rdlasso: Regression Discontinuity with High-Dimensional Data

Marianna Nitti, Marco Ventura
Sapienza University of Rome

17 September 2026
Italian Stata Conference

Outline

1. Motivations
2. Our Contribution
3. The econometric set up
4. The Stata syntax of `rdlasso`
5. Applications (Sharp & Fuzzy) of `rdlasso`
6. Conclusions

Motivations (1)

- Including many pre-determined covariates can improve the precision of treatment effect estimates in RDD (Calonico et al., 2014, 2019).
- However, in high-dimensional settings ($p > n$), **direct inclusion** of all covariates without selection can lead to:
  - Overfitting
  - Distorted inference

Motivations (2)

Post-Lasso RD

- Kreiss & Rothe (2023) propose a two-step approach:
  - **Step 1:** Local Lasso selection of covariates within the bandwidth
  - **Step 2:** Post-Lasso RD estimation using the selected covariates

- From a theoretical perspective, under an **approximate sparsity** condition (i.e., only a small subset of covariates is relevant near the cutoff), the estimator behaves like the standard local linear estimator in terms of bias and variance.

Our Contribution

- We develop `rdlasso`, a new Stata command for Regression Discontinuity Designs with high-dimensional covariates.
- **Main features:**
  - Implements the two-step procedure of Kreiss & Rothe (2023)
  - Supports both sharp and fuzzy RD designs
  - Uses Stata's lasso command to select covariates
  - Uses rdrobust for bandwidth selection, estimation, and inference
- **Advantages for users:**
  - Single fully Stata-native command
  - Automatic covariate and bandwidth selection
  - Different tuning-parameter selection methods
  - Standard rdrobust output and stored results

The econometric set up (1)

- **Sample:** (Y_i, X_i, W_i), with Y_i outcome, X_i running variable, W_i covariates.
- **Treatment:** $T_i = I(X_i \geq 0)$.
- **Observed outcome:** $Y_i = Y_i(T_i)$.
- **Parameter of interest (ATE at threshold):**

$$\tau_Y = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = 0]$$

- **Identification via continuity of conditional expectations:**

$$\tau_Y = \lim_{x \downarrow 0} \mathbb{E}[Y \mid X = x] - \lim_{x \uparrow 0} \mathbb{E}[Y \mid X = x]$$

The econometric set up (2)

Baseline local linear estimator

- Estimate the effect using local linear regressions on each side of the cutoff
- Kernel weights and bandwidth h restrict attention to observations close to the threshold

$$\hat{\tau}_{h,Base} = \mathrm{e}_2^\top \arg\min_\theta \sum_{i=1}^{n} K_h(X_i)\left[Y_i - V_i^\top \theta\right]^2$$

The econometric set up (3)

Covariate-adjusted estimator

$$\hat{\tau}_{h,CCFT} = e_2^\top \arg\min_{(\theta,\gamma)} \sum_{i=1}^{n} K_h(X_i)\left[Y_i - V_i^\top \theta - W_i^\top \gamma\right]^2$$

- Covariates may improve precision
- In high-dimensional settings (p large or $p > n$) the estimator may be undefined
- Risk of overfitting and poor finite-sample inference
- This motivates the two-step procedure of Kreiss & Rothe (2023)

The econometric set up (4)

Step 1 — Local Lasso selection (bandwidth b, penalty λ)

$$(\hat{\theta}_n, \hat{\gamma}_n) = \arg\min_{\theta,\gamma} \sum_{i=1}^{n} K_b(X_i)\left[Y_i - V_i^\top\theta - W_i^\top\gamma\right]^2 + \lambda \sum_{k=1}^{p_n} |\gamma_k|$$

- Select covariates predictive of the outcome within the bandwidth

Step 2 — Post-Lasso RD (bandwidth h)

$$\hat{\tau}_h(\hat{J}_n) = e_2^\top \arg\min_{\theta,\gamma} \sum_{i=1}^{n} K_h(X_i)\left[Y_i - V_i^\top\theta - W_i(\hat{J}_n)^\top\gamma\right]^2$$

- Asymptotically normal with standard RD inference

The econometric set up (5)

Fuzzy Regression Discontinuity

- Treatment not perfectly assigned $\rightarrow$ probability jump at the cutoff
- Local Average Treatment Effect (LATE) identified as the ratio of the discontinuities in the outcome and treatment variables:
$$\tau_{\text{fuzzy}} = \frac{\tau_Y}{\tau_T}$$
- **Implementation:** run the same two-step procedure twice (once for Y_i, once for T_i); standard errors via delta method

The Stata syntax of `rdlasso` (1)

Syntax

```
rdlasso varlist [if] [in] [, t(varname)
z(varname) c(real) fuzzy(string) level(real)
b(real) bfactor(real) h(real) kernel(string)
tpc(string)]
```

Where:

- `varlist(min=3)` must include, in order:
  - outcome variable (Y)
  - running variable (X)
  - one or more covariates (W)

The Stata syntax of `rdlasso` (2)

Options

- `t`(varname) treatment variable (required for fuzzy RD)
- `z`(varname) instrumental variable (required for fuzzy RD)
- `c`(real) cutoff value of the running variable (default: 0)
- `fuzzy`(string) activates fuzzy RD design logic (default: off)
- `level`(real) confidence level for inference (default: 95)
- `b`(real) bandwidth for model selection (Step 1); if omitted, selected via `rdrobust`
- `bfactor`(real) scaling factor applied to *b* (default: 1)
- `h`(real) bandwidth for final RD estimation (Step 2); if omitted, selected via `rdrobust`
- `kernel`(string) kernel function for both steps: triangular (default), epanechnikov, uniform
- `tpc`(string) tuning-parameter selection criterion: plugin(default), cv, adaptive, bic

Example 1: Sharp RD Design (Turkish elections)

Research question: Does Islamic political control affect women's empowerment?[1]

- **Units:** municipalities; mayoral elections (1994)
- **Running variable (X):** vote-share margin between the largest Islamic and secular parties
- **Cutoff:** 0 (Islamic mayor if $X_i \geq 0$, secular otherwise)
- **Treatment (T_i):** indicator of Islamic electoral victory, $T_i = \mathbf{1}(X_i \geq 0)$
- **Outcome (Y):** share of women aged 15–20 who completed high school (2000 census)
- **Covariates (W_i):** baseline municipal covariates + polynomial/interactions (169 candidate covariates)

[1] Meyersson (2014), *Econometrica*.

Example 1: Selected covariates (Local Lasso)

```
.  rdlasso Y X ageshr19 - prov_num_Zonguldak
```

```
 Selected covariates (Y): ageshr19 ageshr60_merkezi hischshr1520m_2 hischshr1520m_lpop1994 his
> chshr1520m_partycount hischshr1520m_sexr prov_num_Konya
--------------------------------------------------------------
 Preliminary Bandwidth (b): 17.23946666597075
 Kernel: triangular
 Cutoff: 0
 Tuning parameter criterion: plugin
```

Example 1: Covariate-adjusted RD estimates

```
Covariate-adjusted sharp RD estimates using local polynomial regression.

     Cutoff c = 0 | Left of c  Right of c            Number of obs =       2629
------------------+-----------------------           BW type       =     Manual
    Number of obs |      2314         315            Kernel        = Triangular
Eff. Number of obs|       349         210            VCE method    =         NN
    Order est. (p)|         1           1
    Order bias (q)|         2           2
       BW est. (h)|    11.285      11.285
       BW bias (b)|    11.285      11.285
         rho (h/b)|     1.000       1.000

Outcome: Y. Running variable: X.
--------------------------------------------------------------------------------
           Method |   Coef.    Std. Err.     z      P>|z|    [95% Conf. Interval]
------------------+-------------------------------------------------------------
     Conventional |  1.9768     .81575    2.4233   0.015    .377934     3.57563
           Robust |     -          -      2.0655   0.039    .127027     4.84401
--------------------------------------------------------------------------------
Covariate-adjusted estimates. Additional covariates included: 7
```

Example 2: Fuzzy RD Design (Colombia – SPP scholarship)

Research question: Does eligibility for the Ser Pilo Paga scholarship increase enrollment in high-quality higher education?[2]

- **Units:** students applying to higher education
- **Running variable (X):** SISBEN wealth index distance from the eligibility cutoff
- **Cutoff:** 0 (eligible if $X_i \geq 0$)
- **Instrument (Z_i):** eligibility indicator
- **Treatment (T_i):** receipt of the scholarship
- **Outcome (Y_i):** enrollment in a high-quality higher education institution
- **Covariates (W_i):** baseline student covariates + polynomial/interactions

[2] Londoño-Vélez et al. (2020), *American Economic Journal: Economic Policy*.

Example 2: Selected covariates (Local Lasso)

```
.  rdlasso Y X icfes_female - icfes_famsize_4,
t(T) z(Z) fuzzy(on)
```

```
Selected covariates (Y): icfes_age
Selected covariates (T): icfes_age
----------------------------------------------------------------
Preliminary Bandwidth (b): 9.027886111773137
Kernel: triangular
Cutoff: 0
Tuning parameter criterion: plugin
----------------------------------------------------------------
```

Example 2: Fuzzy RD estimate

```
-----------------------------------------------------------------
 FUZZY ESTIMATOR
-----------------------------------------------------------------
 RD Effect:                      0.4398
 Robust standard error:          0.0552
 z-statistic:                    7.9665
 p-value:                        0.0000
 95% Confidence Interval:  [   0.3316 ,    0.5480]
```

Summary

- Including covariates in RDD can improve precision, but many controls relative to the sample size may lead to overfitting and unreliable inference
- Local Lasso selection (Kreiss & Rothe, 2023) makes covariate adjustment feasible in high-dimensional settings under approximate sparsity
- `rdlasso` implements this methodology directly in Stata, using `lasso` for covariate selection and `rdrobust` for estimation and inference
- A single command automatically handles both sharp and fuzzy designs

Thank you for your attention!

Marianna Nitti — Marco Ventura
Sapienza University of Rome

marianna.nitti@uniroma1.it
marco.ventura@uniroma1.it

Questions and comments are very welcome.