Aula Prática - Regressão Descontínua

Introdução

A **Regressão Descontínua** (*Regression Discontinuity Design*, RDD) é um método de identificação causal amplamente utilizado em economia e ciências sociais. Sua ideia central é: quando a elegibilidade para um tratamento é determinada por uma variável contínua — chamada de variável de atribuição — ao cruzar um determinado **limiar** (*cutoff*) c, as unidades próximas a esse limiar são comparáveis entre si. Qualquer salto na variável de interesse ao redor do cutoff pode então ser atribuído causalmente ao tratamento.

Nesta aula prática estudamos os dois exemplos da literatura para exemplificar o uso do RDD:

- **RDD Sharp**: Meyersson (2014) analisa se a eleição de prefeitos islâmicos na Turquia aumentou o acesso de mulheres jovens ao ensino médio (dados municipais, eleições de 1994).
- **RDD Fuzzy**: Londoño-Vélez et al. (2020) estudam o efeito do Programa Ser Pilo Paga (SPP) sobre a matrícula no ensino superior na Colômbia.

Esta aula está baseada em Cattaneo, Idrobo e Titiunik (2019) e Cattaneo, Idrobo e Titiunik (2024).

Pacotes

Parte I — RD Sharp

Dados

Os dados contêm informações sobre municípios turcos onde candidatos islâmicos e laicos disputaram as prefeituras nas eleições de 1994. A variável de atribuição é a **margem de votos do partido islâmico** (positiva = vitória islâmica, negativa = derrota), com cutoff em zero.

```
 [1] "X"            "Y"            "T"
"ageshr19"
  [5]  "ageshr60"              "buyuk"
"hischshr1520m"  "i89"
  [9]  "lpop1994"            "merkezi"
"merkezp"        "partycount"
[13]  "prov"               "prov_num"
"sexr"           "shhs"
[17] "subbuyuk"       "vshr_islam1994"
```

A principal variável de resultado de interesse será o percentual de mulhers entre 15 e 20 anos de idade com ensino médio completo em 2000. Além da variável de interesse e da variável de atribuição, a base de dados possui também um conjunto de variáveis pré-determinadas.

```
# A tibble: 3 × 2
  Grupo                              N
  <chr>                          <int>
1 Total                           2629
2 Com prefeito islâmico (X ≥ 0)    315
3 Com prefeito laico (X < 0)      2314
```

Variável	Tipo	Descrição
Y	Resultado	Proporção de mulheres de 15–20 anos com ensino médio completo (2000)
X	Score	Margem de votos do partido islâmico nas eleições de 1994 (cutoff = 0)

Variável	Tipo	Descrição
T	Tratamento	Prefeito islâmico eleito: 1(X ≥ 0)
lpop1994	Covariada	Logaritmo da população municipal em 1994
hischshr1520m	Covariada	Proporção de homens de 15–20 anos com ensino médio (1994)
sexr	Covariada	Razão de sexo da população (mulheres/homens × 100)

Gráfico RD

A função `rdplot()` do pacote `rdrobust` produz o gráfico de RDD: médias dos dados agrupados em bins (“intervalos”) e um polinômio global ajustado separadamente em cada lado do cutoff. O número de bins pode ser fixado manualmente ou selecionado automaticamente pelo pacote.

Dois critérios automáticos comuns são:

- **ES** (*Evenly Spaced*): bins do mesmo tamanho; o número ótimo minimiza o IMSE (erro quadrático médio integrado).
- **MV** (*Mimicking Variance*): bins projetados para imitar a variância dos dados.

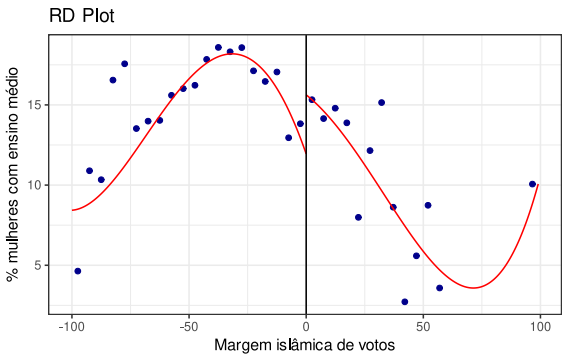


Figura 1: Gráfico RD com 20 bins de cada lado (bins ES fixos)

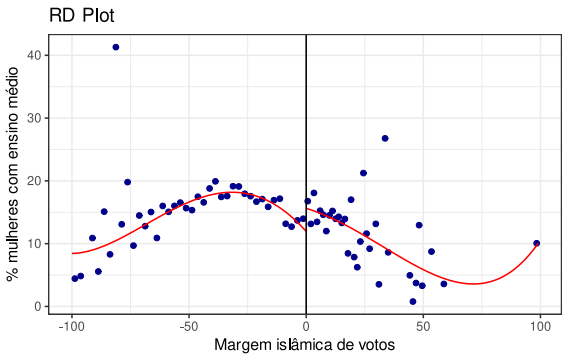


Figura 2: Gráfico RD com seleção ótima de bins (ESMV)

Conforme se observa nos gráficos, independente do método de seleção dos bins, observa-se um salto positivo (descontinuidade) na proporção de mulheres com ensino médio completo em 2000 nas cidades em que o partido islâmico venceu (margem > 0) as eleições para prefeitura de 1994.

Estimação Local

O mais comum é se estimar um modelo RDD usando regressão local: ajusta-se um modelo linear em cada lado do cutoff, usando apenas observações dentro de uma janela de largura h (bandwidth). Os interceptos no cutoff estimam os limites laterais de $\mathbb{E}[Y \mid X]$.

Usando $h = 20$ pontos percentuais, podemos estimar um modelo linear de cada lado do cutoff manualmente:

```
# A tibble: 3 × 2
  Estimativa                Valor
  <chr>                     <dbl>
```

```
1 Limite à esquerda do cutoff 12.6
2 Limite à direita do cutoff  15.5
3 Efeito RD                    2.93
```

O mesmo resultado é obtido ao se utilizar a função `rdrobust()` fixando o kernel uniforme e $h = 20$:

```
Sharp   RD   estimates   using   local
polynomial regression.

Number of Obs.                 2629
BW type                      Manual
Kernel                      Uniform
VCE method                       NN

Number of Obs.                  2314
315
Eff. Number of Obs.              608
280
Order est. (p)             1        1
Order bias  (q)            2        2
BW est. (h)                    20.000
20.000
BW bias (b)                    20.000
20.000
rho (h/b)                       1.000
1.000
Unique Obs.                     2314
315

===============================================================================
              Point   Robust Inference
                  Estimate          z
P>|z|      [ 95% C.I. ]
-------------------------------------------------------------------------------
      RD  Effect      2.927      1.636
0.102    [-0.582 , 6.471]
===============================================================================
```

Seleção de Bandwidth e Estimação Robusta

Na prática, a bandwidth h não é escolhida manualmente. O critério **MSE-ótimo** (`bwselect = "mserd"`) minimiza o erro quadrático médio da estimativa de RDD:

```
Sharp RD estimates using local
polynomial regression.

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                  2314
315
Eff. Number of Obs.              529
266
Order est. (p)                   1        1
Order bias  (q)                  2        2
BW est. (h)                          17.240
17.240
BW bias (b)                          28.576
28.576
rho (h/b)                             0.603
0.603
Unique Obs.                      2311
315

=============================================================================
              Point   Robust Inference
                  Estimate          z
P>|z|      [ 95% C.I. ]
-----------------------------------------------------------------------------
      RD Effect      3.020      1.776
0.076    [-0.309 , 6.276]
=============================================================================
```

A linha **Robust** usa o *bias--corrected* confidence interval (Calonico, Cattaneo e Titiunik, 2014), mais confiável em amostras finitas. O kernel triangular dá maior peso às observações mais próximas do cutoff.

Leitura do output:

- `BW est. (h)`: bandwidth de estimação (observações usadas para calcular $\hat{\tau}$)
- `BW bias (b)`: bandwidth de estimação do viés (para a correção)
- `Eff. N`: número efetivo de observações dentro da bandwidth

Testes de Validação

A validade do design RDD requer três pressupostos principais que podem ser avaliados empiricamente.

Teste de Densidade (McCrary)

Se agentes puderem **manipular** o score (e.g., partidos gerenciando contagem de votos), haveria excesso de observações logo acima de zero. O teste `rddensity()` verifica continuidade da densidade de X no em torno do cutoff:

```
Manipulation testing using local
polynomial density estimation.

Number of obs =       2629
Model =               unrestricted
Kernel =              triangular
BW method =           estimated
VCE method =          jackknife

c = 0                     Left of c
Right of c
Number of obs                 2314
315
Eff. Number of obs             965
301
Order est. (p)                   2
2
Order bias (q)                   3
3
BW est. (h)                 30.539
28.287

Method             T               P
> |T|
Robust                       -1.3937
0.1634


P-values of binomial tests (H0: p=0.5).

Window Length / 2          <c     >=c
P>|T|
0.874                               20
26    0.4614
1.748                               42
49    0.5296
2.622                               70
63    0.6030
3.496                               95
```

```
81    0.3271
4.370                                131
98    0.0342
5.245                                155
112    0.0100
6.119                                183
131    0.0039
6.993                                209
148    0.0015
7.867                                229
160    0.0005
8.741                                257
173    0.0001
```

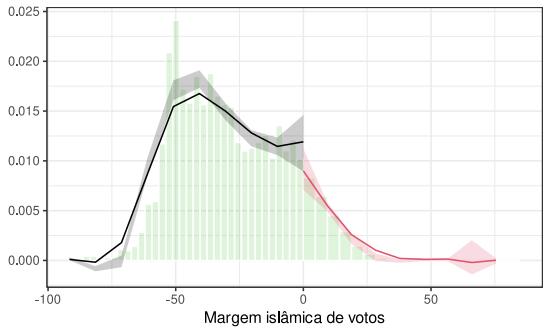


Figura 3: Teste de densidade — H0: sem manipulação da variável de score

```
$Estl
Call: lpdensity

Sample
size                              2314
Polynomial order for point estimation
(p=)    2
Order    of    derivative    estimated
(v=)    1
Polynomial   order   for   confidence
interval (q=)    3
Kernel                         function
triangular
Scaling                          factor
0.880517503805175
Bandwidth                        method
user provided

Use summary(...) to show estimates.
```

```
$Estr
Call: lpdensity

Sample
size                                315
Polynomial order for point estimation
(p=)    2
Order    of    derivative    estimated
(v=)    1
Polynomial   order   for   confidence
interval (q=)    3
Kernel                          function
triangular
Scaling                           factor
0.11986301369863
Bandwidth                         method
user provided

Use summary(...) to show estimates.

$Estplot
```

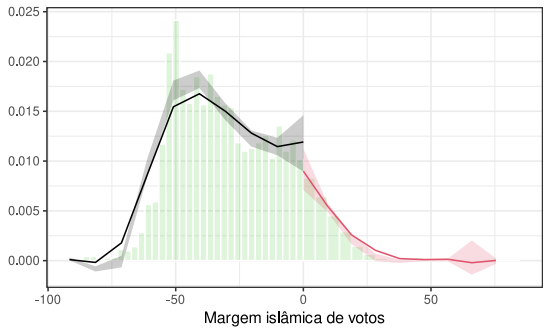


Figura 4: Teste de densidade — H0: sem manipulação da variável de score

Falha em rejeitar H_0 é evidência de ausência de manipulação.

Covariadas Predeterminadas

Variáveis determinadas **antes** do tratamento não deveriam mostrar descontinuidade no cutoff. Rodamos `rdrobust()` para as três covariadas disponíveis e reportamos os resultados na tabela abaixo:

Teste de continuidade — covariadas predeterminadas

Variável	Coef.	Std. Err.	p-valor	IC 95% inf.	IC 95% sup.
Log. população (1994)	0.0604	0.3288	0.9991	-0.6441	0.6448
Prop. homens 15–20 c/ ensino médio (1994)	1.5526	1.6887	0.3579	-1.7572	4.8624
Razão de sexo (mulheres/homens × 100)	2.2216	2.1866	0.3096	-2.0640	6.5072

Como podemos verificar, não há nenhuma covariada pré-determinada com efeito significativo, o que consiste em evidência positiva para validade do desenho de pesquisa.

Cutoff Placebo

Cutoffs "falsos" não deveriam produzir descontinuidades, afinal, os tratamentos não são determinados por valores em cutoffs falsos. Testamos o cutoff verdadeiro ($c = 0$) e quatro cutoffs placebo — dois à esquerda ($c \in \{-10, -5\}$, usando apenas municípios com $X < 0$) e dois à direita ($c \in \{5, 10\}$, usando apenas municípios com $X > 0$). Somente o cutoff verdadeiro deveria apresentar efeito significante, que é justamente o que o gráfico abaixo mostra (ou seja, um efeito marginalmente significativo no cuttof zero).

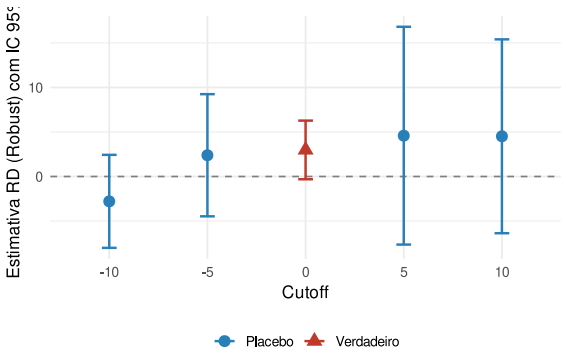


Figura 5: Cutoffs placebo — estimativa RD Robust com IC 95%

Parte II — RD Fuzzy

Dados

O Programa Ser Pilo Paga (SPP) subsidia a matrícula no ensino superior de estudantes de baixa renda na Colômbia. A elegibilidade é determinada pelo índice SISBEN: alunos com índice abaixo de um cutoff eram elegíveis. A variável `X1` é o índice SISBEN normalizado no cutoff ($X1 > 0$ = elegível, $X1 < 0$ = não elegível).

O design é **fuzzy com não adesão unilateral**: nenhum aluno não elegível ($X1 < 0$) recebe o SPP, mas apenas ~62% dos elegíveis ($X1 > 0$) efetivamente o recebem.

```
[1] "X1"           "T"            "D"
"Y"
[5]   "icfes_female"      "icfes_age"
"icfes_urm"     "icfes_stratum"
[9] "icfes_famsize"
```

A principal variável de resultado de interesse será a dummy que indica se o aluno se matriculou ou não no ensino superior. Além da variável de interesse e da variável de atribuição, a base de dados possui também outras variáveis de controle.

```
# A tibble: 4 × 2
  Estatística                  Valor
  <chr>                        <dbl>
1          Observações         totais
23132
```

```
2     Elegíveis     (X1     >     0)
15423
3  Taxa  de  recebimento  –  elegíveis
0.594
4 Taxa de recebimento – não elegíveis
0
```

Variável	Tipo	Descrição
Y	Resultado	Matrícula em instituição de ensino superior (0/1)
X1	Score	Índice SISBEN normalizado no cutoff (X1 > 0 = elegível para SPP)
T	Elegibilidade	Elegível para o SPP: 1(X1 ≥ 0)
D	Tratamento recebido	Recebeu o subsídio SPP efetivamente (0/1)
icfes_female	Covariada	Sexo feminino (0/1)
icfes_age	Covariada	Idade do candidato
icfes_urm	Covariada	Pertence a minoria racial ou étnica (0/1)
icfes_stratum	Covariada	Estrato socioeconômico

Variável	Tipo	Descrição
		(1 = mais pobre, 6 = mais rico)
icfes_famsize	Covariada	Tamanho da família

Motivação e Parâmetros

No design **fuzzy**, a regra de elegibilidade não determina perfeitamente o tratamento recebido. Distinguimos:

- $T_i = \mathbf{1}(X_{1i} \geq 0)$: **elegibilidade** (tratamento atribuído)
- D_i: **recebimento efetivo** do SPP (tratamento recebido)

Os três parâmetros de interesse são:

$$\tau_{FS} = \lim_{x \to 0^+} \mathbb{E}[D_i \mid X_{1i} = x] - \lim_{x \to 0^-} \mathbb{E}[D_i \mid X_{1i} = x] \quad \text{(Primeiro Estágio)}$$

$$\tau_{ITT} = \lim_{x \to 0^+} \mathbb{E}[Y_i \mid X_{1i} = x] - \lim_{x \to 0^-} \mathbb{E}[Y_i \mid X_{1i} = x] \quad \text{(Intenção de Tratar)}$$

$$\tau_{FRD} = \frac{\tau_{ITT}}{\tau_{FS}} \quad \text{(Efeito RD Fuzzy = LATE)}$$

τ_{FRD} é o **efeito médio local do tratamento** (LATE): o efeito do SPP para os **compliers** — aqueles que recebem o SPP quando elegíveis, mas não o receberiam se não fossem elegíveis.

Gráficos RD

```
[1] "Mass points detected in the running variable."
```

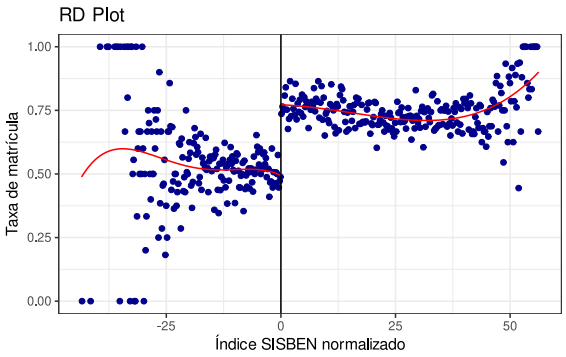


Figura 6: Gráfico RD — Taxa de matrícula no ensino superior (resultado Y)

```
[1] "Mass points detected in the
running variable."
[1] "Warning: not enough variability
in the outcome variable below the
threshold"
```

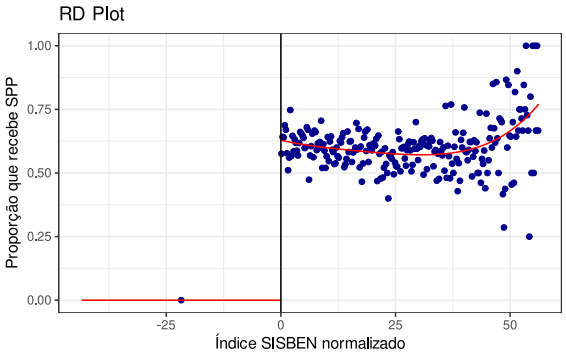


Figura 7: Gráfico RD — Recebimento do SPP (primeiro estágio D)

O gráfico do primeiro estágio confirma o não-cumprimento unilateral: nenhum aluno com $X1 < 0$ recebe o SPP, e há um salto acentuado em $X1 = 0$ para ~62%.

Primeiro Estágio e Efeito ITT

A seguir mostramos as estimativas do primeiro estágio e do ITT separadamente para entender/visualizar cada componente:

```
Sharp RD estimates using local
polynomial regression.

Number of Obs.                23132
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 7709
15423
Eff. Number of Obs.            6594
7457
Order est. (p)                    1
1
Order bias (q)                    2
2
BW est. (h)                  18.496
18.496
BW bias (b)                  28.982
28.982
```

```
rho (h/b)                         0.638
0.638
Unique Obs.                        3644
9274

=============================================================================
              Point   Robust Inference
                  Estimate          z
P>|z|      [ 95% C.I. ]
-----------------------------------------------------------------------------
      RD Effect      0.625     43.136
0.000     [0.595 , 0.652]
=============================================================================
```

```
Sharp   RD   estimates   using   local
polynomial regression.

Number of Obs.                23132
BW type                       mserd
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 7709
15423
Eff. Number of Obs.            3877
3908
Order est. (p)              1      1
Order bias  (q)             2      2
BW est. (h)                    9.041
9.041
BW bias (b)                   14.402
14.402
rho (h/b)                      0.628
0.628
Unique Obs.                    3644
9274

=============================================================================
              Point   Robust Inference
                  Estimate          z
P>|z|      [ 95% C.I. ]
-----------------------------------------------------------------------------
      RD Effect      0.269     10.051
0.000     [0.221 , 0.328]
=============================================================================
```

O **primeiro estágio** $\hat{\tau}_{FS} \approx 0.62$ indica que a elegibilidade aumenta a probabilidade de receber o SPP em 62 pp. O **ITT** $\hat{\tau}_{ITT} \approx 0.27$ indica o efeito da elegibilidade (independentemente de receber) sobre a matrícula.

Estimativa RD Fuzzy

Para obter o efeito causal para os **compliers** (τ_{FRD}), basta adicionar `fuzzy = D` ao `rdrobust()`. Internamente, o estimador é análogo a um MQ2E local:

```
Fuzzy RD estimates using local
polynomial regression.

Number of Obs.                23132
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 7709
15423
Eff. Number of Obs.            3877
3908
Order est. (p)                    1        1
Order bias  (q)                   2        2
BW est. (h)                   9.041
9.041
BW bias (b)                  14.402
14.402
rho (h/b)                     0.628
0.628
Unique Obs.                    3644
9274

First-stage estimates.

=============================================================================
              Point   Robust Inference
                  Estimate          z
P>|z|      [ 95% C.I. ]
=============================================================================
      Rd Effect      0.619     29.917
0.000     [0.575 , 0.656]
=============================================================================

Treatment effect estimates.
```

```
=============================================================================
             Point   Robust Inference
                 Estimate          z
P>|z|      [ 95% C.I. ]
-----------------------------------------------------------------------------
      RD Effect      0.435     11.023
0.000     [0.366 , 0.524]
=============================================================================
```

O pacote reporta os resultaods em duas partes:

1. **Primeiro Estágio**: estimativa de $\hat{\tau}_{FS}$ (salto em D)
2. **Estimativa RD Fuzzy**: $\hat{\tau}_{FRD} = \hat{\tau}_{ITT}/\hat{\tau}_{FS}$, o efeito causal do SPP sobre a matrícula para os cumprintes

Como $\hat{\tau}_{FRD} > \hat{\tau}_{ITT}$, o efeito para quem efetivamente recebe o programa é maior do que o efeito médio de ser elegível.

Testes de Validação

Os mesmos testes de validação que valem no RDD Sharp são aplicáveis ao RDD Fuzzy. Além disso, é importante lembrar que o RDD Fuzzy é, em última instância, um estimar de variáveis instrumentais. Assim, é necessário que o primeiro estágio seja forte para obtenção de estimativas causais. Do contrário, teremos problemas de instrumentos fracos.

Atividade

1. RD Sharp — Interpretação

Usando o output de `summary(rdr_sharp)`, interprete:

a. Qual é o efeito estimado da eleição do partido islâmico sobre a proporção de mulheres com ensino médio? Use a linha **Robust**.
b. O resultado é estatisticamente significante ao nível de 5%?
c. Qual é a bandwidth MSE-ótima selecionada? Quantas observações são utilizadas efetivamente?

2. Gráfico RD

Compare os dois gráficos gerados com `rdplot()`:

a. Por que o número de bins importa para a visualização do efeito?
b. O gráfico com 20 bins de cada lado ou o ESMV comunica melhor a descontinuidade? Justifique.

3. Validação do Design Sharp

Examine os três testes de validação (densidade, covariada, cutoff placebo):

a. Os resultados sustentam a validade do design RD neste contexto? Explique para cada teste.
b. O que concluiríamos se o teste de densidade rejeitasse H_0?

4. RD Fuzzy — Interpretação

Com base nas estimativas da Parte II:

a. Por que $\hat{\tau}_{FRD} > \hat{\tau}_{ITT}$? Dê a intuição econômica.
b. Para qual subpopulação específica vale a estimativa $\hat{\tau}_{FRD}$?
c. O que significa "não-cumprimento unilateral" neste contexto? Como isso aparece nos dados?

5. Extensão — Sharp

Usando os dados de Meyersson, escolha outra variável disponível em `dados_sharp` (use `names(dados_sharp)` para ver as opções) e realize um **teste de covariada predeterminada**. Interprete o resultado: a variável escolhida é contínua no cutoff?

6. Extensão — Fuzzy

Adapte os códigos sobre testes de validação do RDD Sharp para o exemplo discutido na seção de RDD Fuzzy.