RDD, inferência causal e um exemplo em R

Pedro Cavalcante

2018-12-20

Uma das coisas que mais me fascinam em econometria é inferência causal, a arte de separar o sinal do ruído. Boa parte do trabalho de economistas ~~sérios~~ que estudam temas aplicados é conseguir inferir relações causais e não meramente correlações de dados que não são laboratoriais. É difícil controlar todas as variáveis possíveis que afetem performance de alunos - não podemos designar pais atenciosos (!) - e impossível observar dois Brasis, um em que vigora uma regra X e outro em que não vigora.

Somos, nesse sentido, muito limitados em nossas ambições. O melhor com que podemos sonhar é conduzir estudos caríssimos em que um tratamento é designado aleatoriamente entre participantes, como fazem em estudos clínicos para drogas novas. Um exemplo interessante é Nyqvist *et al.* (2018, AEJ: Applied Economics), que montou uma loteria para pacientes com HIV no Lesoto. A nossa capacidade de controlar covariáveis relevantes também é limitada a depender do contexto. Em um laboratório é razoavelmente fácil controlar os fatores relevantes para o comportamento de duas pessoas jogando o Jogo do Ultimato, não é tão simples dizer para alunos (ou mesmo equipe) de uma escola que metade da turma vai receber um pagamento em dinheiro, aulas de música ou ficar em uma turma menor.

Por isso no mundo real usualmente dependemos de quasi-experimentos, ou experimentos naturais. O resgate de cubanos de Mariel como em Card e DiNardo (2000, AER), a colonização européia de boa parte do mundo como em Acemoglu *et. al* (2001, AER) e a divisão da América do Sul no Tratado de Tordesillas como em Fujiwara *et. al* (2017) exploram esse tipo de evento como fontes de variação exógena em algum tipo de variável: oferta de trabalho, instituições políticas e presença de escravos, respectivamente. Tendo isso em mente, qualquer variação em outras variáveis que seja explicada por essa variação exógena que identificamos pode ser crivelmente atribuída ao efeito causal que a primeira variável tem. Isso é, em termos bem amplos, o que chamamos de Variáveis Instrumentais. É uma técnica bem popular de inferência causal.

Esses exemplos, no entanto, parecem muito grandiosos, históricos. Podem ser menores e normalmente regras implementadas por burocratas são fontes valiosas de variações agudas. Aqui entra o RDD.

É plausível que dia de nascimento seja relevante para renda e escolaridade? Muito pouco a princípio, a menos que - por exemplo - um pai precise esperar um ano para matricular seu filho numa escola pública porque ele nasceu um dia depois da data limite para o ano. Essa é a ideia de McCrary e Royer (2011, AER). Esse tipo de evento não causa só variação plausivelmente exógena na escolaridade entre crianças, mas variação aguda. Uma *descontinuidade*, por assim dizer. O gráfico abaixo, tirado do paper, ilustra isso:

Você pode entender o efeito de tratamento da regra como a diferença dos limites laterais desse polimônio estimado no ponto da descontnuidade. Chamamos essa técnica de *Regression Descontinuity Design* (RDD).

Leitor, RDD

RDD, leitor

A história dessa técnica é curiosa: foi proposta por dois psicólogos educacionais, Campbell e Thistlewaite (1960) para avaliar o efeito de ganhar competições científicas nos hábitos de atendência em pós-graduação de alunos. Você pode ler mais sobre a história dela clicando aqui se quiser.

Tá, mas e mão na massa?

Já falei demais sem escrever uma linha de código. Vamos ver a magia acontecer. Vou replicar um exemplo do incrível livro *Causal Inference: The Mixtape*, do Scott Cunningham (que tem uma maravilhosa conta no twitter), disponível de graça no site dele. Três pacotes trazem ferramentas de R para estimar e brincar com RDDs:

Figure 1. Education at Motherhood, by Day of Birth: Native First-Time Mothers 23 Years Old and Younger

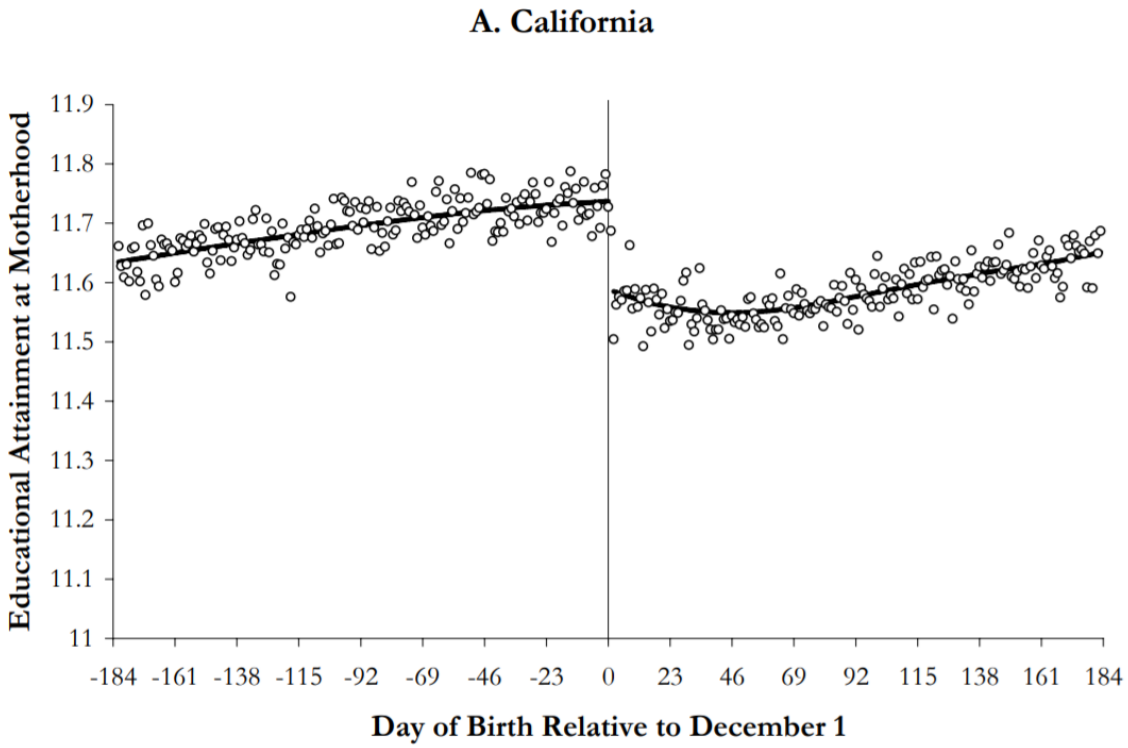


Figure 1:

- O primeiro é `rdrobust`, que implementa o estimador proposto em Calonico, Cattaneo e Titiunik (Econometrica, 2014), um RDD com intervalos de confiança menos sensíveis à variações no bandwidth selecionado. Eles fazem isso usando um estimador novo para o erro-padrão. O pacote também tem um port para Stata publicado no Stata Journal, além da versão em R - que saiu no Journal of Statistical Software.
- O segundo é `rddtools`, que traz algumas ferramentas de inferência, implementações variadas e uma base de dados interessante.
- O terceiro é `rdd`. O menos completo de todos, traz ferramentas para visualização e uma implementação, completamente *barebones*, mas entrega o que precisa ser feito. Talvez seja bom para quem está começando R, porque é de longe o mais simples.

Vamos replicar parte das regressões em Lee, Moretti e Butler (2004, QJE). Lembre-se de instalar o pacote `mixtape` com o comando `devtools::install_github('johnson-shuffle/mixtape')`, já que o autor não submeteu ao CRAN. Nele estão as bases de dados com exemplos. Vamos usar a base `lmb_data`, com dados eleitorais a nível de distrito. `score` é o ADA Score, uma medida de 0 (muito conservador) a 100 (muito progressista) de cada legislador e `demvoteshare` é a fração dos votos no distrito para legisladores democratas.

A pergunta que estamos nos fazendo é: eleitores elegem ou afetam políticas públicas? Se eles afetam, então entende-se que pressão competitiva por voto induz convergência política - assim como no jogo de Hotelling farmácias se agrupam em Copacabana (desculpa para quem não é do Rio, não resisti). No entanto, se eles *elegem*, então entende-se que políticos não conseguem crivelmente se comprometer com plataformas específicas. Eleições então são mecanismos que revelam preferências sociais de maneira bem clara, quem ganha melhor satisfaz esses desejos.

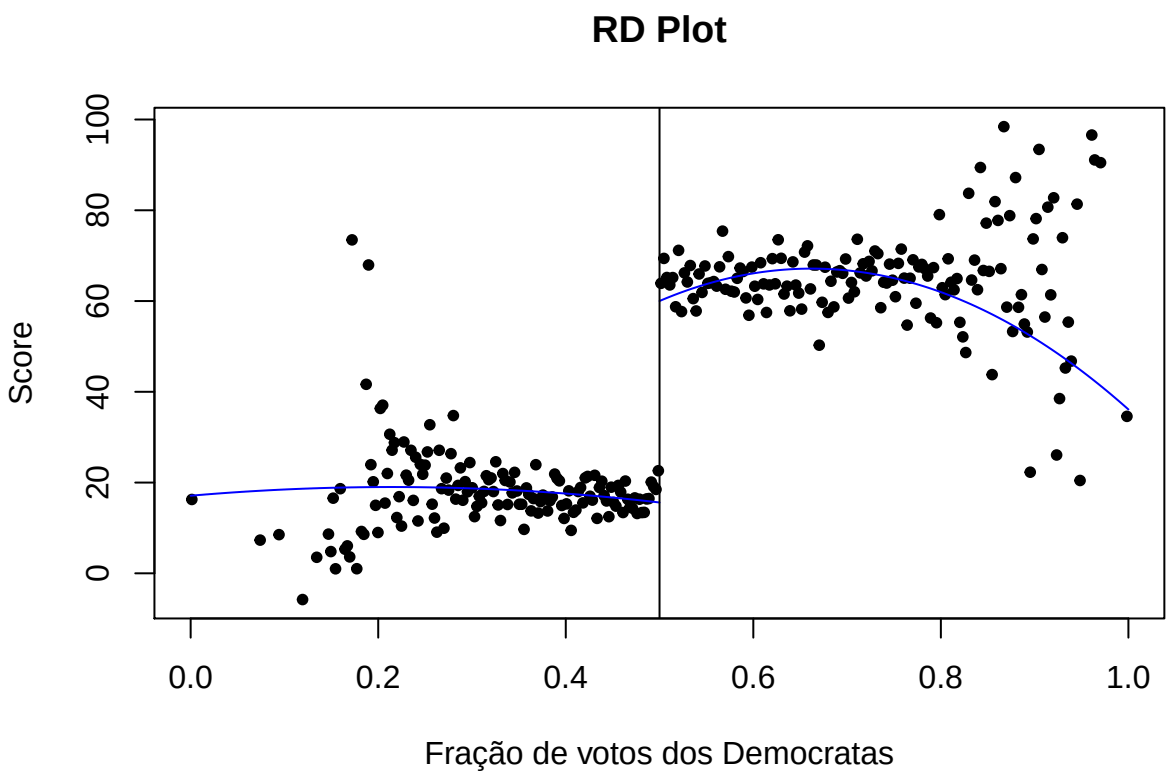
```
library(rdrobust)
library(mixtape)
library(tidyverse)

data("lmb_data")
lmb_data = as.tibble(lmb_data)
head(lmb_data)
```

```
## # A tibble: 6 x 178
##   state district incmbncy demvote repvote  year congress occupanc name
##   <dbl>    <dbl> <dbl+lb>   <dbl>   <dbl> <dbl>    <dbl>    <dbl> <chr>
## 1 1            1 -1        127802  103294  1948       81        0 RIBI~
## 2 1            1 -1        127802  103294  1948       81        0 RIBI~
## 3 1            1 " 1"      134258   96251  1950       82        0 RIBI~
## 4 1            1 " 1"      134258   96251  1950       82        0 RIBI~
## 5 1            1 " 1"      148935  112526  1954       84        0 DODD
## 6 1            1 " 1"      148935  112526  1954       84        0 DODD
## # ... with 169 more variables: eq_Dwhip <dbl>, eq_Rwhip <dbl>,
## #   eq_Dlead <dbl>, eq_Rlead <dbl>, vote <dbl>, republic <dbl>,
## #   party <dbl>, office <dbl>, icpsr_id <dbl>, whip_D <dbl>, whip_R <dbl>,
## #   vote_tot <dbl>, demvoteshare <dbl>, dembin <dbl>,
## #   lagdemvoteshare <dbl>, clusterid <dbl>, pooleyear <dbl>,
## #   redistrict <dbl>, poolename <chr>, aclu_vs <dbl>, acu_vs <dbl>,
## #   aca_vs <dbl>, ada_vs <dbl>, afbf_vs <dbl>, afge_vs <dbl>,
## #   afscme_vs <dbl>, aft_vs <dbl>, asc_vs <dbl>, bfw_vs <dbl>,
## #   bctd_vs <dbl>, carter_vs <dbl>, ccus1_vs <dbl>, ccus2_vs <dbl>,
## #   cwla_vs <dbl>, cv_vs <dbl>, cvvf_vs <dbl>, sane_vs <dbl>,
## #   cfnfmp_vs <dbl>, aflcio_vs <dbl>, cfsca_vs <dbl>, cfsce_vs <dbl>,
## #   cfscd_vs <dbl>, cfscs_vs <dbl>, ccause_vs <dbl>, cw_vs <dbl>,
## #   cc_vs <dbl>, cfa_vs <dbl>, ike_vs <dbl>, pfpike_vs <dbl>,
## #   pdpike_vs <dbl>, ford_vs <dbl>, fcnl_vs <dbl>, lbj_vs <dbl>,
```

```
## #   pfplbj_vs <dbl>, pdplbj_vs <dbl>, jfk_vs <dbl>, pfpjfk_vs <dbl>,
## #   pdpjfk_vs <dbl>, lfr_vs <dbl>, lcv_vs <dbl>, lwv_vs <dbl>,
## #   ll_vs <dbl>, lfs_vs <dbl>, nasc_vs <dbl>, ncsc_vs <dbl>, nea_vs <dbl>,
## #   nfo_vs <dbl>, nfu_vs <dbl>, nfib_vs <dbl>, ntu_vs <dbl>,
## #   nwpc_vs <dbl>, nr_vs <dbl>, nixon_vs <dbl>, reagan_vs <dbl>,
## #   ripon_vs <dbl>, twr_vs <dbl>, uaw_vs <dbl>, umw_vs <dbl>,
## #   firstyear <dbl>, adayear <dbl>, nomada <dbl>, realada <dbl>,
## #   demvs2 <dbl>, demvs3 <dbl>, demvs4 <dbl>, lagdemvs2 <dbl>,
## #   lagdemvs3 <dbl>, lagdemvs4 <dbl>, dem2 <dbl>, lagdem2 <dbl>,
## #   lagaclu <dbl>, lagacu <dbl>, lagaca <dbl>, lagafbf <dbl>,
## #   lagafge <dbl>, lagafscme <dbl>, lagaft <dbl>, lagasc <dbl>,
## #   lagbfw <dbl>, lagbctd <dbl>, ...
```

Agora procuramos uma descontinuidade, que sabemos existir quando democratas ganham:

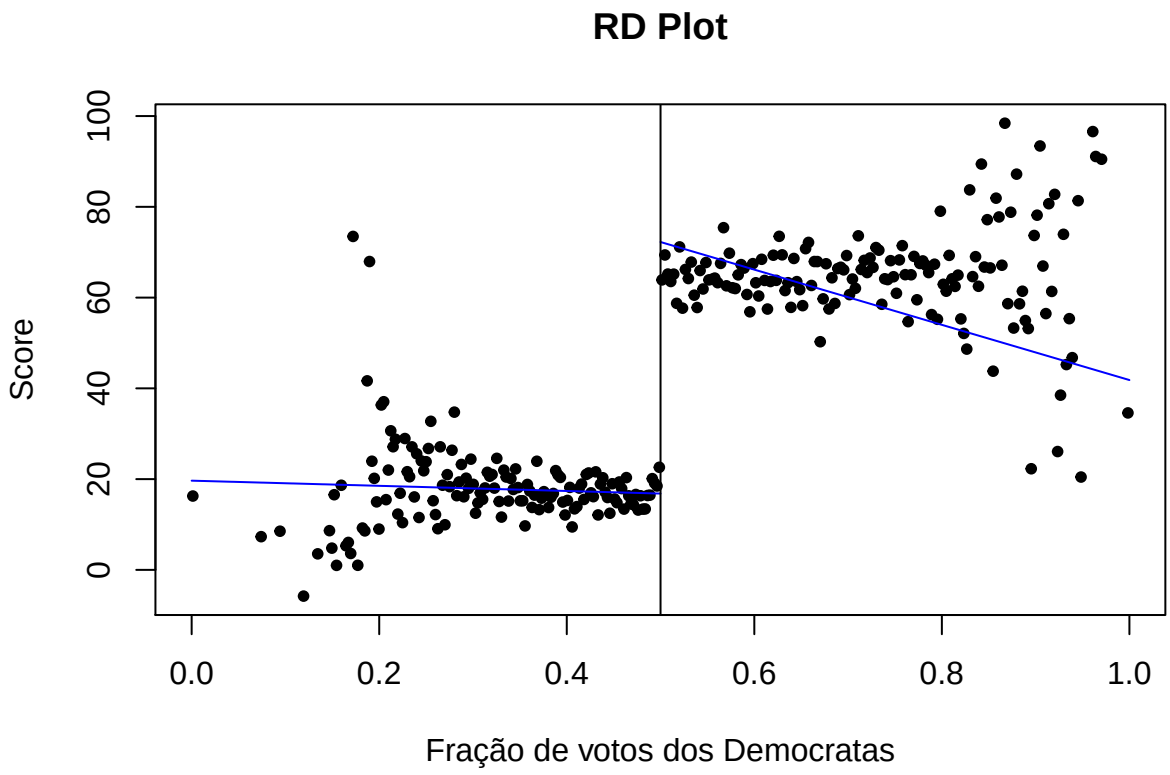
```
rdrobust::rdplot(y = lmb_data$score,
       x = lmb_data$demvoteshare,
       p = 2, # grau do polinômio
       c = .5, # onde está o cut-off
       x.label = "Fração de votos dos Democratas",
       y.label = "Score")
```



Podemos tentar repetir com um polinômio linear:

```
rdrobust::rdplot(y = lmb_data$score,
       x = lmb_data$demvoteshare,
       p = 1, # grau do polinômio
       c = .5, # onde está o cut-off
```

```
       x.label = "Fração de votos dos Democratas",
       y.label = "Score")
```



Podemos só estimar o RDD sem o auxílio gráfico. Observe que o print padrão da função não é completo e vai omitir informações importantes como p-valor. É de bom tom armazenar o modelo em um objeto e pedir o sumário dele.

```
rdd1 = rdrobust(y = lmb_data$score,
       x = lmb_data$demvoteshare,
       p = 2,
       c = .5,
       kernel = "triangular")

summary(rdd1)
```

```
## Call: rdrobust
##
## Number of Obs.                13577
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 5480        8097
## Eff. Number of Obs.            3171        2950
## Order est. (p)                    2           2
## Order bias  (p)                   3           3
## BW est. (h)                   0.135       0.135
```

```
## BW bias (b)                  0.184       0.184
## rho (h/b)                    0.730       0.730
##
## ===============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ===============================================================================
##   Conventional    46.201     1.466    31.507     0.000    [43.327 , 49.075]
##         Robust         -         -    27.412     0.000    [42.714 , 49.293]
## ===============================================================================
```

O próximo passo - central - é repetir essa estimação para legisladores democratas e depois somente para legisladores republicanos. Se o efeito da competição domina, então observaríamos ambos os partidos indo mais para a esquerda em distritos que tem maioria Democrata.

Deixo ao leitor fazer essa parte e tirar suas próprias conclusões.