Roteiro de Estudo

Miguel Sallum

22/06/2021

Questão 1

Use GPA3.RAW for this exercise. The data set is for 366 student-athletes from a large university for fall and spring semesters. [A similar analysis is in Maloney and McCormick (1993), but here we use a true panel data set.] Because you have two terms of data for each student, an unobserved effects model is appropriate. The primary question of interest is this: Do athletes perform more poorly in school during the semester their sport is in season?

(a) Use pooled OLS to estimate a model with term GPA (trmgpa) as the dependent variable. The explanatory variables are spring, sat, hsperc, female, black, white, frstsem, tothrs, crsgpa, and season. Interpret the coefficient on season. Is it statistically significant?

```r
data("gpa3")

pdata.frame(gpa3, index = 732) %>%
  plm(trmgpa ~ spring + sat + hsperc + female + black + white + frstsem +
        tothrs + crsgpa + season, data = ., model = "pooling") %>%
  summary()
```

```
## Warning in pdata.frame(gpa3, index = 732): column 'id' overwritten by id index
```

```
## Pooling Model
##
## Call:
## plm(formula = trmgpa ~ spring + sat + hsperc + female + black +
##     white + frstsem + tothrs + crsgpa + season, data = ., model = "pooling")
##
## Balanced Panel: n = 732, T = 1, N = 732
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -1.848992 -0.331324  0.019153  0.380015  1.579236
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept) -1.75284744  0.34790488 -5.0383 5.943e-07 ***
## spring      -0.05800662  0.04803681 -1.2075   0.22762
## sat          0.00169844  0.00014942 11.3671 < 2.2e-16 ***
## hsperc      -0.00866104  0.00103628 -8.3578 3.280e-16 ***
## female       0.35040133  0.05185242  6.7577 2.894e-11 ***
## black       -0.25414949  0.12292159 -2.0676   0.03904 *
## white       -0.02331462  0.11739542 -0.1986   0.84263
## frstsem     -0.03465848  0.07603448 -0.4558   0.64865
## tothrs      -0.00033894  0.00072672 -0.4664   0.64108
```

```
## crsgpa       1.04786549  0.10411440 10.0646 < 2.2e-16 ***
## season      -0.02729036  0.04904604 -0.5564   0.57809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    420.3
## Residual Sum of Squares: 219.58
## R-Squared:      0.47756
## Adj. R-Squared: 0.47032
## F-statistic: 65.907 on 10 and 721 DF, p-value: < 2.22e-16
```

Questão 2

The purpose of this exercise is to compare the estimates and standard errors obtained by correctly using 2SLS with those obtained using inappropriate procedures. Use the data file WAGE2.RAW.

(a) Use a 2SLS routine to estimate the equation

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 black + u$$

where sibs is the IV for educ. Report the results in the usual form

```r
data("wage2")

  ivreg(log(wage) ~ educ + exper + tenure + black |
          sibs + exper + tenure + black, data = wage2) %>%
  summary()
```

```
##
## Call:
## ivreg(formula = log(wage) ~ educ + exper + tenure + black | sibs +
##     exper + tenure + black, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8176 -0.2403  0.0139  0.2567  1.3225
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.215976   0.543451   9.598  < 2e-16 ***
## educ         0.093632   0.033719   2.777  0.00560 **
## exper        0.020922   0.008388   2.494  0.01279 *
## tenure       0.011548   0.002740   4.215 2.74e-05 ***
## black       -0.183329   0.050136  -3.657  0.00027 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3848 on 930 degrees of freedom
## Multiple R-Squared: 0.1685,  Adjusted R-squared: 0.165
## Wald test: 24.92 on 4 and 930 DF,  p-value: < 2.2e-16
```

Questão 3

Use the data in HTV.RAW for this exercise:

(a) Run a simple OLS regression of log(wage) on educ. Without controlling for other factors, what is the 95% confidence interval for the return to another year of education?

```r
data("htv")

htv %>%
  lm(log(wage) ~ educ, .) %>%
  confint( 'educ', level = .95)
```

```
##           2.5 %   97.5 %
## educ 0.08843358 0.114289
```

(b) Now, add to the simple regression model in part (a) a quadratic in experience and a full set of regional dummy variables for current residence and residence at age 18. Also include the urban indicators for current and age 18 residences. What is the estimated return to a year of education?

```r
htv %>%
  lm(log(wage) ~ educ + exper + exper^2 + ne + nc + west + urban +
       ne18 + nc18 + urban18 +west18, .) %>%
  summary()
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + exper^2 + ne + nc + west +
##     urban + ne18 + nc18 + urban18 + west18, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17845 -0.30434  0.03943  0.31318  1.71808
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.162741   0.200078  -0.813   0.4162
## educ         0.134477   0.009045  14.868  < 2e-16 ***
## exper        0.046925   0.007134   6.578 7.07e-11 ***
## ne          -0.017967   0.086188  -0.208   0.8349
## nc           0.001180   0.071016   0.017   0.9867
## west         0.025978   0.080848   0.321   0.7480
## urban        0.209471   0.041700   5.023 5.84e-07 ***
## ne18         0.164642   0.086744   1.898   0.0579 .
## nc18         0.001450   0.072720   0.020   0.9841
## urban18      0.128948   0.048847   2.640   0.0084 **
## west18      -0.028056   0.086499  -0.324   0.7457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5284 on 1219 degrees of freedom
## Multiple R-squared:  0.2144, Adjusted R-squared:  0.208
## F-statistic: 33.28 on 10 and 1219 DF,  p-value: < 2.2e-16
```

(c) Estimate the model from part (b) by IV, using ctuit as an IV for educ. How does the confidence interval for the return to education compare with the OLS CI from part (b)?

```r
htv %$%
  ivreg(log(wage) ~ educ + exper + exper^2 + ne + nc + west + urban +
       ne18 + nc18 + urban18 +west18 |
```

```
      ctuit + exper + exper^2 + ne + nc + west + urban +
      ne18 + nc18 + urban18 +west18) %>%
  summary()
```

```
##
## Call:
## ivreg(formula = log(wage) ~ educ + exper + exper^2 + ne + nc +
##     west + urban + ne18 + nc18 + urban18 + west18 | ctuit + exper +
##     exper^2 + ne + nc + west + urban + ne18 + nc18 + urban18 +
##     west18)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38625 -0.35296  0.03416  0.37996  1.74426
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.90654    2.66935  -1.089   0.2764
## educ         0.27382    0.13546   2.021   0.0435 *
## exper        0.12233    0.07352   1.664   0.0964 .
## ne           0.03840    0.10891   0.353   0.7244
## nc           0.03971    0.08614   0.461   0.6449
## west        -0.05801    0.12017  -0.483   0.6294
## urban        0.22543    0.04813   4.683 3.14e-06 ***
## ne18         0.06999    0.13196   0.530   0.5959
## nc18        -0.04592    0.09180  -0.500   0.6170
## urban18      0.27036    0.14714   1.837   0.0664 .
## west18       0.03937    0.11495   0.343   0.7320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5775 on 1219 degrees of freedom
## Multiple R-Squared: 0.0615,  Adjusted R-squared: 0.0538
## Wald test: 9.759 on 10 and 1219 DF,  p-value: 7.884e-16
```

Questão 4

For this exercise, we use JTRAIN.RAW to determine the effect of the job training grant on hours of job training per employee. The basic model for the three years is:

$$hrsemp_{it} = \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} + \beta_2 grant_{i,t-1} + \beta_3 log(employ_{it}) + a_i + u_{it}$$

(a) Estimate the equation using first differencing.

```
data("jtrain")

pdata.frame(jtrain) %>%
  plm(hrsemp ~ d88 + d89 + grant + grant_1 + log(employ),
      data = ., model = "fd") %>%
  summary()
```

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = hrsemp ~ d88 + d89 + grant + grant_1 + log(employ),
```

```
##     data = ., model = "fd")
##
## Unbalanced Panel: n = 3, T = 127-134, N = 390
## Observations used in estimation: 387
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -157.41567  -11.75941   -0.27834   11.30032  152.19533
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept)  -0.0061314   1.6344980 -0.0038 0.9970089
## grant        16.6353946   5.6176572  2.9613 0.0032546 **
## grant_1     -13.3854667   8.6738846 -1.5432 0.1236101
## log(employ)  -4.2604781   1.0978787 -3.8806 0.0001226 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    423940
## Residual Sum of Squares: 395950
## R-Squared:      0.066029
## Adj. R-Squared: 0.058713
## F-statistic: 9.02564 on 3 and 383 DF, p-value: 8.6751e-06
```

Questão 5

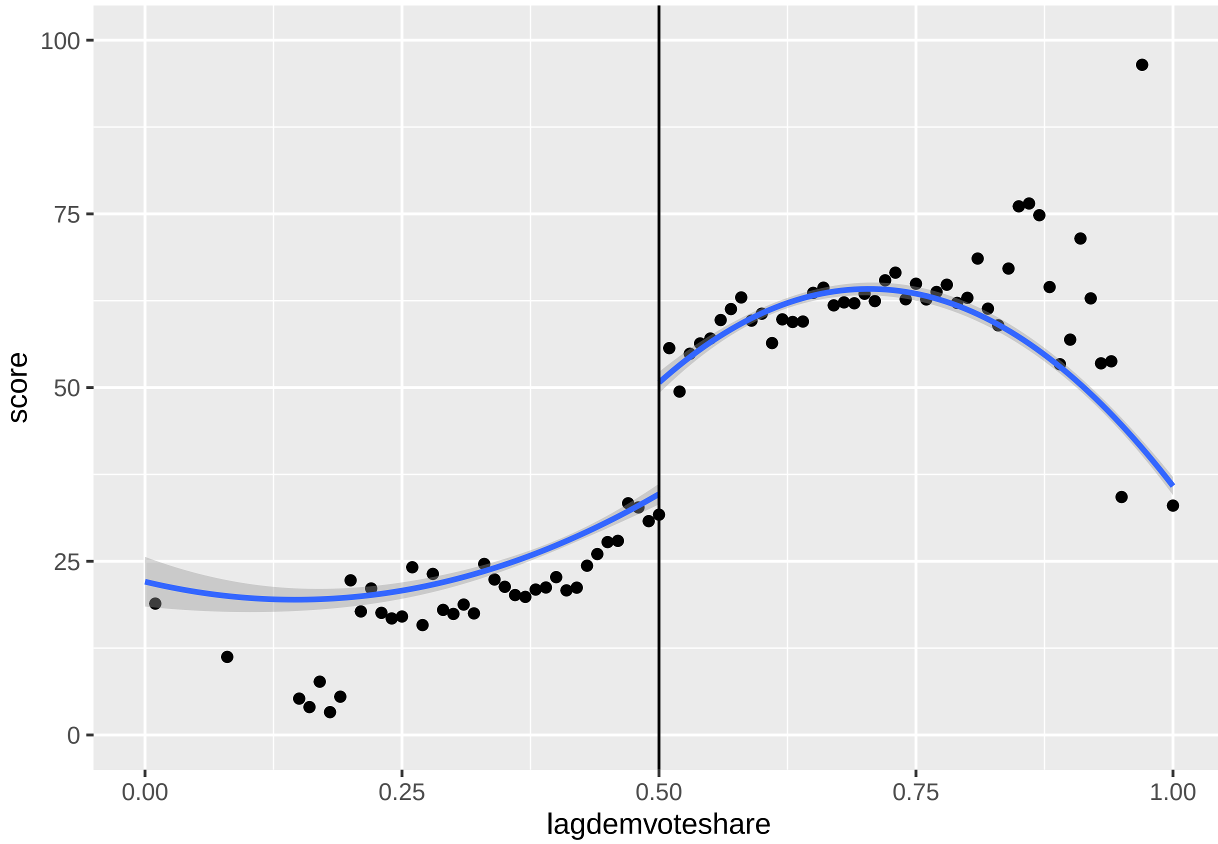
Replicação do gráfico RDD do livro *"Causal Inference: The Mixtape"*

```r
read_data <- function(df)
{
  full_path <- paste("https://raw.github.com/scunning1975/mixtape/master/",
                     df, sep = "")
  df <- read_dta(full_path)
  return(df)
}

lmb_data <- read_data("lmb-data.dta")

#aggregating the data
categories <- lmb_data$lagdemvoteshare

demmeans <- split(lmb_data$score, cut(lmb_data$lagdemvoteshare, 100)) %>%
  lapply(mean) %>%
  unlist()

agg_lmb_data <- data.frame(score = demmeans, lagdemvoteshare = seq(0.01,1, by = 0.01))

#plotting
lmb_data <- lmb_data %>%
  mutate(gg_group = case_when(lagdemvoteshare > 0.5 ~ 1, TRUE ~ 0))

ggplot(lmb_data, aes(lagdemvoteshare, score)) +
  geom_point(aes(x = lagdemvoteshare, y = score), data = agg_lmb_data) +
  stat_smooth(aes(lagdemvoteshare, score, group = gg_group), method = "lm",
```

```
               formula = y ~ x + I(x^2)) +
  xlim(0,1) + ylim(0,100) +
  geom_vline(xintercept = 0.5)
```

```
## Warning: Removed 1093 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 15 rows containing missing values (geom_point).
```

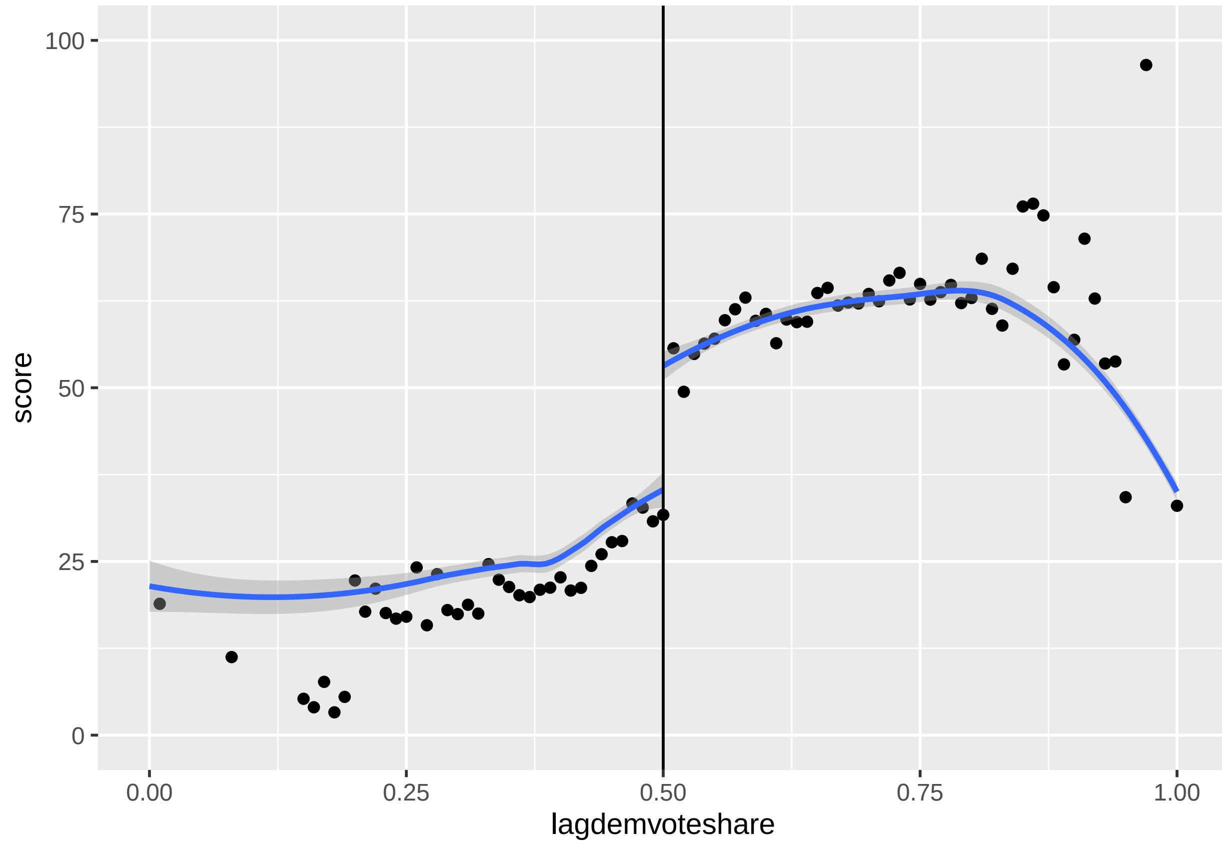



```
ggplot(lmb_data, aes(lagdemvoteshare, score)) +
  geom_point(aes(x = lagdemvoteshare, y = score), data = agg_lmb_data) +
  stat_smooth(aes(lagdemvoteshare, score, group = gg_group), method = "loess") +
  xlim(0,1) + ylim(0,100) +
  geom_vline(xintercept = 0.5)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

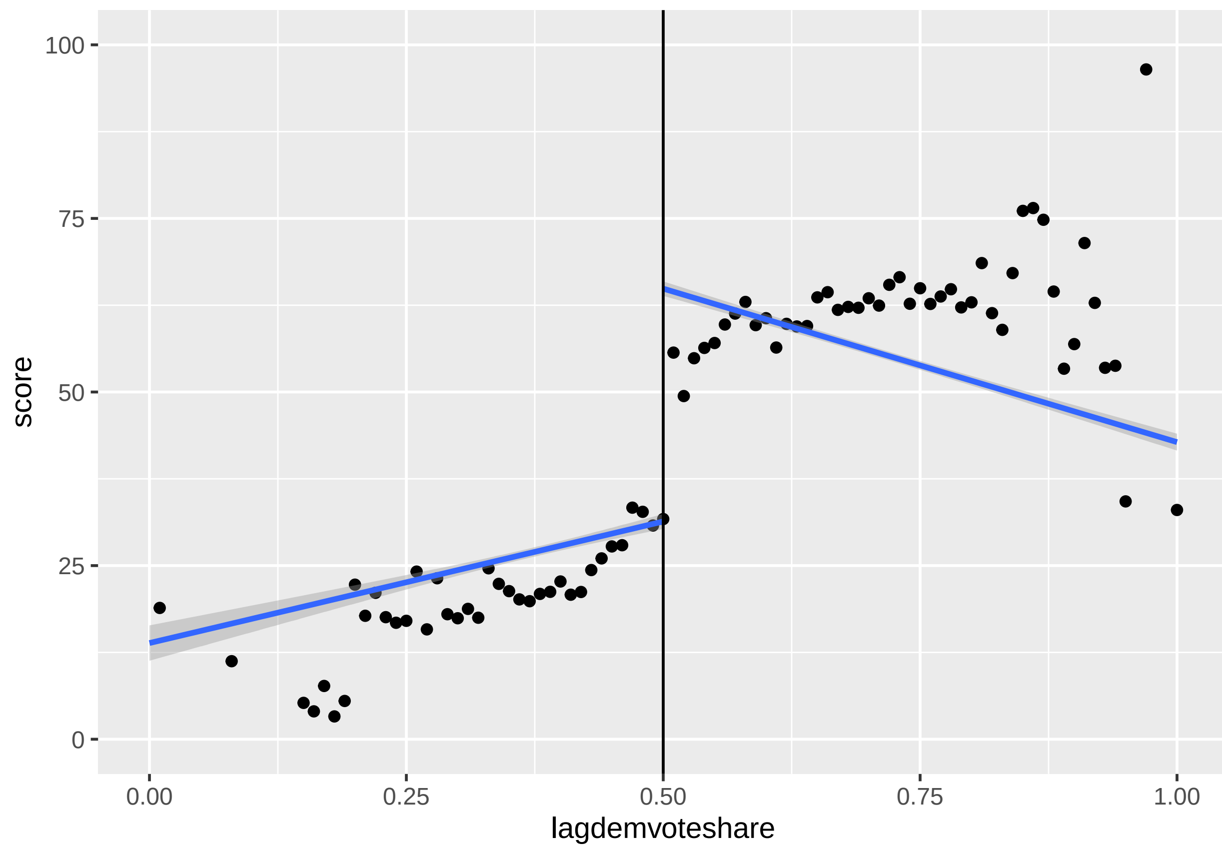
```
## Warning: Removed 1093 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 15 rows containing missing values (geom_point).
```




```
ggplot(lmb_data, aes(lagdemvoteshare, score)) +
  geom_point(aes(x = lagdemvoteshare, y = score), data = agg_lmb_data) +
  stat_smooth(aes(lagdemvoteshare, score, group = gg_group), method = "lm") +
  xlim(0,1) + ylim(0,100) +
  geom_vline(xintercept = 0.5)
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 1093 rows containing non-finite values (stat_smooth).

## Warning: Removed 15 rows containing missing values (geom_point).
```



Questão 6

Estimando DIFF IN DIFF: Replicação do seguinte exercício. https://www.princeton.edu/~otorres/DID101R.pdf

```
read_dta("http://dss.princeton.edu/training/Panel101.dta") %>%
  mutate(treatment_time = as.numeric(year >= 1994),
         treated = as.numeric(country > 4),
         did = treated*treatment_time) %>%
  lm(y ~ treatment_time + treated + did, .) %>%
  summary()
```

```
##
## Call:
## lm(formula = y ~ treatment_time + treated + did, data = .)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.768e+09 -1.623e+09  1.167e+08  1.393e+09  6.807e+09
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.581e+08  7.382e+08   0.485   0.6292
## treatment_time  2.289e+09  9.530e+08   2.402   0.0191 *
## treated        1.776e+09  1.128e+09   1.575   0.1200
## did           -2.520e+09  1.456e+09  -1.731   0.0882 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.953e+09 on 66 degrees of freedom
## Multiple R-squared:  0.08273,    Adjusted R-squared:  0.04104
## F-statistic: 1.984 on 3 and 66 DF,  p-value: 0.1249
```

```
read_dta("http://dss.princeton.edu/training/Panel101.dta") %>%
  mutate(treatment_time = as.numeric(year >= 1994),
         treated = as.numeric(country > 4)) %>%
  lm(y ~ treatment_time*treated, .) %>%
  summary()
```

```
##
## Call:
## lm(formula = y ~ treatment_time * treated, data = .)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.768e+09 -1.623e+09  1.167e+08  1.393e+09  6.807e+09
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.581e+08  7.382e+08   0.485   0.6292
## treatment_time         2.289e+09  9.530e+08   2.402   0.0191 *
## treated                1.776e+09  1.128e+09   1.575   0.1200
## treatment_time:treated -2.520e+09  1.456e+09  -1.731   0.0882 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.953e+09 on 66 degrees of freedom
## Multiple R-squared:  0.08273,    Adjusted R-squared:  0.04104
## F-statistic: 1.984 on 3 and 66 DF,  p-value: 0.1249
```

Questão 7

Replicação do seguinte exercício RDD: http://erikgahner.dk/slides/2015-aas/12-rdd.pdf

```
data("house")

house %>%
  rdd_data(y=y, x=x, cutpoint=0, data=.) %>%
  rdd_reg_lm()
```

```
## ### RDD regression: parametric ###
##  Polynomial order:  1
##  Slopes:  separate
##  Number of obs: 6558 (left: 2740, right: 3818)
##
##  Coefficient:
##    Estimate Std. Error t value  Pr(>|t|)
## D 0.1182314  0.0056799  20.816 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
house %>%
  rdd_data(y=y, x=x, cutpoint=0, data=.) %>%
```

```r
  rdd_reg_lm(., bw = rdd_bw_ik(.))
```

```
## ### RDD regression: parametric ###
##  Polynomial order:  1
##  Slopes:  separate
##  Bandwidth:  0.2938561
##  Number of obs: 3200 (left: 1594, right: 1606)
##
##  Coefficient:
##    Estimate Std. Error t value  Pr(>|t|)
## D 0.0823378  0.0080236  10.262 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

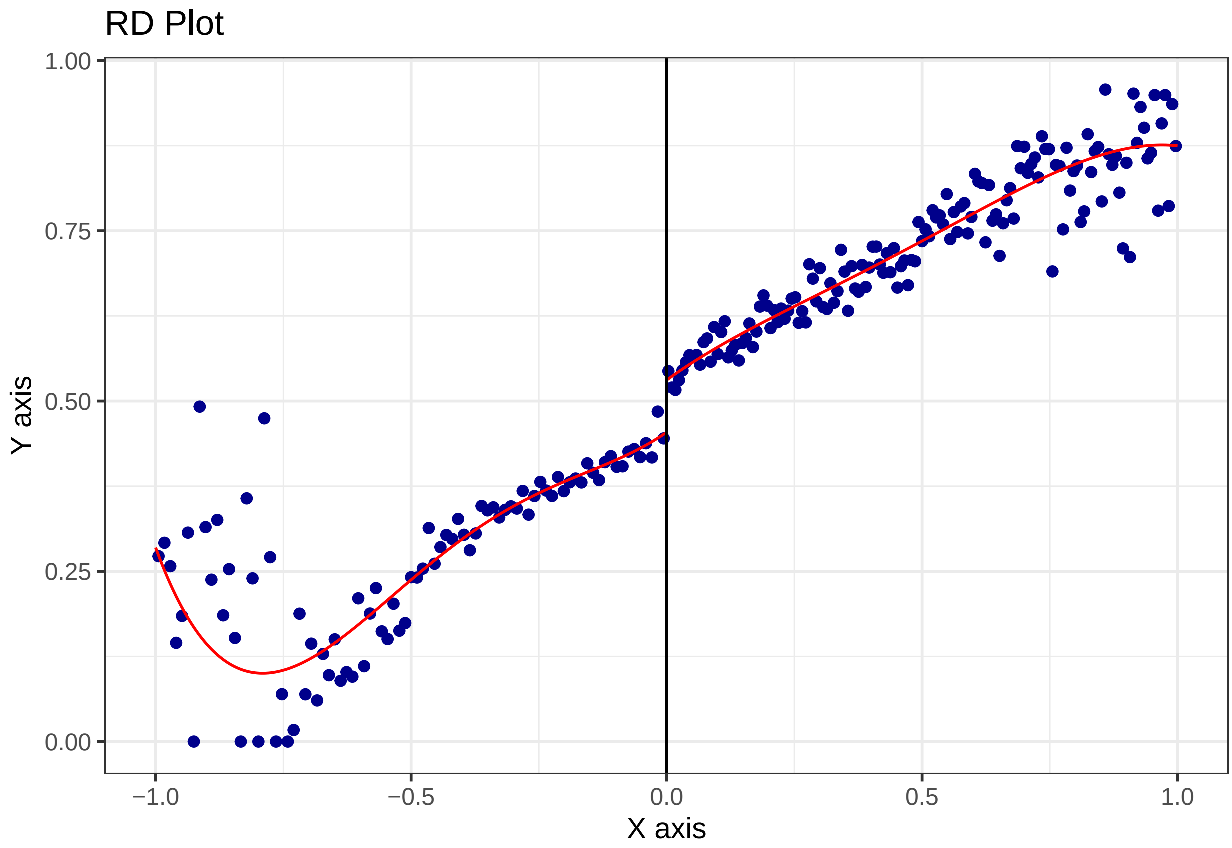
```r
house %$%
  rdrobust(y, x)
```

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

```
## Call: rdrobust
##
## Number of Obs.                 6558
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 2740         3818
## Eff. Number of Obs.             789          817
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   0.136        0.136
## BW bias (b)                   0.240        0.240
## rho (h/b)                     0.565        0.565
## Unique Obs.                    2108         2581
```

```r
house %$%
  rdplot(y, x)
```

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


RD Plot
1.00
0.75
0.50
0.25
0.00
Y axis
−1.0
−0.5
0.0
0.5
1.0
X axis