Lab #18 - More on Regression Discontinuity

Econ 224

November 8th, 2018

Part I: Handling Non-linearity

A problem with the linear RD model that we used in our last lab is that it can be "fooled" by a non-linear trend. As we discussed last time, RD relies on the fact that any discontinuity in the relationship between X and Y at the cutoff c indicates a causal effect of D on Y. This is because we assume that both $\mathbb{E}[Y_0|X]$ and $\mathbb{E}[Y_1|X]$ are continuous when X is close to c. A rapid change in Y near the cutoff is *not* the same thing as a discontinuity and provides no evidence of a causal effect. But a linear RD model can have a hard time distinguishing such a non-linear relationship from a genuine discontinuity. You will explore this problem along with a simple solution to it in the following exercise.

Exercise 1

This exercise relies on the following simulation code:

```
set.seed(1234)
n <- 100
x <- runif(n)
y <- pnorm(x, 0.5, 0.1) + rnorm(n, sd = 0.1)
D <- 1 * (x >= 0.5)
```

(a) What is the RD cutoff in the simulation design?
(b) What are $E[Y_0|X]$ and $E[Y_1|X]$ in the simulation design?
(c) What is the true value of the RD causal effect in the simulation design?
(d) Use the data from the simulation experiment to fit a linear RD model. Summarize your results. Do you find evidence of a causal effect of D on Y? Calculate a 95% confidence interval for this effect.
(e) Make a plot of your results from (d) along with $E[Y_0|X]$ and $E[Y_1|X]$. Comment on your findings.
(f) Building on your derivations in Lab 17, figure out how to fit a *quadratic* RD model in R; rather than fitting two different linear relationships, fit two different *quadratic* relationships.
(g) How do your results from (f) change if you use a quadratic rather than linear RD specification?

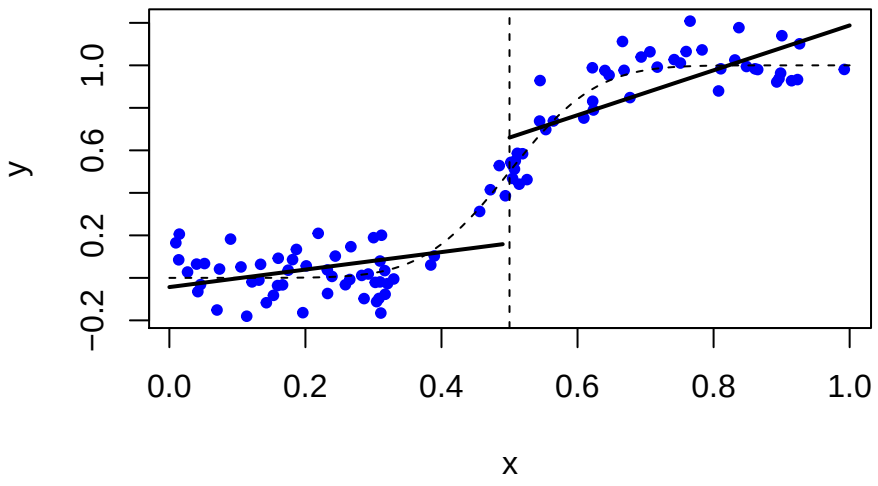
Solutions

(a) The cutoff is 0.5.
(b) These two functions are identical: `pnorm`$(x, 0.5, 0.1)$ i.e. the CDF of a normal RV with mean 0.5 and standard deviation 0.1.
(c) Zero since $E[Y_0|X] = E[Y_1|X]$.

```
# Linear RD results
xtilde <- x - 0.5
rd1 <- lm(y ~ D + xtilde + xtilde:D)
summary(rd1)
```

```
Call:
lm(formula = y ~ D + xtilde + xtilde:D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25014 -0.10356 -0.01141  0.10083  0.37186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.16249    0.04554   3.568 0.000563 ***
D            0.49758    0.05745   8.661 1.11e-13 ***
xtilde       0.41180    0.14794   2.784 0.006476 **
D:xtilde     0.64377    0.20280   3.174 0.002018 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1378 on 96 degrees of freedom
Multiple R-squared:  0.9092,    Adjusted R-squared:  0.9064
F-statistic: 320.4 on 3 and 96 DF,  p-value: < 2.2e-16
```

```r
# Plot of linear RD along with E[Y0|X] and E[Y1|X] which are the same!
plot(x, y, pch = 20, col = 'blue')
xseq <- seq(0, 1, 0.01)
points(xseq, pnorm(xseq, 0.5, 0.1), type = 'l', lty = 2)
abline(v = 0.5, lty = 2)
x_before <- seq(0, 0.5 - 0.01, 0.01)
y_before <- predict(rd1, data.frame(xtilde = x_before - 0.5,
                                    D = 1 * (x_before >= 0.5)))
x_after <- seq(0.5, 1, 0.01)
y_after <- predict(rd1, data.frame(xtilde = x_after - 0.5,
                                   D = 1 * (x_after >= 0.5)))
points(x_before, y_before, type = 'l', lwd = 2)
points(x_after, y_after, type = 'l', lwd = 2)
```




```r
# Quadratic RD results
rd2 <- lm(y ~ D + xtilde + I(xtilde^2) + xtilde:D + I(xtilde^2):D)
summary(rd2)
```

```
Call:
lm(formula = y ~ D + xtilde + I(xtilde^2) + xtilde:D + I(xtilde^2):D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20652 -0.06123 -0.00373  0.06043  0.25875

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.42341    0.05252   8.062 2.37e-12 ***
D               0.09839    0.06137   1.603    0.112
xtilde          2.89289    0.40472   7.148 1.87e-10 ***
I(xtilde^2)     4.60454    0.72469   6.354 7.45e-09 ***
D:xtilde        0.65788    0.54328   1.211    0.229
D:I(xtilde^2) -10.43831    1.09033  -9.574 1.48e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09913 on 94 degrees of freedom
Multiple R-squared:  0.954, Adjusted R-squared:  0.9516
F-statistic: 390.1 on 5 and 94 DF,  p-value: < 2.2e-16
```

Part II: Empirical Exercise

In this part you will apply what you have learned above to data from the MLDA example from MM. Before beginning the following exercises, first download the MLDA dataset `AEJfigs.dta` from the Mastering 'Metrics website under "Chapter 4" and use an appropriate package to convert this file and load it in R. The only two variables you will need for your analysis are `agecell` which gives age in years (with a decimal point since the ages are binned) and `all` which gives mortality rates per 100,000 individuals.

Exercise

(a) Use a linear RD model to estimate the causal effect of legal access to alcohol on death rates. Plot your results and carry out appropriate present appropriate statistical inference. Discuss your findings.
(b) Repeat (a) using a *quadratic* rather than linear specification. Compare and contrast your findings.
(c) RD analysis is fundamentally *local* in nature: the mortality rates of individuals far from the cutoff should not inform us about the causal effect for 21 year olds. Check the sensitivity of your results from parts (a) and (b) by restricting your sample to ages between 20 and 22, inclusive. Discuss your findings.

```
# Load data
library(tidyverse)
library(haven)
mlda <- read_dta('~/econ224/labs/mlda.dta')
mlda
```

```
# A tibble: 50 x 19
   agecell   all allfitted internal internalfitted external externalfitted
     <dbl> <dbl>     <dbl>    <dbl>          <dbl>    <dbl>          <dbl>
 1    19.1  92.8      91.7     16.6           16.7     76.2           75.0
```

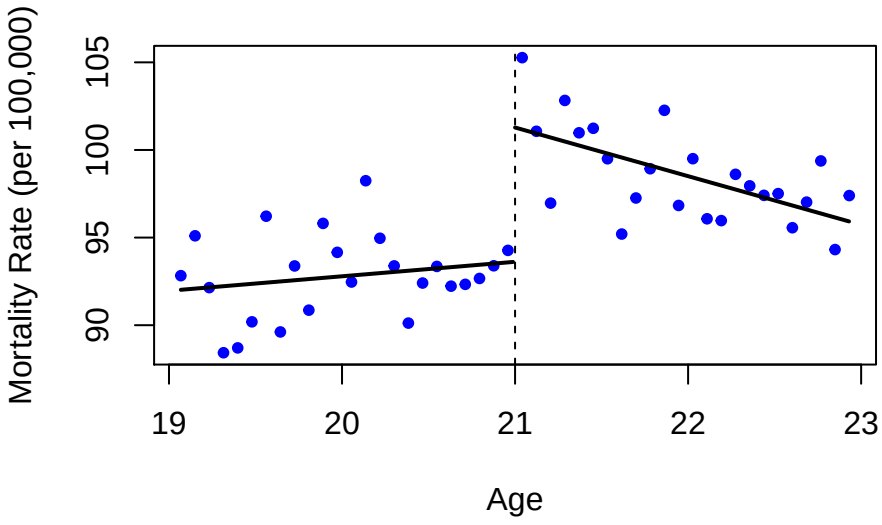
```
 2    19.2  95.1      91.9     18.3           16.9     76.8           75.0
 3    19.2  92.1      92.0     18.9           17.1     73.2           75.0
 4    19.3  88.4      92.2     16.1           17.3     72.3           74.9
 5    19.4  88.7      92.3     17.4           17.4     71.3           74.9
 6    19.5  90.2      92.5     17.9           17.6     72.3           74.9
 7    19.6  96.2      92.6     16.4           17.8     79.8           74.8
 8    19.6  89.6      92.7     16.0           17.9     73.6           74.8
 9    19.7  93.4      92.8     17.4           18.1     75.9           74.7
10    19.8  90.9      92.9     18.3           18.2     72.6           74.6
# ... with 40 more rows, and 12 more variables: alcohol <dbl>,
#   alcoholfitted <dbl>, homicide <dbl>, homicidefitted <dbl>,
#   suicide <dbl>, suicidefitted <dbl>, mva <dbl>, mvafitted <dbl>,
#   drugs <dbl>, drugsfitted <dbl>, externalother <dbl>,
#   externalotherfitted <dbl>
```

```r
# Center age around the cutoff and create treatment indicator
mlda <- mlda %>% mutate(age = agecell - 21,
                        over21 = 1 * (agecell >= 21))

# Linear RD model
linear <-  lm(all ~ over21 + age + age:over21, mlda)
summary(linear)
```

```

Call:
lm(formula = all ~ over21 + age + age:over21, data = mlda)

Residuals:
   Min     1Q Median     3Q    Max
-4.368 -1.787  0.117  1.108  5.341

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.6184     0.9325 100.399  < 2e-16 ***
over21        7.6627     1.3187   5.811  6.4e-07 ***
age           0.8270     0.8189   1.010  0.31809
over21:age   -3.6034     1.1581  -3.111  0.00327 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.283 on 44 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.6677,    Adjusted R-squared:  0.645
F-statistic: 29.47 on 3 and 44 DF,  p-value: 1.325e-10
```

```r
# Function for making an RD plot
make_RD_plot <- function(reg, dat, inc = 0.01) {
  plot(all ~ agecell, dat, xlab = 'Age',
       ylab = 'Mortality Rate (per 100,000)',
       pch = 20, col = 'blue')
  abline(v = 21, lty = 2)
  x_min <- min(dat$agecell)
  x_max <- max(dat$agecell)
```

```
  x_before <- seq(x_min, 21 - inc, inc)
  x_after <- seq(21, x_max, inc)
  y_before <- predict(reg, data.frame(age = x_before - 21,
                                      over21 = 1 * (x_before >= 21)))
  y_after <- predict(reg, data.frame(age = x_after - 21,
                                     over21 = 1 * (x_after >= 21)))
  points(x_before, y_before, type = 'l', lwd = 2)
  points(x_after, y_after, type = 'l', lwd = 2)
}

make_RD_plot(linear, mlda)
```

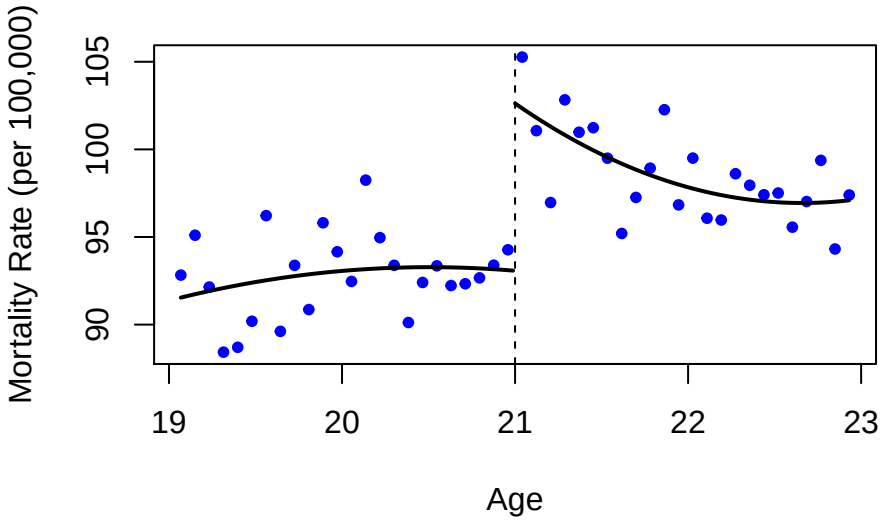



```
# Quadratic RD Model
quadratic <-  lm(all ~ over21 + age + I(age^2) +
                   age:over21 + I(age^2):over21, mlda)
summary(quadratic)
```

```
Call:
lm(formula = all ~ over21 + age + I(age^2) + age:over21 + I(age^2):over21,
    data = mlda)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3343 -1.3946  0.1849  1.2848  5.0817

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     93.0729     1.4038  66.301  < 2e-16 ***
over21           9.5478     1.9853   4.809 1.97e-05 ***
age             -0.8306     3.2901  -0.252    0.802
I(age^2)        -0.8403     1.6153  -0.520    0.606
over21:age      -6.0170     4.6529  -1.293    0.203
over21:I(age^2)  2.9042     2.2843   1.271    0.211
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.285 on 42 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.6821,    Adjusted R-squared:  0.6442
F-statistic: 18.02 on 5 and 42 DF,  p-value: 1.624e-09
```

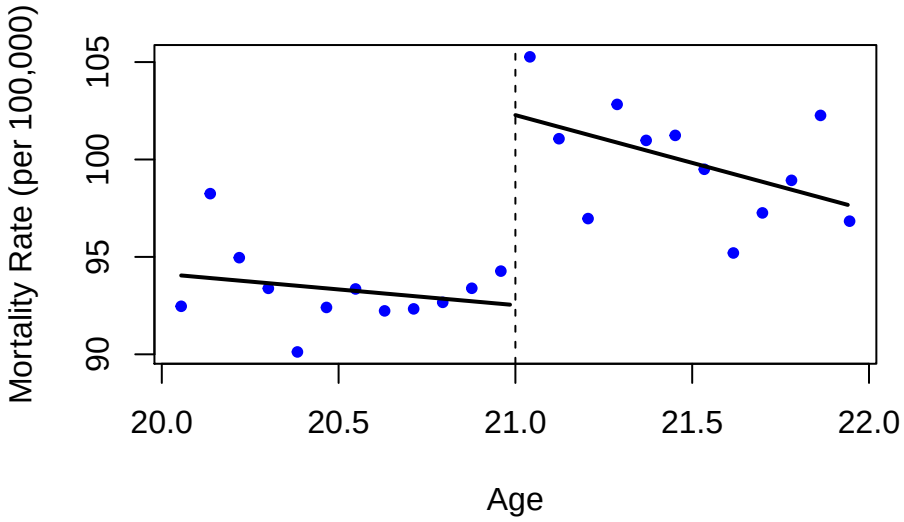
```
make_RD_plot(quadratic, mlda)
```




```
# Sensitivity Analysis: what changes if we restrict to ages 20-22?
mlda_subset <- mlda %>% filter(agecell >= 20 & agecell <= 22)

linear2 <-  lm(all ~ over21 + age + age:over21, mlda_subset)
summary(linear2)
```

```
Call:
lm(formula = all ~ over21 + age + age:over21, data = mlda_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3038 -0.9132 -0.1746  1.1758  4.3307

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   92.524      1.370  67.550  < 2e-16 ***
over21         9.753      1.937   5.035 6.34e-05 ***
age           -1.612      2.407  -0.669    0.511
over21:age    -3.289      3.405  -0.966    0.346
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.366 on 20 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7161,    Adjusted R-squared:  0.6735
F-statistic: 16.82 on 3 and 20 DF,  p-value: 1.083e-05
```

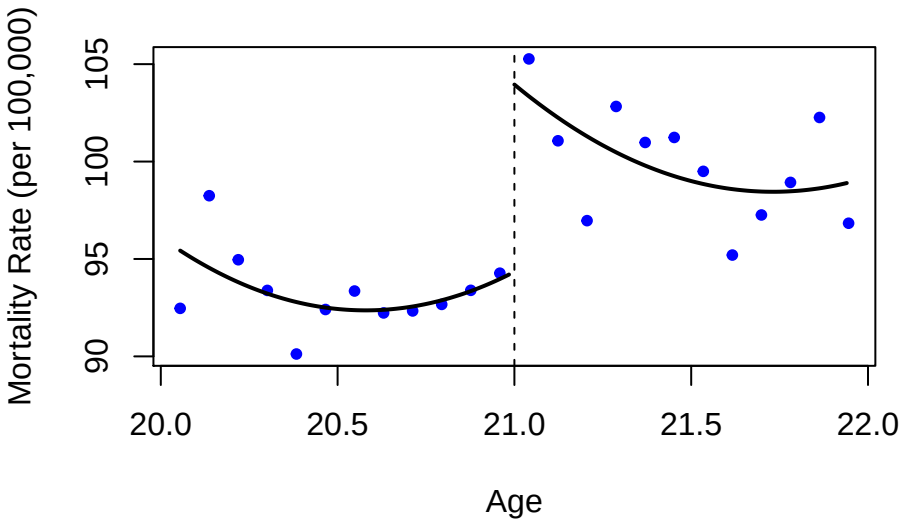
```
make_RD_plot(linear2, mlda_subset)
```




```
quadratic2 <-  lm(all ~ over21 + age + I(age^2) +
                    age:over21 + I(age^2):over21, mlda_subset)
summary(quadratic2)
```

```
Call:
lm(formula = all ~ over21 + age + I(age^2) + age:over21 + I(age^2):over21,
    data = mlda_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3270 -1.1898  0.1008  1.1612  3.7022

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     94.3403     2.0464  46.101   <2e-16 ***
over21           9.6111     2.8940   3.321   0.0038 **
age              9.3988     9.6193   0.977   0.3415
I(age^2)        11.1633     9.4514   1.181   0.2529
over21:age     -24.4478    13.6038  -1.797   0.0891 .
over21:I(age^2)  -0.8742    13.3663  -0.065   0.9486
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.333 on 18 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7517,    Adjusted R-squared:  0.6827
F-statistic:  10.9 on 5 and 18 DF,  p-value: 6.056e-05
```

```
make_RD_plot(quadratic2, mlda_subset)
```


Mortality Rate (per 100,000)
90
95
100
105
20.0
20.5
21.0
21.5
22.0
Age