0 < \Pr[D_i = 1 | X_i = x] < 1$$

- With a SRD, the propensity score is only 0 or 1 and so positivity is violated.
  - $\rightsquigarrow$ we can't use ignorability directly.
- Thus, we need to extrapolate from the treated to the control group and vice versa.

Extrapolation and smoothness

- Remember the quantity of interest here is the effect at the threshold:

$$\begin{aligned}\tau_{SRD} &= E[Y_i(1) - Y_i(0)|X_i = c] \\ &= E[Y_i(1)|X_i = c] - E[Y_i(0)|X_i = c]\end{aligned}$$

- But we don't observe $E[Y_i(0)|X_i = c]$ ever due to the design, so we're going to extrapolate from $E[Y_i(0)|X_i = c - \varepsilon]$.
- Extrapolation, even at short distances, requires smoothness in the functions we are extrapolating.

Continuity of the CEFs

Assumption 1: Continuity

The functions

$$E[Y_i(0)|X_i = x] \qquad \text{and} \qquad E[Y_i(1)|X_i = x]$$

are continuous in x.

- This continuity implies the following:

$$\begin{aligned} E[Y_i(0)|X_i = c] &= \lim_{x \uparrow c} E[Y_i(0)|X_i = x] && \text{(continuity)} \\ &= \lim_{x \uparrow c} E[Y_i(0)|D_i = 0, X_i = x] && \text{(SRD)} \\ &= \lim_{x \uparrow c} E[Y_i|X_i = x] && \text{(consistency/SRD)} \end{aligned}$$

- Note that this is the same for the treated group:

$$E[Y_i(1)|X_i = c] = \lim_{x \downarrow c} E[Y_i|X_i = x]$$

Identification results

- Thus, under the consistency assumption, the sharp RD assumption, and the continuity assumption, we have:

$$
\begin{aligned}
\tau_{SRD} &= E[Y_i(1) - Y_i(0)|X_i = c] \\
&= E[Y_i(1)|X_i = c] - E[Y_i(0)|X_i = c] \\
&= \lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x]
\end{aligned}
$$

- Note that each of these is identified at least with infinite data, as long as X_i has positive density around the cutpoint
- Why? With arbitrarily high N, we'll get an arbitrarily good approximations to the expectation of the line
- How to estimate these nonparametrically is difficult as we'll see (endpoints are a big problem)

What can go wrong?

- If the potential outcomes change at the discontinuity for reasons other than the treatment, then smoothness will be violated.
- For instance, if people sort around threshold, then you might get jumps other than the one you care about.
- If things other than the treatment change at the threshold, then that might cause discontinuities in the potential outcomes.

2/ Estimation in the SRD

Graphical approaches

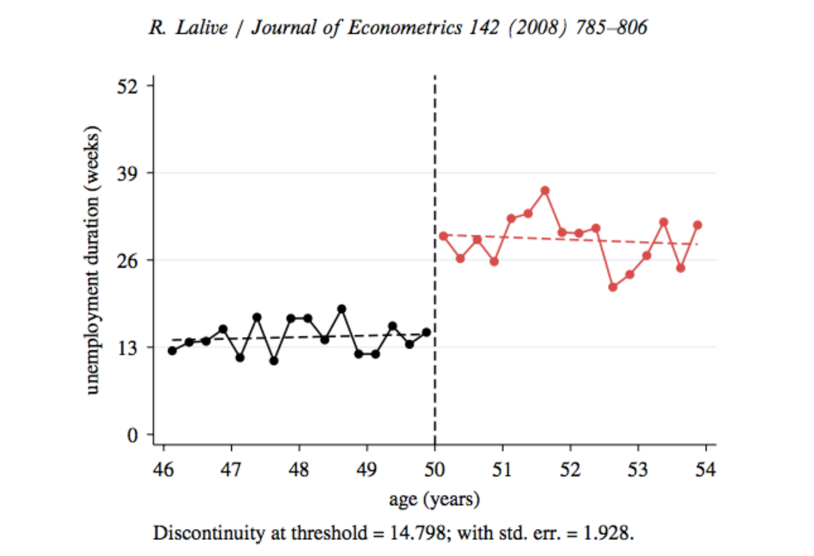
- Simple plot of mean outcomes within bins of the forcing variable:

$$\overline{Y}_k = \frac{1}{N_k} \sum_{i=1}^{N} Y_i \cdot \mathbb{I}(b_k < X_i \leq b_{k+1})$$

  where N_k is the number of units within bin k and b_k are the bin cutpoints.
- Obvious discontinuity at the threshold?
- Are there other, unexplained discontinuities?
- As Imbens and Lemieux say:

  > *The formal statistical analyses discussed below are essentially just sophisticated versions of this, and if the basic plot does not show any evidence of a discontinuity, there is relatively little chance that the more sophisticated analyses will lead to robust and credible estimates with statistically and substantially significant magnitudes.*

Example from RD on extending unemployment



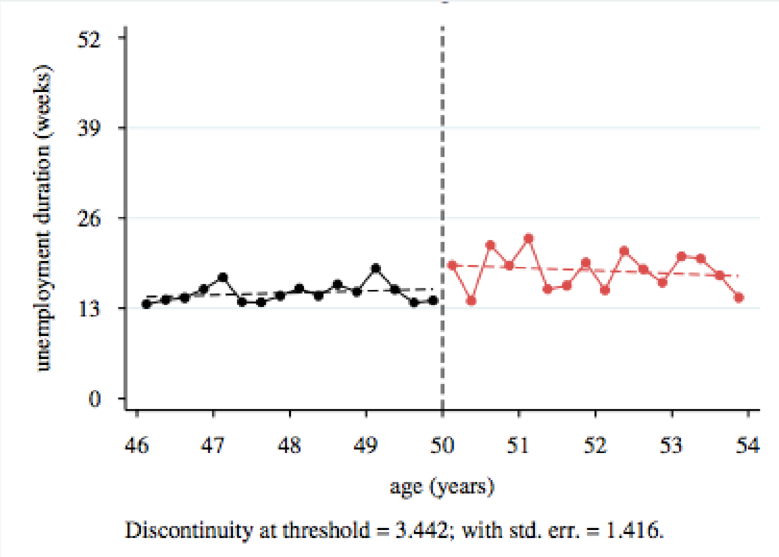
Other graphs to include

- Next, it's a good idea to plot covariates by the forcing variable to see if these covariates also jump at the discontinuity.
- Same binning strategy:

$$\overline{Z}_{km} = \frac{1}{N_k} \sum_{i=1}^{N} Z_{im} \cdot \mathbb{I}(b_k < X_i \leq b_{k+1})$$

- Intuition: our key assumption is that the potential outcomes are smooth in the forcing variable.
- Discontinuities in covariates unaffected by the threshold could be indications of discontinuities in the potential outcomes.
- Similar to balance tests in matching

Checking covariates at the discontinuity



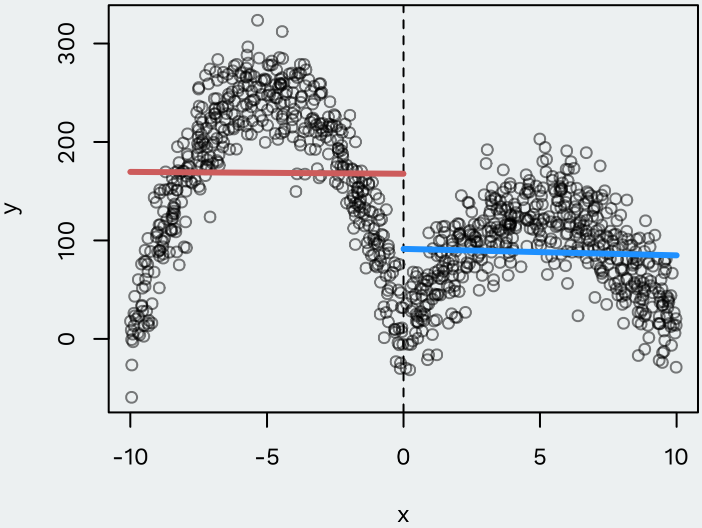
General estimation strategy

- The main goal in RD is to estimate the limits of various CEFs such as:

$$\lim_{x \uparrow c} E[Y_i | X_i = x]$$

- It turns out that this is a hard problem because we want to estimate the regression at a single point and that point is a boundary point.
- As a result, the usual kinds of nonparametric estimators perform poorly.
- In general, we are going to have to choose some way of estimating the regression functions around the cutpoint.
- Using the entire sample on either side will obviously lead to bias because those values that are far from the cutpoint are clearly different than those nearer to the cutpoint.
- → restrict our estimation to units close to the threshold.

Example of misleading trends



Nonparametric and semiparametric approaches

- Let's define

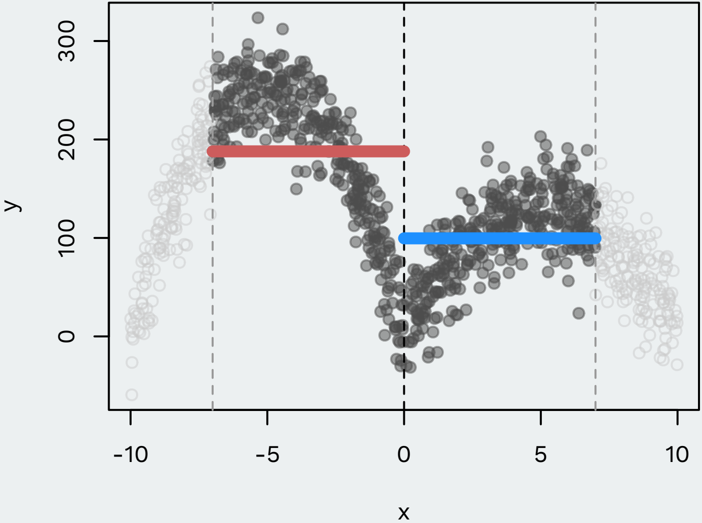
$$\mu_+(x) = \lim_{z \downarrow x} E[Y_i(1)|X_i = z]$$
$$\mu_-(x) = \lim_{z \uparrow x} E[Y_i(0)|X_i = z]$$

- For the SRD, we have $\tau_{SRD} = \mu_+(c) - \mu_-(c)$.
- One nonparametric approach is to estimate nonparametrically $\mu_-(x)$ with a uniform kernel:

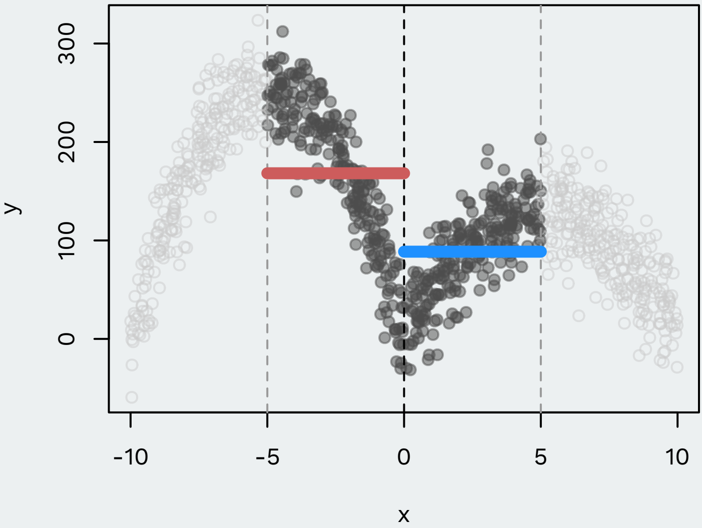
$$\widehat{\mu}_-(c) = \frac{\sum_{i=1}^N Y_i \cdot \mathbb{I}\{c - h \leq X_i < c\}}{\sum_{i=1}^N \mathbb{I}\{c - h \leq X_i < c\}}$$

- h is a bandwidth parameter, selected by you.
- Basically, calculate means among units no more than h away from the threshold.

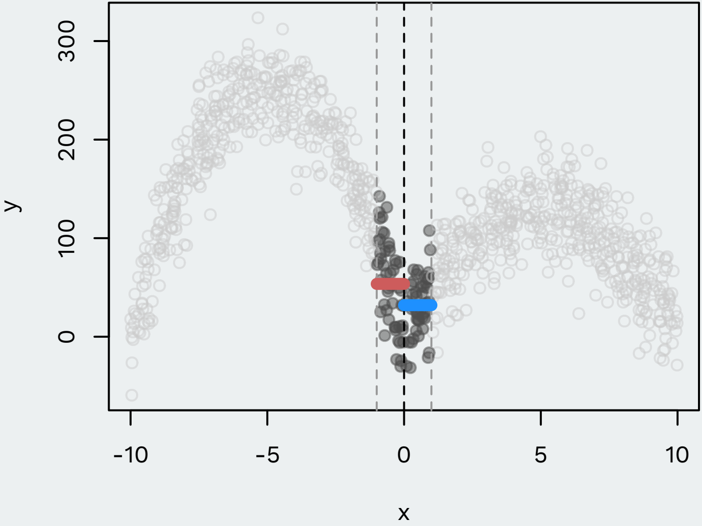
Bandwidth equal to 7



Bandwidth equal to 5



Bandwidth equal to 1



Local averages

- Estimate mean of Y_i when $X_i \in [c, c+h]$ and when $X_i \in [c-h, c)$.
- Reformulate uniform kernel approach as regression on those units less than h away from c:

$$(\widehat{\alpha}, \widehat{\tau}) = \arg\min_{\alpha,\tau} \sum_{i:X_i \in [c-h, c+h]} (Y_i - \alpha - \tau D_i)^2$$

- Predictions about Y_i are locally constant on either side of the cutoff.
- Here, $\widehat{\tau}_{SRD} = \widehat{\tau}$.
- Downside: large bias as the we increase the bandwidth.

Local linear regression

- Instead of a local constant, we can use a local linear regression.
- Run a linear regression of Y_i on $X_i - c$ in the group $X_i \in [c-h, c)$:

$$(\widehat{\alpha}_-, \widehat{\beta}_-) = \arg\min_{\alpha,\beta} \sum_{i: X_i \in [c-h, c)} (Y_i - \alpha - \beta(X_i - c))^2$$

- Same regression for group with $X_i \in [c, c+h]$:

$$(\widehat{\alpha}_+, \widehat{\beta}_+) = \arg\min_{\alpha,\beta} \sum_{i: X_i \in [c, c+h]} (Y_i - \alpha - \beta(X_i - c))^2$$

- Our estimate is

$$\begin{aligned}\widehat{\tau}_{SRD} &= \widehat{\mu}_+(c) - \widehat{\mu}_-(c) \\ &= \widehat{\alpha}_+ + \widehat{\beta}_+(c - c) - \widehat{\alpha}_- - \widehat{\beta}_-(c - c) \\ &= \widehat{\alpha}_+ - \widehat{\alpha}_-\end{aligned}$$

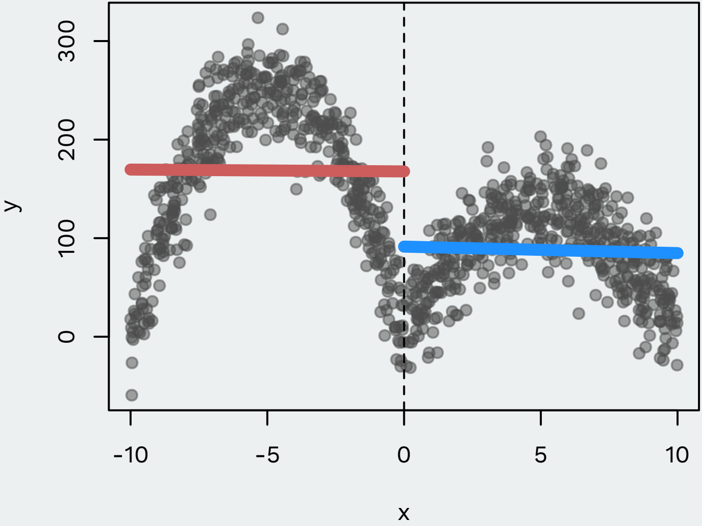
More practical estimation

- We can estimate this local linear regression by dropping observations more than h away from c and then running the following regression:

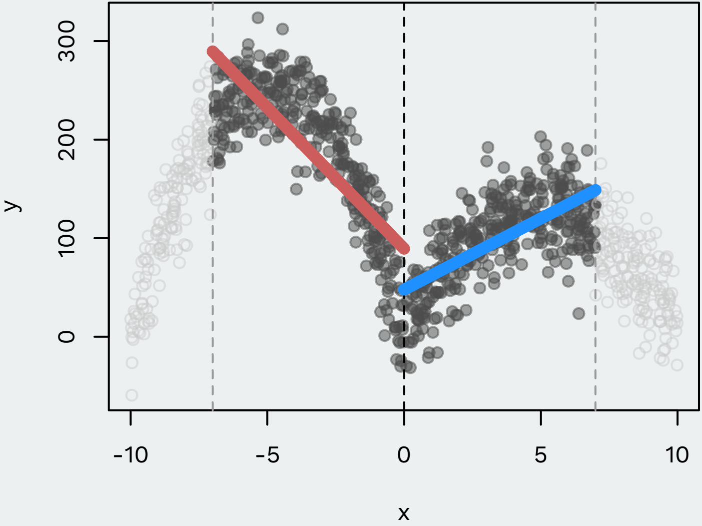
$$Y_i = \alpha + \beta(X_i - c) + \tau D_i + \gamma(X_i - c)D_i + \eta_i$$

- Here we just have an interaction term between the treatment status and the forcing variable.
- Here, $\widehat{\tau}_{SRD} = \widehat{\tau}$ which is the coefficient on the treatment.
- Yields numerically the same as the separate regressions.

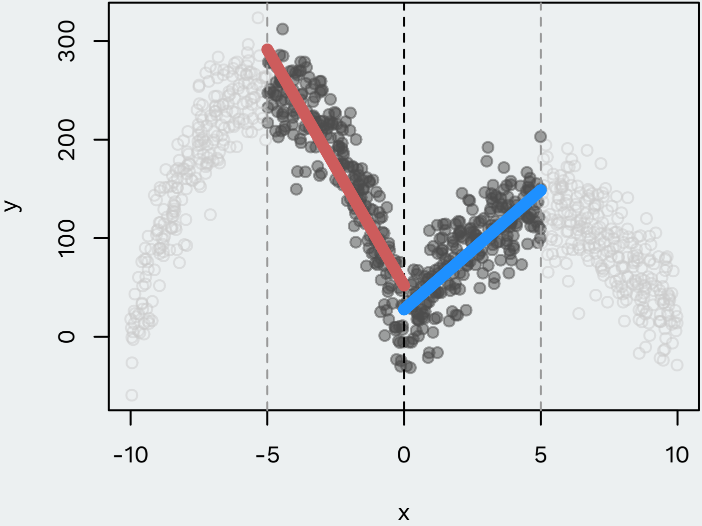
Bandwidth equal to 10 (Global)



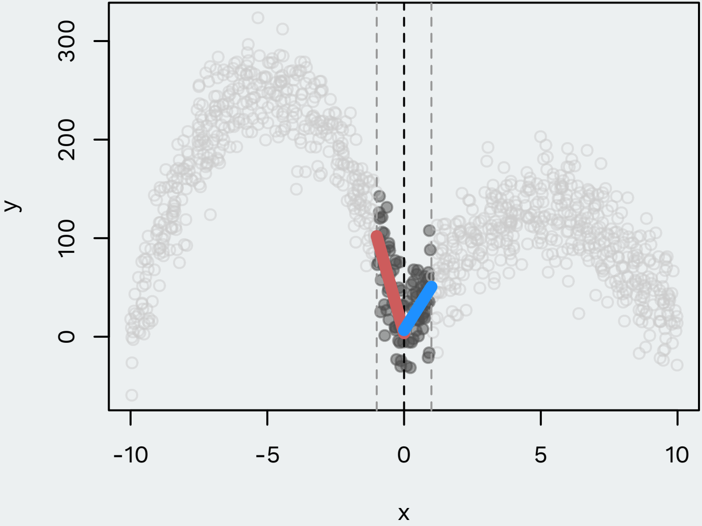
Bandwidth equal to 7



Bandwidth equal to 5



Bandwidth equal to 1



Odds and ends for the SRD

- Standard errors: robust standard errors from local OLS are valid.
- Covariates: shouldn't matter, but can include them for increased precision.
- ALWAYS REPORT MODELS WITHOUT COVARIATES FIRST
- You can include polynomials of the forcing variable in the local regression. Let $\tilde{X}_i = X_i - c$

$$Y_i = \alpha + \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i^2 + \tau D_i + \gamma_1 \tilde{X}_i D_i + \gamma_2 \tilde{X}_i^2 D_i + \eta_i$$

- Make sure that your effects aren't dependent on the polynomial choice.

3/ Fuzzy Regression Discontinuity Designs

Setup

- With fuzzy RD, the treatment assignment is no longer a deterministic function of the forcing variable, but there is still a discontinuity in the probability of treatment at the threshold:

Assumption FRD

$$\lim_{x \downarrow c} \Pr[D_i = 1 | X_i = x] \neq \lim_{x \uparrow c} \Pr[D_i = 1 | X_i = x]$$

- In the sharp RD, this is also true, but it further requried the jump in probability to be from 0 to 1.
- Fuzzy RD is often useful when the a threshold encourages participation in program, but does not actually force units to participate.

Fuzzy RD in a graph

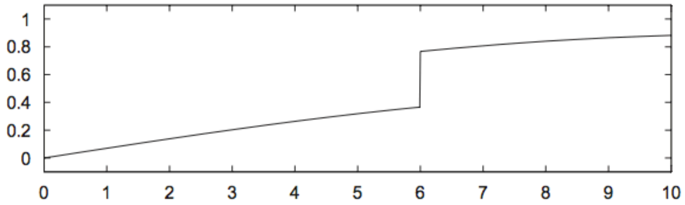


Fig. 3. Assignment probabilities (FRD).

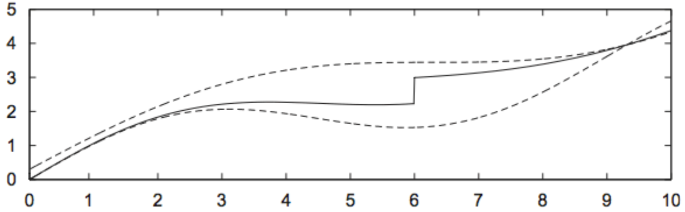


Fig. 4. Potential and observed outcome regression (FRD).

Fuzzy RD is IV

- Forcing variable is an instrument:
  - affects Y_i, but only through D_i (at the threshold)
- Let $D_i(x)$ be the potential value of treatment when we set the forcing variable to x, for some small neighborhood around c.
- $D_i(x) = 1$ if unit i would take treatment when X_i was x
- $D_i(x) = 0$ if unit i would take control when X_i was x

Fuzzy RD assumptions

Assumption 2: Monotoncity

There exists ε such that $D_i(c+e) \geq D_i(c-e)$ for all $0 < e < \varepsilon$

- No one is discouraged from taking the treatment by crossing the threshold.

Assumption 3: Local Exogeneity of Forcing Variable

In a neighborhood of c,

$$\{\tau_i, D_i(x)\} \perp\!\!\!\perp X_i$$

- Basically, in an ε-ball around c, the forcing variable is randomly assigned.

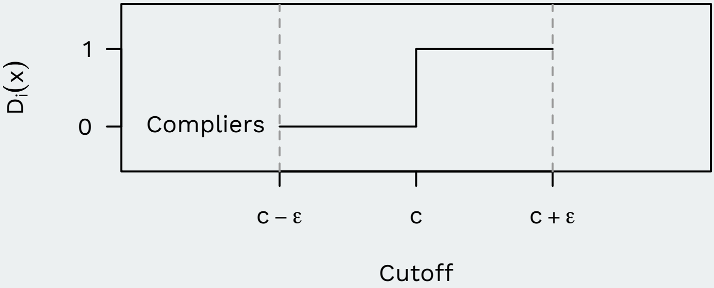
Compliance in Fuzzy RDs

- Compliers are those i such that for all $0 < e < \varepsilon$:

$$D_i(c + e) = 1 \quad \text{and} \quad D_i(c - e) = 0$$

- Think about college students that get above a certain GPA are encouraged to apply to grad school.
- Compliers would:
  - apply to grad school if their GPA was just above the threshold
  - not apply to grad school if their GPA was just below the threshold
- We don't get to see their compliance status because due to the fundamental problem of causal inference
- Could also think about this as changing the threshold instead of changing X_i

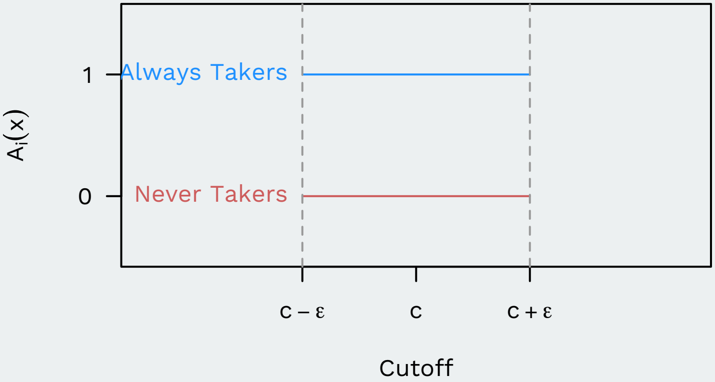
Compliance graph



- Compliers would not take the treatment if they had $X_i = c$ and we increased the cutoff by some small amount
- These are compliers at the threshold

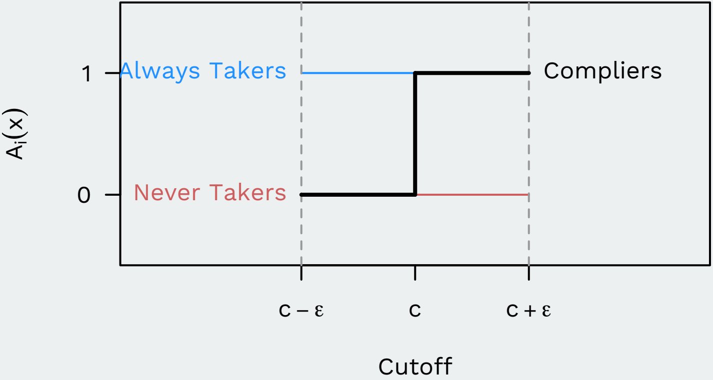
Compliance groups

- Compliers: $D_i(c + e) = 1 \quad \text{and} \quad D_i(c - e) = 0$
- Always-takers: $D_i(c + e) = D_i(c - e) = 1$
- Never-takers: $D_i(c + e) = D_i(c - e) = 0$



Compliance groups

- Compliers: $D_i(c + e) = 1 \quad \text{and} \quad D_i(c - e) = 0$
- Always-takers: $D_i(c + e) = D_i(c - e) = 1$
- Never-takers: $D_i(c + e) = D_i(c-e) = 0 $



LATE in the Fuzzy RD

- We can define an estimator that is in the spirit of IV:

$$\tau_{FRD} = \frac{\lim_{x\downarrow c} E[Y_i|X_i = x] - \lim_{x\uparrow c} E[Y_i|X_i = x]}{\lim_{x\downarrow c} E[D_i|X_i = x] - \lim_{x\uparrow c} E[D_i|X_i = x]}$$

$$= \frac{\text{effect of threshold on } Y_i}{\text{effect of threshold on } D_i}$$

- Under the FRD assumption, continuity, consistency, monotonicity, and local exogeneity, we can write that the estimator is equal to the effect at the threshold for compliers.

$$\tau_{FRD} = \lim_{e\downarrow 0} E[\tau_i|D_i(c+e) > D_i(c-e)]$$

Proof

- To prove this, we'll look at the discontinuity in Y_i in a window around the threshold and then shrink that window:

$$E[Y_i|X_i = c + e] - E[Y_i|X_i = c - e]$$

- First, remember that by consistency,

$$\begin{aligned} Y_i &= Y_i(1)D_i + Y_i(0)(1 - D_i) \\ &= Y_i(0) + (Y_i(1) - Y_i(0))D_i \\ &= Y_i(0) + \tau_i D_i \end{aligned}$$

- Plug this into the CEF of the outcome:

$$\begin{aligned} E[Y_i|X_i = c + e] &= E[Y_i(0) + \tau_i D_i|X_i = c + e] \\ &= E[Y_i(0) + \tau_i D_i(c + e)] \end{aligned}$$

- Thus, we can write the difference around the threshold as:

$$E[Y_i|X_i = c+e] - E[Y_i|X_i = c-e] = E[\tau_i(D_i(c+e) - D_i(c-e))]$$

Proof (cont)

- Let's break this expectation apart using the law of iterated expectations:

$$E[\tau_i(D_i(c+e) - D_i(c-e))] =$$

$$\begin{aligned}
&E[\tau_i \times (D_i(c+e) - D_i(c-e)) \,|\, \text{complier}\ ] \times \Pr[\text{complier}] \\
&+E[\tau_i \times (D_i(c+e) - D_i(c-e)) \,|\, \text{defier}] \times \Pr[\text{defier}] \\
&+E[\tau_i \times (D_i(c+e) - D_i(c-e)) \,|\, \text{always}] \times \Pr[\text{always}] \\
&+E[\tau_i \times (D_i(c+e) - D_i(c-e)) \,|\, \text{never}] \times \Pr[\text{never}]
\end{aligned}$$

$$= E[\tau_i \,|\, \text{complier}] \times \Pr[\text{complier}]$$

Proof (cont)

- So far, we've shown that the outcome jump at the discontinuity is the LATE times the probability of compliance:

$$E[Y_i|X_i = c+e] - E[Y_i|X_i = c-e] = E[\tau_i \,|\, \text{complier}] \times \Pr[\text{complier}]$$

- What is the probability of compliance though?

$$\begin{aligned}
\Pr[\text{complier}] &= \Pr[D_i(c+e) - D_i(c-e) = 1] \\
&= E[D_i(c+e) - D_i(c-e)] \\
&= E[D_i(c+e)] - E[D_i(c-e)] \\
&= E[D_i(c+e)|X_i = c+e] - E[D_i(c-e)|X_i = c-e] \\
&= E[D_i|X_i = c+e] - E[D_i|X_i = c-e]
\end{aligned}$$

- Thus,

$$\frac{E[Y_i|X_i = c+e] - E[Y_i|X_i = c-e]}{E[D_i|X_i = c+e] - E[D_i|X_i = c-e]} = E[\tau_i \,|\, D_i(c+e) > D_i(c-e)]$$

Misc notes

- Taking the limit as $e \to 0$, we've shown that:

$$\begin{aligned}\tau_{FRD} &= \frac{\lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x]}{\lim_{x \downarrow c} E[D_i | X_i = x] - \lim_{x \uparrow c} E[D_i | X_i = x]} \\ &= \lim_{e \downarrow 0} E[\tau_i | D_i(c+e) > D_i(c-e)]\end{aligned}$$

- Note that the FRD estimator emcompasses the SRD estimator because with a sharp design:

$$\lim_{x \downarrow c} E[D_i | X_i = x] - \lim_{x \uparrow c} E[D_i | X_i = x] = 1$$

- A note on external validity: obsviously, FRD puts even more restrictions on the external validity of our estimates because not only are we discussing a LATE, but also the effect is at the threshold. That might give us pause about generalizing other populations for the both the SRD and FRD.

Estimation in FRD

- Remember that we had:
$$\tau_{FRD} = \frac{\lim_{x\downarrow c} E[Y_i|X_i = x] - \lim_{x\uparrow c} E[Y_i|X_i = x]}{\lim_{x\downarrow c} E[D_i|X_i = x] - \lim_{x\uparrow c} E[D_i|X_i = x]}$$
- We can estimate the numerator using the SRD approaches we just outlined, $\widehat{\tau}_{SRD}$.
- For the denominator, we simply apply the local linear regression to the D_i:
$$(\widehat{\alpha}_{dL}, \widehat{\beta}_{dL}) = \arg\min_{\alpha,\beta} \sum_{i:X_i\in[c-h,c)} (D_i - \alpha - \beta(X_i - c))^2$$
$$(\widehat{\alpha}_{dR}, \widehat{\beta}_{dR}) = \arg\min_{\alpha,\beta} \sum_{i:X_i\in[c,c+h]} (D_i - \alpha - \beta(X_i - c))^2$$
- Use this to calculate the effect of threshold on D_i: $\widehat{\tau}_d = \widehat{\alpha}_{dR} - \widehat{\alpha}_{dL}$
- Calculate ratio estimator:
$$\widehat{\tau}_{FRD} = \frac{\widehat{\tau}_{SRD}}{\widehat{\tau}_d}$$

More practical FRD estimation

- The ratio estimator above is equivalent to a TSLS approach.
- Use the same specification as above with the following covariates:

$$V_i = \begin{pmatrix} 1 \\ \mathbb{I}\{X_i < c\}(X_i - c) \\ \mathbb{I}\{X_i \geq c\}(X_i - c) \end{pmatrix}$$

- First stage:

$$D_i = \delta_1' V_i + \rho \mathbb{I}\{X_i \geq c\} + \nu_i$$

- Second stage:

$$Y_i = \delta_2' V_i + \tau D_i + \eta_i$$

- Thus, being above the threshold is treated like an instrument, controlling for trends in X_i.

4/ Bandwidth selection

How to choose the bandwidth

- The bandwidth, h, is a tuning parameter that you set.
- h controls the bias-variance tradeoff:
  - High h: high bias, low variance (more data points, farther from the cutoff)
  - Low h: low bias, high variance (fewer data points, closer to the cutoff)
- Bias-variance tradeoff captured in the mean-square error of the estimator:

$$MSE(h) = \mathbb{E}[(\widehat{\tau}_h - \tau_{SRD})^2] = \underbrace{(\mathbb{E}[\widehat{\tau}_h] - \tau_{SRD})^2}_{\text{bias}^2} + \underbrace{\mathbb{V}[\widehat{\tau}_h]}_{\text{variance}}$$

- Given the setup we need to minimize the MSE of these two estimators:

$$MSE_+(h) = \mathbb{E}\left[(\widehat{\mu}_+(c,h) - \mathbb{E}[Y_i(1)|X_i = c])^2\right]$$
$$MSE_-(h) = \mathbb{E}\left[(\widehat{\mu}_-(c,h) - \mathbb{E}[Y_i(0)|X_i = c])^2\right]$$

Choosing the optimal bandwidth

- Goal: choose a value of h that minimizes the MSE of our CEF estimators.
  - But that requires knowing the true CEFs, $\mathbb{E}[Y_i(d)|X_i]$.
- Two ways to handle this situation:
  1. Use cross validation to choose h that produces the best fit for the CEFs.
  2. Solve for the optimal bandwidth in terms of MSE and estimate that bandwidth.

Model fit and model selection

- Think a bivariate regression context and let h be the order of the polynomial that we should include in the model:

$$\mathbb{E}[Y_i|X_i = x] = \beta_0 + \sum_{k=1}^{h} \beta_k x^k$$

- How many orders of the polynomial should we include? How do we compare models?
  - More polynomials will always fit a particular dataset better.
  - But this could lead to overfitting for this particular dataset.
  - We could test our model on a separate dataset to get a sense of the MSE.

Cross validation

- Cross validation in general:
  1. Randomly split the data into a training set and a validation set, S of size m.
  2. Use the training set to estimate $\mu(x, h) = \mathbb{E}[Y_i|X_i = x]$ for many values of h.
  3. Estimate the MSE of each choice of h using data in the validation set:

$$\widehat{MSE}(h) = \frac{1}{m} \sum_{i \in S} (Y_i - \widehat{\mu}(X_i))^2$$

  4. Choose the value of h that produces the lowest $\widehat{MSE}(h)$

Flavors of cross-validation

- K-fold cross-validation:
  1. Randomly split data into K subsets.
  2. For one subset k, use S_k as the validation set and S_{-k} as the test set.
  3. Calculate the MSE for many values of h: $\widehat{MSE}^k(h)$
  4. Repeat 2-3 for all $k = 1, \ldots, K$
  5. Average across K cross-validations:

$$\widehat{MSE}(h) = \frac{1}{K}\sum_{k=1}^{K} MSE^k(h)$$

  6. Choose the h that minimizes $\widehat{MSE}(h)$
- Leave one out cross-validation: the above procedure with $K = N$.

CV for RDD

- Run the SRD model for a given h:

$$\underset{(\alpha,\beta,\tau,\gamma)}{\arg\min} \frac{1}{N_h} \sum_{i:X_i \in (c-h, c+h)} (Y_i - \alpha + \beta(X_i - c) + \tau D_i + \gamma(X_i - c) D_i)$$

- Perform K-fold CV with this regression to choose h.
- Problem: minimizes error across many values of $X_i = x$ but we only care about $X_i = c$.
  - Partial solution: only consider bandwidths that contain less than 50% of data.
  - Still a problem.

Optimal bandwidth selection

- Imbens and Kalyanaraman derive an approximation to the asymptotic MSE for each value of h.
  - The optimal bandwidth depends on the density of the forcing variable at c, the variance of Y_i around c, and the curvature of the CEFs at c.
- IK procedure:
  1. Choose initial bandwidth h_1 and calculate conditional variances on either side of c and the density of X_i at c
  2. Choose another initial bandwidth h_2 to calculate the 2nd derivative of $\mu_+(c)$ and $\mu_-(c)$.
  3. Add a small regularization penalty that ensures h isn't "too big" in finite samples.
- IK procedure depends on a kernel to weight units differently depending on how far they are from the cutoff.