Regressão Discontínua

Raphael Corbi

Universidade de São Paulo

Maio 2015

Inferência Causal

- grande objetivo da econometria aplicada

1. **Dados experimentais (padrão ouro)**
   1.1 mais confiável, mais caro, mais difícil de implementar
2. **Dados não-experimentais I**
   2.1 problema da seleção
   2.2 seleção em observáveis (matching, DID, controle sintético)
3. **Dados não-experimentais II (quasi-experimentais)**
   3.1 seleção em não-observáveis (IV, DID, RDD)

Problema de Seleção

- Pergunta: **Pronto-Socorro torna pacientes mais saudáveis?**
- hospitais oferecem tratamento, mas também contato com doentes
- dados da National Health Interview Survey
- 2 perguntas: foi ao pronto-socorro? condição de saúde 1 a 5?

	Sample size	Mean Health Status	Std Error
No Hospital	90,049	3.93	0,003
Hospital	7,774	3.21	0.014
Difference	—	0.72	0.000

Problema de Seleção - formalizando...

- dummy de tratamento $D_i = 0, 1$
- Resultado Potencial $Y_0 i, Y_1 i$

$$Y_i = \begin{cases} Y_{1,i} & if \ D_i = 1 \\ Y_{0,i} & if \ D_i = 0 \end{cases}$$

$$Y_i = Y_{0,i} + (Y_{1,i} - Y_{0,i}) D_i$$

- porém não observamos ambos estados para o mesmo indivíduo

Problema de Seleção - formalizando...

$$\underbrace{E[Y_i|D_i=1] - E[Y_i|D_i=0]}_{\text{diferença observada em saúde média}} = \underbrace{E[Y_{1,i}|D_i=1] - E[Y_{0,i}|D_i=1]}_{\text{efeito tratamento médio nos tratados}}$$

$$+ \underbrace{E[Y_{0,i}|D_i=1] - E[Y_{0,i}|D_i=0]}_{\text{viés de seleção}}$$

- viés de seleção mascara o efeito causal do tratamento
- grande objetivo da pesquisa empírica é superar este viés

Inferência Causal

- como superar então o viés?

1. **Dados experimentais (padrão ouro)**
   - 1.1 não há problema de seleção
   - 1.2 tratamento aleatório $= \{Y_{0,i}, Y_{1,i}\} \perp\!\!\!\perp D_{0,i}$
2. **Dados não-experimentais I**
   - 2.1 Hipótese da Independência Condicional (HIC)
   - 2.2 HIC $= \{Y_{0,i}, Y_{1,i}\} \perp\!\!\!\perp D_{0,i}|X_i$
3. **Dados não-experimentais II (quasi-experimentais)**
   - 3.1 hipótese IV: parte exógena da variação de X identificada por Z
   - 3.2 hipótese RDD: tratamento é 'localmente' aleatório

Regressão descontínua desde 1960

- Thistlethwaite Campbell (1960)
  - prêmios excelência acadêmica em publicações

- set-up não experimental
- regra de alocação descontínua, baseada em uma variável contínua
- forcing (running) variable
- abaixo do cutoff = bom controle para acima do cutoff
- método esquecido até fim dos 1990's

Regressão descontínua nos 2000's

- primeiras aplicações (educação):
  - financial aid (van der Klaauw (2002))
  - class size (Angrist Lavy (1999))
  - school districts (Black (1999))
- outros tópicos:
  - efeito de seguridade social sobre a oferta de trabalho
  - Medicaid sobre saúde
  - reforço e desempenho escolar
  - economia política (median voter models)
  - sindicalização sobre salário e emprego

Vantagens da Regressão descontínua

- RD requer hipótese relativamente fracas para identificação
  - Hahn, Todd, and van der Klaauw (2001)

- inferência causal é potencialmente mais crível que DID e IV
  - Lee (2008)
  - RDD não assume que o tratamento é "aleatório"
  - tal variação aleatória é consequência da incapacidade dos agentes de controlar com precisão a *forcing variable* próximo ao cutoff
  - logo, não é só mais uma estratégia de identificação

Objetivo desta aula

- sumário do conhecimento atual de RDD
  - quando (não) vale, pontos fortes e fracos
- guia prático de implementação para um paper aplicado
  - técnica ainda não faz parte da maioria dos livros-texto
- discussão de questões de *identificação, interpretação e estimação*

Principais pontos da aula:

1. H1: indivíduos não controlam precisamente X (running variável)
   1.1 mecanismo gera incentivos para indivíduos tentarem controlar X

2. como consequência, a variação ao redor da discontinuidade é aleatória (come se fosse gerada por experimento aleatório)
   2.1 indivíduos têm a mesma probabilidade de ter um X t.q. T=0 ou 1

3. diferença crucial em relação ao IV:
   3.1 IV: instrumento é aleatório por hipótese (difícil de justificar)
   3.2 RDD: aleatoriedade como consequência do controle impreciso

Principais pontos da aula:

4. RDD podem ser analisados e testados como experimentos aleatórios
   - 4.1 teste das co-variadas acima/abaixo do cutoff
   - 4.2 IV: temos que assumir a exogeneidade condicional

5. exposição gráfica do efeito causal (transparência)

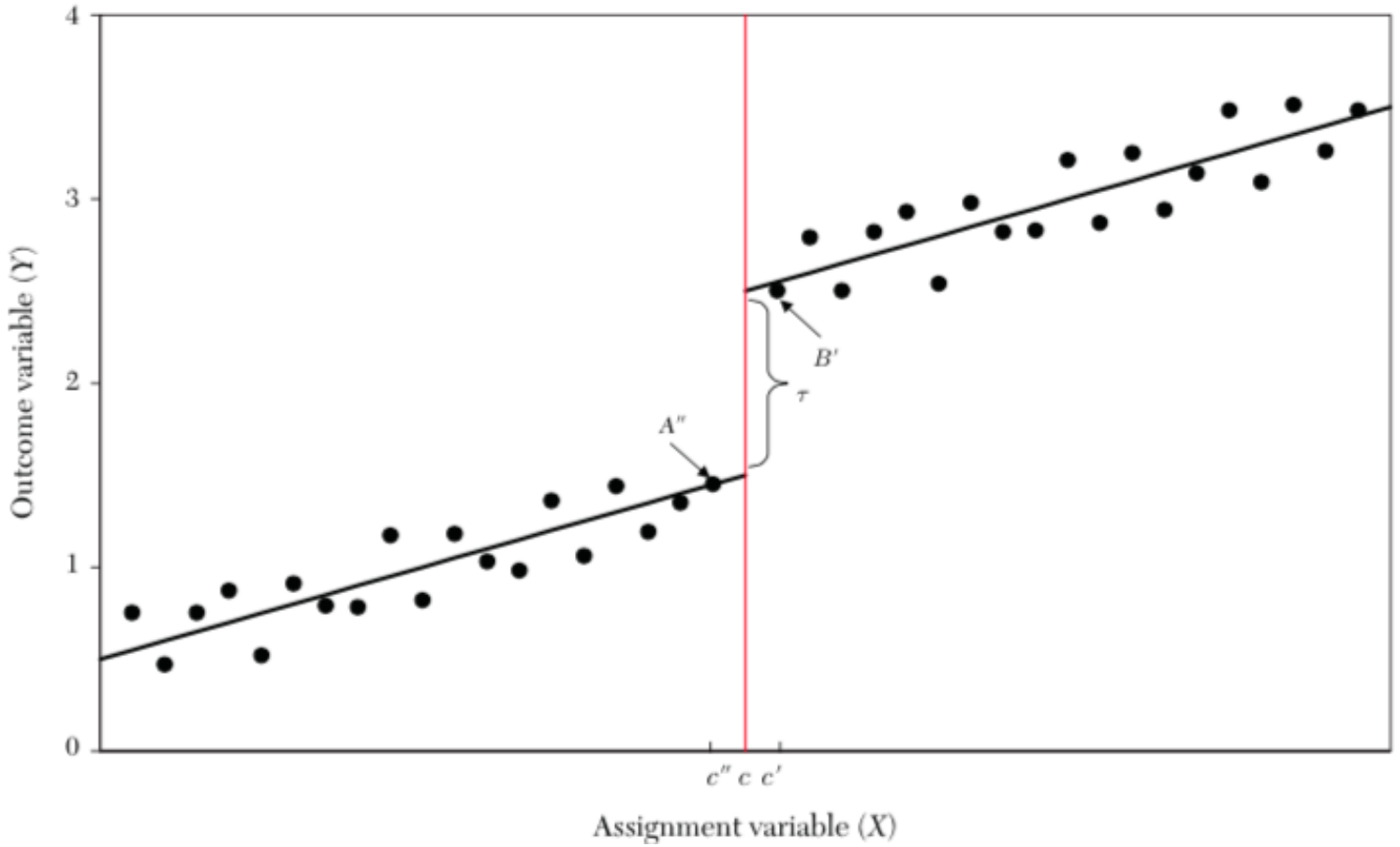
6. forma funcional paramétrica e não-paramétrica (complementos)

RDD set-up

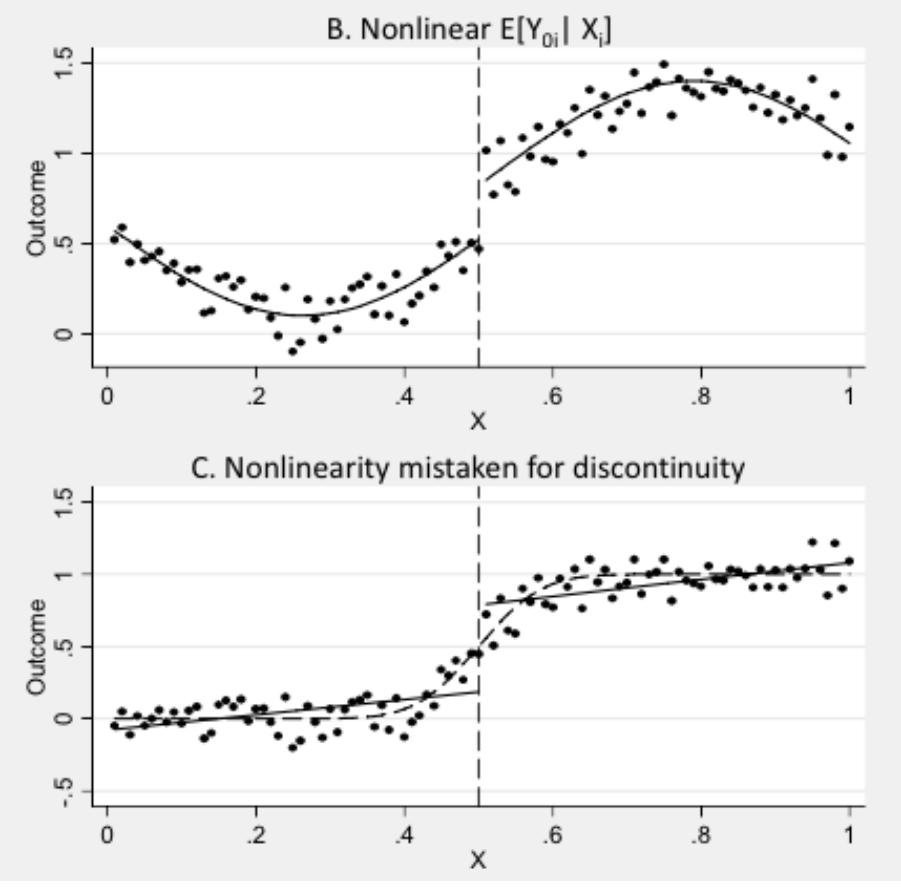
- impacto dos prêmios de excelência acadêmica sobre resultados futuros
- prêmios (D=1) somente para alunos com notas $X > c$
- H2: resultados são funções não-contínuas somente devido ao prêmio
- importância da formal funcional

$$Y = \alpha + D\tau + X\beta + \epsilon$$

RD Linear



Importância da formas funcionais

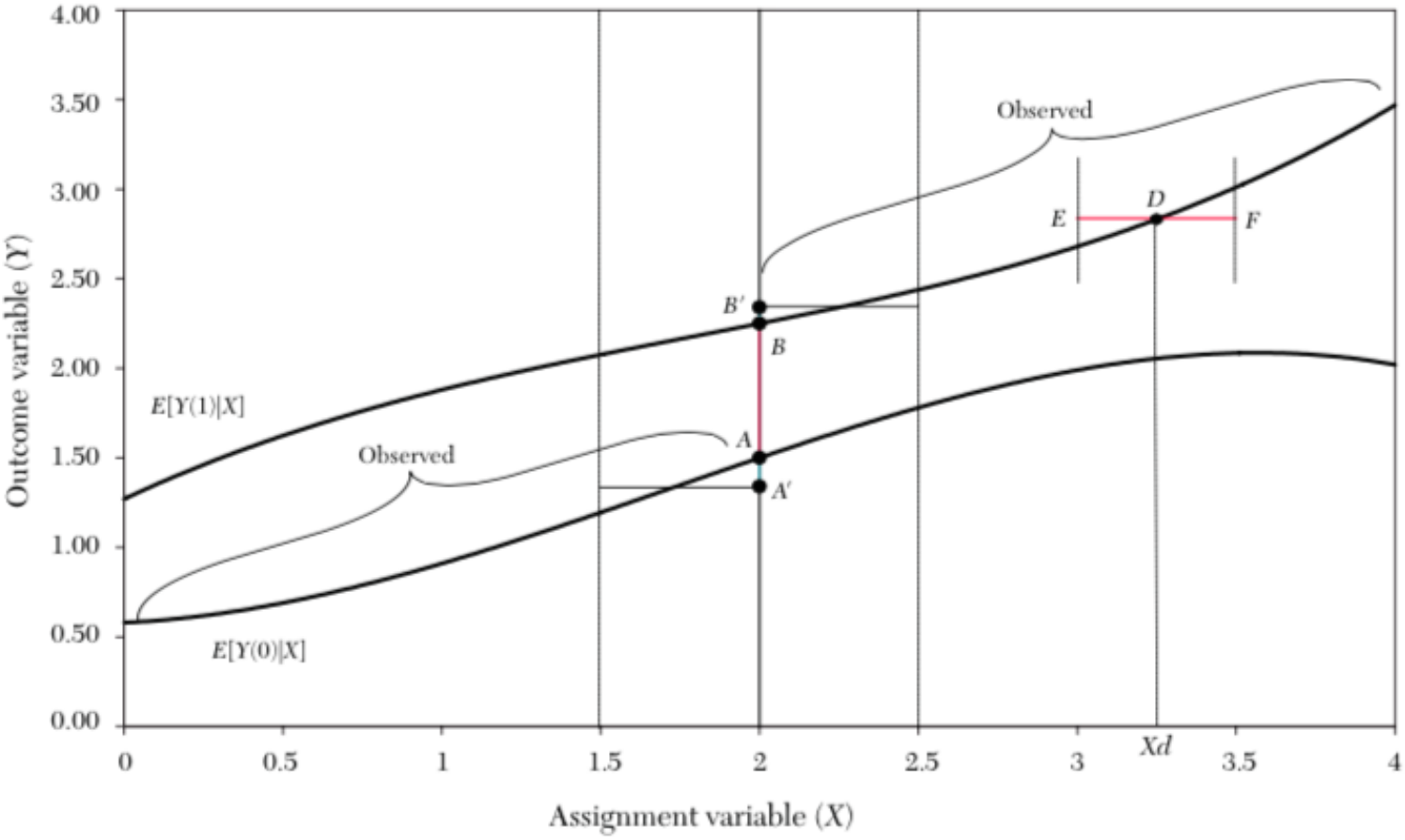


RDD e a abordagem dos resultados potenciais

- Hahn, Todd, and van der Klaauw (2001)
- formalização do RDD na linguagem da literatura de tratamento
- H1: todos outros fatores são *contínuos* em X

- abordagem dos resultados potenciais:
  - para cada i, existem Y_{0i} e Y_{1i} não-observáveis
  - $[Y_{0i} - Y_{1i}]$ = efeito causal do tratamento
  - podemos pensar em duas relações entre Y e X
  - $E[Y_{0i}|X]$ e $E[Y_{1i}|X]$ contínuos dado H1
  - porém somente observamos Y_{0i} caso $X < c$ ou Y_{1i} caso $X > c$
  - observamos: $Y_i = Y_{0i} + DY_{1i}$

Importância da hipótese de continuidade



RDD e a abordagem dos resultados potenciais

$$\begin{aligned} B - A &= lim_{\epsilon \to 0^+} E[Y_i | X_i = c + \epsilon] - lim_{\epsilon \to 0^-} E[Y_i | X_i = c + \epsilon] \\ &= E[Y_{1i} - Y_{0i} | X_i = c] \end{aligned}$$

Identificação e Interpretação

- Comparando IV e RDD:

1. Em qual contexto o RDD é plausível? (hipóteses de identificação)
2. Existe algum meio de testar tais hipóteses?
3. As estimativas de RDD são generalizáveis?

Identificação e Interpretação: **RDD (1)**

1. Em qual contexto o RDD é plausível? (hipóteses de identificação)
   - quando todos fatores não observados forem contínuos em X
2. Existe algum meio de testar tais hipóteses?
   - não podemos testar a condição *necessária* de continuidade
3. As estimativas de RDD são generalizáveis?
   - não pois o efeito estimado vale somente para a sub-população ao redor do cutoff

Identificação e Interpretação: **Variáveis Instrumentais**

1. Em qual contexto o RDD é plausível? (hipóteses de identificação)
   - quando o instrumento for não-correlacionado com o erro da eq. estrutural

2. Existe algum meio de testar tais hipóteses?
   - não podemos testar a condição *necessária* de exclusão

3. As estimativas de RDD são generalizáveis?
   - não pois o efeito estimado vale somente para a sub-população afetada pelo instrumento

Identificação e Interpretação: **RDD (2)**

1. Em qual contexto o RDD é plausível? (hipóteses de identificação)
   - quando X tiver um componente estocástico (indivíduos não têm controle preciso), a variação do tratamento ao redor da discontinuidade será aleatória

2. Existe algum meio de testar tais hipóteses?
   - como num experimento aleatório, a distribuição de covariadas não pode mudar abruptamente ao redor do cutoff

3. As estimativas de RDD são generalizáveis?
   - RDD pode ser interpretado como média ponderada do efeito tratamento, com pesos de probabilidade de estar próximos ao cutoff

Experimento Aleatório com seleção não-aleatória

- considere a seguinte formulação do RDD:

$$Y = D\tau + W\delta_1 + U$$
$$D = 1[X \geq c]$$
$$X = W\delta_2 + V$$

- parece um modelo de dummy endogeno, mas observamos X:
  - W pode ser endógeno, desde que determinado antes de V
  - elementos de δ_1, δ_2 podem ser zero
  - não há imposição sobre as correlações de W,U,V
- note que heterogeneidade individual em Y é completamente determinado por (W,U)

Densidade de X sobre diferentes hipóteses

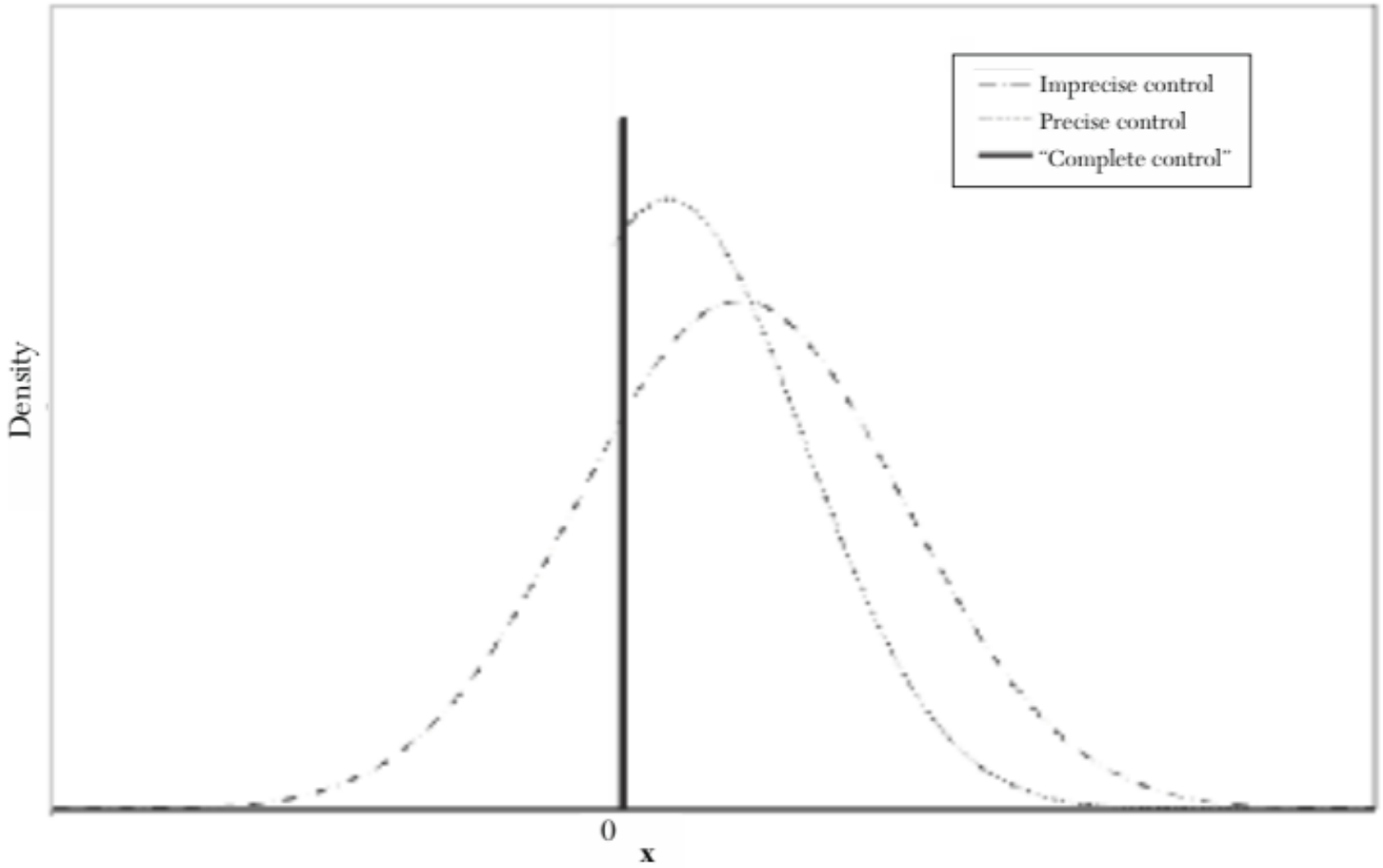


Figure 4. Density of Assignment Variable Conditional on $W = w, U = u$

Controle Impreciso de X

- um indivíduo consegue via esforço deslocar a curva para a direita
- mas estamos somente assumindo que ele não tem controle total

- **Definição**: Indivíduos tem controle impreciso sobre X quando condicional W=w e U=u, a densidade de V é contínua **(lousa)**
- **Aleatorização Local**: Se indivíduos tem controle impreciso sobre X, o tratamento é aleatório ao redor do cutoff

Consequências da Aleatorização Local

1. Identificação do Efeito Tratamento **(lousa)**

2. Teste da Validade do RDD
   - (i) testar se $Pr[W = w|X = x]$ é contínuo em x no cutoff (similar ao teste de balanceamento do experimento aleatório)
   - (ii) testar se a densidade de X é contínua no cutoff

3. Irrelevância das covariadas **(lousa)**
   - natureza aleatória do tratamento náo é condicional
   - inclusão das co-assiadas pre-determinadas reduz variância

4. Generalização: RDD como tratamento médio ponderado **(lousa)**
   - no caso de efeito de tratamento heterogêneo, RDD não se aplica somente a população ao redor do cutoff

Fuzzy RD

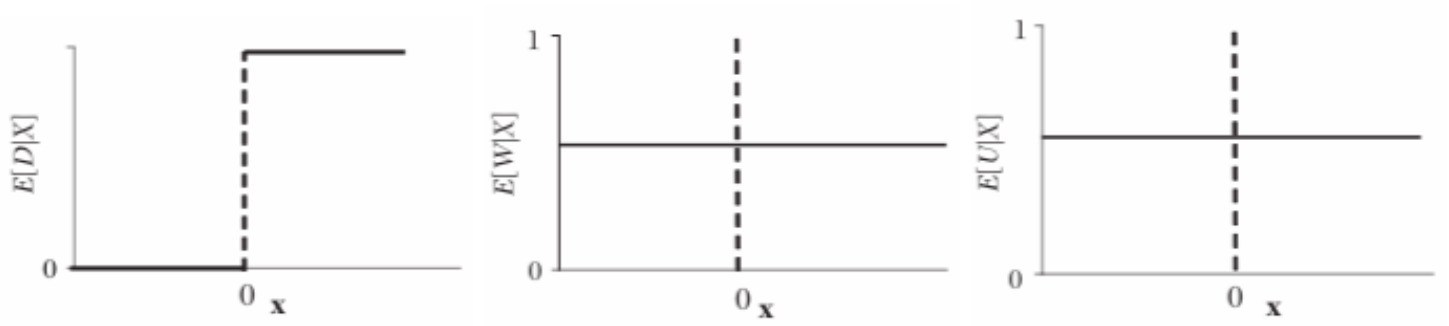
1. RDD com imperfect compliance
   - no sharp RD, probabilidade de tratamento pula de 0 para 1 no cutoff
   - no fuzzy RD, probabilidade de tratamento muda menos de 1
   - $lim_{\epsilon \to 0^+} prob(D = 1|X = c + \epsilon) \neq lim_{\epsilon \to 0^-} prob(D = 1|X = c + \epsilon)$

- discontinuidade entre Y e X não é mais o efeito tratamento médio
- temos que ajustar pela fração induzida pelo cutoff (Wald IV)

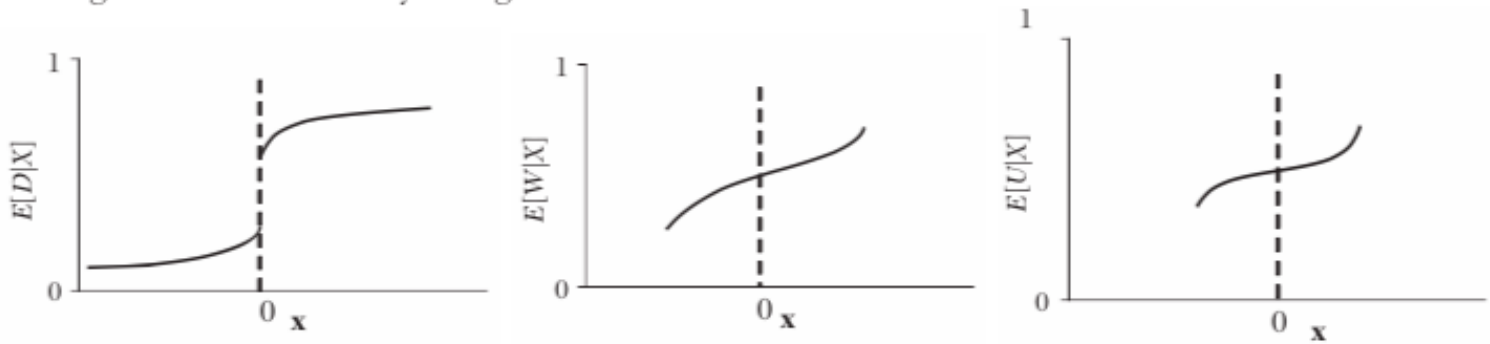
$$\tau_F = \frac{lim_{\epsilon \to 0^+} E[Y_i|X_i = c + \epsilon] - lim_{\epsilon \to 0^-} E[Y_i|X_i = c + \epsilon]}{lim_{\epsilon \to 0^+} E[D_i|X_i = c + \epsilon] - lim_{\epsilon \to 0^-} E[D_i|X_i = c + \epsilon]}$$

Comparando RDD com outras estratégias de avaliação...

A. Randomized Experiment

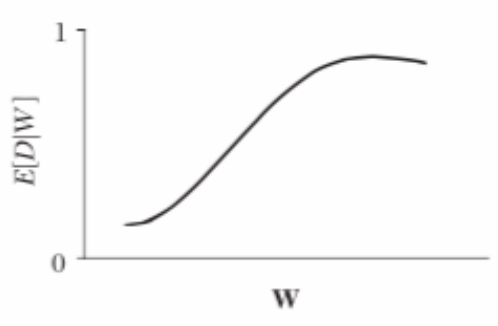


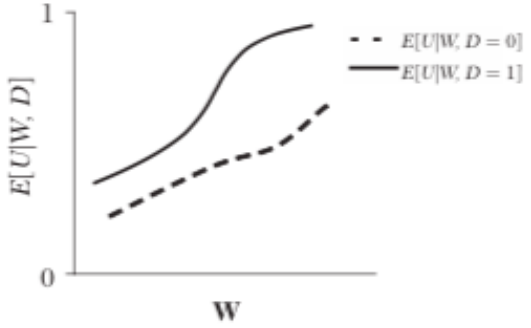
B. Regression Discontinuity Design



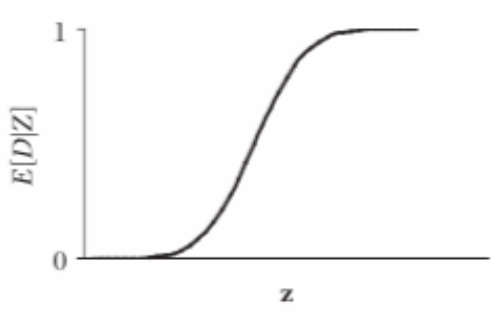
Comparando RDD com outras estratégias de avaliação...

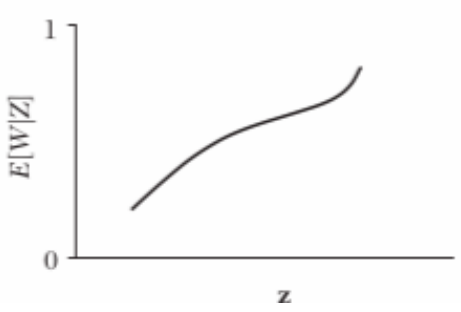
C. Matching on Observables

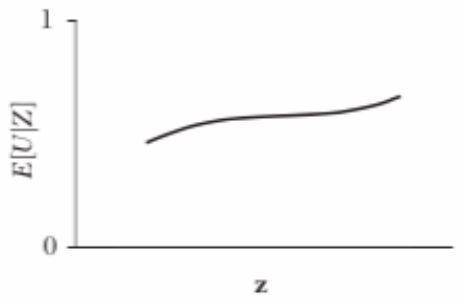




D. Instrumental Variables







Apresentação, Estimação e Inferência

1. Análise Gráfica
2. Análise de Regressão (especificação, (não) paramétrica)
3. Questões Práticas de Implementação
   - covariadas, running variable discreta, testes de validação
4. Guia Prático de Implementação

Análise Gráfica

- (i) dividir X em intervalos e plotar contra média de Y
- (ii) plotar a regressão prevista $E[Y|X]$

- vantagens da análise gráfica:
  - visualizar a forma funcional
  - regressão não-paramétrica (bom guia para a análise paramétrica)
  - o jump da idéia da magnitude do efeito tratamento
  - checar se existem outros jumps fora do cutoff

- bandwidth ótimo vs inspeção visual?

Intervalo de 2%

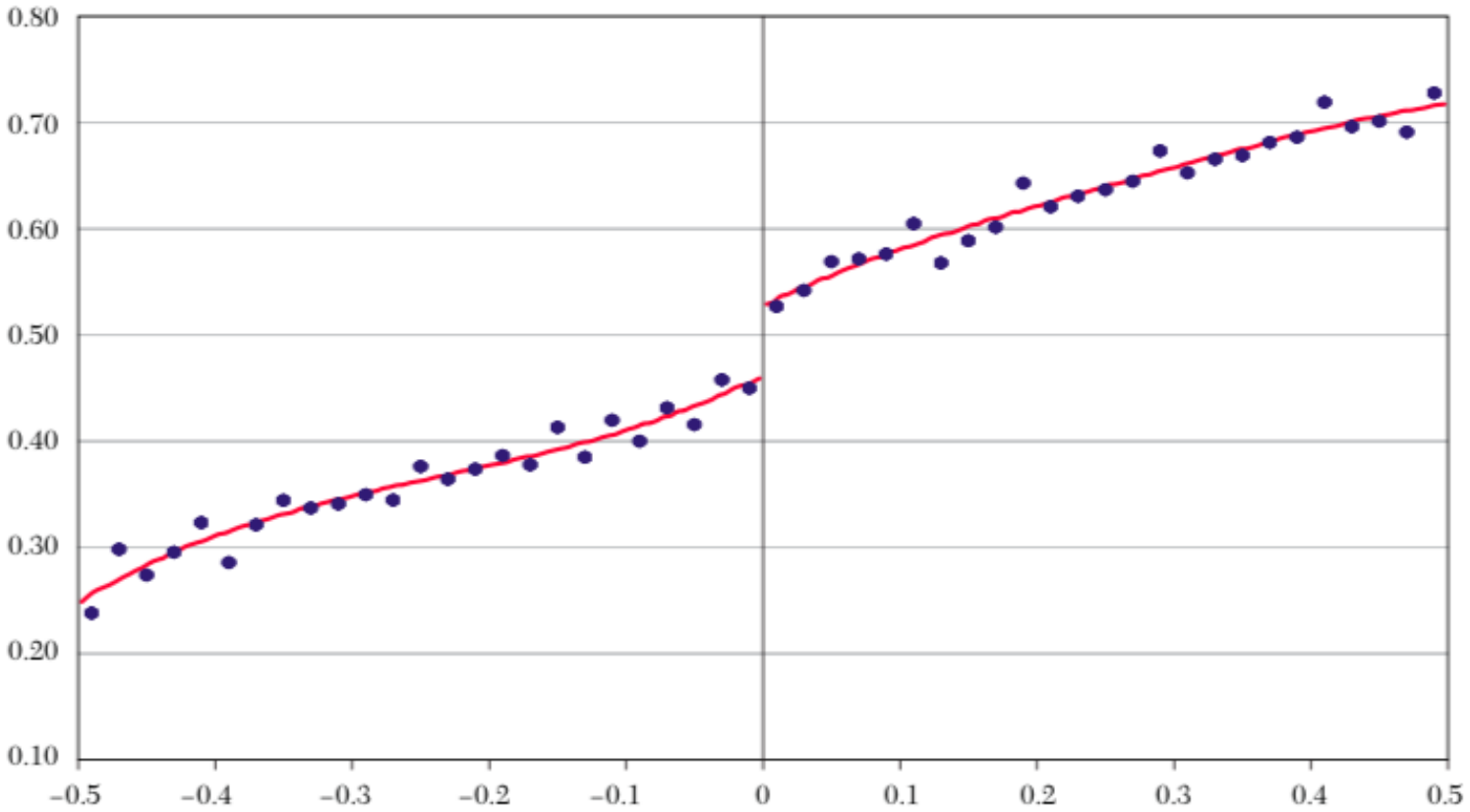


Figure 6. Share of Vote in Next Election, Bandwidth of 0.02 (50 bins)

Intervalo de 1%

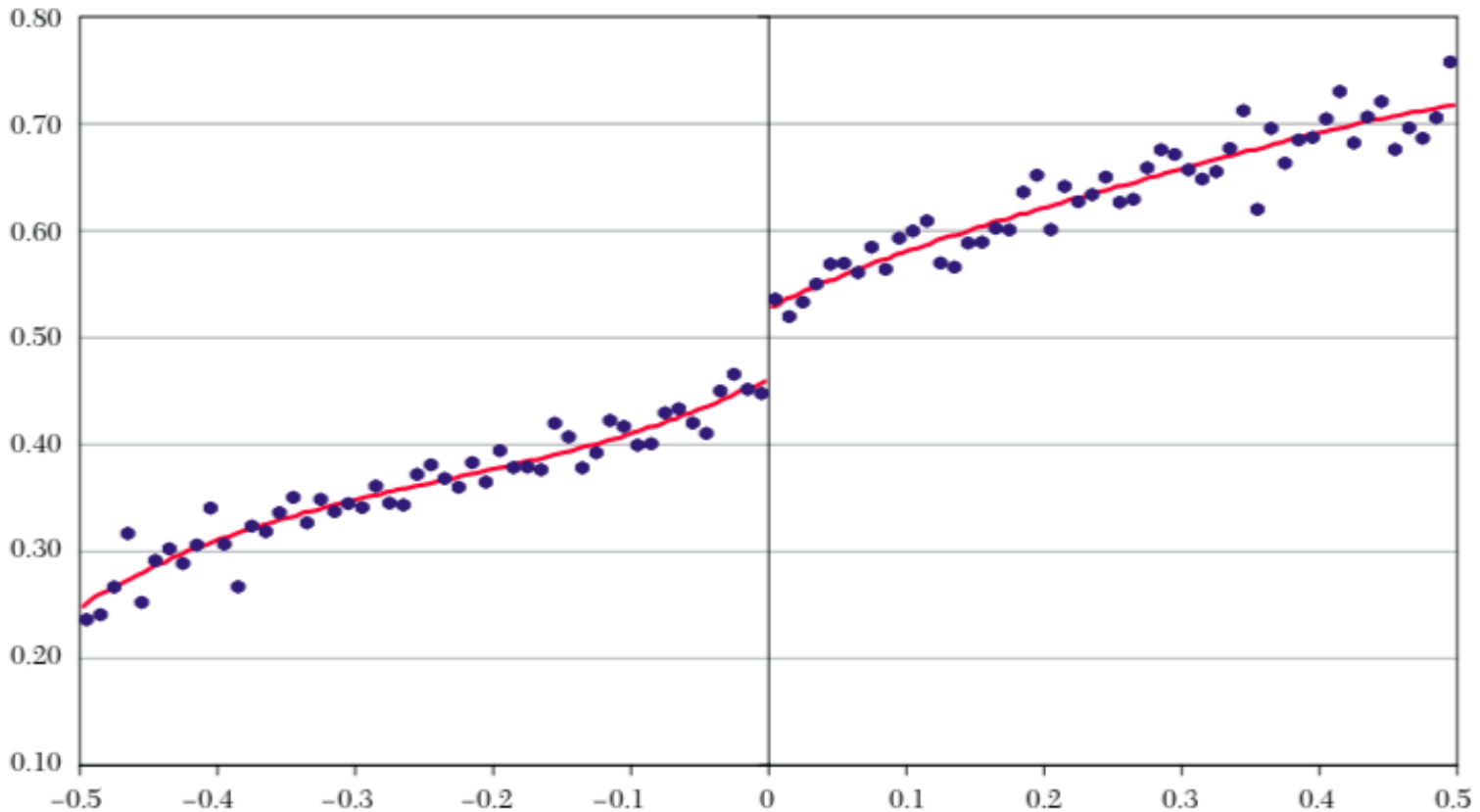


Figure 7. Share of Vote in Next Election, Bandwidth of 0.01 (100 bins)

Intervalo de 0.5%

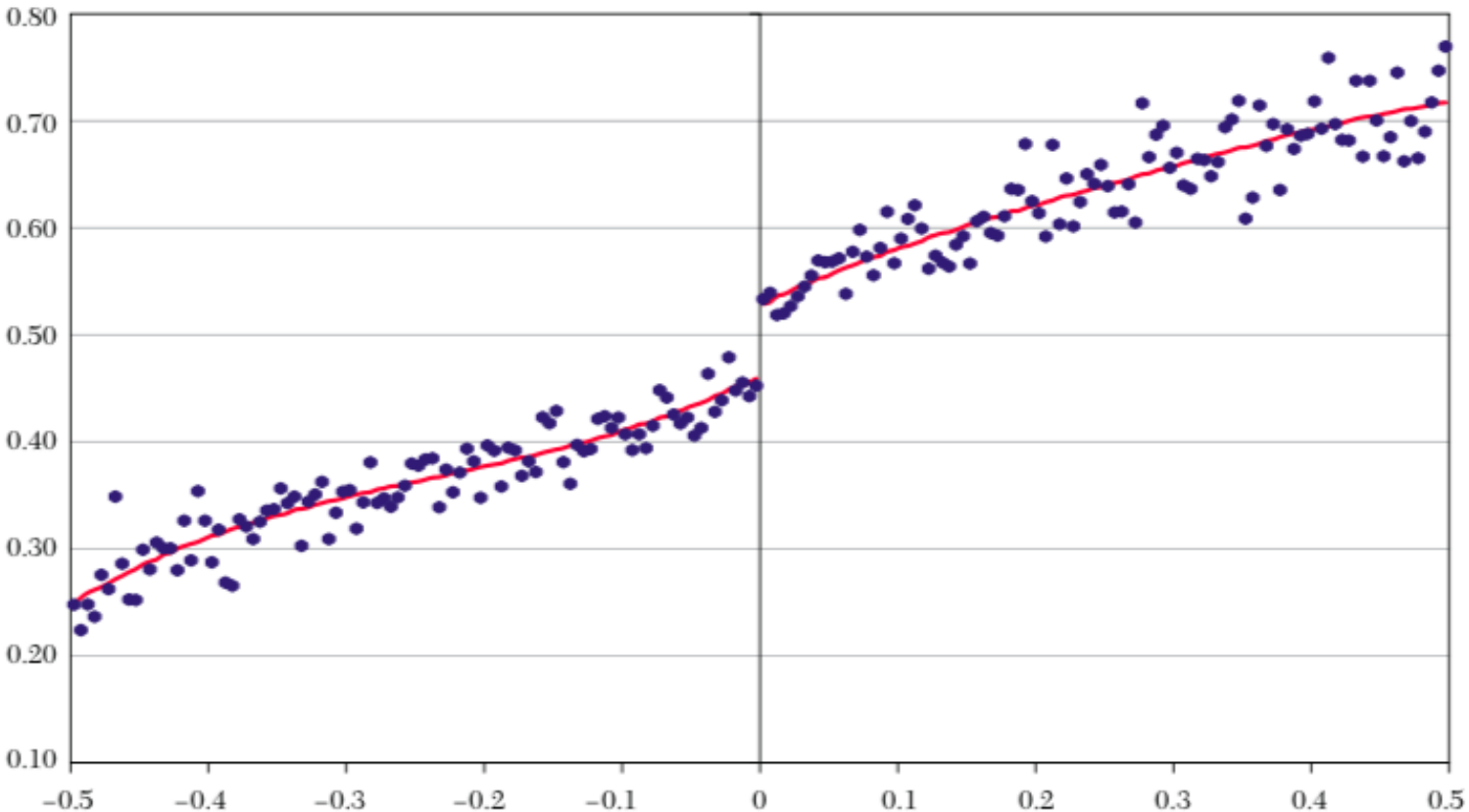


Figure 8. Share of Vote in Next Election, Bandwidth of 0.005 (200 bins)

Análise de Regressão

$$Y = \alpha + D\tau + X\beta + \epsilon$$

- até agora focamos no caso de regressão paramétrica linear
- porém a hipótese de linearidade não é garantida
- já vimos que isso é particularmente importante no caso de RDD

- qual a solução? trade-off entre eficiência e viés...

Análise de Regressão

1. Regressão Paramétrica com polinómios flexíveis (RD Global)
   - usamos todas as observações, mesmo longe do cutoff
   - polinomios de até 3a-5a ordem (Imbens e Gelman (2014))
   - (i) muito peso para observações longe do cutoff
   - (ii) resultados sensíveis a ordem do polinómio
   - (iii) inferência ruim - Pr(erro tipo I) aumenta

2. Regressão Não-Paramétrica Local (RD Local)
   - usamos observações próximas do cutoff (ineficiência)
   - viés caso inclinação não-nula

3. Combinar tendência linear (1) com regressão local (2)
   - Hahn, Todd, and van der Klaauw (2001)

Teste de Validação

- hipótese: indivíduos não controlam precisamente X
- não podemos testar pois somente observamos uma realização, mas podemos testar a densidade agregada (McCrary (2008))

- problema: alguns pulam pra cima, alguns pra baixo...
- pode não ser um problema. ex: professor empurrando aluno aleatoriamente

Teste de McCrary

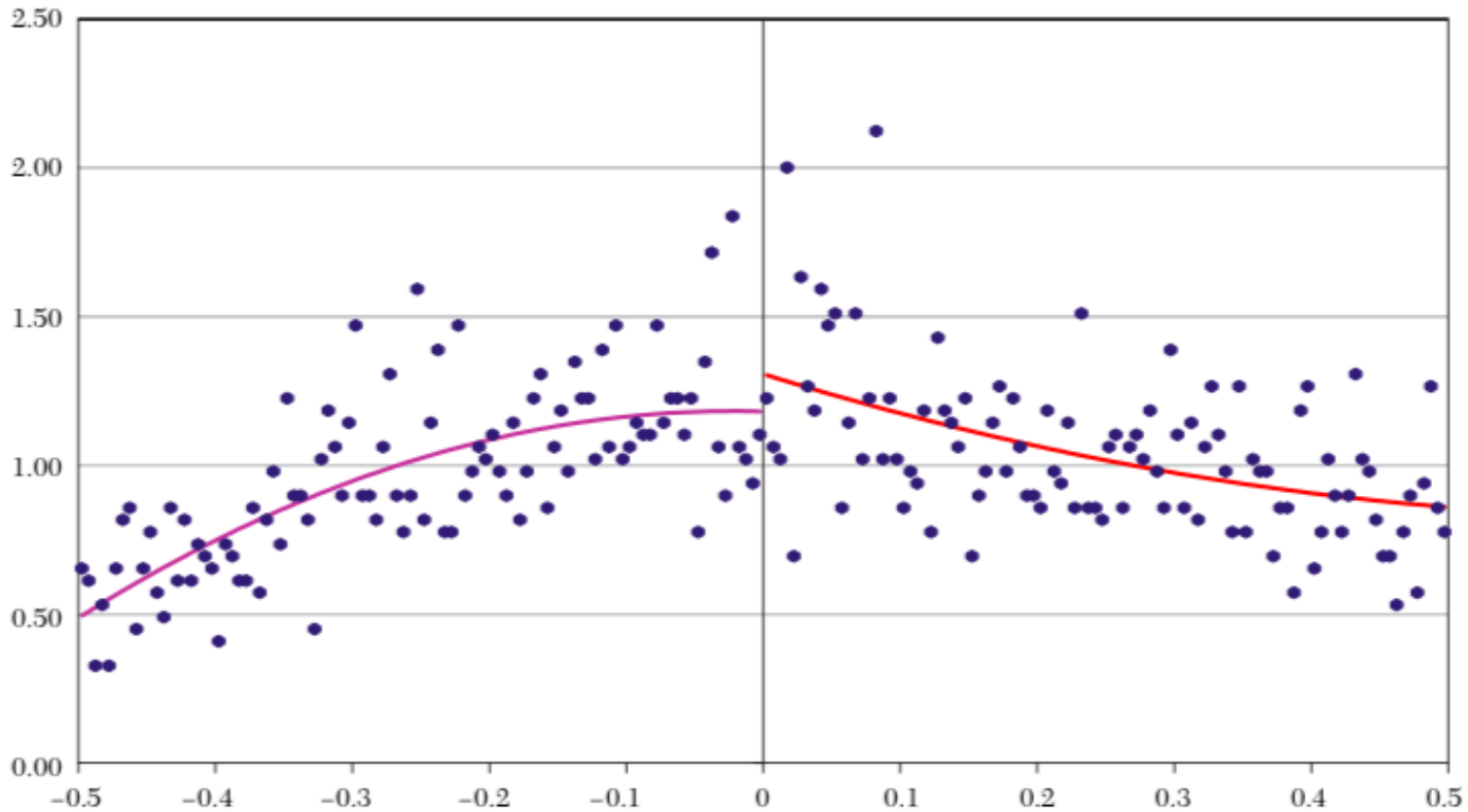


Figure 16. Density of the Forcing Variable (Vote Share in Previous Election)

Inspecionando as Covariadas

- teste alternativo da validade do RDD
- covariadas são localmente balanceadas?
  - (i) replicar o gráfico da descontinuidade para cada W
  - (ii) replicar a regressão principal para cada W

- como W é determinado antes do tratamento, descontinuidade não pode afetar W caso RD seja válido
- caso W seja grande, alguns coeficientes serão significantes
- usar SUR com teste χ^2 de significância conjunta

Checklist

1. To assess the possibility of manipulantion of the assignment variable, show its distribution.
2. Present the main RD graph using binned local averages
3. Graph a benchmark polynomial specification
4. Explore the sensitivity of the results to a range of bandwidths, and a range of orders to the polynomial.
5. Conduct a parallel RD analysis on the baseline covariates
6. Explore the sensitivity of the results to the inclusion of baseline covariates.