Tutorial 11: Causal Inference Techniques - DiD and RDD

Jan Vogler (jan.vogler@duke.edu)

November 10, 2016

Today's Agenda

1. Descriptive summary statistics
2. Differences-in-differences
3. Regression discontinuity

1. Descriptive summary statistics

In the past tutorials, we have learned how to turn our regression results into a table for LaTeX through the package "stargazer". Sometimes you might want to show some descriptive summary statistics, too. Fortunately, there is a package our there that allows one to easily display descriptive summary statistics for variables in LaTeX.

Let us use our LDC dataset to illustrate this.

```
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/ps630_f16/")
library(foreign)
LDC = read.dta("LDC_IO_replication.dta")
```

The package we will use to display summary statistics is called "reporttools".

install.packages("reporttools")

```
library(reporttools)
```

```
## Loading required package: xtable
```

```
varsLDC <- LDC[, c("newtar", "fdignp", "polityiv_update2", "gdp_pc_95d")]
capLDC <- "Descriptive Statistics: LDC dataset"
tableContinuous(vars = varsLDC, cap = capLDC, lab = "tab: cont1", longtable = F,
    prec = 2)
```

The output code (which is hidden here to save space) can easily be used in LaTeX.

2. Differences-in-differences

What is differences-in-differences (diff-in-diff)?

When we use a diff-in-diff design, we have two groups of units and we are interested in how one of the two groups is affected by an "exogeneous" treatment. In order to estimate the effect of the treatment, we make the so-called "parallel trends" assumption, i.e. we believe that the variable we are interested in moves would

have moved in the same fashion if it was not for the treatment. The "control group is not affected by the treatment. The name"differences-in-differences" stems from the fact that we observe a variable that moves over time (has differences over time) and we are interested in how this movement is different between units.

Note that although I use experimental language, such as the terms "exogeneous", "treatment" and "control group", a diff-in-diff does not meet the same rigorous standards as a controlled, randomized experiment.

The following content is from Kevin Goulding.

The original can be found here:

https://thetarzan.wordpress.com/2011/06/20/differences-in-differences-estimation-in-r-and-stata/

The following plot illustrates the concept of differences-in-differences:

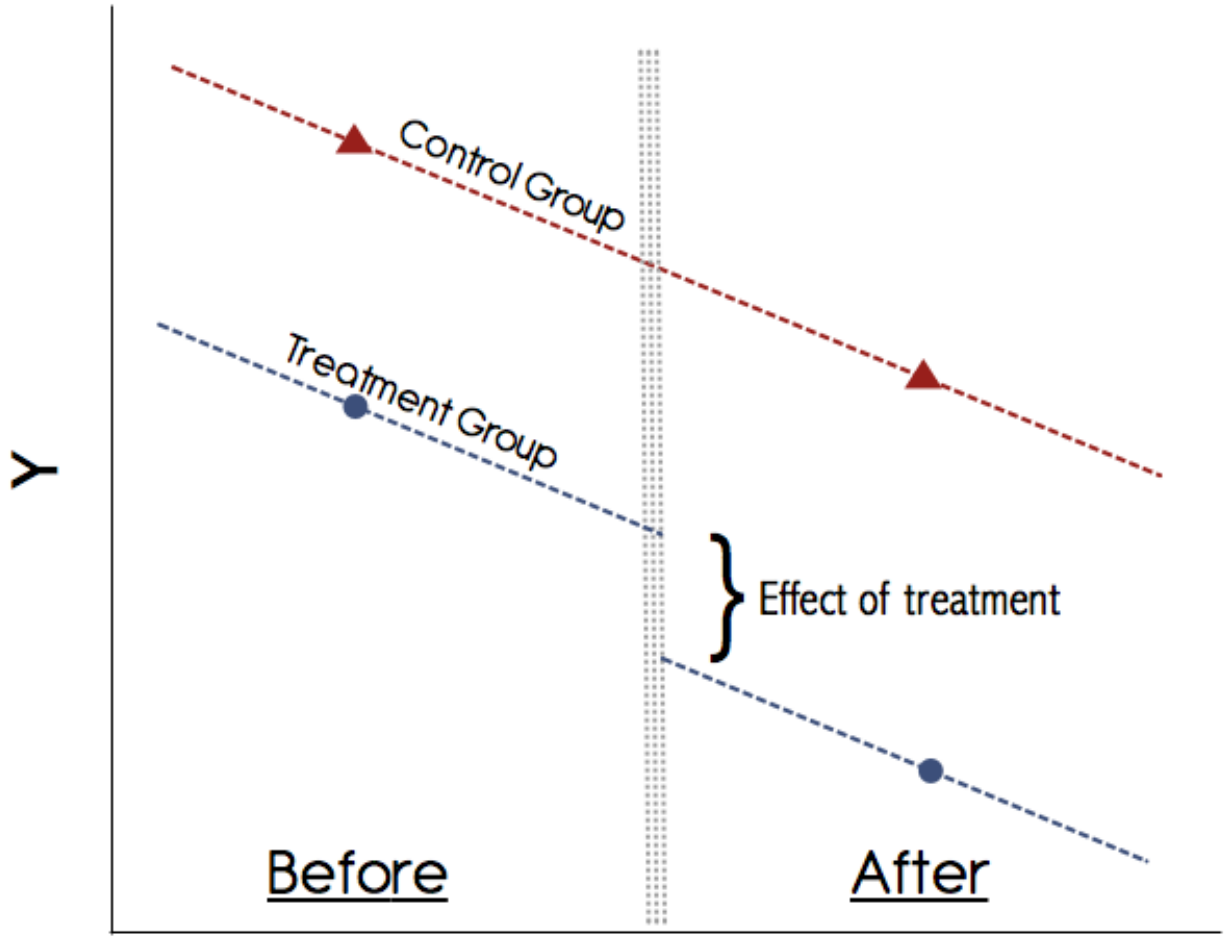


Figure 1: Differences-in-differences

In order to illustrate differences-in-differences, we use a dataset that shows the effect of the *Earned Income Tax Credit (EITC)* that was introduced in 1993 and benefited women with children. The question we are interested in is: how did the EITC affect the employment status of women? Let us first load the dataset.

```
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/ps630_f16/W11/")
eitc = read.dta("eitc.dta")
summary(eitc)
```

```
##      state            year          urate           children
```

```
##  Min.   :11.00   Min.   :1991   Min.   : 2.600   Min.   :0.000
##  1st Qu.:31.00   1st Qu.:1992   1st Qu.: 5.700   1st Qu.:0.000
##  Median :56.00   Median :1993   Median : 6.800   Median :1.000
##  Mean   :54.52   Mean   :1993   Mean   : 6.762   Mean   :1.193
##  3rd Qu.:81.00   3rd Qu.:1995   3rd Qu.: 7.700   3rd Qu.:2.000
##  Max.   :95.00   Max.   :1996   Max.   :11.400   Max.   :9.000
##     nonwhite          finc            earn            age
##  Min.   :0.0000   Min.   :     0   Min.   :     0   Min.   :20.00
##  1st Qu.:0.0000   1st Qu.:  5123   1st Qu.:     0   1st Qu.:26.00
##  Median :1.0000   Median :  9637   Median :  3332   Median :34.00
##  Mean   :0.6007   Mean   : 15255   Mean   : 10432   Mean   :35.21
##  3rd Qu.:1.0000   3rd Qu.: 18659   3rd Qu.: 14321   3rd Qu.:44.00
##  Max.   :1.0000   Max.   :575617   Max.   :537881   Max.   :54.00
##        ed              work           unearn
##  Min.   : 0.000   Min.   :0.000   Min.   :  0.000
##  1st Qu.: 7.000   1st Qu.:0.000   1st Qu.:  0.000
##  Median :10.000   Median :1.000   Median :  2.973
##  Mean   : 8.806   Mean   :0.513   Mean   :  4.823
##  3rd Qu.:11.000   3rd Qu.:1.000   3rd Qu.:  6.864
##  Max.   :11.000   Max.   :1.000   Max.   :134.058
```

Our dependent variable is "work" - a binary variable indicating employment with 1.

We need a minimum of four groups to conduct a diff-in-diff analysis. Those groups are separated according to two categories - time point 1 and time point 2 to estimate the trend. And treatment group and control group to estimate the effect of the treatment. The EITC went into effect in the year 1994, meaning that the time before 1994 is our time point 1 and 1994 onwards is time point 2.

```
# We turn our logical vector into a numeric (1/0) The value '1' stands for
# 1994 onwards
eitc$post93 = as.numeric(eitc$year >= 1994)
```

As stated above, the EITC benefits women with children, so let us subset the data and only look at people with children.

```
# We create a new binary variable The value '1' stands for someone with 1 or
# more child(ren)
eitc$anykids = as.numeric(eitc$children >= 1)
```

Next we look at the means of the four possible combinations.

```
a = sapply(subset(eitc, post93 == 0 & anykids == 0, select = work), mean)
b = sapply(subset(eitc, post93 == 0 & anykids == 1, select = work), mean)
c = sapply(subset(eitc, post93 == 1 & anykids == 0, select = work), mean)
d = sapply(subset(eitc, post93 == 1 & anykids == 1, select = work), mean)
```

Finally, we compute the differences-in-differences by using the above four means.

```
(d - c) - (b - a)
```

```
##       work
## 0.04687313
```

$$work = \beta_0 + \delta_0 post93 + \beta_1 anykids + \delta_1 (anykids \times post93) + \varepsilon$$

Figure 2: Regression

How does this capture the causal effect?

Another way to look at this result is to take into account the following regression.

```
eitc$p93kids.interaction = eitc$post93 * eitc$anykids
reg1 = lm(work ~ post93 + anykids + p93kids.interaction, data = eitc)
summary(reg1)
```

```
##
## Call:
## lm(formula = work ~ post93 + anykids + p93kids.interaction, data = eitc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5755 -0.4908  0.4245  0.5092  0.5540
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.575460   0.008845  65.060  < 2e-16 ***
## post93              -0.002074   0.012931  -0.160  0.87261
## anykids             -0.129498   0.011676 -11.091  < 2e-16 ***
## p93kids.interaction  0.046873   0.017158   2.732  0.00631 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4967 on 13742 degrees of freedom
## Multiple R-squared:  0.0126, Adjusted R-squared:  0.01238
## F-statistic: 58.45 on 3 and 13742 DF,  p-value: < 2.2e-16
```

How would we interpret this output? What does it tell us about the four groups we created?

For more on diff-in-diff see the following sources:

Angrist & Pischke (2015) *Mastering 'Metrics*, Chapter 5

Angrist & Pischke (2009) *Mostly Harmless Econometrics*, Chapter 5.2

For a more comprehensive discussion of the above example, see the following website:

https://thetarzan.wordpress.com/2011/05/24/surviving-graduate-econometrics-with-r-difference-in-difference-estimation-2-of-

3. Regression discontinuity designs

Regression discontinuity designs are powerful tools for causal inference where a real randomized experiment is not feasible (like diff-in-diff). They make use of the existence of a cut-off point which is associated with only minor changes in a continuous/interval variable that are expected to lead to significant substantive effects between the affected units.

You were asked to read the article **Hidalgo & Nichter (2015) Voter Buying - Shaping the Electorate through Clientelism**. In this article, the authors show how the "random" introduction of voter auditing

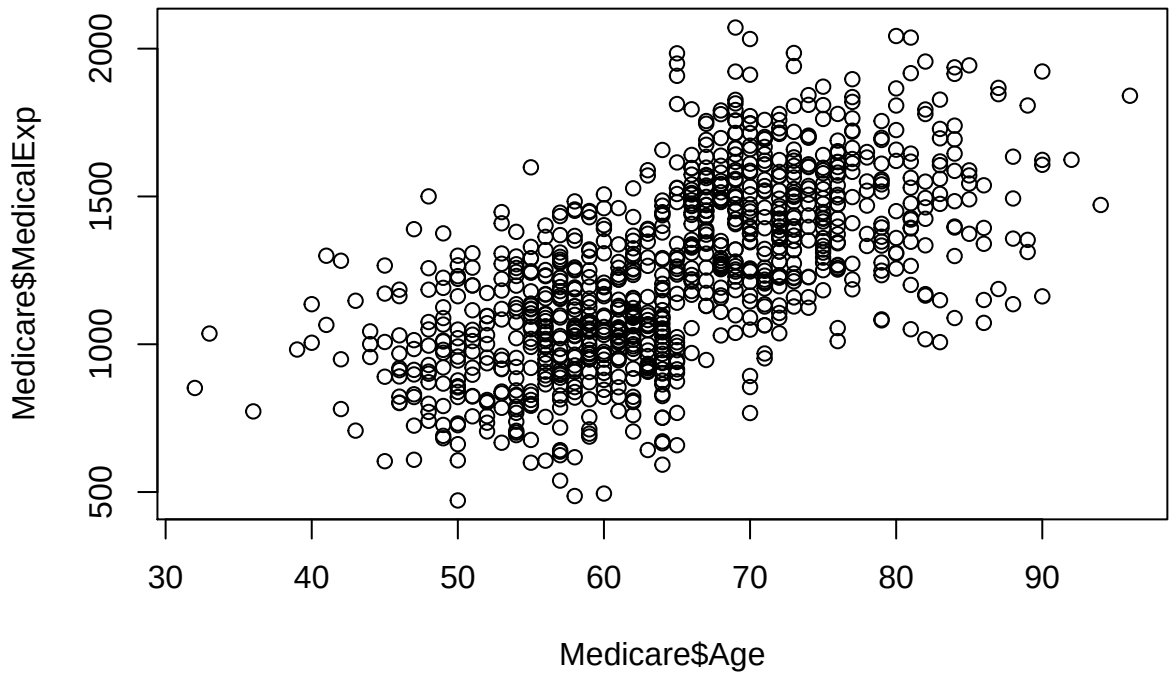
can affect the extent to which parties can buy votes. This is a good example of how applied political research can make use of as-if-random real-world discontinuities.

In order to illustrate RDD, let us first load an artificial dataset that contains medical expenditures by age for US citizens. We will make use of a discontinuity at the age of 65. In the United States, the medicare program helps citizens at and above the age of 65 to finance their medical expenditures. Although citizens at the age of 64 can be expected to not substantially differ from citizens at the age of 65 in terms of their medical needs, the fact that medicare supports citizens at the age of 65 indicates that there could be a discontinuity in medical expenditures at this cutoff point.

```
load("Medicare.Rdata")
summary(Medicare)
```

```
##       Age          MedicalExp       covariate1       covariate2
##  Min.   :32.00   Min.   : 471.6   Min.   : 69.44   Min.   : 66.38
##  1st Qu.:57.00   1st Qu.:1014.8   1st Qu.: 93.15   1st Qu.: 93.37
##  Median :64.00   Median :1234.0   Median :100.32   Median :100.23
##  Mean   :64.78   Mean   :1241.3   Mean   :100.06   Mean   : 99.88
##  3rd Qu.:72.00   3rd Qu.:1464.2   3rd Qu.:106.77   3rd Qu.:106.44
##  Max.   :96.00   Max.   :2071.4   Max.   :135.19   Max.   :126.70
```

```
plot(Medicare$Age, Medicare$MedicalExp)
```



The plot does not really tell us what is going. There might or might not be a discontinuity at the age of 65. Let us use some RDD tools to figure this out.

The following content is based on code from Matthieu Stigler. Note that the R commands in the original instructions are based on an older version of the package. The modified commands are below.

The original can be found here: https://github.com/MatthieuStigler/RDDtools

We will use a package called "rddtools" to conduct this analysis.

install.packages("rddtools")

```
library(rddtools)
```

```
## Warning: package 'rddtools' was built under R version 3.3.2

## Loading required package: AER

## Loading required package: car

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

## Loading required package: np

## Warning: package 'np' was built under R version 3.3.2

## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-2)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
```

"rddtools" requires us to declare our data an rdd object so that it can apply the correct tools.

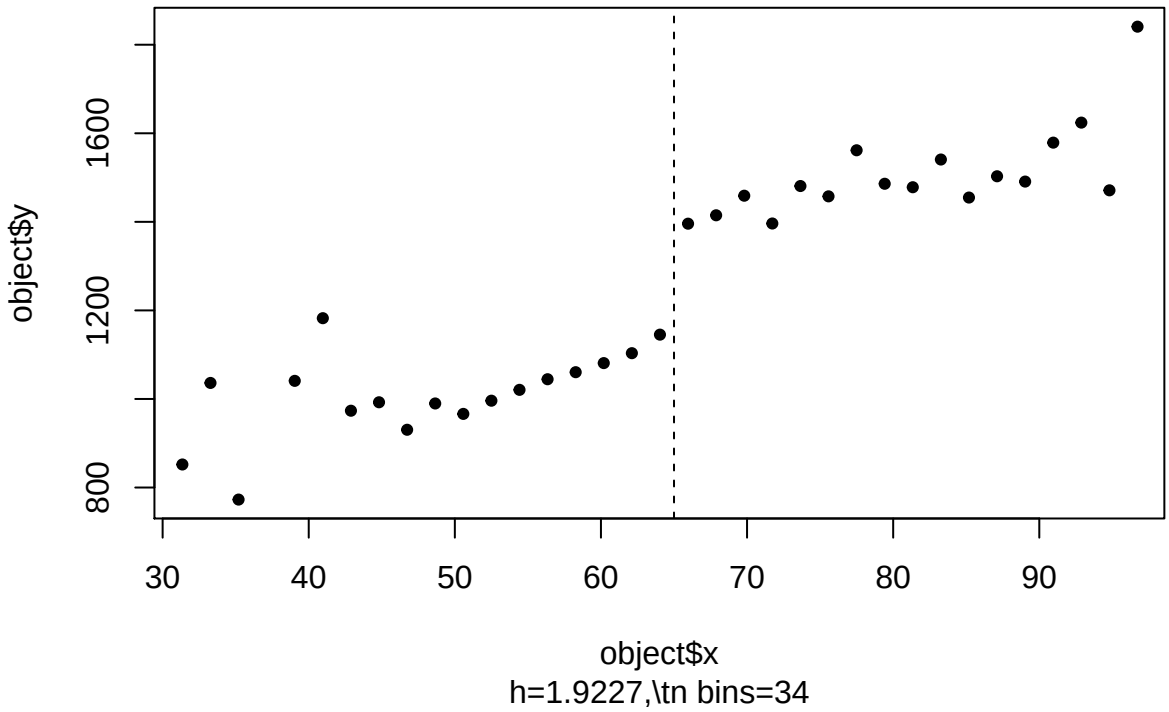
```
medicare_rdd <- rdd_data(y = Medicare$MedicalExp, # Dependent variable
                         x = Medicare$Age, # Key independent variable
                         covar = Medicare, # Dataset containing control variables
                         cutpoint = 65) # Our cutpoint
```

We can now apply "summary" and "plot" to get an overview of the data.

```
summary(medicare_rdd)
```

```
## ### rdd_data object ###
##
## Cutpoint: 65
## Sample size:
##  -Full : 1000
##  -Left : 503
##  -Right: 497
## Covariates: yes
```

```
plot(medicare_rdd)
```



The plot shows the average values in each of the bins, where each bin. Unlike our scatterplot, it gives us a clear impression of whether or not there might be a discontinuity. Indeed, it appears that there is some discontinuity in our data.

Now, let us turn to parametric regression estimation. In this regression approach, we assume that we know the function that describes the data. Linear regression is a parametric regression because we assume a linear relationship of form y = a + bx. The "rddtools" package allows us to include polynomials of different order in our regression to estimate non-linear regressions. Even with a discontinuity, there might be non-linear relationships.

```
reg_para <- rdd_reg_lm(rdd_object = medicare_rdd, order = 1, bw = 10)  # Bandwidth of the regression di
reg_para
```

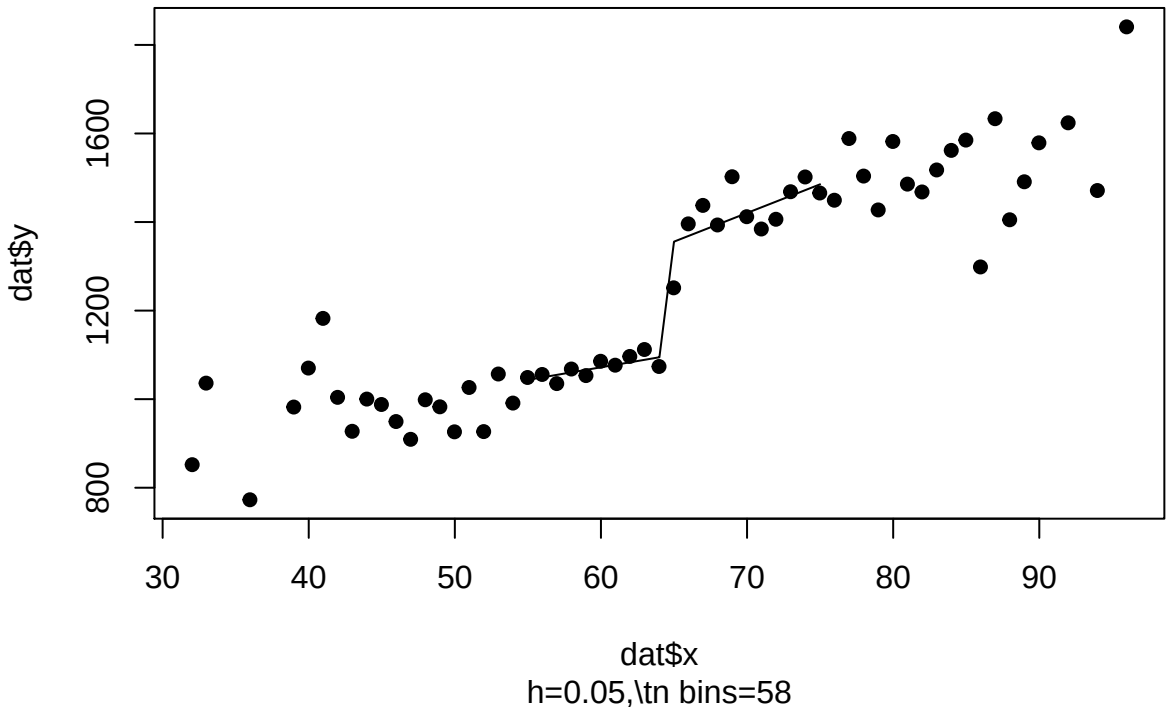
```
## ### RDD regression: parametric ###
##  Polynomial order:  1
```

```
##  Slopes:  separate
##  Bandwidth:  10
##  Number of obs: 690 (left: 344, right: 346)
##
##  Coefficient:
##   Estimate Std. Error t value  Pr(>|t|)
## D  255.496     33.554  7.6145 8.793e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How would we interpret this result?

Now let us plot the regression.

```
plot(reg_para)
```



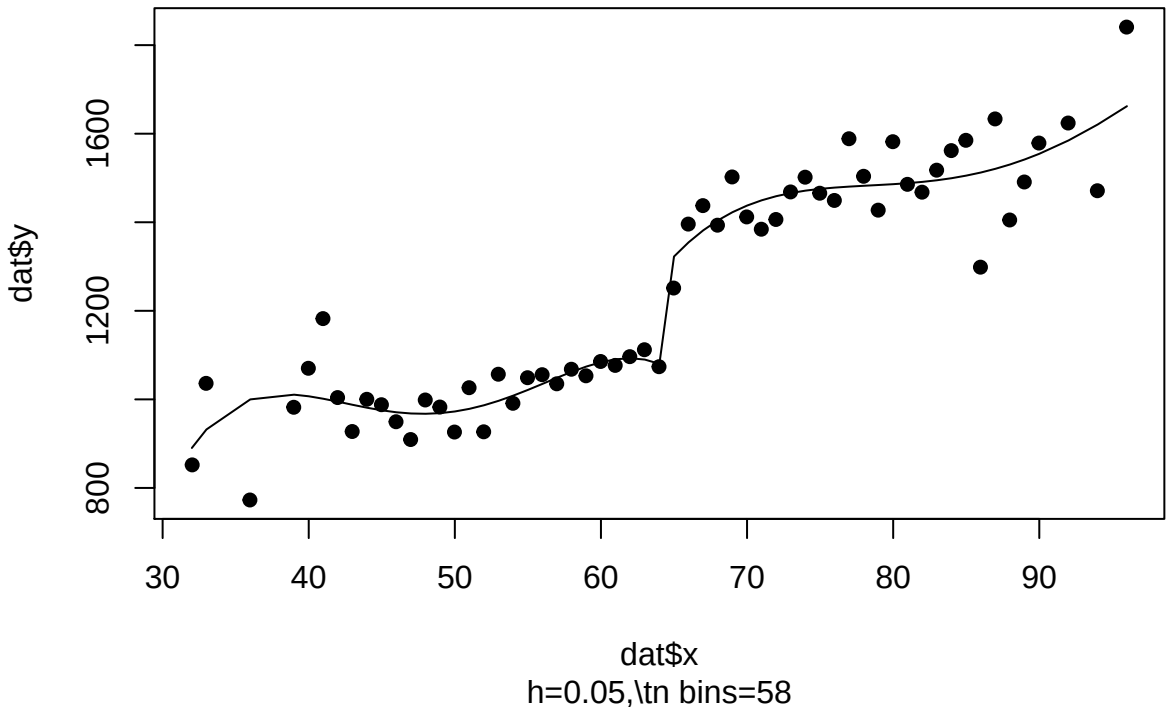
Alternatively, let us use a regression with a higher-order polynomial.

```
reg_para2 <- rdd_reg_lm(rdd_object = medicare_rdd, order = 4)
reg_para2
```

```
## ### RDD regression: parametric ###
##  Polynomial order:  4
##  Slopes:  separate
##  Number of obs: 1000 (left: 503, right: 497)
##
```

```
##  Coefficient:
##   Estimate Std. Error t value  Pr(>|t|)
## D  260.007     52.539  4.9488 8.772e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(reg_para2)
```



Note: if we have control variables, we can include those control variables in the following way:

```
reg_para3 <- rdd_reg_lm(rdd_object = medicare_rdd, covariates = "covariate1 + covariate2",
    order = 4)
reg_para3
```

```
## ### RDD regression: parametric ###
##  Polynomial order:  4
##  Slopes:  separate
##  Number of obs: 1000 (left: 503, right: 497)
##
##  Coefficient:
##   Estimate Std. Error t value  Pr(>|t|)
## D  260.117     52.583  4.9468 8.864e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(reg_para3)
```

```
##
## Call:
## lm(formula = y ~ ., data = dat_step1, weights = weights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -665.65 -147.91   -9.58  149.31  658.69
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.098e+03  1.100e+02   9.989  < 2e-16 ***
## D            2.601e+02  5.258e+01   4.947 8.86e-07 ***
## x           -2.204e+01  2.117e+01  -1.041   0.2981
## `x^2`       -4.678e+00  3.083e+00  -1.517   0.1295
## `x^3`       -2.461e-01  1.627e-01  -1.512   0.1308
## `x^4`       -3.918e-03  2.738e-03  -1.431   0.1528
## x_right      5.675e+01  2.724e+01   2.083   0.0375 *
## `x^2_right`  1.879e+00  4.144e+00   0.453   0.6503
## `x^3_right`  3.415e-01  2.284e-01   1.495   0.1352
## `x^4_right`  2.948e-03  4.027e-03   0.732   0.4644
## covariate1  -4.335e-01  7.053e-01  -0.615   0.5389
## covariate2   7.570e-02  7.139e-01   0.106   0.9156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 220.7 on 988 degrees of freedom
## Multiple R-squared:  0.4706, Adjusted R-squared:  0.4647
## F-statistic: 79.84 on 11 and 988 DF,  p-value: < 2.2e-16
```

The following command computes the "optimal bandwidth" of the RDD, following a procedure by Imbens, Guido and Karthik Kalyanaraman (2012).

```
bw_ik <- rdd_bw_ik(medicare_rdd)
bw_ik
```

```
##    h_opt
## 14.77028
```

Non-parametric regression refers to regressions where we do not make any assumptions about the function that links our variables.

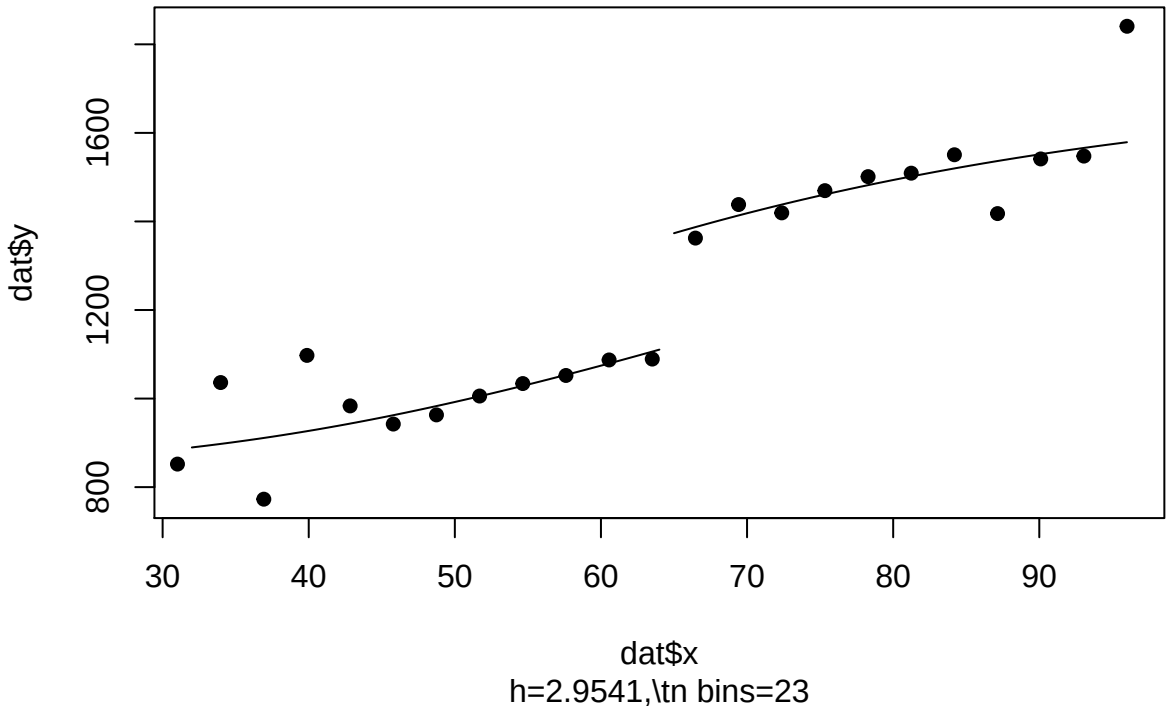
```
reg_nonpara <- rdd_reg_np(rdd_object = medicare_rdd, bw = bw_ik)
print(reg_nonpara)
```

```
## ### RDD regression: nonparametric local linear###
##  Bandwidth:  14.77028
##  Number of obs: 824 (left: 413, right: 411)
##
##  Coefficient:
```

```
##   Estimate Std. Error z value  Pr(>|z|)
## D  247.776     33.827  7.3249 2.391e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

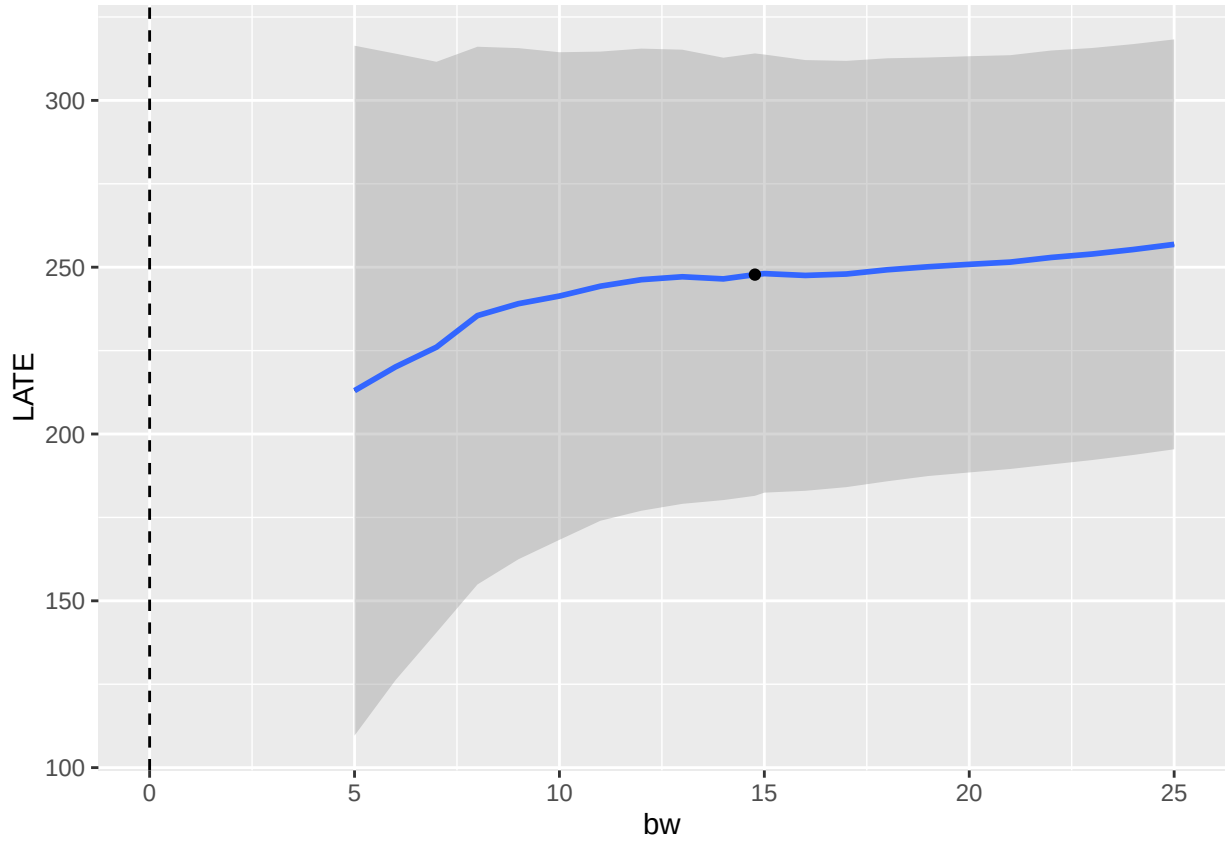
Next, we plot the nonparametric regression.

```
plot(x = reg_nonpara)
```



Let us apply a regression sensitivity test. This plot shows us how our results change if we use different bandwidths for estimating the differences. In the example below we look at bandwidths 5 to 25.

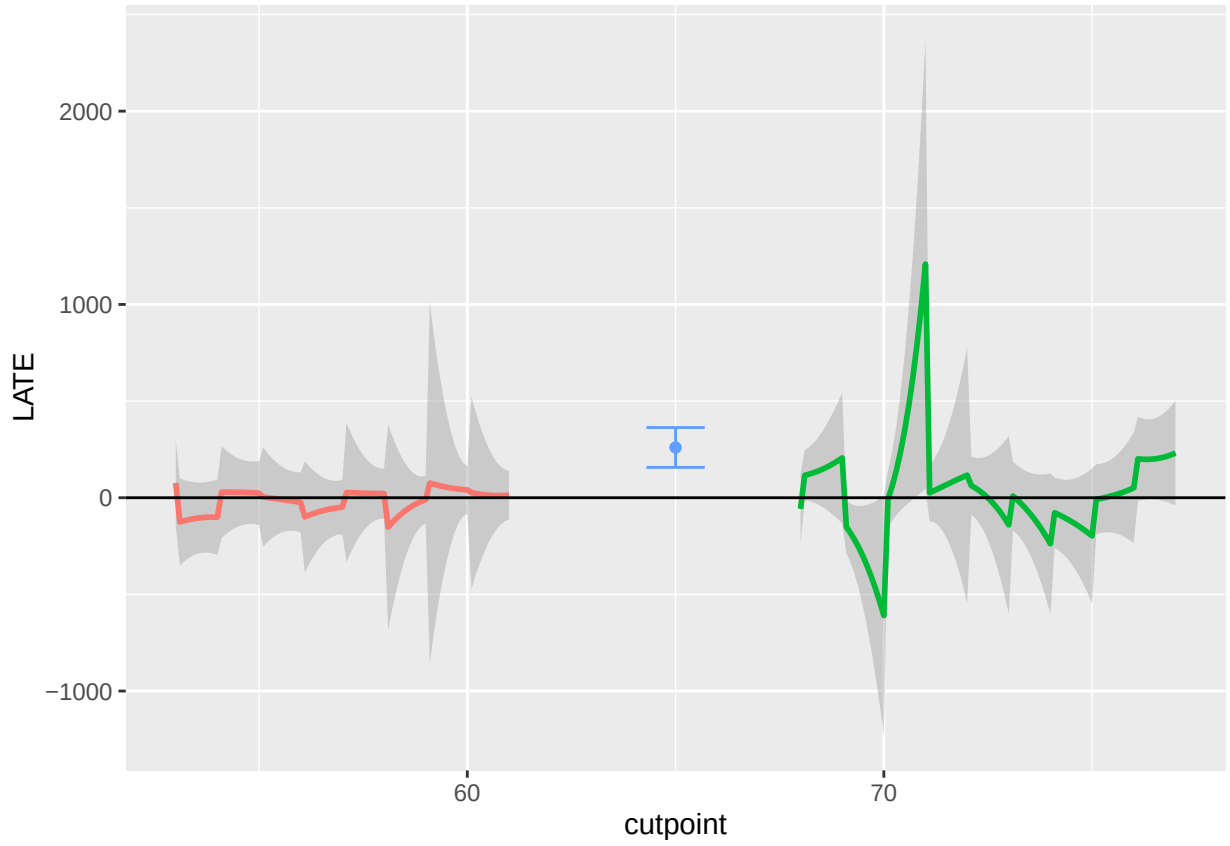
```
plotSensi(reg_nonpara, from = 5, to = 25, by = 1)
```



How would we interpret these results?

Another way to check the robustness of our results is a placebo test. Here we check for other values in the regression discontinuity whether they would yield any significant results.

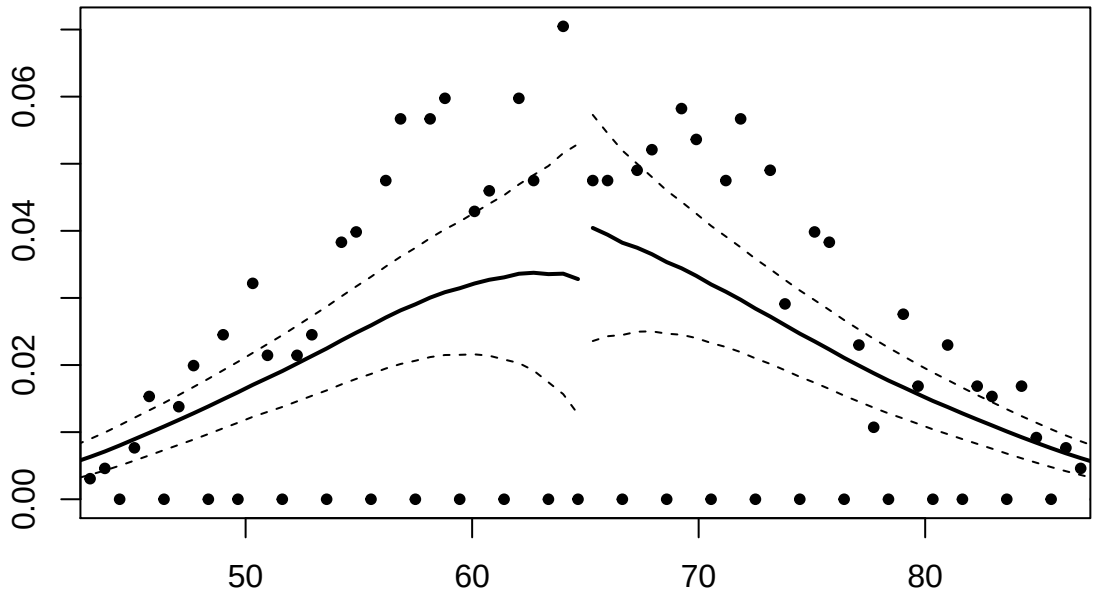
```
plotPlacebo(reg_para2)
```



How would we interpret this plot?

Another problem that might arise in regression discontinuity designs is that there might be some units that should not have been treated but were treated. For example, assume in our case that some 63- or 64-year olds who are in need of medical treatment deliberately claim that they are 65 year old to receive treatment. In this case, we would observe an unnatural discontinuity in the density of values around the threshold. The following command allows us to conduct such a test:

```
dens_test(reg_nonpara)
```




```
##
##  McCrary Test for no discontinuity of density around cutpoint
##
## data:  reg_nonpara
## z-val = 1.6753, p-value = 0.09388
## alternative hypothesis: Density is discontinuous around cutpoint
## sample estimates:
## Discontinuity
##     0.2243895
```

How would we interpret these results? Considering the p-value, should we have more or less confidence in our estimates?

Finally, we can check whether or not our discontinuity comes from covariates. In our example of medicare expenditures, we could assume that several factors can contribute to the increase in medical expenditures. For example, many people stop working at the age of 65. Maybe people who do not work anymore become bored and go to the doctor more frequently to talk to someone. If we have data on whether or not people are still in employment, we could use this covariate test to identify whether or not the employment status after 65 primarily contributes to the spike in medical expenditures.

The following test shows us whether there is a significant difference in the covariates around the threshold with a bandwidth of 5.

```
covarTest_mean(medicare_rdd, bw = 5)
```

```
##            mean of x mean of y Difference statistic  p.value
```

```
## Age        62.21264  67.60697  5.394321   -33.40622  4.295048e-114
## MedicalExp 1088.108  1402.745  314.6375   -13.12917  1.29602e-32
## covariate1 100.4661  99.70858  -0.757559  0.7700209  0.441775
## covariate2 99.6914   100.5774  0.8859947  -0.8563658 0.3923748
```

How would we interpret this output?

RDD replication dataset: Hidalgo & Nichter (2015)

If you would like to see another implementation of RDD based on the article that you have read, please check out the data and replication files that have been made available by **Hidalgo & Nichter (2015) Voter Buying - Shaping the Electorate through Clientelism**. The replication files include both the data and the code, making it possible for you to follow their steps one by one. It can be found here:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9OOLQ7

Note that their code is very complex, so you should only look at it when you plan to implement RDD yourself.