Desenho de Regressão Discontínua

Manoel Galdino

2024-05-07

Características-chave da RDD

A Regressão Discontínua (RDD) é caracterizada por uma variável contínua X_i, que determina quem recebe tratamento, denotado por T_i (1 se tratado). Por convenção, X é chamada de "running variable", "assignment variable" ou "forcing variable".

Determinação do Tratamento

Em um desenho RDD *sharp*, uma unidade é tratada se $X_i \geq c$ e não tratada se $X_i < c$. Assim, T_i é uma função determinística de X_i: $T_i = f(X_i)$. A *running variable* determina completamente quem recebe tratamento.

Fuzzy RDD

- Pode acontecer do ponto da regra não determinar quem recebe ou não o tratamento, mas apenas a probabilidade de receber o tratamento.
- Nesse caso, a regra serve como variável instrumental ao redor do ponto de corte.
- Tudo se passa como se houvesse always-takers e/ou never-takers ao redor do ponto de limiar.
- Ex.: regra de voto determina número de cadeiras. Mas migração partidária altera o número. Então quem fica abaixo do número mínimo em um distrito pode ter cadeiras naquele distrito via migração partidária. São always-takers.

Observação e Corte

É essencial observar X e conhecer o **ponto de corte** ou **limiar** c.

Identificação em RDD

Uma das suposições da RDD é que ela requer a continuidade da variável X para identificação, embora, na prática, alguns estudos de RDD tenham usado *running variables* discretas. A continuidade de X é necessária porque a identificação ocorre no limite.

Estimativa dos Efeitos do Tratamento

A comparação de $\lim_{x \to c} E[Y_i|X_i = x]$ com $\lim_{x \leftarrow c} E[Y_i|X_i = x]$ fornece uma estimativa dos efeitos do tratamento (note a direção das setas). Esta comparação é equivalente a: $\lim_{x \to c} E[Y_i|X_i = x, T_i = 0]$ e $\lim_{x \leftarrow c} E[Y_i|X_i = x, T_i = 1]$, uma vez que, neste exemplo, à direita de c todos recebem tratamento; à esquerda, ninguém recebe. Portanto:

- $\lim_{x \to c} E[Y_i|X_i = x] \approx E[Y_{0i}|X_i = c]$
- $\lim_{x \leftarrow c} E[Y_i|X_i = x] \approx E[Y_{1i}|X_i = c]$

Suposição de continuidade

- A suposição de continuidade é tão crítica que vale discutirmos um pouco mais sobre ela.
- Se há continuidade, isso significa que, na auência do ponto de corte c, x (e outras covariáveis) não devem apresentar descontinuidade.
- Ex.: Suponha que estamos interessados em estudar o efeito da incumbência sobre a chance de reeleição futura ou riqueza futura desses políticos.
- Habilidades e carisma são variáveis que devem influenciar tanto a chance de serem incumbentes como os resultados de interesse. Em um RDD, podemos usar *close elections* para estimar o efeito. E a suposição de continuidade requer que carisma e habilidades não tenham descontinuidade no *cut off* de 50%. Na verdade, apenas o resultado eleitoral é descontínuo no *cut off*, que vai de não-eleito para eleito.

Suposições na RDD

Suposição de Não-manipulação com Precisão

A identificação dos efeitos do tratamento na RDD baseia-se na premissa de que X atua como um aleatorizador ao redor de c. Imagine que X seja uma variável aleatória uniforme usada para atribuir tratamento. Se $X \geq c$, uma unidade recebe tratamento. Na RDD, X tem o mesmo papel, exceto que não assumimos que X é independente do resultado Y. Na maioria das aplicações, X e Y são correlacionados de alguma forma.

Problemas de Manipulação

No entanto, se c não for arbitrário ou tiver uma relação determinística com Y, ou se as unidades puderem — com precisão — determinar seus escores X e, assim, escolher receber tratamento ou não, então X ao redor de c não se comporta mais como um aleatorizador — há alguma forma de auto-seleção que poderia depender de variáveis não observáveis.

Testabilidade da Suposição de não-Manipulação

Em parte, isso é testável. As unidades não pareceriam semelhantes perto de c e haveria um "acúmulo" próximo a c. No entanto, não podemos descartar a manipulação com precisão apenas com dados — devemos argumentar isso com conhecimento do assunto (é uma restrição de exclusão).

Estimação em RDD

Problema de Complete Overlapping

Um problema chave na estimação em RDD estrita é a completa falta de sobreposição.
Em matching, dicustimos como a ausência de sobreposição gerava problemas de extrapolação.
Sobreposição requer que $0 < P(D_i = 1|X_i) < 1$ para o domínio de X_i. No domínio da *running vairable* X_i, isso claramente não é satisfeito. Em RDD estrita, temos $P(D_i = 1|X_i < c) = 0$ e $P(D_i = 1|X_i \geq c) = 1$.

Dependência de Extrapolação

Devido à falta de sobreposição, dependemos de extrapolação para estimar os efeitos do tratamento. Dito de outra forma, podemos não ser capazes de estimar corretamente os efeitos do tratamento se errarmos a forma funcional $Y_i = f(X_i)$. Novamente, essa foi uma motivação para usar matching.
O problema é que nunca sabemos se acertamos, então a especificação do modelo é uma questão chave na estimação RDD.

Métodos de Estimação

O problema sugere a necessidade de um método de estimação não paramétrico. Utilizaremos métodos paramétricos, não paramétricos (ou semiparamétricos) para tentar abordar essas questões.

Identificação no Limite

A identificação dos efeitos do tratamento ocorre no limite, à medida que $X_i \to c$. Quanto mais usarmos observações distantes de c em X, mais dependeremos de extrapolação e das suposições sobre a forma funcional.

Trade-off de Viés-Variância

- **Mais perto de c:** Melhor em termos de precisão, mas pode haver uma amostra insuficiente. Resulta em menos viés, mas mais variância.
- **Mais distante de c:** Dependemos menos de extrapolação, mas introduzimos mais viés, mesmo com menor variância.

Métodos de Largura de Banda Ótima

A ideia é restringir a estimativa a uma janela ao redor de $X_i = c$, que pode ter tamanhos diferentes à esquerda ou à direita. Estes métodos buscam equilibrar a precisão das estimativas minimizando viés e variância conforme a proximidade do ponto de corte c.

Regras arbitrárias

Atribuição de "coisas" a partir de regras com pontos de cortes

Bolsa família: a partir de certa renda

Educação: aprovação no ensino superior a partir de certa nota de corte

Espacial: polítia pública para donos de áreas em abaixo ou acima de certas áreas.

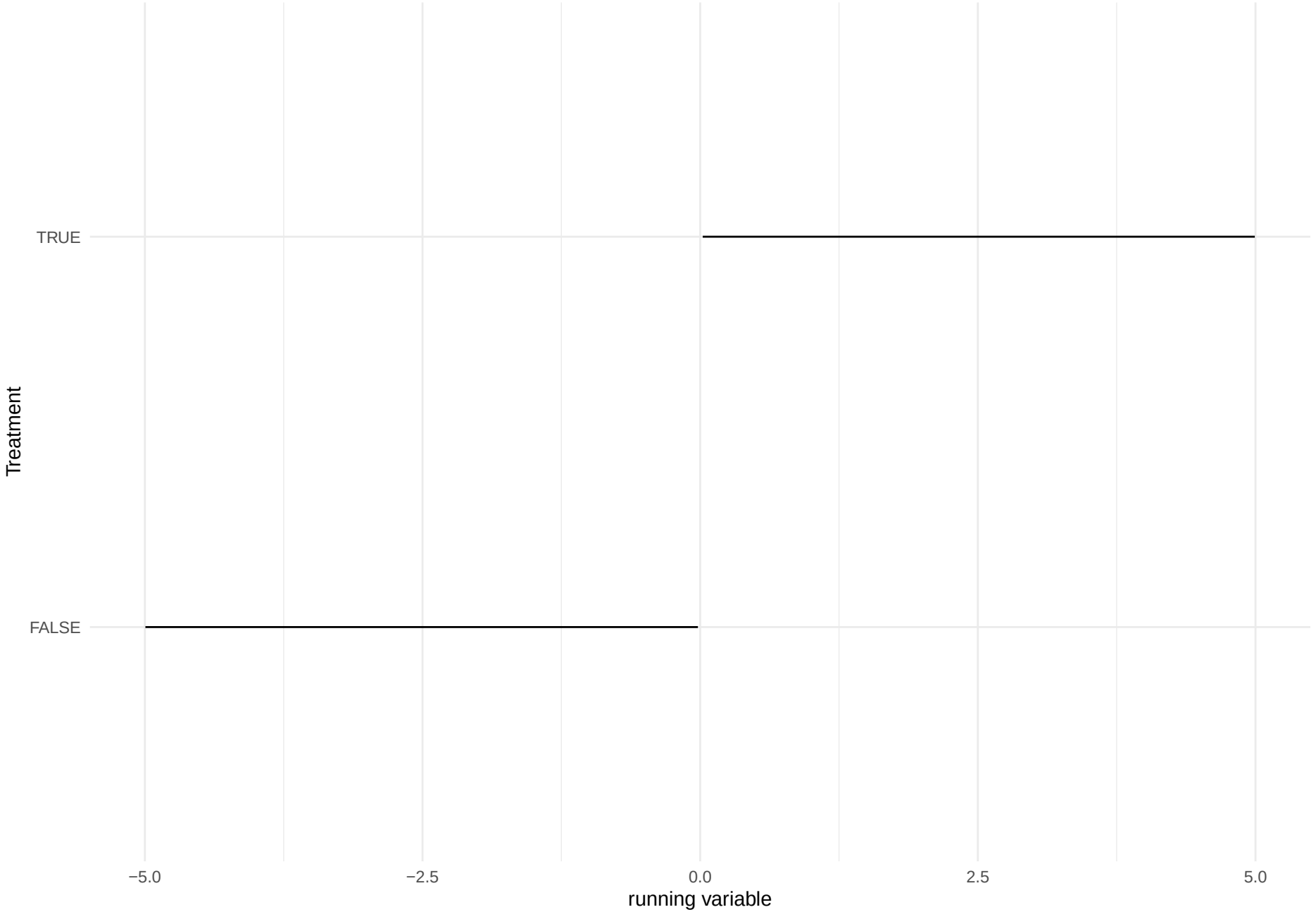
Data: regras para aposentadoria, idade para entrar na escola, data para perdão de dívida: Desenrola: "...cujas dívidas tenham sido incluídas no cadastro de inadimplentes no período entre 1º de janeiro de 2019 e 31 de dezembro de 2022".

Política: regras de número de vereadores, regras de população para ter segundo turno, regras para ter biometria etc.

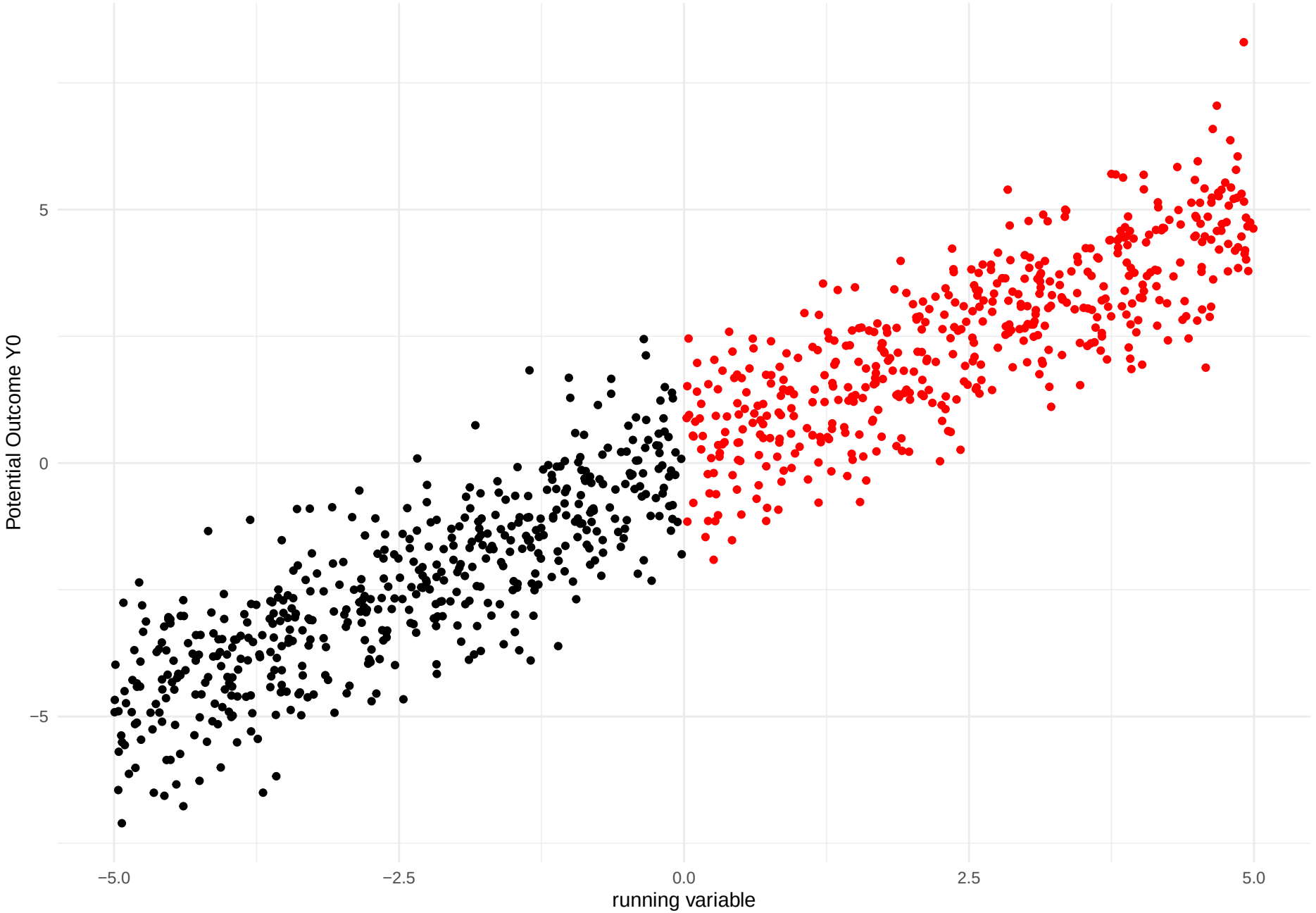
Simulação

```r
## Basic RD Model
set.seed(123)
N <- 1000 # number of observations
X <- runif (N , -5,5)
Y0 <- rnorm ( n =N , mean =X , sd=1) # control potential outc
Y1 <- rnorm ( n =N , mean = X+2, sd=1) # treatment potential
#You only get treatment if X>0
Treatment <- ( X >= 0)
# What we observe
Y = Y1* Treatment + Y0*(1- Treatment )
```

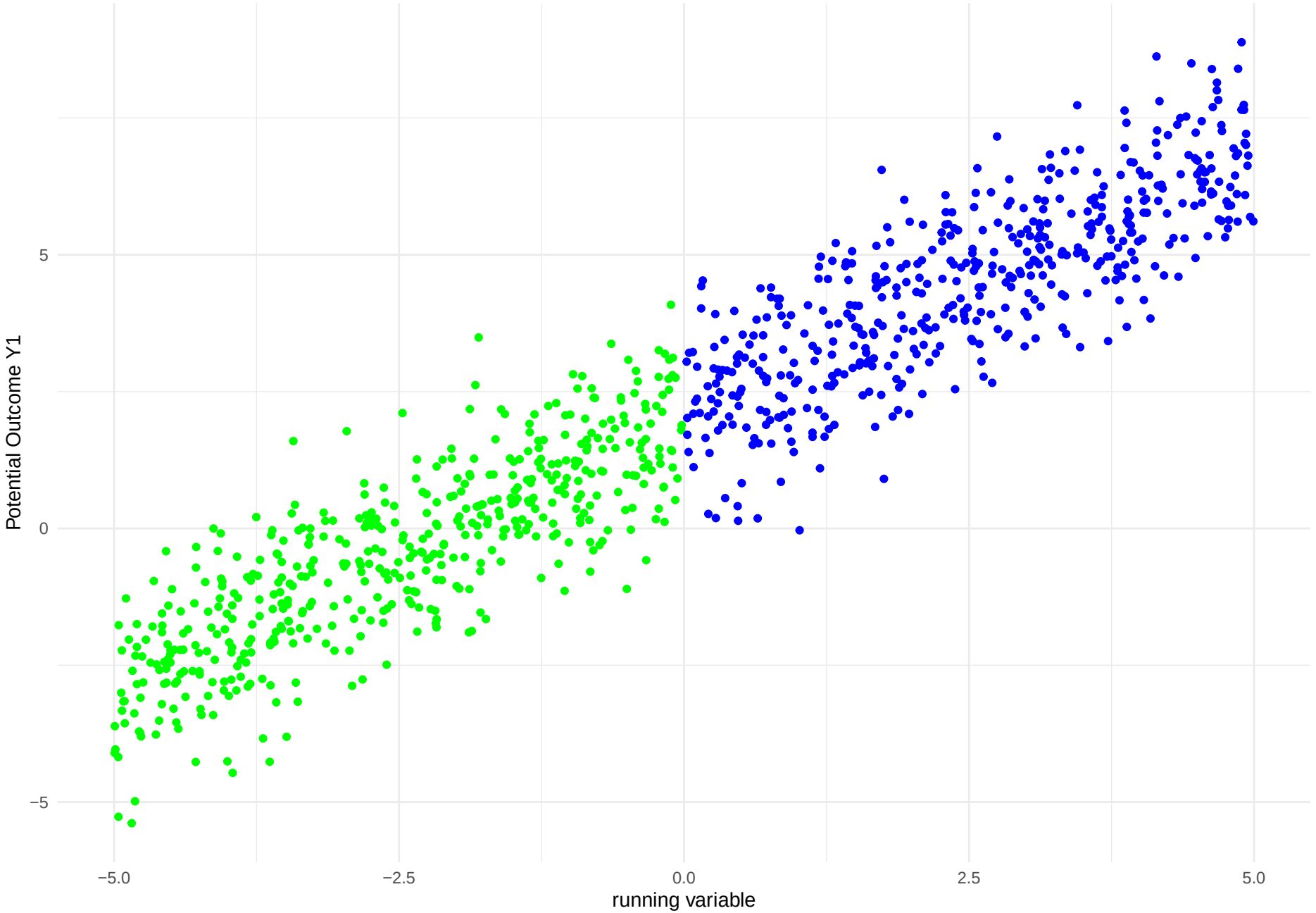
Simulação - Treatment assignment



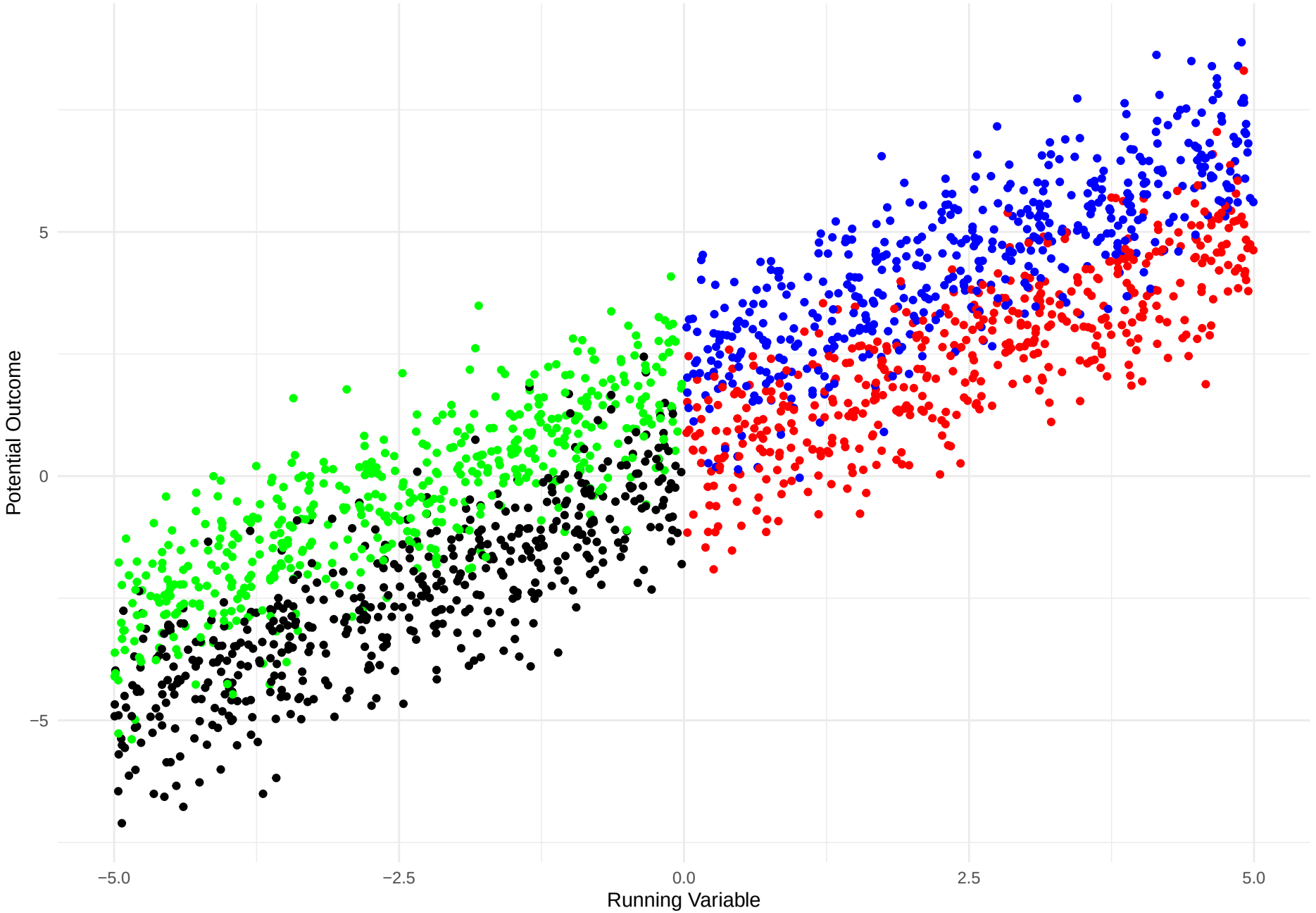
Simulação - Potential Outcomes Y0



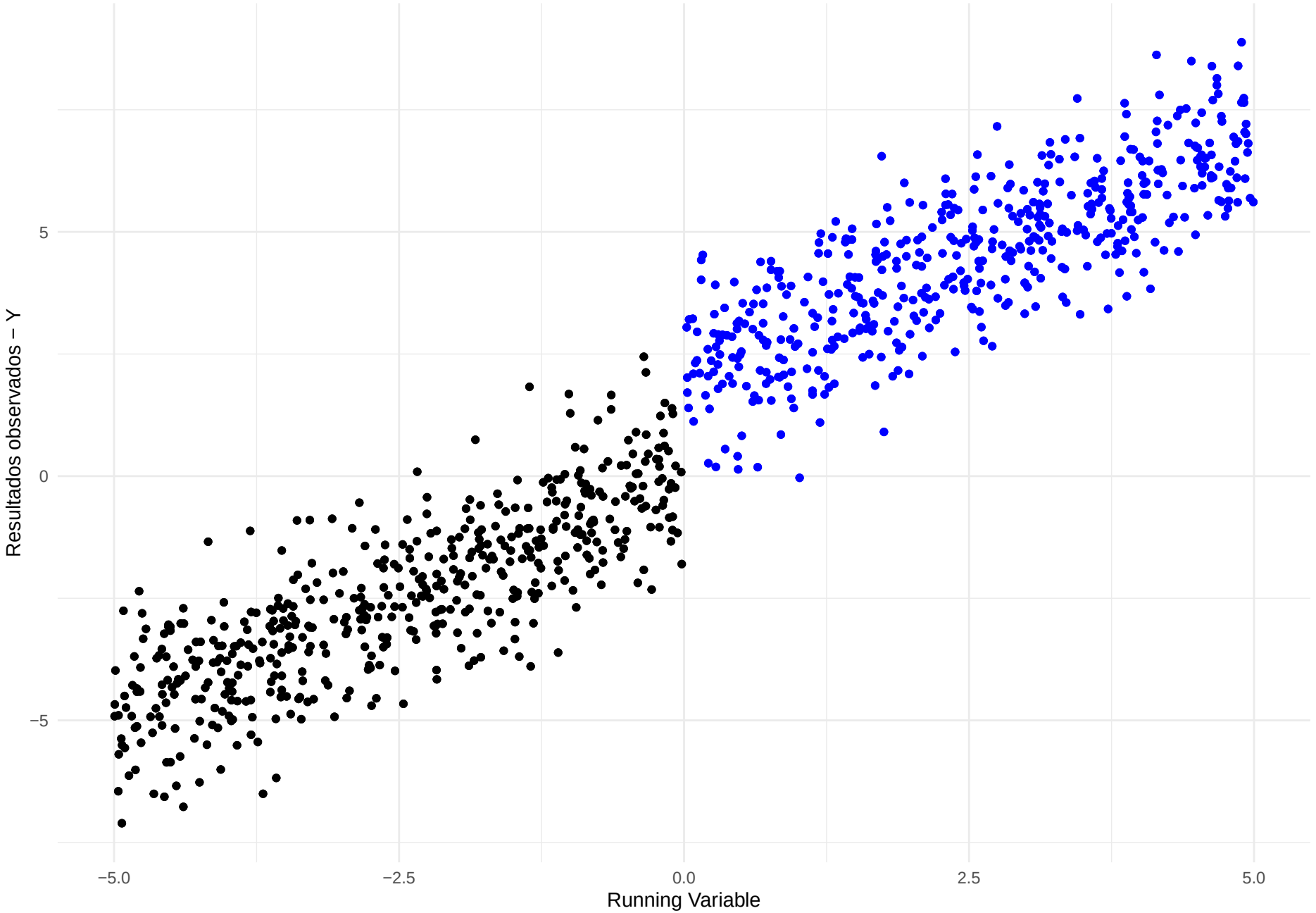
Simulação - Potential Outcomes Y1



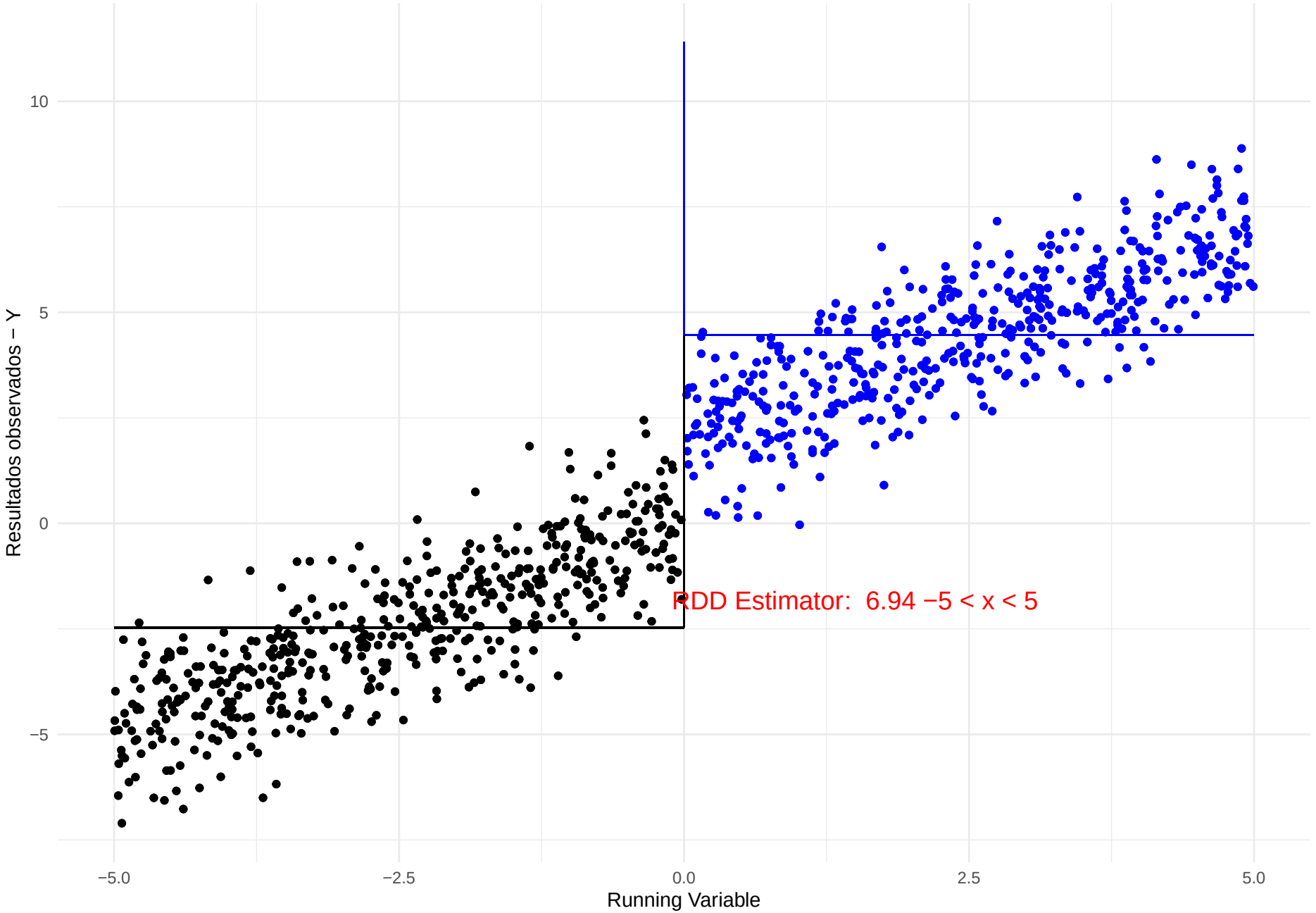
Simulação - Potential Outcomes Y1 e Y0



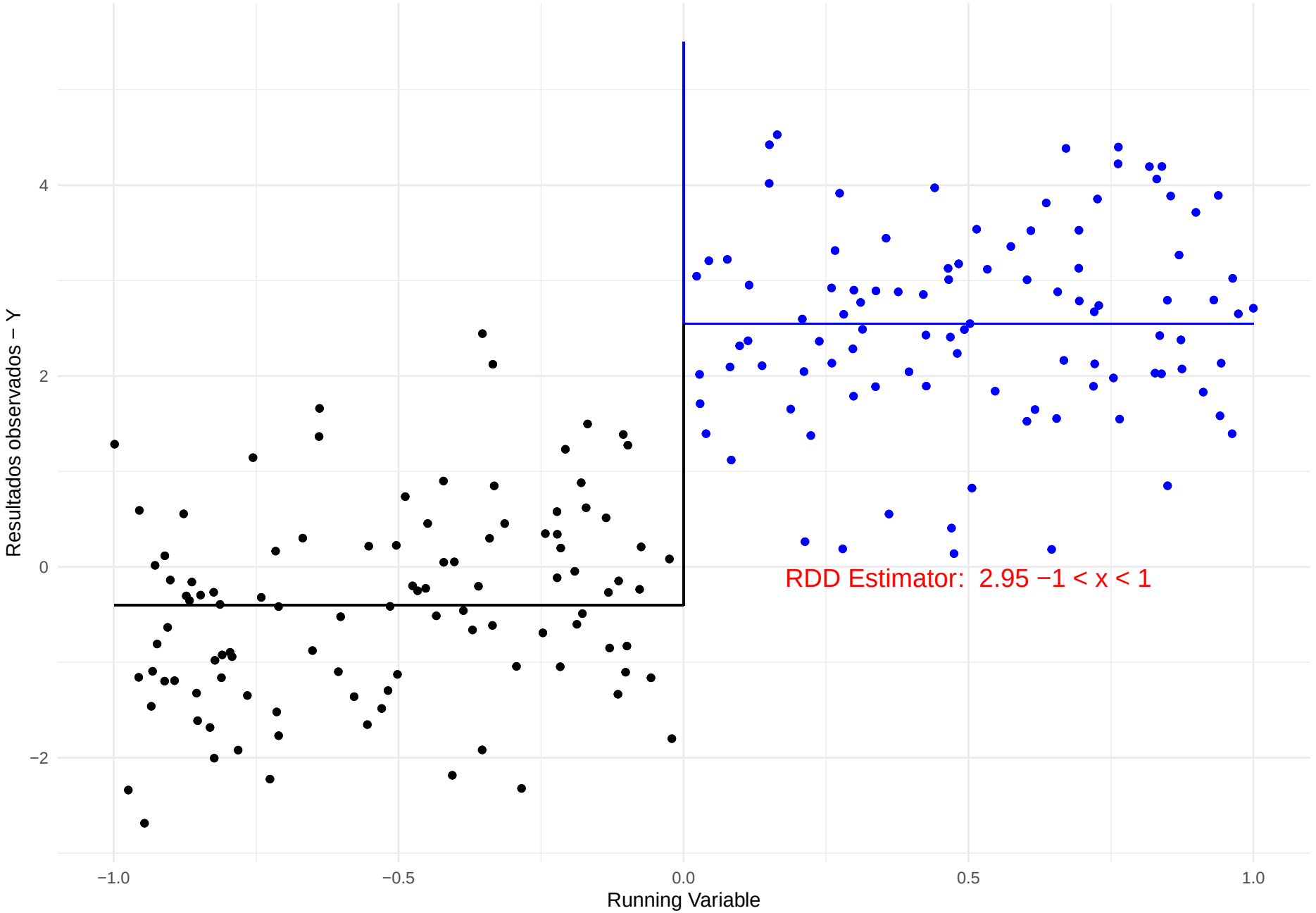
Simulação - Y observado



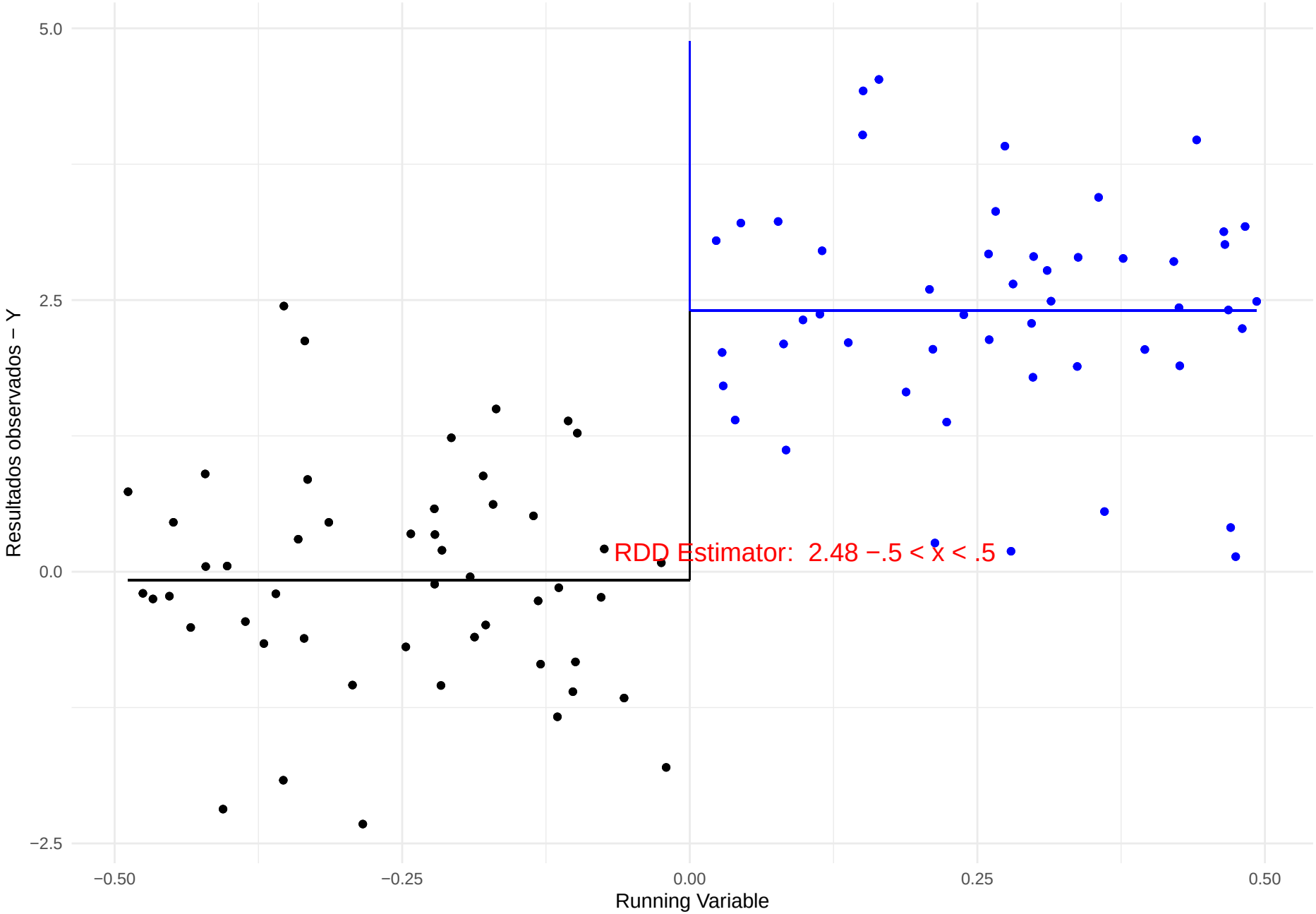
Simulação - Estimativa (1)



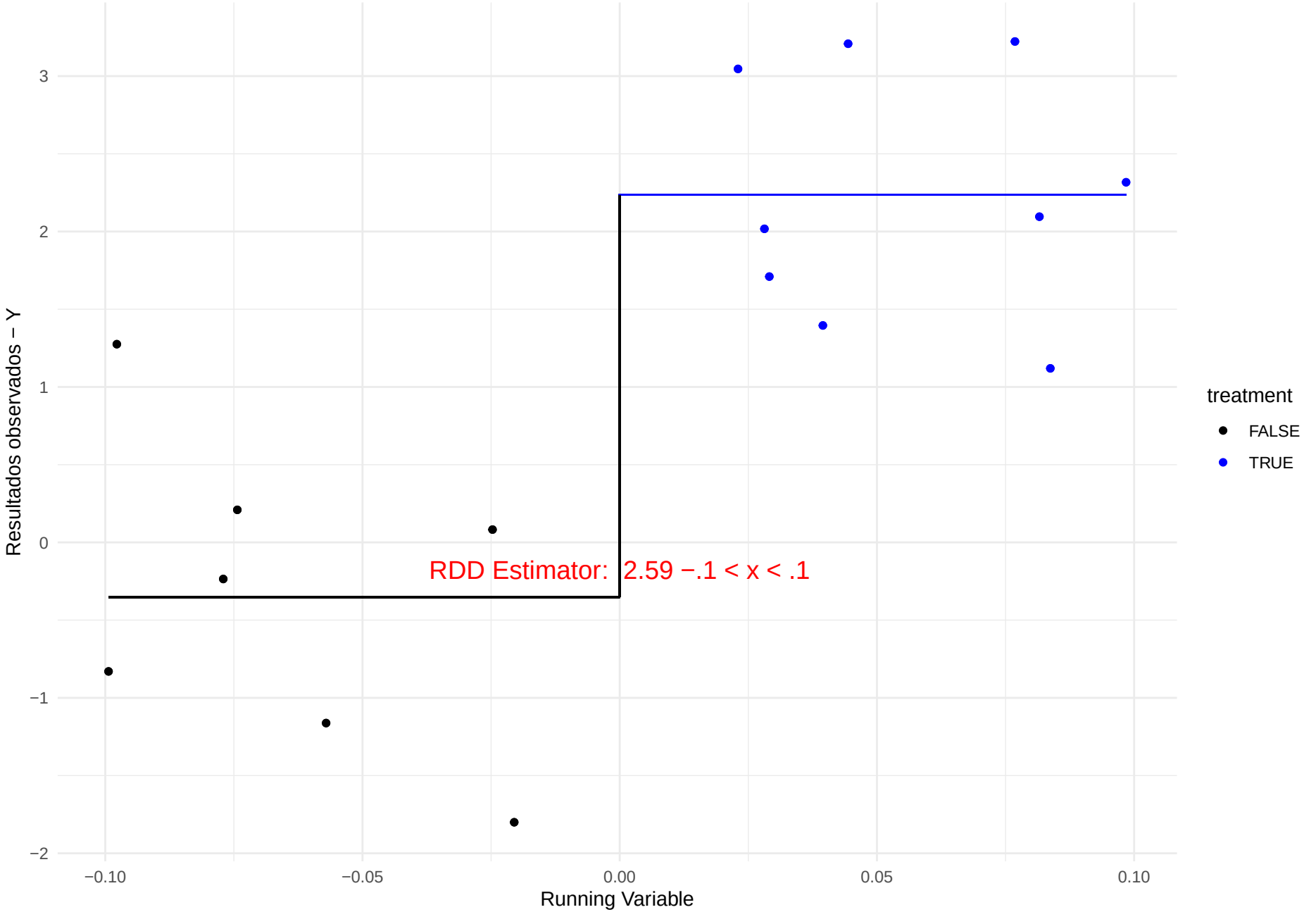
Simulação - Estimativa (2)



Simulação - Estimativa (3)



Simulação - Estimativa (4)

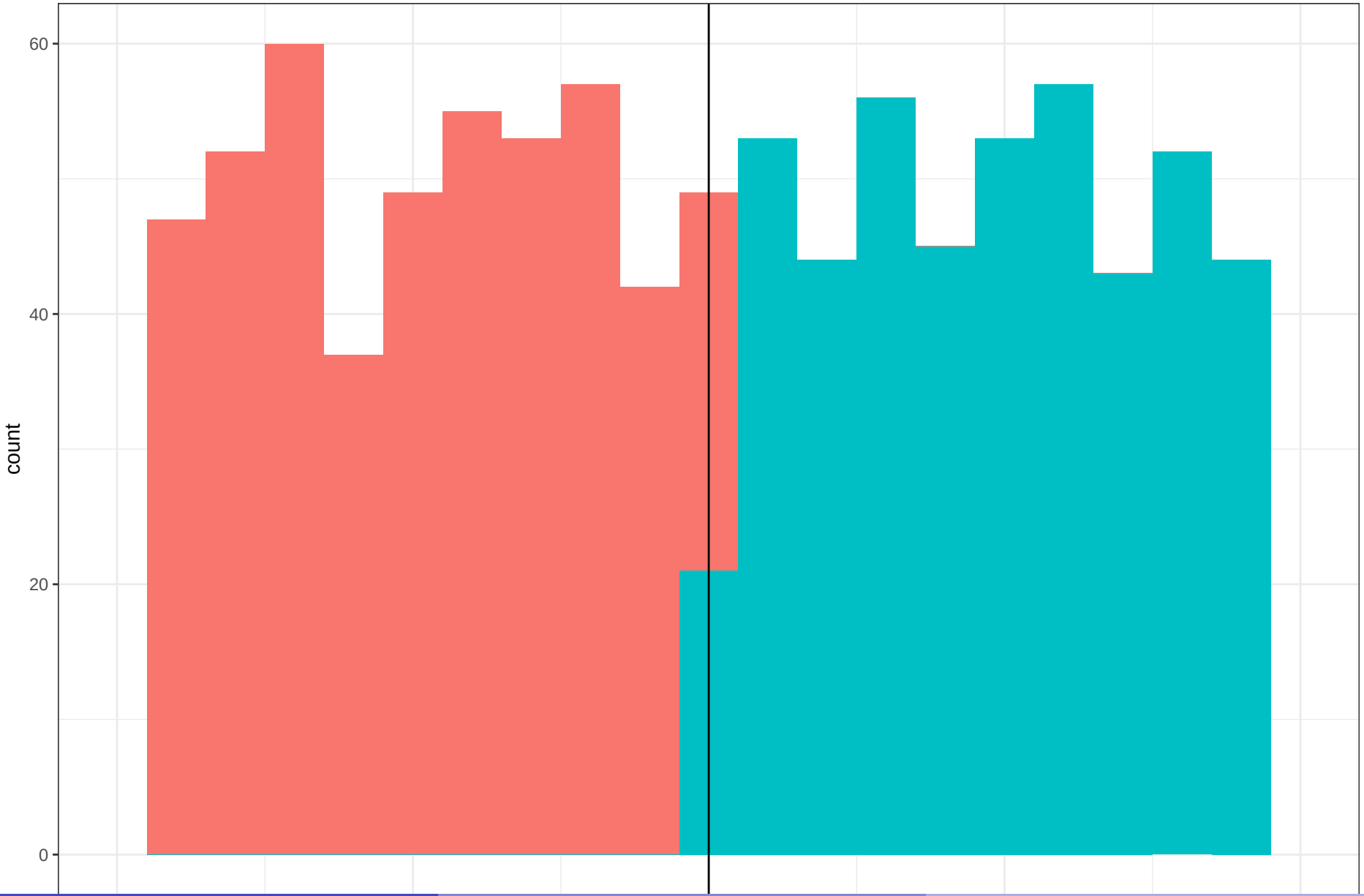


Identificação fácil vs Estimação difícil

- Identificação de RDD (com n infinito) é bem robusto
- Problema é que a estimação depende de extrapolação
- Extrapolação é um problema difícil

Suposição de não-manipulação

– Densidade (não deve ter diferença se não há seleção)



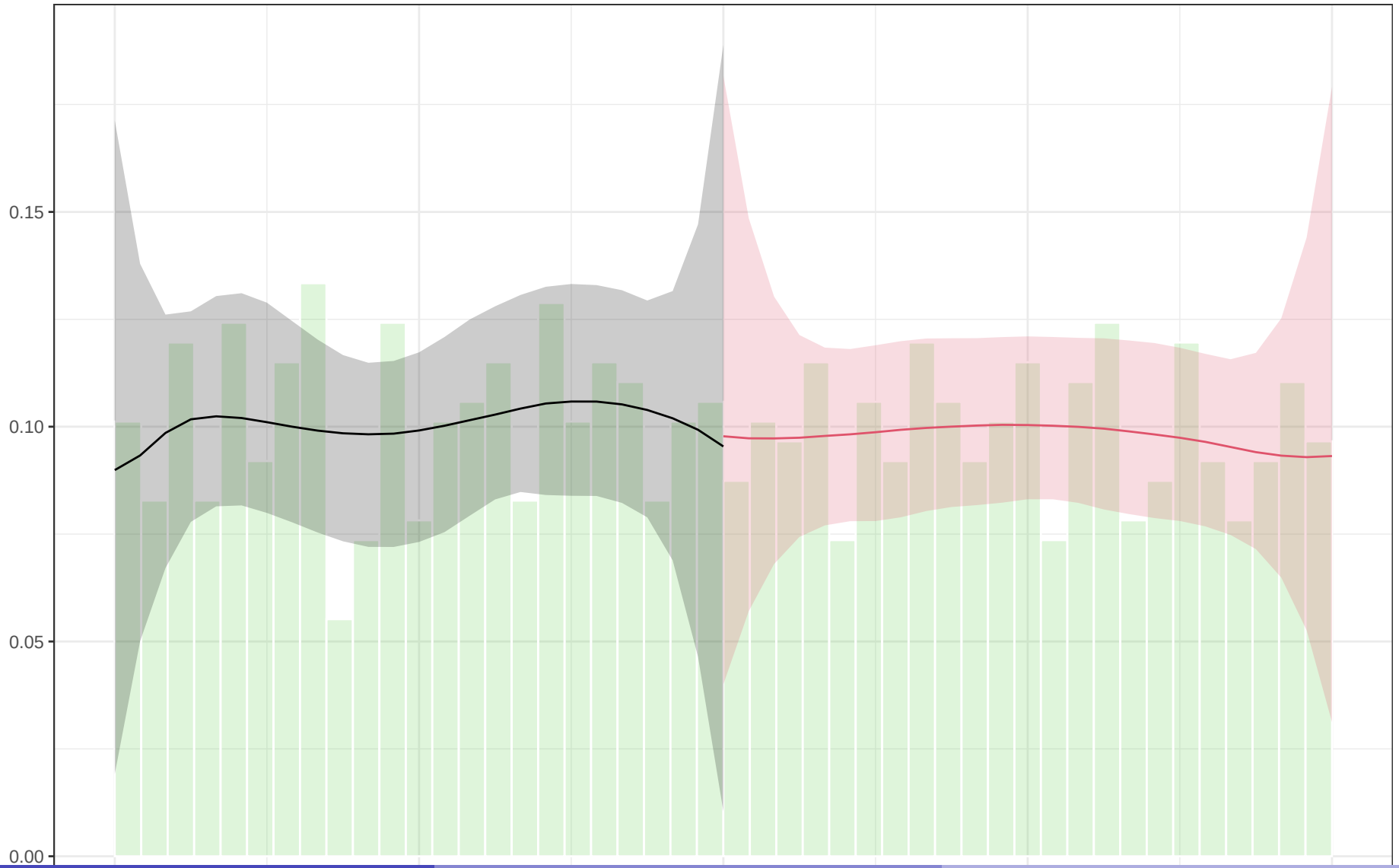
teste formal da densidade

```r
rdd <- rddensity(X = df$x, vce="jackknife")
summary_part1 <- summary(rdd)[1:4]   # Assuming you want to sho
```

```
##
## Manipulation testing using local polynomial density estimat
##
## Number of obs =       1000
## Model =               unrestricted
## Kernel =              triangular
## BW method =           estimated
## VCE method =          jackknife
##
## c = 0                 Left of c           Right of c
## Number of obs         507                 493
## Eff. Number of obs    191                 219
## Order est. (p)        2                   2
## Order bias (q)        3                   3
## BW est. (h)           1.823               2.251
```

Plot da densidade

```
rdplotdensity(rdd, df$x, plotRange = c(-5, 5), plotN = 25, CIu
```



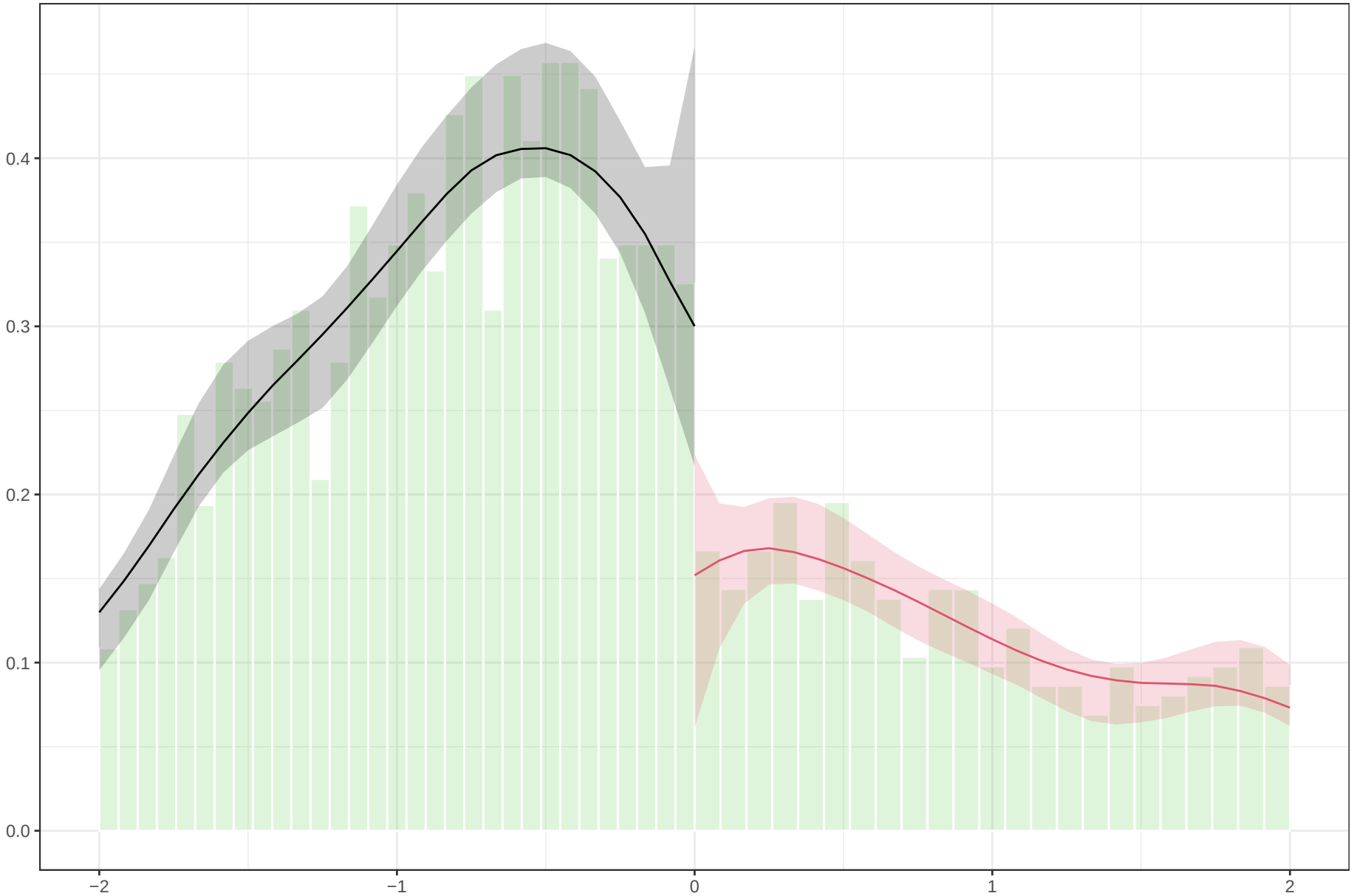
Densidade descontínua - código

```
# Generate a random sample with a density discontinuity at 0
set.seed(42)
x <- rnorm(2000, mean = -0.5)
x[x > 0] <- x[x > 0] * 2

# Estimation
rdd <- rddensity(X = x)
summary(rdd)

# Density plot (from -2 to 2 with 25 evaluation points at eac
plot1 <- rdplotdensity(rdd, x, plotRange = c(-2, 2), plotN =
```

Densidade descontínua - results

```
##
## Manipulation testing using local polynomial density estimat
##
## Number of obs =       2000
## Model =               unrestricted
## Kernel =              triangular
## BW method =           estimated
## VCE method =          jackknife
##
## c = 0                 Left of c           Right of c
## Number of obs         1394                606
## Eff. Number of obs    469                 207
## Order est. (p)        2                   2
## Order bias (q)        3                   3
## BW est. (h)           0.607               0.632
##
## Method                T                   P > |T|
## Robust                -2.5923             0.0095
```

Densidade descontínua - plot



Regressão RDD

```r
library(rdrobust)
# Assuming the cutoff is at x=0
basic_model <- rdrobust(y = df$y, x = df$x, c = 0)
summary(basic_model)
```

```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 1000
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 507          493
## Eff. Number of Obs.            123          109
## Order est. (p)                   1            1
## Order bias  (q)                  2            2
## BW est. (h)                  1.172        1.172
## BW bias (b)                  2.151        2.151
```

Regressão RDD - summary

```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                  1000
## BW type                        mserd
## Kernel                  Triangular
## VCE method                        NN
##
## Number of Obs.                   507          493
## Eff. Number of Obs.              123          109
## Order est. (p)                     1            1
## Order bias  (q)                    2            2
## BW est. (h)                    1.172        1.172
## BW bias (b)                    2.151        2.151
## rho (h/b)                      0.545        0.545
## Unique Obs.                      507          493
##
## ============================================================
##         Method     Coef. Std. Err.          z     P>|z|
```

Exemplo de Descontinuidade