Empirical Methods for Policy Evaluation

Second Part

Matteo Bobba

Toulouse School of Economics (TSE)

TSE PhD Program (MRes)
Fall 2024

Outline and Readings for this Section (3 Classes)

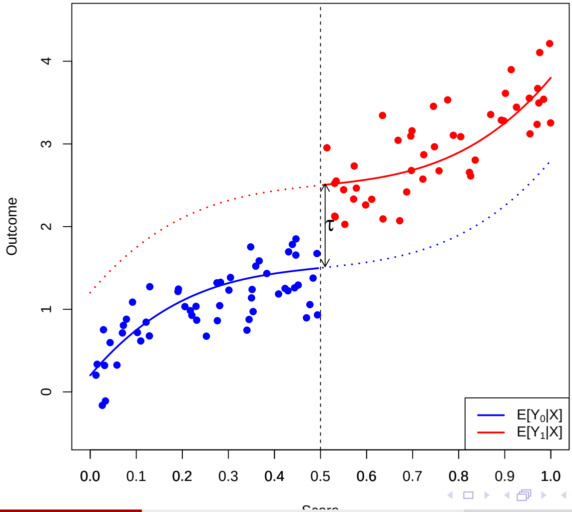
- Regression Discontinuity Designs
  - Local randomization approach (Cattaneo-Idrobo-Titiunik: Book 2, Section 2)
  - Continuity-based approach (CIT: Book 1)
  - RD extensions (CIT: Book 2, Sections 3, 4 and 5)
- RDDs and Empirical Matching Models
  - **Bobba, Ederer, Leon-Ciliotta, Neilson, and Nieddu (2024)**

The Local Randomization Approach

Setup and Notation

- Potential outcomes: $Y_i(1), Y_i(0)$, with $\tau_i = Y_i(1) - Y_i(0)$
- Continuous running variable (score): X_i
- Treatment indicator: $D_i = D_i(X_i) = 1$ if treated, 0 otherwise
- Observed outcome: $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$
- RD exploits a discontinuity in $P[D_i = 1|X_i]$ at some cutoff c
- Sharp design (will extend this later): $P[D_i = 1|X_i] = \mathbb{I}(X_i \geq c)$

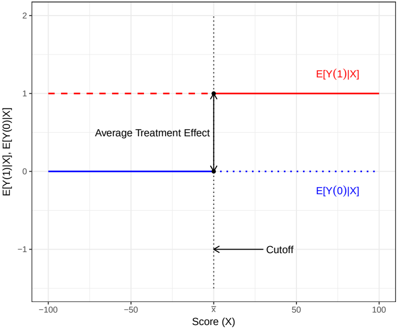
Graphical Intuition



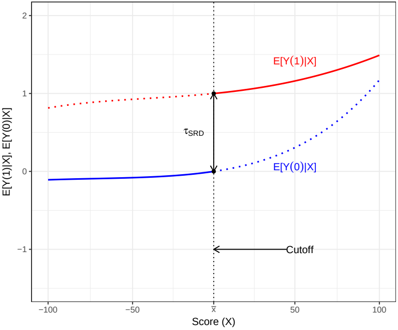
RD as a Randomized Experiment

- Idea: close enough to the cutoff, some units were "lucky"
- Treatment as if randomly assigned in a window around c if:
  - Units do not have **exact** control of their score
  - There is a random chance element in score each unit receives
  - Probability of this random "error" does not jump at the cutoff
- Example: each unit assigned a score $X_i \sim U[0,1]$, $D_i = \mathbb{I}(X_i \geq c)$, then $P[D_i = 1] = 1 - c$

RD as a Randomized Experiment



(a) Randomized Experiment



(b) RD Design

Local Randomization Approach to RD

- There is a window $W_0 = [c - w, c + w]$ in which:
  1. Probability distribution of X_i is unrelated to individual characteristics

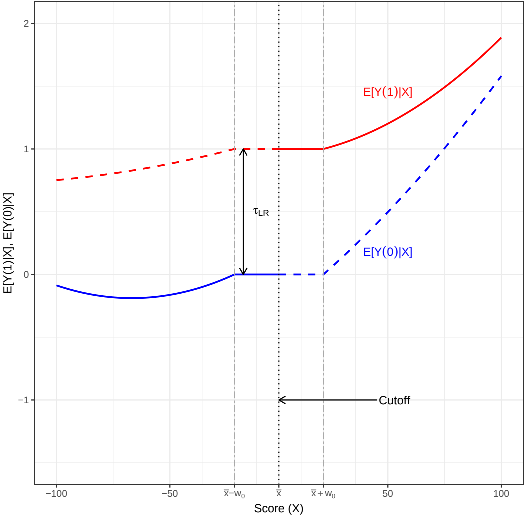
$$P[X_i \leq x | X_i \in W_0] = F_0(x), \quad \forall\, i$$

  2. Potential outcomes not affected by value of the score:

$$Y_i(d, x) = Y_i(d)$$

- Potential outcomes are a constant function of the score

Local Randomization RD: Intuition



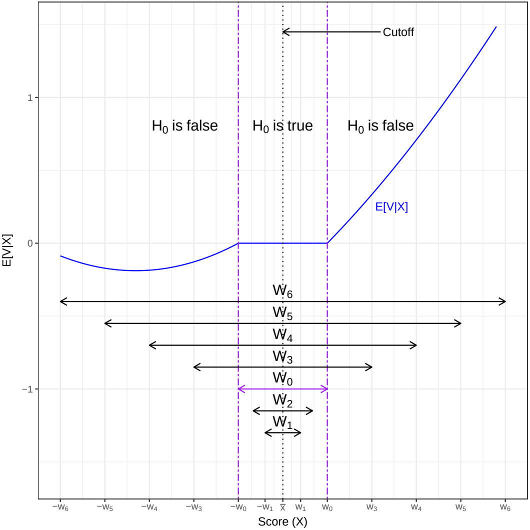
Window Selection: A Data Driven Method

- Under random assignment, covariates should be balanced:

$$P[V_i \leq v | D_i = 1] = P[V_i \leq v | D_i = 0]$$

- Can use this idea as a windows selection criterion:
  - Find window in which all covariates are balanced
- Iterative procedure:
  1. Choose a test statistic (diff. means, Kolmogorov-Smirnov,...)
  2. Choose an initial "small" window $W_0^{(1)} = [c - w_{(1)}, c + w_{(1)}]$
  3. Test null that covariates are balanced above and below c
  4. Enlarge slightly the window and repeat until null hypothesis is rejected

Window Selection Procedure



Estimation and inference

- Once W_0 is found, proceed as in a randomized experiment

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$$

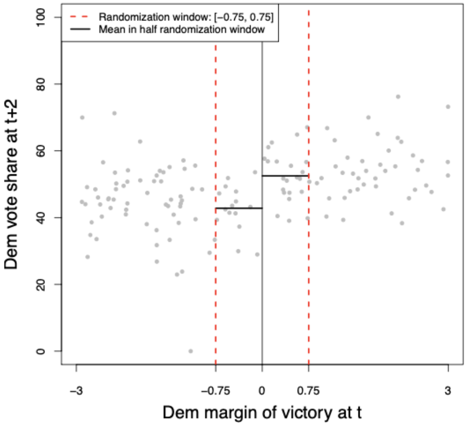
- Covariate-balance criterion may yield windows with few obs
- Inference based on large-sample approximations may not be reliable
- Alternative approach: randomization inference

Software Implementations

- Cattaneo, Titiunik and Vazquez-Bare (Stata Journal, 2016)
- `rdlocrand` package: statistical inference and graphical procedures for RDD employing local randomization methods
  - `rdwinselect`: window selection
  - `randinf`: randomization inference
  - `rdsensitivity`: sensitivity analysis
  - `rdrbounds`: Rosenbaum bounds

Example: Incumbency Advantage in U.S. Senate

- Y_i = election outcome at $t+1$ ($=1$ if party wins)
- D_i = election outcome at t ($=1$ if party wins)
- X_i = margin of victory at t ($c=0$)



The Continuity-based Approach

Identification (Hahn, Todd, and Van der Klauw, 2001)

1. (sharp design): $D_i = \mathbb{I}(X_i \geq c)$
2. (smoothness): $\mathbb{E}[Y_i(0)|X_i = x]$, $\mathbb{E}[Y_i(1)|X_i = x]$ continuous at $x = c$

Then,

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

- Treatment effect only (nonparametrically) identified at the cutoff
  - Only point of overlap (in the limit)
  - We actually have zero observations at $X_i = c$

Identification

- Naive difference in means:

$$\begin{aligned}\Delta(h) &= \mathbb{E}\{Y_i \mid X_i \in [c, c+h]\} - \mathbb{E}\{Y_i \mid X_i \in [c-h, c)\} \\ &= \mathbb{E}\{Y_i(1) \mid X_i \in [c, c+h]\} - \mathbb{E}\{Y_i(0) \mid X_i \in [c-h, c)\} \\ &= \mathbb{E}\{\tau_i \mid X_i \in [c, c+h]\} + \text{Bias}(h)\end{aligned}$$

where $\text{Bias}(h) = E\{Y_i(0) \mid X_i \in [c, c+h]\} - E\{Y_i(0) \mid X_i \in [c-h, c)\}$

- If $\mathbb{E}[Y_i(d)|X_i = x]$ is continuos at $x = c$ for $d = 0, 1$, then:

$$\lim_{h \downarrow 0} \Delta(h) = \mathbb{E}[\tau_i | X_i = c]$$

Estimation: Overview

1. Global:
   - Estimate a p-th-order polynomial on full sample
   - Sensitive to misspecification
   - Erratic behavior at boundary points
2. "Flexible parametric":
   - Estimate a polynomial within an ad-hoc bandwidth
   - Sensitive to misspecification and bandwidth choice
3. Nonparametric local polynomial:
   - Data-driven bandwidth selection
   - Accounts for misspecification when performing inference

Global Parametric Approach

- Parametric assumption on conditional expectations, e.g.

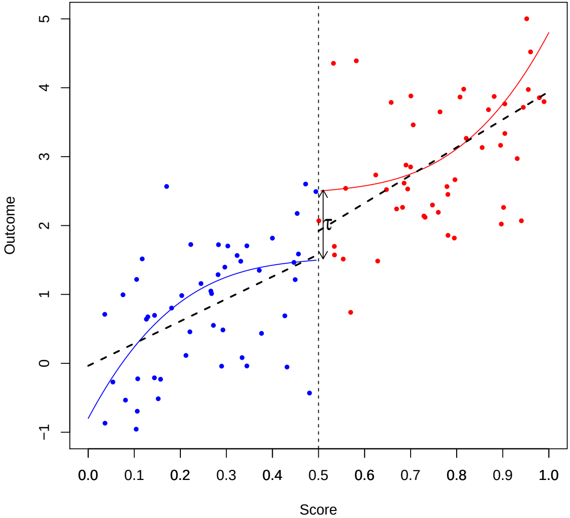
$$\mathbb{E}[Y_i(d)|X_i] = \alpha_d + \beta_d(X_i - c)$$

- This implies

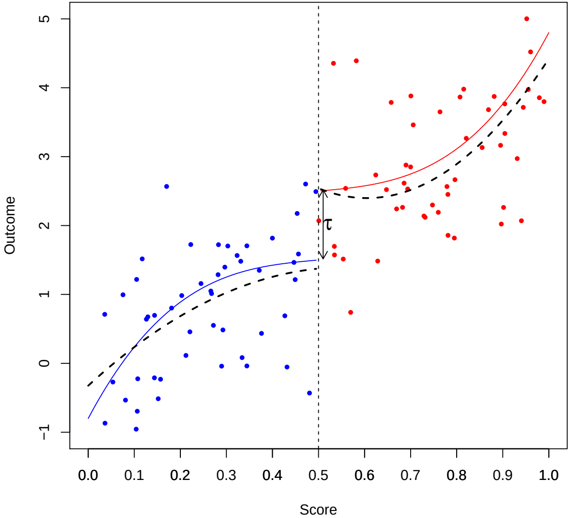
$$\mathbb{E}[Y_i|X_i] = \alpha_0 + (\alpha_1 - \alpha_0)D_i + \beta_0(X_i - c) + (\beta_1 - \beta_0)(X_i - c)D_i + u_i$$

- Easily estimated by OLS on full sample
- Coefficient $\alpha_1 - \alpha_0$ recovers the treatment effect at the cutoff

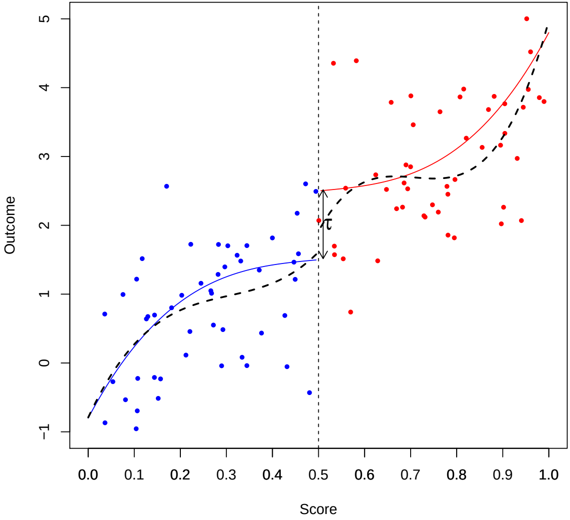
Global Parametric Approach: $p = 1$



Global Parametric Approach: $p = 2$



Global Parametric Approach: $p = 3$



Local Linear Regression

- Suppose $c = 0$ (otherwise, use $X_i - c$)
- Choose some bandwidth $h > 0$ and estimate by OLS:

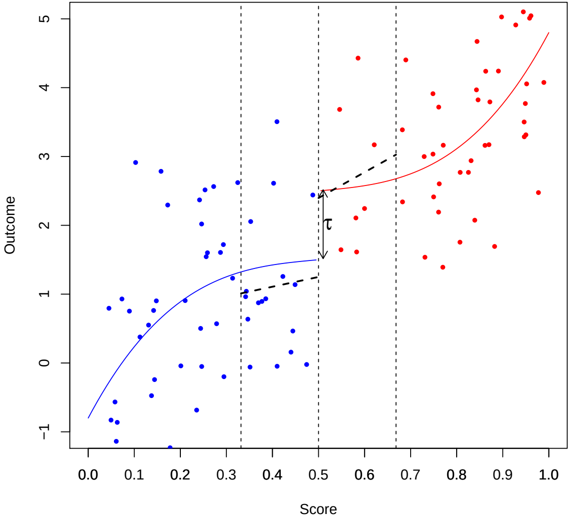
$$(\hat{\alpha}^+, \hat{\beta}^+) = \underset{(\alpha,\beta)}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 \mathbb{I}(0 \leq X_i \leq h)$$

$$(\hat{\alpha}^-, \hat{\beta}^-) = \underset{(\alpha,\beta)}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 \mathbb{I}(-h \leq X_i < 0)$$

- Estimated treatment effect at the cutoff:

$$\hat{\tau} = \hat{\alpha}^+ - \hat{\alpha}^-$$

Local Linear Regression: Graphical Intuition



Local Linear Regression: OLS Estimands

- By standard OLS algebra:

$$\hat{\beta}^+ = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X}_h)\mathbb{I}(0 \leq X_i \leq h)}{\sum_{i=1}^n X_i(X_i - \bar{X}_h)\mathbb{I}(0 \leq X_i \leq h)}$$

$$\hat{\alpha}^+ = \bar{Y}_h - \bar{X}_h\hat{\beta}^+$$

where

$$\bar{X}_h = \frac{\sum_{i=1}^n X_i\mathbb{I}(0 \leq X_i \leq h)}{\sum_{i=1}^n \mathbb{I}(0 \leq X_i \leq h)}$$

$$\bar{Y}_h = \frac{\sum_{i=1}^n Y_i\mathbb{I}(0 \leq X_i \leq h)}{\sum_{i=1}^n \mathbb{I}(0 \leq X_i \leq h)}$$

Local Linear Regression: Bias

- It can be shown that (analogous result for $E[\hat{\alpha}^-|\mathbf{X}]$):

$$E[\hat{\alpha}^+|\mathbf{X}] = \mu_1(0) + h^2\mathcal{B}_+ + o_p(h^2)$$

- $\mu_1(x) = E[Y_i(1)|X_i = x]$
- $\mathcal{B}_+$ is a constant that depends on:
    1. The curvature of $\mu_1(x)$
    2. The kernel function
    3. The order of polynomial, p
- If $h = 0$ the estimator would be unbiased
- Smaller h implies small bias but fewer observations: more variance

Variance

- Similarly, it can be shown that (analogous result for $V[\hat{\alpha}^{-}|\mathbf{X}]$):

$$V[\hat{\alpha}^{+}|\mathbf{X}] = \frac{\mathcal{V}_{+}}{nh} + o_p(h)$$

- $\mathcal{V}_{+}$ is a constant that depends on:
  1. $V[Y_i(1)|X_i = 0]$
  2. The density of the score variable at the cutoff
  3. The kernel function
  4. The order of polynomial, p

- Decreasing the variance requires $nh \to \infty$

MSE

- Therefore:

$$E[\hat{\tau}|\mathbf{X}] - \tau = h^2\mathcal{B} + o_p(h^2)$$
$$V[\hat{\tau}|\mathbf{X}] = \frac{\mathcal{V}}{nh} + o_p(h)$$

- Mean-squared error (MSE):

$$\begin{aligned}\text{MSE}(\hat{\tau}) &= Bias(\hat{\tau})^2 + V[\hat{\tau}] \\ &= h^4\mathcal{B}^2 + \frac{\mathcal{V}}{nh}\end{aligned}$$

Bandwidth Selection

- Trade-off in bandwidth selection: bias vs variance
- MSE-optimal bandwidth:

$$
\begin{aligned}
h^*_{\mathsf{MSE}} &= \operatorname*{argmin}_h \mathsf{MSE}(\hat{\tau}) \\
&= \left(\frac{\mathcal{V}}{4\mathcal{B}^2}\right)^{1/5} n^{-1/5}
\end{aligned}
$$

- MSE-optimal bandwidth is proportional to $n^{-1/5}$

Inference

- In this case $V[\hat{\tau}|\mathbf{X}] = O_p(n^{-1}h^{-1})$ so:

$$\sqrt{nh}(\hat{\tau} - \tau) \rightarrow_{\mathcal{D}} \mathcal{N}(0, \Omega)$$

- But recall that $h^*_{\text{MSE}} \propto n^{-1/5}$ so the Normal approximation will have a bias
- Two alternatives:
  - Undersmoothing: use a "smaller" bandwidth
  - Bias correction

Bias Correction (Calonico, Cattaneo and Titiunik, ECMA 2014)

- CCT propose the following de-biasing approach:

$$\sqrt{nh}(\hat{\tau} - \tau) = \sqrt{nh}(\hat{\tau} - \mathbb{E}[\hat{\tau}|\mathbf{X}]) + \sqrt{nh}B_n$$

- Bias correction:

$$\sqrt{nh}(\hat{\tau} - \tau - B_n) = \sqrt{nh}(\hat{\tau} - \mathbb{E}[\hat{\tau}|\mathbf{X}]) \to_{\mathcal{D}} \mathcal{N}(0, \Omega)$$

- But the bias is unknown, so we need to estimate it:

$$\sqrt{nh}(\hat{\tau} - \tau - \hat{B}_n) \to_{\mathcal{D}} \mathcal{N}(0, \Omega + \Sigma)$$

where Σ accounts for the estimation of the bias

Assessing the Validity of the Continuity-based Approach

- Density discontinuity tests
- Continuity away from the cutoff
- Testing for discontinuities in covariates / placebo outcomes

Density discontinuity tests

- RDDs can be invalid if individuals manipulate X_i
- Manipulation can imply sorting on one side of the cutoff
- Test whether the density of X_i is continuous around c
- McCrary (2008), Cattaneo, Jansson and Ma (2018)

Continuity away from the cutoff

- Identification relies on continuity of $\mathbb{E}[Y_i(d)|X_i]$
- Can estimate $\mathbb{E}[Y_i(0)|X_i]$ for controls, $\mathbb{E}[Y_i(1)|X_i]$ for treated
- Check continuity away from the cutoff (graphically or formally)

Continuity in covariates / placebo outcomes

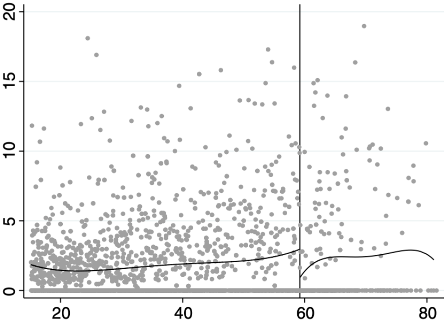
- Some variables should reveal no treatment effect:
  - Outcomes not targeted by treatment (placebo outcomes)
  - Exogenous or predetermined covariates
- Estimate an RD effect on these variables
- Finding a non-zero effect suggests an invalid RDD:
  - Existence of other (unobserved) treatments at the cutoff
  - Selection

Software Implementations

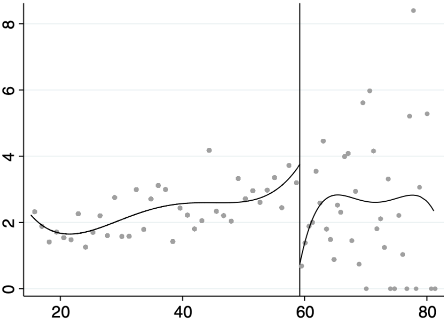
- Calonico, Cattaneo, Farrell, and Titiunik (Stata Journal, 2017)
- `rdrobust` package: estimation, inference and graphical analysis
  - `rdbwselect`: bandwidth selection procedures for local polynomial RD
  - `rdplot`: data-driven regression discontinuity plots
  - `rddensity`: manipulation testing
  - `rdpower`: power and sample size calculations for RD designs

Example: Impact of Head Start on Child Mortality

- Federal program that provides health and social services for children aged 5-9
  - HS assistance for 300 counties based on poverty index ($X_i \geq 59.19$)
  - Y_i = county-level mortality rates per 100,000



(a) Scatter Plot, Raw Data, $N^{-} = 2,455$, $N^{+} = 290$



(b) RD Plot, ES, and MV, $J_{-} = 37$, $J_{+} = 38$

Example: Impact of Head Start on Child Mortality

- Running variable manipulation falsification approach
  - Non-parametric test for continuity of the PDF of X_i near the cutoff

	Density tests				
	h_-	h_+	N_W^-	N_W^+	p-value
Method					
Unrestricted, 2-h	10.151	9.213	351	221	0.788
Unrestricted, 1-h	9.213	9.213	316	221	0.607
Restricted (1-h)	13.544	13.544	482	255	0.655

Notes: (i) Cutoff is $\bar{r} = 59.1984$ and $W = [\bar{r} - h,\ \bar{r} + h]$ denotes the symmetric window around the cutoff used for each choice of bandwidth; (ii) Density test p-values are computed using Gaussian distributional approximation to bias-corrected local-linear polynomial estimator with triangular kernel and robust standard errors; (iii) column "Method" reports unrestricted inference with two distinct estimated bandwidths ("U, 2-h"), unrestricted inference with one common estimated bandwidth ("U, 1-h"), and restricted inference with one common estimated bandwidth ("R, 1-h"). See Cattaneo, Jansson, and Ma (2016a, 2016b) for methodological and implementation details.

Example: Impact of Head Start on Child Mortality

- Ludwig and Miller (QJE, 2007): flexible parametric RD
  - $\hat{\tau}_{\{p=4,\text{full sample}\}} = -3.065$, p-value $= 0.005$
  - $\hat{\tau}_{\{p=1,h=18\}} = -1.198$, p-value $= 0.071$
  - $\hat{\tau}_{\{p=1,h=9\}} = -1.895$, p-value $= 0.055$
- Cattaneo, et al (JPAM, 2017): robust bias-corrected non-parametric RD
  - $\hat{\tau}_{\{p=0,\hat{h}_{MSE}=3.24\}} = -2.114$, robust p-value $= 0.037$
  - $\hat{\tau}_{\{p=0,h=9\}} = -1.059$, robust p-value $= 0.048$
  - $\hat{\tau}_{\{p=1,\hat{h}_{MSE}=6.81\}} = -2.409$, robust p-value $= 0.042$

RD Extensions

Fuzzy RD

- Sharp RD: score perfectly determines treatment status
  - All units scoring above the cutoff receive the treatment
  - $D_i = \mathbb{I}(X_i \geq c)$
  - Probability of treatment jumps from 0 to 1 at c
- Fuzzy RD: imperfect compliance
  - Some units below c may be treated or vice versa
  - Jump in probability at c may be < 1 (but > 0)

Intention-to-treat (ITT) parameter

- ITT: effect of being assigned to treatment
- Sharp RD design on the treatment assignment variable

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

- Under some continuity assumptions,

$$\tau_{\text{ITT}} = \mathbb{E}[(\underbrace{Y_i(1) - Y_i(0)}_{\tau_i})(\underbrace{D_{1i} - D_{0i}}_{\substack{=\ \ 1 \text{ for compliers} \\ =-1 \text{ for defiers} \\ =\ \ 0 \text{ for always/never takers}}})|X_i = c]$$

The Monotonicity Assumption

- We will rule out the presence of defiers:

$$P[\text{defier}|X_i = c] = 0$$

- This assumption is called *monotonicity*, since it implies that:

$$D_{1i} \geq D_{0i}, \quad \forall\, i$$

- Intuition: $X_i \geq c$ does not decrease the probability of treatment

Intention-to-treat (ITT) Parameter

- $D_{1i} - D_{0i} = 1$ for compliers, 0 for always-takers and never-takers
- Then

$$\tau_{\text{ITT}} = \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c, D_{1i} > D_{0i}]}_{\text{ATE on compliers: LATE}} \times \underbrace{P[D_{1i} > D_{0i}|X_i = c]}_{\text{prop of compliers}}$$

- ITT can be ≈ 0 even if LATE is large
- But still a policy relevant parameter:
  - Effect of offering the treatment

First Stage

- First stage: effect of treatment assignment on treatment status:

$$\tau_{\mathsf{FS}} = \lim_{x \downarrow c} \mathbb{E}[D_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[D_i | X_i = x]$$

- Under monotonicity,

$$\tau_{\mathsf{FS}} = P[D_{1i} > D_{0i} | X_i = c] = P[\text{complier} | X_i = c]$$

- First stage identifies the proportion of compliers at the cutoff

Recovering the ATE on Compliers

- Instrument D_i with $\mathbb{I}(X_i \geq c)$

$$\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c, D_{1i} > D_{0i}] = \frac{\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[D_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[D_i|X_i = x]}$$

- Fuzzy RD parameter is "doubly local":
  - At the cutoff
  - On the subpopulation of compliers

Estimation in Fuzzy Designs

- ITT and FS are sharp RD estimators
- The FRD parameter can be estimated using two-stage least squares
- Can adapt all previous tools to this case
  - Data driven bandwidth selection
  - Local polynomial estimation
  - Robust bias-corrected inference

Multicutoff and Multiscore RD

1. Multiple cutoffs:
    - Cutoffs change across regions, time periods, etc
    - All units receive the same treatment when they exceed their cutoff
2. Cumulative cutoffs:
    - Treatment is multivalued
    - Different dosage of treatment depending on value of X_i
    - E.g. $D_i = \mathbb{I}(X_i \leq c_1) + 2\mathbb{I}(c_1 < X_i \leq c_2)$
3. Multiple scores:
    - Treatment assigned based on multiple running variables
    - E.g. scholarship if both math and language scores above a cutoff

RD with Multiple Cutoffs

- Common empirical approach: pooling
  - $C_i \in \mathcal{C}$ (random) cutoff faced by unit i
  - Discrete cutoffs: $\mathcal{C} = \{c_0, c_1, ..., c_J\}$
  - Re-centered running variable: $\tilde{X}_i = X_i - C_i$
  - Pooled estimand:

$$\tau^p = \lim_{x \downarrow 0} \mathbb{E}[Y_i | \tilde{X}_i = x] - \lim_{x \uparrow 0} \mathbb{E}[Y_i | \tilde{X}_i = x]$$

Identification under the Pooling Approach

- If the CEFs and $f_{X|C}(x|c)$ are continuous at the cutoffs,

$$\tau^p = \sum_{c \in \mathcal{C}} \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c, C_i = c] \omega(c)$$

- Where

$$\omega(c) = \frac{f_{X|C}(c|c) P[C_i = c]}{\sum_{c \in \mathcal{C}} f_{X|C}(c|c) P[C_i = c]}$$

Exploiting Multiple Cutoffs

- Two drawbacks of the pooling approach:
  - Policy relevance: combines TEs *for different populations*
  - Discards variation that can identify parameters of interest
- What are the parameters of interest in this context?
- Potential CEFs:

$$\mu_d(x,c) = \mathbb{E}[Y_i(d)|X_i = x, C_i = c], \qquad d \in \{0,1\}$$

- (Conditional) ATE:

$$\tau(x,c) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x, C_i = c] = \mu_1(x,c) - \mu_0(x,c)$$

RD with Cumulative Cutoffs: Parameters

- Multivalued treatment $D_i \in \{d_1, d_2, \ldots, c_J\}$
- Effect of switching to one dosage to the next one:

$$\tau_j = \mathbb{E}[Y_i(d_j) - Y_i(d_{j-1})|X = c_j]$$

- Under continuity assumptions,

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

RD with Cumulative Cutoffs: Estimation and Inference

- Can use robust bias-corrected techniques cutoff by cutoff
- Unit i is "control" for some units, "treated" for others
- Bandwidth selection:
  - Ensure bandwidths are non-overlapping or
  - Joint estimation accounting for overlap

RD with Multiple Scores

- Bivariate score: $\mathbf{X}_i = (X_{1i}, X_{2i})$
- Suppose treatment is assigned if both scores exceed a cutoff:

$$D_i = \mathbb{I}(X_{1i} \geq b_1)\mathbb{I}(X_{2i} \geq b_2)$$

- Multidimensional RD parameter:

$$\tau(\mathbf{b}) = \mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{b}], \quad \mathbf{b} \in \mathcal{B}$$

  - ATE at each point in the boundary set $\mathcal{B}$

RD with Multiple Scores: Identification

- Under continuity assumptions,

$$\tau(\mathbf{b}) = \lim_{\substack{d(\mathbf{x},\mathbf{b})\to 0 \\ \mathbf{x}\in\mathcal{B}_t}} \mathbb{E}[Y_i|\mathbf{X}_i = \mathbf{x}] - \lim_{\substack{d(\mathbf{x},\mathbf{b})\to 0 \\ \mathbf{x}\in\mathcal{B}_c}} \mathbb{E}[Y_i|\mathbf{X}_i = \mathbf{x}]$$

  - $\mathcal{B}_t =$ treated region
  - $\mathcal{B}_c =$ control region

- Need to define a notion of distance $d(\mathbf{x}, \mathbf{b})$

RD with Multiple Scores: Estimation

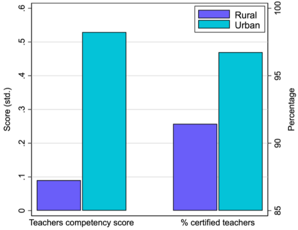
- Estimating a whole curve of $\tau(\mathbf{b})$ may not be feasible
- Alternative approach: pooling
  - Define distance measure $d(\cdot, \cdot)$
  - Normalize running variable as distance to closest boundary point
  - Run RD on (unidimensional) normalized running variable $\tilde{X}_i$

Bobba, Ederer, Leon-Ciliotta, Neilson, and Nieddu (2024)

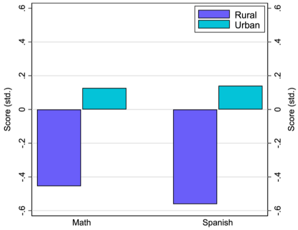
Teacher Compensation and Structural Inequality: Evidence from Centralized Teacher School Choice in Peru

- Rich admin dataset on nation-wide allocation of public teachers in Peru
  - Document large urban-rural gap in teacher quality and student test scores
- RD-based evidence of teacher wage bonuses in remote and rural locations
  - Teacher competency $\uparrow$ by 0.39σ + student test scores $\uparrow$ by $0.23\text{-}0.32\sigma$
- Model of teacher school choice/value added to study aggregate policy effects
  - Possible to close urban-rural gap by leveraging match effects
  - Framework to design cost-effective wage policy for equity/efficiency objective

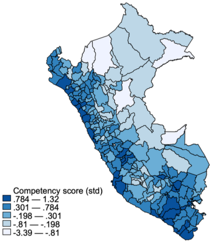
Inequality of Education Inputs and Output



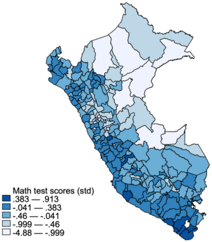
a) Teacher Competency by Urban/Rural



b) Student Achievement by Urban/Rural

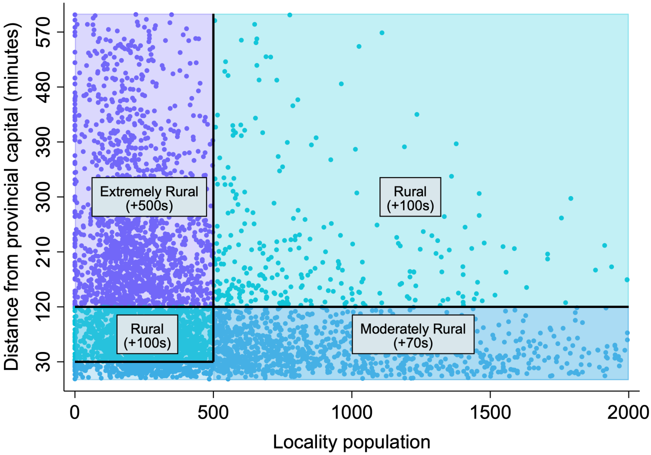


c) Teacher Competency by Province



d) Student Achievement (Math) by Province

The Rural Wage Bonus Policy



- From Rural to Extremely Rural wages increase by $\approx 1/4$ of base salary

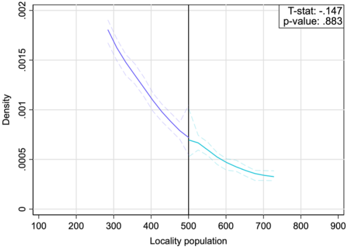
RD Design

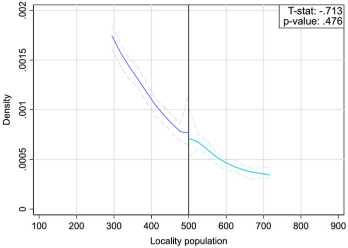
- Let $y_i(d)$ be the potential outcome of teacher i (or student taught by i)
- $D_{(i)j} = \mathbb{I}(pop_j \leq pop_c)\mathbb{I}(time_j \geq time_c)$ for high-paying vs. low-paying j
- This sharp and multiscore RD design relies on:

  Cont $\mathbb{E}(Y_i(d) \mid \mathbf{X}_{(i)j} = \mathbf{x})$ is continuos in $\mathbf{x}$, $\forall d \in \{0, 1\}$

- The plausibility of this assumption can be (indirectly) checked in the data

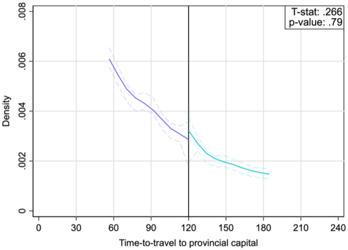
Density Tests Around Extremely Rural Cutoff

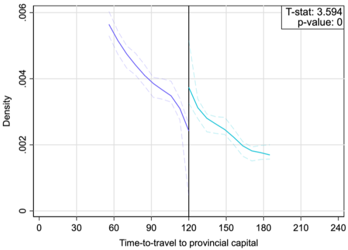




a. Population (2016)

b. Population (2018)





Sharp RD Along Population Cutoff

- We rely on pop-based assignment rule for rural schools with dist>30min

  $\Rightarrow$ Weighted average increase in wages of 11%

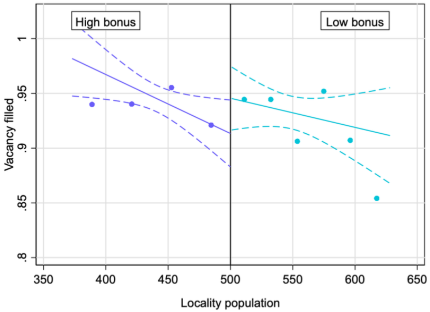
- Given continuity of potential outcomes around the cutoff

$$y_{ijt} = \gamma_0 + \gamma_1 \mathbf{1}(pop_{jt} < pop_c) + g(pop_{jt}, pop_c) + \delta_t + u_{ijt}$$

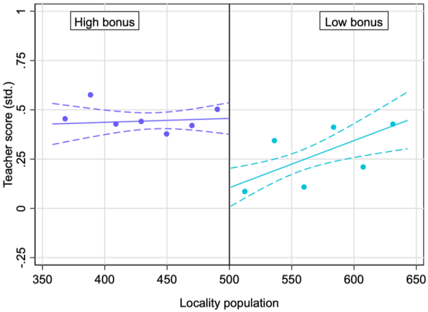
  - $g(\cdot)$: flexible polynomial on population of the locality of school j
  - δ_t: indicator for year of assignment
  - u_{jt}: error term, clustered at the school-year level

$\Rightarrow$ Estimate γ_1 non-parametrically within MSE-optimal bandwidths

Rural Bonus and Teacher Choices over Job Postings



a) **Vacancy Filled**



b) **Competency Score**

	(1) Vacancy filled	(2) Preferences	(3) Teacher Score (Std.)
High Bonus	-0.043	0.103	0.386
	(0.040)	(0.035)	(0.137)
Bandwidth	127.521	157.452	141.447
Schools	715	850	764
Observations	1851	2080	1870

Rural Bonus and Student Achievement

Panel A: Dependent Variable is Spanish Test (z-score)			
	(1)	(2)	(3)
	Vacancy	No vacancy	All
High Bonus	0.395	-0.004	0.232
	(0.152)	(0.127)	(0.088)
Bandwidth	107.818	148.920	105.822
Schools	264	451	832
Observations	4635	6773	16681
Panel B: Dependent Variable is Math Test (z-score)			
	(1)	(2)	(3)
	Vacancy	No vacancy	All
High Bonus	0.579	0.067	0.317
	(0.193)	(0.143)	(0.105)
Bandwidth	85.848	155.174	95.638
Schools	220	470	764
Observations	3939	7039	15363

Takeaways from RD Analysis

1. Rural bonus shifted supply of teachers towards disadvantaged areas
    - No effect on the probability of creating new matches
    - Inflow of more competent teachers (0.4σ)
2. Students perform better in schools that pay higher wages
    - Only in schools that attracted better quality teachers (0.4-0.6σ)
    - No effort response to higher wages for incumbent teachers
    - No effect on teachers' retention or composition of teaching staff

What is the Rationale of the Model?

- The RD evidence is limited on the local effect of the rural wage bonus
  - What is the overall effect of the policy on urban-rural gaps in edu. outcomes?
  - Can we characterize more effective teacher-school allocations?
  - Can we achieve those with alternative wage schedules?
- An empirical matching model of teachers and schools
  - A discrete choice framework with heterogenous preferences over wages and non-wage amenities
  - A potential outcome framework that maps teacher sorting across schools into the distribution of student achievement

Wages, Preferences, and Equilibrium

- Teacher i's utility from school j (off-platform $j = p$) + outside option $j = 0$:

$$U_{ijt} = \underbrace{w_{jt}}_{\text{wage}} + \underbrace{\alpha_i^{-1}(u(a_{jt}, x_{it}) + \epsilon_{ijt})}_{\text{non-pecuniary amenities}},$$

$$U_{ipt} = \alpha_i^{-1}(x'_{it}\beta_p + \epsilon_{ipt}),$$

$$U_{i0t} = \alpha_i^{-1}(\beta_i + \epsilon_{i0t}),$$

- Serial dictatorship $\Rightarrow$ discrete choice with observed choice sets

$$\mu^*_w(i,t) = \arg\max_{j \in \Omega(s_{it})} U_{ijt},$$

Teachers Value Added

- Student l's potential outcome when matched with teacher i:

$$Y_{lij} = + \underbrace{c'_{jt}\beta}_{\text{school/classroom effect}} + \underbrace{z'_{lt}\bar{\delta}}_{\text{student ability}} + \underbrace{z'_{lt}(\delta_i - \bar{\delta})}_{\text{teacher ATE + match effects}} + \nu_{lij}$$

- We allow teachers VA δ_i to correlate with their choices $\theta_i = (\log \alpha_i, \beta_i)$

$$(\theta_i, \delta_i)|x_{it} \sim \mathcal{N}\left[\begin{pmatrix} x'_{1it}\gamma^{\theta} \\ x'_{2it}\gamma^{\delta} \end{pmatrix}, \begin{pmatrix} \Sigma_{\theta,\theta} & \Sigma_{\theta,\delta} \\ \Sigma_{\delta,\theta} & \Sigma_{\delta,\delta} \end{pmatrix}\right]$$

$\Rightarrow$ Use teacher characteristics to inform the prior and reduce variance

$\Rightarrow$ Link teacher effectiveness with observed measures of teacher quality

Data and Identification

- We identify choice parameters using data on realized matches + choice sets
  - Repeated choice data help identify the distribution of random coefficients θ_i
  - Wages vary **only** with observables $\Rightarrow$ residual variation is RD effect
- We identify the achievement prod. function using teacher-classroom data
  - Estimate teacher effectiveness as fix effects δ_i
  - Use variation in observables x_{2it} to recover $(\gamma^{\theta}, \Sigma_{\delta,\delta})$
- We identify Cov(TVA, random coeffs) by linking assignments with teacher-classroom data
  - Conditional on knowing $\Sigma_{\delta,\delta}$ we can recover $\Sigma_{\theta,\delta}$

Estimation

- We flexibly parametrize the non-wage component of the choice model as:

$$u(a_{jt}, x_{it}, \boldsymbol{\theta}) = \underbrace{x'_{it}\Gamma_1 q_{jt}}_{\text{amenities}} + \underbrace{x'_{it}\Gamma_2 d_{ijt}}_{\text{moving costs}} + \underbrace{x'_{it}\Gamma_3 m_{ij}}_{\text{match effects}} + \underbrace{\kappa_j}_{\text{unobs. amenities}}$$

- Estimation in two steps (see Appendix D.2 for details)
  1. Estimate the parameters of the achievement prod. function
  2. Estimate $(\boldsymbol{\Gamma}, \gamma, \Sigma)$ by maximizing the log-likelihood function:

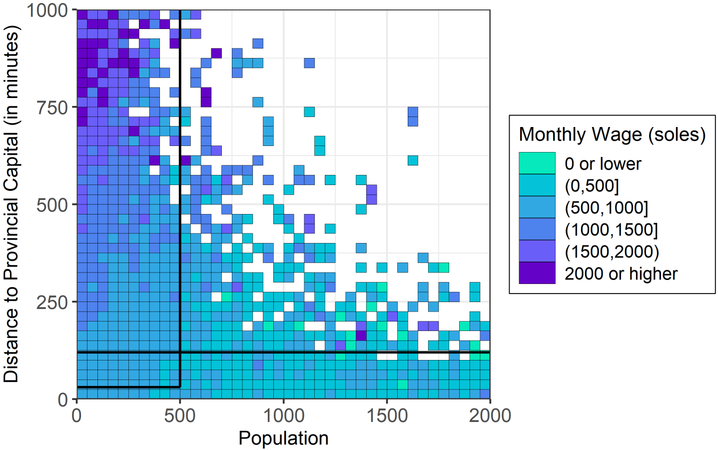
$$L(\boldsymbol{\Gamma}, \gamma, \Sigma) = \sum_{i=1}^{n_w} \sum_{t:\{\mu^*(i,t)\neq\emptyset\}} \log \mathbb{P}\left((\mu^*(i,t))_{t=1}^{T}, \hat{\delta}_i | x_i, \boldsymbol{w}, \boldsymbol{a}, \Omega(s_{it}) \right),$$

Monthly Willingness to Pay for Non-Wage Characteristics

	Mean		10% Quantile		90% Quantile	
	Soles	% Wage	Soles	% Wage	Soles	% Wage
	(1)	(2)	(3)	(4)	(5)	(6)
Amenities, Infrastructure and Remoteness						
Amenity/Infrastructures	200	10	30	2	440	22
Closer to Home by 1km						
$0 \leq$ Distance < 20	200	10	33	2	443	22
$20 \leq$ Distance < 100	113	6	23	1	243	12
Distance ≥ 100	20	1	3	0	43	2
Ethnolinguistic Proximity						
Same Language: Spanish	2,777	139	393	20	6,180	309
Same Language: Quechua	986	49	303	15	1,929	96
Same Language: Aymara	3,264	163	656	33	6,976	349
Teaching Conditions						
No Border	406	20	-97	-5	1,122	56
No Multigrade	962	48	147	7	2,121	106
No Single Teacher	1,758	88	120	6	4,123	206

- Non-wage attributes induce vertical+horizontal differentiation across schools

Rural vs. Urban Non-Pecuniary Utility Differences



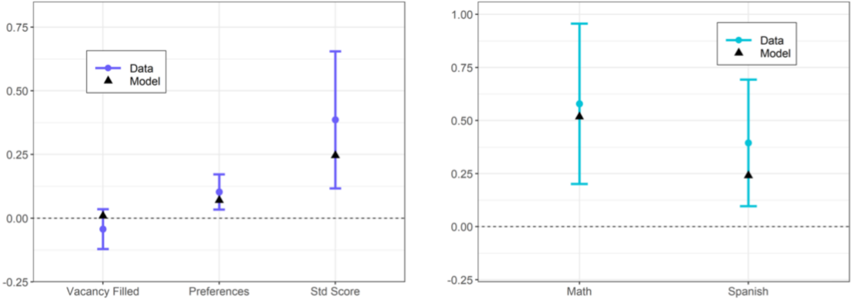
- Utility differences are merely compensated by the wage bonus policy

Standard Deviation of TVA Coefficients

	Math	Spanish
	(1)	(2)
ATE	0.465 (0.006)	0.408 (0.006)
Lagged Score	0.145 (0.005)	0.150 (0.005)
Lagged Score^2	0.049 (0.004)	0.061 (0.003)
Female	0.098 (0.010)	0.083 (0.013)
Quechua - Aymara	0.040 (0.030)	0.067 (0.019)
Age	0.115 (0.007)	0.110 (0.008)

- One SD increase in TVA $\Rightarrow \uparrow$ in test scores by 0.44-0.50 SD
- Significant match effects on lagged measures of student achievement
- 12-18% of variance in TVA explained by teachers' comparative advantage

Model Fit



a. Sorting **b. Value Added**

- Estimated model replicates the RD evidence induced by the rural wage bonus
- Good fit on moments away from the pop. threshold (urban-rural gaps, etc.)

Counterfactual 1: Aggregate Effects of the Rural Bonus

- Predict teachers' choices over schools with and without rural wage bonus
  - Simulate U_{ijt} from estimated parameters and a random draw of ϵ_{ijt} and θ_i
- Compute the stable matching eq. using the teacher-proposing DA algorithm
- Predict the distribution of teacher value-added without and with rural bonus
  - Use the mean of the posterior distribution of δ_i (see Appendix D.3)

Counterfactual 1: Aggregate Effects of the Rural Bonus

	Status Quo	No Rural Bonus	Policy Effect
	(1)	(2)	(3)
Panel A: Total Value Added			
Urban-Rural Gap	0.077	0.164	-0.087
Urban	0.024	0.059	-0.036
Rural	-0.053	-0.105	0.052
Moderately Rural	-0.033	-0.055	0.022
Rural	-0.111	-0.049	-0.063
Extremely Rural	0.067	-0.099	0.166
Panel B: Match Effects			
Urban	-0.007	0.002	-0.009
Rural	0.008	0.001	0.007

- Small average effects on TVA, mostly concentrated in very remote schools
- Rural bonus does not induce sorting based on comparative advantages

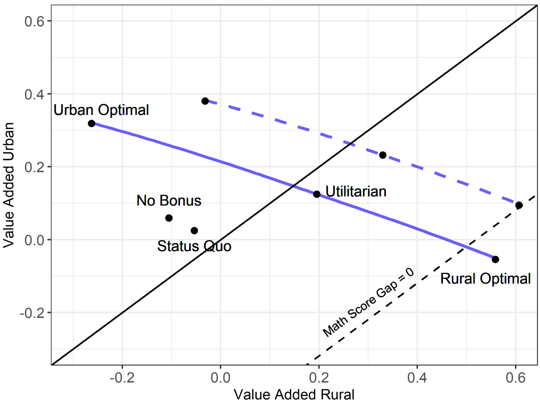
Counterfactual 2: Gains from Teachers' Reallocation

- We consider the following the linear program:

$$\max_{\mu} \sum_{i=1}^{n_w} \sum_{j=1}^{n_m} \pi_j \bar{z}_j' \hat{\delta}_i \mathbb{I}\{\mu(i) = j\}$$

- $\bar{z}_j' \hat{\delta}_i$ is the predicted (shrunken) average TVA for teacher i in school j
- Solution $\mu^\star(\pi_j)$ depends on weight put on students in school j

Counterfactual 2: Gains from Teachers' Reallocation



- Match effects loom large for efficiency (esp. drawing from outside option)
- No trade-off between equity and efficiency

Counterfactual 3: Optimal Wage Policy

- Policymaker can set priorities and wages in each school such that:

$$\min_{w} \sum_{j} w_j, \text{ s.t. } \begin{cases} \max_{i \in \mu(j)} z'_{lt}\delta_i \geq c_j, \ \forall j \in \mathcal{S} & \text{(C1)} \\ \mu \text{ is stable given } w \text{ and using } z'_{lt}\delta_i \text{ as priorities} & \text{(C2)} \end{cases}$$

- For a fixed wage, schools strictly rank teachers according to $z'_{lt}\delta_i$
- Otherwise, the allocation with the lower wage is always strictly preferred
- A stable set of contracts always exists in this counterfactual economy
  - Each school $j \in \mathcal{S}$ bids upward until (C1) is satisfied
  - Outcome is (μ, w) that satisfies (C1)-(C2) while minimizing total wage bill

Counterfactual 3: Optimal Wage Policy

	Status Quo	Teacher Value Added Threshold				
		$c = -0.4$	$c = -0.3$	$c = -0.2$	$c = -0.1$	$c = 0$
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Teacher Value Added						
Urban	0.055	0.036	0.035	0.019	-0.009	-0.058
Rural	-0.048	0.015	0.076	0.133	0.197	0.258
Moderately Rural	0.025	0.007	0.058	0.040	0.127	0.203
Rural	-0.154	-0.060	0.034	0.094	0.117	0.199
Extremely Rural	-0.022	0.080	0.131	0.225	0.296	0.357
Panel B: Match Effects						
Urban	0.019	0.017	0.018	0.018	0.013	0.022
Rural	0.040	0.063	0.111	0.137	0.180	0.191
Moderately Rural	0.008	0.002	0.031	0.022	0.065	0.089
Rural	0.039	0.085	0.141	0.107	0.154	0.161
Extremely Rural	0.070	0.106	0.168	0.218	0.247	0.300
Panel C: Monthly Total Cost (in Soles)						
% Base Wage	0.111	0.086	0.140	0.234	0.379	0.621
Mean Bonus per School	223	171	279	467	759	1,242
SD Bonus per School	220	407	576	839	1,184	1,698

- It's possible to close the urban-rural gap in TVA at a small cost
- Optimal policy induces teachers to sort on their comparative advantage

Wrapping up

- Policy reform largely increased teacher compensation in remote schools
- Attracted higher quality teachers and substantially improved student learning
- An estimated model show large gains from teacher reallocation
- Current policy falls short in closing urban-rural gaps in achievement
- Counterfactual policies leverage info on teachers prefs and effectiveness