# Advanced Quantitative Techniques

# (Class 9)

Gregory M. Eirich
QMSS

## 1. Regression Discontinuity Analysis

1. Regression Discontinuity Analysis

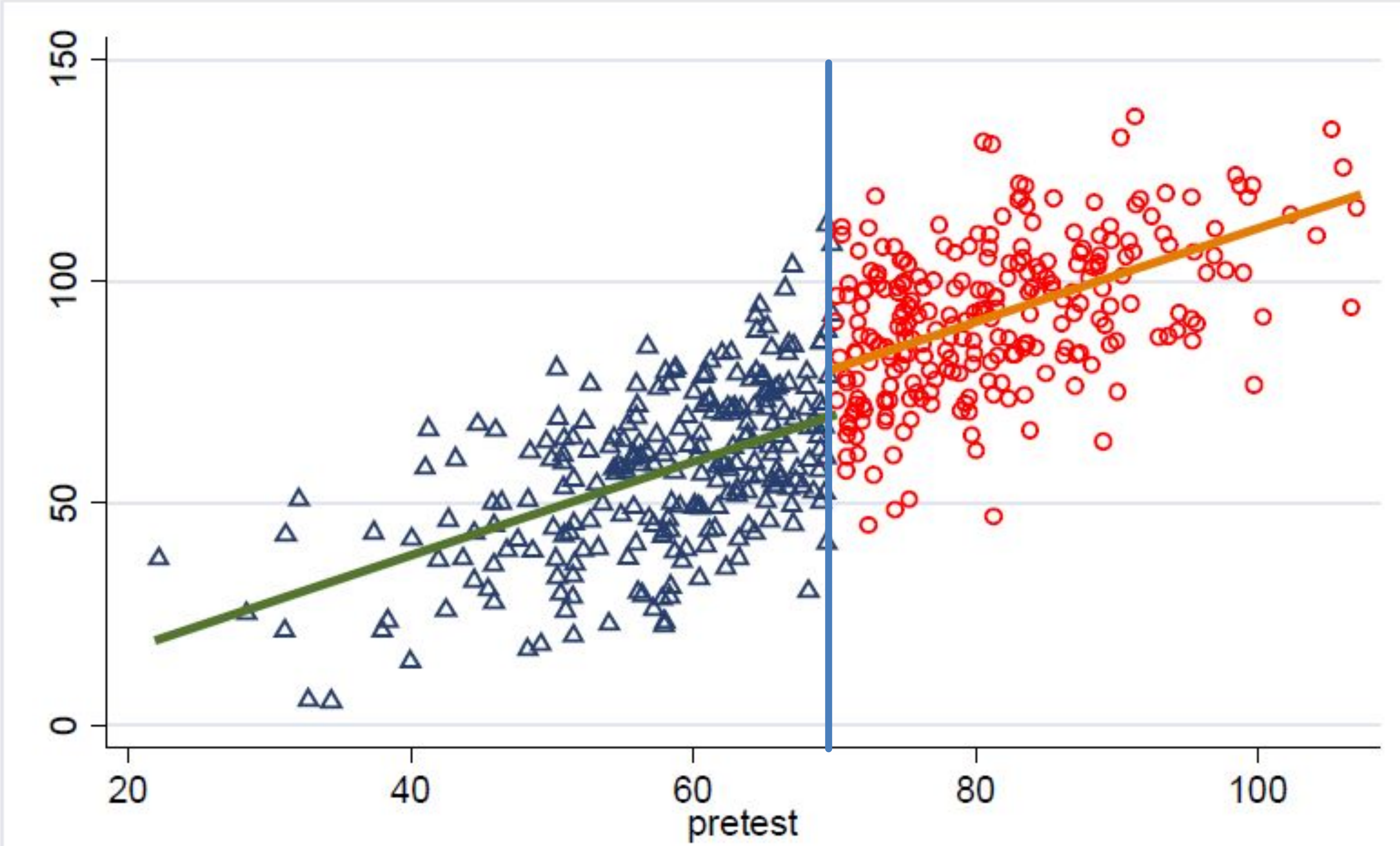
(P.S., A couple of these slides are borrowed from my former TA, Emma Garcia – Thanks!)

# What is regression discontinuity?

- Assignment to treatment is decided solely on values of one measured variable, called the “forcing variable (X)”
- Can be thought as a natural experiment, where treatment affects those individuals close to the cut-off as if this were a randomized experiment or quasi experimental design, where we look at the impact of one treatment/intervention (D) on one outcome (Y)

# What is regression discontinuity?

- – Because the cut-off is arbitrary, those individuals just on either side of it are essentially equivalent
- – But we can look at the consequences for those on the "right" side of the cut-off (who therefore receive the treatment) vs. those on the "wrong" side of the cut-off who receive nothing




Example:

- Students scoring above the cutoff participate in a math enrichment program. Post/pre-test scores represented.

*

# Issues of causal inference

ESSAYS ON URBAN SCHOOL ORGANIZATION:

EVIDENCE FROM CHICAGO PUBLIC SCHOOLS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE IRVING B. HARRIS

GRADUATE SCHOOL OF PUBLIC POLICY STUDIES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

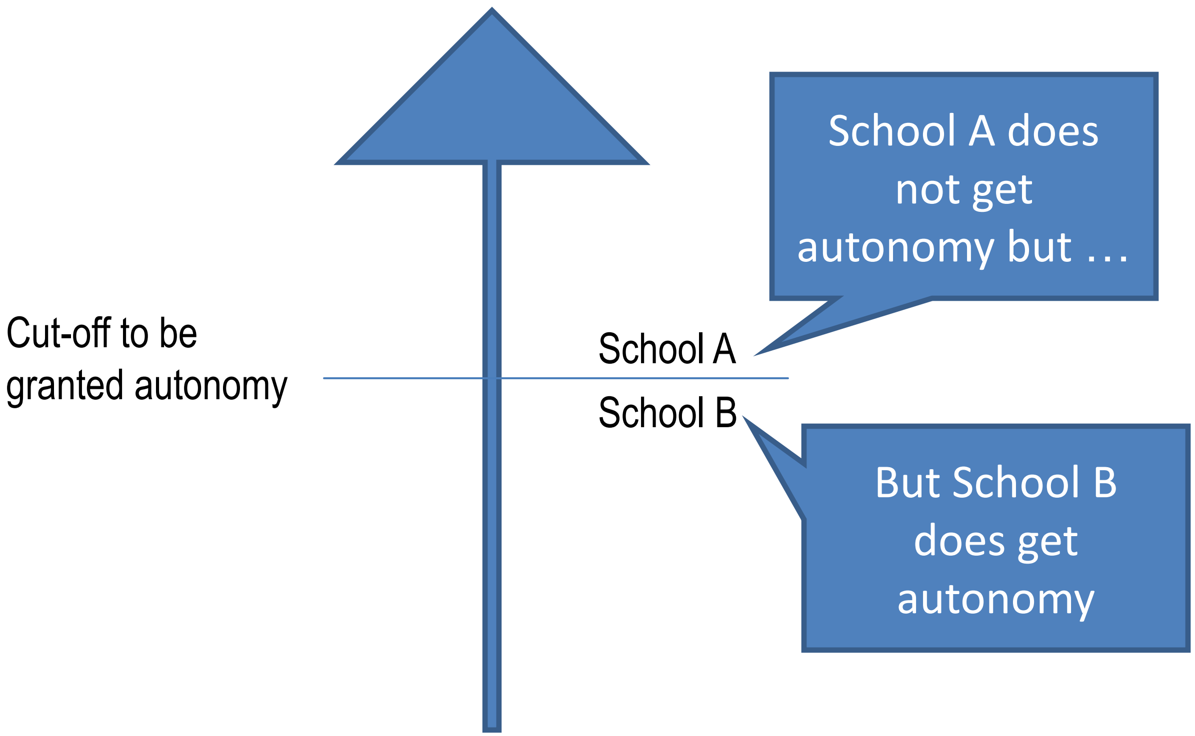
MATTHEW PHILIP STEINBERG

# The substantive issue

- Does greater autonomy improve school performance?  Evidence from a regression discontinuity analysis in Chicago

# Designing a quasi-experiment

- Are schools any different if they are separated by 1 point?




What happens to School A vs. School B in the following years?

# The methodological issue …

- He performed a regression discontinuity analysis, which looks at roughly equivalent schools to see how they performed in the years that followed their separate treatments

# The substantive conclusion

- Greater school autonomy poses a short-term risk to school performance. Notably, for schools at the discontinuity margin, receipt of greater autonomy adversely impacted math (but not reading) achievement after the first year.
- Then there were relative performance improvements after two years of autonomy.

# In conclusion

- We want to try to mimic experiments as much as we can to get true apples to apples comparisons

# More examples of RDs

- Anything test-based: e.g., National Merit Semifinalists
- Order on a sign-up list (queue)
- Rankings for wines, colleges, etc.
- More …

# Our example today:

Does turning 65 (not 64, not 66, but **65**) lead to an increase in the utilization of health services?

Why would it matter?

Can we see it?

# Our measure of utilization:

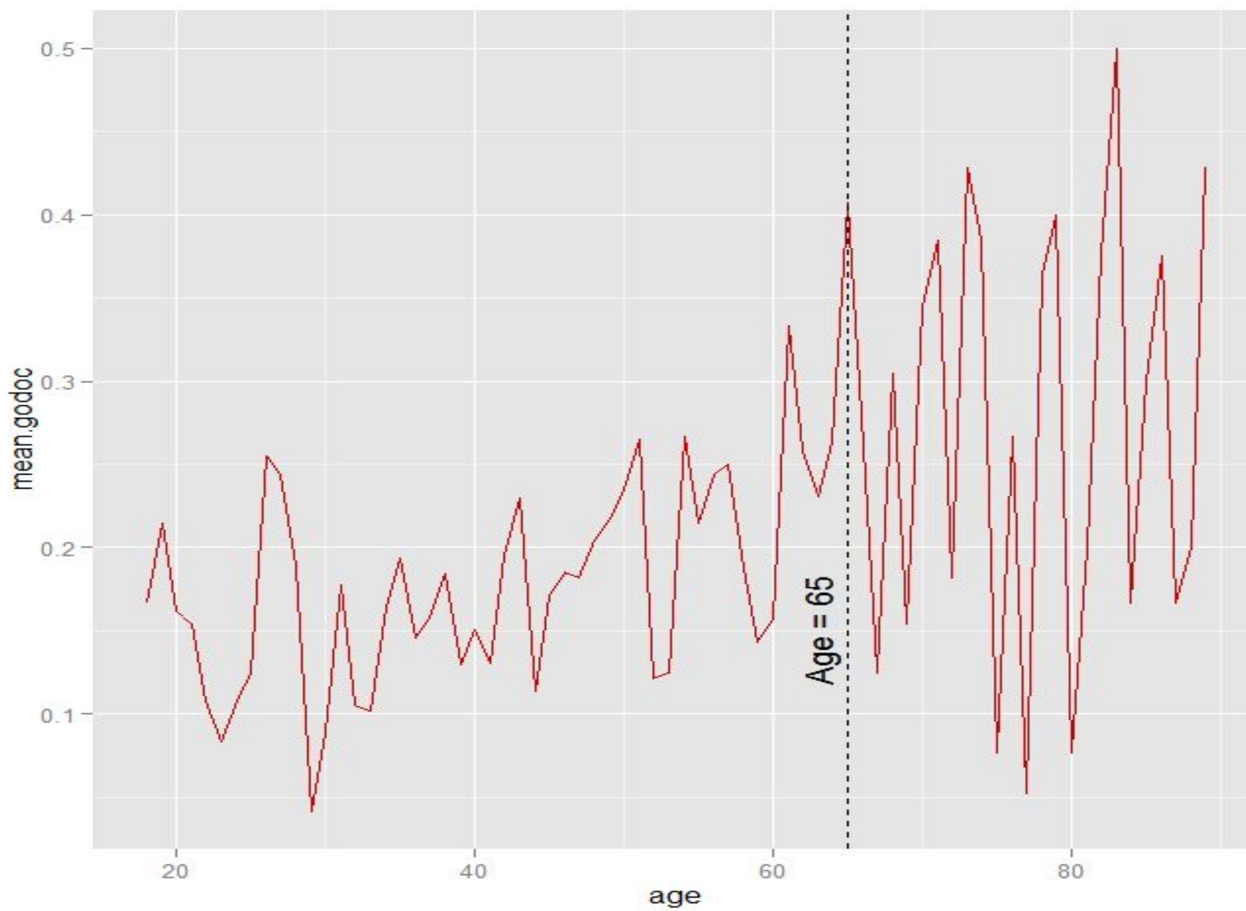
Now I'm going to ask you about things you did during the last seven days. I'm only interested in what you did during the last seven days. From last (DAY OF WEEK) to today did you... a. Go to see a doctor or receive medical treatment at a clinic or hospital?

```
library(QMSS)
library(ggplot2)
library(plyr)
```

```
> Tab(sub$godoc)
  Count   Pct Cum.Pct
1   453 18.99   18.99
2  1932 81.01  100.00

> sub$n.godoc <- mapvalues(sub$godoc, from = 1:2, to = 1:0)

> with(sub, table(godoc, n.godoc))
     n.godoc
godoc    0    1
    1    0  453
    2 1932    0
```

# Our example today:

Does turning 65 (not 64, not 66, but **65**) lead to an increase in the utilization of health services?



# How did I do that?

```
by.age <- ddply(sub, "age", summarize, mean.godoc = mean(n.godoc))
g_by.age <- ggplot(by.age, aes(x = age, y = mean.godoc)) + geom_line(color = "red3")
g_by.age
 # add vertical line at age = 65
g_by.age + geom_vline(xintercept = 65, lty = 2)
 # label the line
label_65 <- annotate("text", x = 63, y = 0.15, label = "Age = 65", angle = 90)
g_by.age + geom_vline(xintercept = 65, lty = 2) + label_65
```

# What discontinuity do we have?

## Sharp RD

* When assignment to treatment (D) is perfectly determined by the forcing variable (X)
* That is: X=D

* E.g., When you get a certain score on the PSATs (X), you are automatically labeled a National Merit Semifinalist (D)

# What discontinuity do we have?

## Fuzzy RD

* When assignment to treatment (D) is NOT perfectly determined by the forcing variable (X).

* That is: X ≈ D

* E.g., When you reach a certain score on a test (X), you become *eligible* to do an enrichment program (D), but some students will opt not to do it

# How to model Sharp RD?

- We can model this parametrically
- Specifically, we can run a piecewise regression with a break in the line at the cut-off point

$$Y = \alpha_{before\_cutoff} + \beta X_{before\_cutoff} + \alpha_{after\_cutoff} + \beta X_{after\_cutoff} + \varepsilon$$

- We can add polynomials to allow for ramping up and decaying effects relative to the cut-off point too

*

# Our measure of utilization:

I want to make it so that my lines are relative to age 65, so to do that, I need to recode for new slopes and new intercepts:

```
# make the slopes for younger than 65 and older than 65
sub$ageY <- ifelse(sub$age >= 65, 0, sub$age - 65)
sub$ageO <- ifelse(sub$age < 65, 0, sub$age - 65)


# make the intercepts for for younger than 65 and older than 65
sub$intY <- ifelse(sub$age >= 65, 0, 1)
sub$intO <- ifelse(sub$age < 65, 0, 1)
```

# What does this recoding look like?

```
. findit tablist // and install it … // STATA STUFF

. tablist age inty into agey ageo, sort(v)

  +-----------------------------------------------------------------------------+
  |      age   inty   into   agey   ageo   _Freq_   _Perc_   _CFreq_   _CPerc_ |
  |-----------------------------------------------------------------------------|
  |       18      1      0    -47      0       49     0.29        49      0.29 |
  |       19      1      0    -46      0      184     1.08       233      1.37 |
  |       20      1      0    -45      0      221     1.30       454      2.67 |
    [output omitted]
  |       61      1      0     -4      0      194     1.14     13781     81.16 |
  |       62      1      0     -3      0      194     1.14     13975     82.30 |
  |-----------------------------------------------------------------------------|
  |       63      1      0     -2      0      195     1.15     14170     83.45 |
  |       64      1      0     -1      0      161     0.95     14331     84.39 |
  |       65      0      1      0      0      179     1.05     14510     85.45 |
  |       66      0      1      0      1      188     1.11     14698     86.56 |
  |       67      0      1      0      2      181     1.07     14879     87.62 |
  |-----------------------------------------------------------------------------|
  |       68      0      1      0      3      184     1.08     15063     88.71 |
  |       69      0      1      0      4      160     0.94     15223     89.65 |
     [output omitted]
  | 89 or ol      0      1      0     24       73     0.43     16981    100.00 |
  +-----------------------------------------------------------------------------+
```

# Now what?

Inty is the predicted value for someone almost 65 years old; they have a 22.97% chance of having seen a doctor in the last week

```
> summary(lm.godoc)

Call:
lm(formula = n.godoc ~ 0 + intY + intO + ageY + ageO, data = sub)

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
intY 0.2296747  0.0204181  11.249  < 2e-16 ***
intO 0.2714238  0.0333853   8.130 6.83e-16 ***
ageY 0.0022365  0.0007296   3.065   0.0022 **
ageO 0.0006480  0.0030106   0.215   0.8296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3899 on 2381 degrees of freedom
Multiple R-squared:  0.2008,Adjusted R-squared:  0.1995
F-statistic: 149.6 on 4 and 2381 DF,  p-value: < 2.2e-16
```

# Now what?

Into is the predicted value for someone who just turned 65 years old; they have a 27.14% chance of having seen a doctor in the last week

```
> summary(lm.godoc)

Call:
lm(formula = n.godoc ~ 0 + intY + intO + ageY + ageO, data = sub)

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
intY 0.2296747  0.0204181  11.249  < 2e-16 ***
intO 0.2714238  0.0333853   8.130 6.83e-16 ***
ageY 0.0022365  0.0007296   3.065   0.0022 **
ageO 0.0006480  0.0030106   0.215   0.8296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3899 on 2381 degrees of freedom
Multiple R-squared:  0.2008,Adjusted R-squared:  0.1995
F-statistic: 149.6 on 4 and 2381 DF,  p-value: < 2.2e-16
```

# Now what?

Agey is the slope for anyone under 65 years old; for each year older they get (short of 65), they increase their chances by 0.22% of having seen a doctor in the last week

```
> summary(lm.godoc)

Call:
lm(formula = n.godoc ~ 0 + intY + intO + ageY + ageO, data = sub)

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
intY 0.2296747  0.0204181  11.249  < 2e-16 ***
intO 0.2714238  0.0333853   8.130 6.83e-16 ***
ageY 0.0022365  0.0007296   3.065   0.0022 **
ageO 0.0006480  0.0030106   0.215   0.8296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3899 on 2381 degrees of freedom
Multiple R-squared:  0.2008,Adjusted R-squared:  0.1995
F-statistic: 149.6 on 4 and 2381 DF,  p-value: < 2.2e-16
```

# Now what?

Ageo is the slope for anyone 65 years old and older; for each year older they get, they decrease their chances by 0.06% of having seen a doctor in the last week (thought this is not stat. sig.)

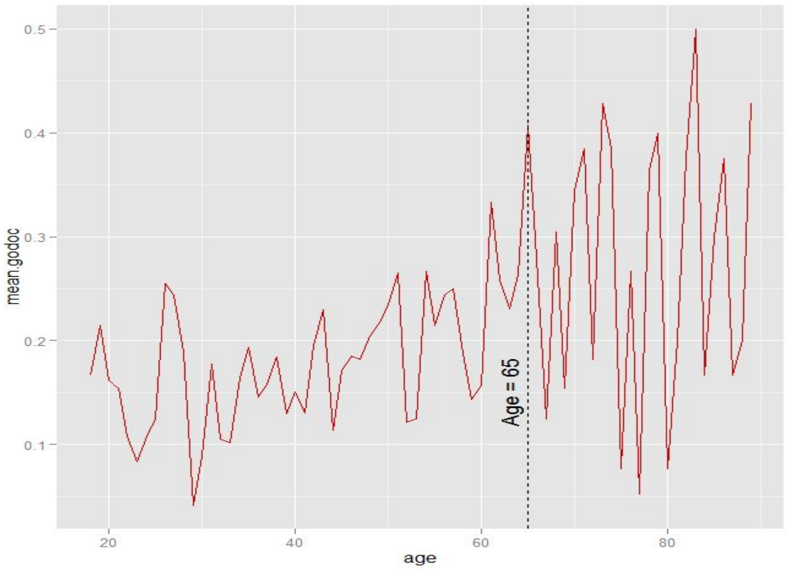
```
> summary(lm.godoc)

Call:
lm(formula = n.godoc ~ 0 + intY + intO + ageY + ageO, data = sub)

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
intY 0.2296747  0.0204181  11.249  < 2e-16 ***
intO 0.2714238  0.0333853   8.130 6.83e-16 ***
ageY 0.0022365  0.0007296   3.065   0.0022 **
ageO 0.0006480  0.0030106   0.215   0.8296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3899 on 2381 degrees of freedom
Multiple R-squared:  0.2008,Adjusted R-squared:  0.1995
F-statistic: 149.6 on 4 and 2381 DF,  p-value: < 2.2e-16
```

# What does this recoding look like?

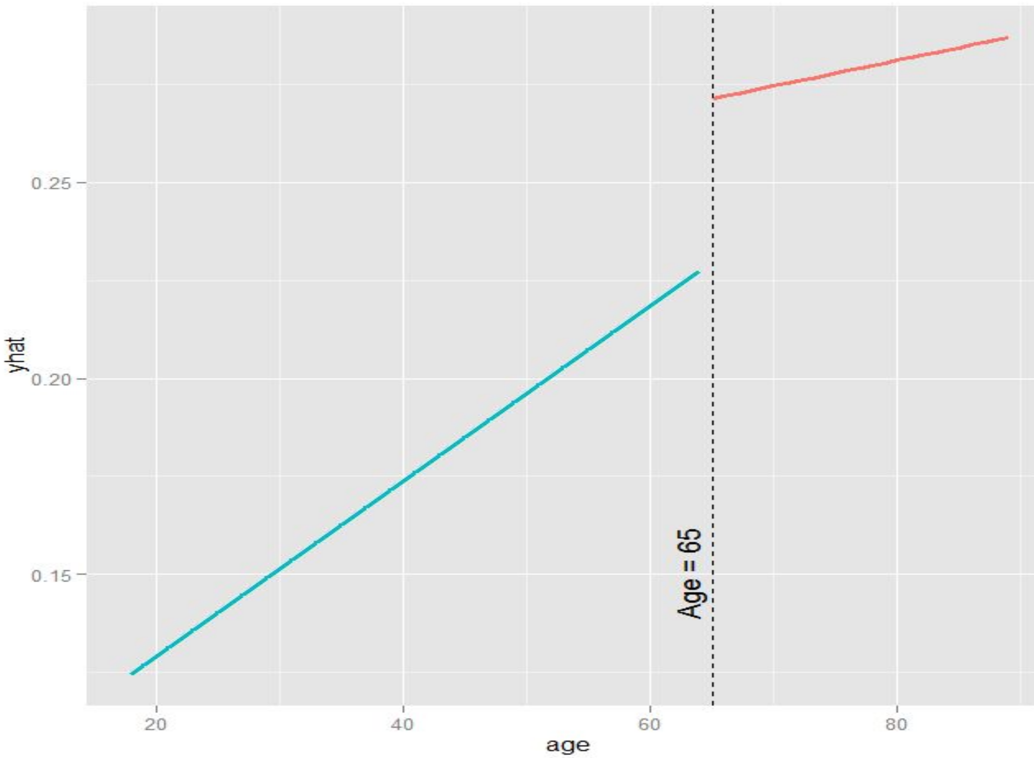
```
# regression model
lm.godoc <- lm(n.godoc ~ 0 + intY + intO + ageY + ageO, data = sub) # the 0 tells R not to add an intercept (we're using the intercept we created above)
summary(lm.godoc)
sub$yhat <- predict(lm.godoc) # get fitted values

# look at fitted values by age
ddply(sub, "age", summarize, yhat = mean(yhat), freq = length(age))
```

```
# look at fitted values by age
> ddply(sub, "age", summarize, yhat = mean(yhat), freq = length(age))
   age      yhat freq
1   18 0.1245614    6
….
46  63 0.2252018   26
47  64 0.2274383   19
48  65 0.2714238   27
49  66 0.2720718   23
50  67 0.2727198   24
...
72  89 0.2869758    7
```

# From this initial picture …




# … to this one




```
 twoway(line yhat age if age
<65, sort) (line yhat age if age
>=65, sort), xline(65) // STATA
code
```

# How did I do that?

```
# plot the discontinuity
no_legend <- theme(legend.position = "none")
g_disc <- ggplot(sub, aes(x = age, y = yhat, group = intY, color = factor(intY))) +
no_legend
g_disc + geom_line(size = 1.25) + geom_vline(xintercept = 65, lty = 2) + label_65
```

# Another way to model this ...

Net of the usually slope on age (0.25% increases per year of age), those who are 65 have a 16.97% greater chance of going to the doctor than all of the other age groups combined

```
> sub$spike65 <- ifelse(sub$age == 65, 1, 0)
> lm.godoc2 <- lm(n.godoc ~ age + spike65, data = sub)
> summary(lm.godoc2)

Call:
lm(formula = n.godoc ~ age + spike65, data = sub)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0743530  0.0228641   3.252  0.00116 **
age         0.0025127  0.0004757   5.282 1.39e-07 ***
spike65     0.1697303  0.0759908   2.234  0.02560 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3895 on 2382 degrees of freedom
Multiple R-squared:  0.01506,   Adjusted R-squared:  0.01423
F-statistic: 18.21 on 2 and 2382 DF,  p-value: 1.422e-08
```

# Maybe I just cherry-picked this age …

## Let's include 2 years before 65 and 2 years after.

```
>   # create indicator each age between 63 and 67
> for(i in 63:67){
+   name <- paste0("age",i)
+   sub <- within(sub, assign(name, age == i))
+ }

> lm.godoc3 <- lm(n.godoc ~ age + age63 + age64 + age65 + age66 + age67, data =
sub)
> summary(lm.godoc3)

Call:
lm(formula = n.godoc ~ age + age63 + age64 + age65 + age66 +
    age67, data = sub)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0720696  0.0231578   3.112  0.00188 **
age          0.0025830  0.0004901   5.270 1.49e-07 ***
age63TRUE   -0.0040279  0.0774072  -0.052  0.95851
age64TRUE    0.0257778  0.0902923   0.285  0.77529
age65TRUE    0.1674443  0.0761223   2.200  0.02793 *
age66TRUE    0.0183235  0.0823581   0.222  0.82395
age67TRUE   -0.1201290  0.0807373  -1.488  0.13691
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Maybe I just cherry-picked this age …

We see that the coefficient on age65 did not change much, and the rest of the variables are not statistically significant

```
> summary(lm.godoc3)

Call:
lm(formula = n.godoc ~ age + age63 + age64 + age65 + age66 +
    age67, data = sub)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0720696  0.0231578   3.112  0.00188 **
age          0.0025830  0.0004901   5.270 1.49e-07 ***
age63TRUE   -0.0040279  0.0774072  -0.052  0.95851
age64TRUE    0.0257778  0.0902923   0.285  0.77529
age65TRUE    0.1674443  0.0761223   2.200  0.02793 *
age66TRUE    0.0183235  0.0823581   0.222  0.82395
age67TRUE   -0.1201290  0.0807373  -1.488  0.13691
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3897 on 2378 degrees of freedom
Multiple R-squared:  0.01604,	Adjusted R-squared:  0.01356
F-statistic: 6.462 on 6 and 2378 DF,  p-value: 9.003e-07
```

# Why is this a bad RD example?

# Discontinuities along what dimensions?

# What do we think?

- It looks like something does happen at age 65

# How to model Sharp RD?

**Non-Parametric approach**

Basically, this means that we calculate an optimal bandwidth (or distance to the cutoff), using weights - Kernel functions, local linear regression (LOWESS) or other sophisticated methods - to compute mean outcome differences.

# Non-parametric estimation …

```
> # RDestimate() from rdd package
> # install.packages("rdd")
> library(rdd)
>
> rd.godoc <- RDestimate(n.godoc ~ age, data = sub, cutpoint = 65)
> summary(rd.godoc)

Call:
RDestimate(formula = n.godoc ~ age, data = sub, cutpoint = 65)

Type:
sharp

Estimates:
           Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
LATE       3.493      173           0.11995   0.1800      0.6665   0.5051
Half-BW    1.746       69           0.14425   0.1415      1.0195   0.3079
Double-BW  6.985      315           0.04279   0.1151      0.3718   0.7100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F-statistics:
           F       Num. DoF  Denom. DoF  p
LATE       1.6444  3         169         0.3621
Half-BW    0.7542  2          66         0.9488
Double-BW  0.7336  3         311         0.9347
```

# Non-parametric estimation …

The program chooses the best size for the “bandwidths” across values of X, age, at almost 4 year chunks (3.87). But to see how sensitive our results are to bandwidth choice, R also runs RD with bandwidths half that size and then also double that size

```
Estimates:
           Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
LATE       3.493      173           0.11995   0.1800      0.6665   0.5051
Half-BW    1.746       69           0.14425   0.1415      1.0195   0.3079
Double-BW  6.985      315           0.04279   0.1151      0.3718   0.7100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

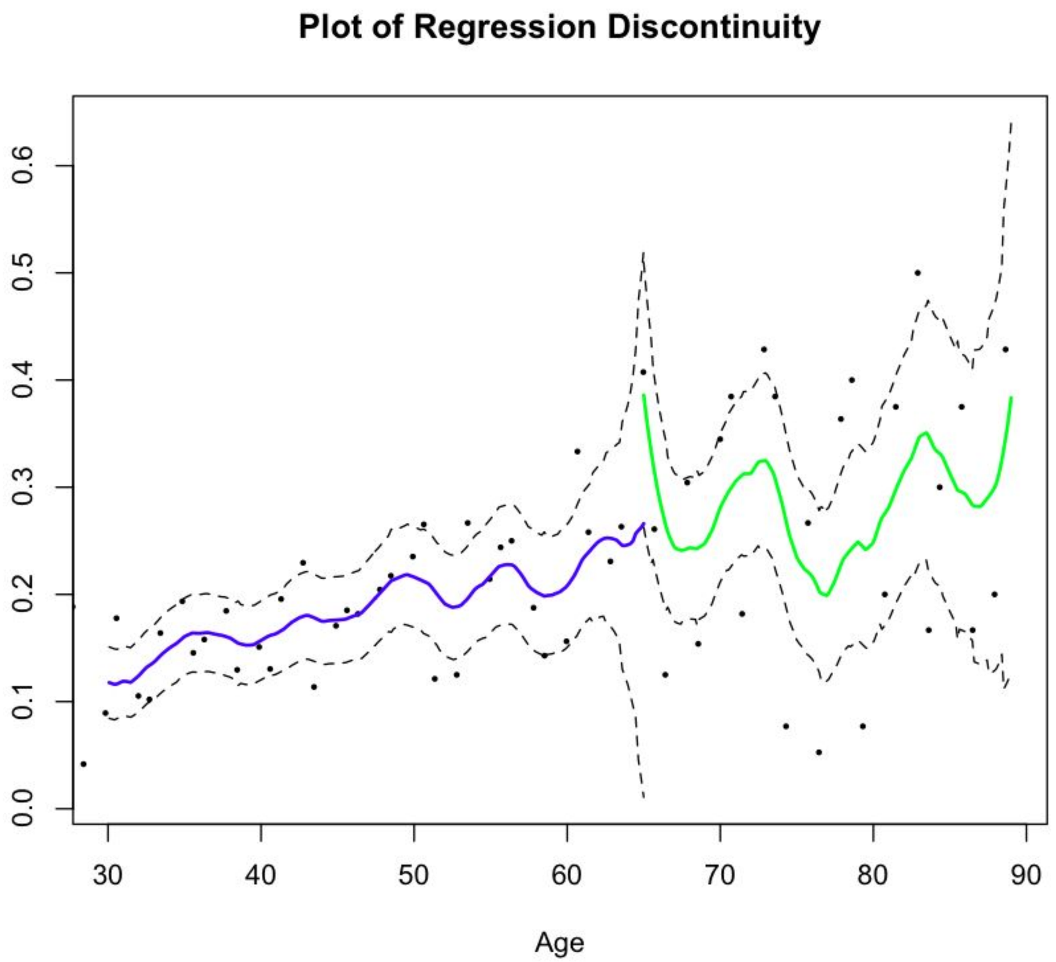
# Non-parametric estimation …

Our results indicate that the discontinuity between the bandwidth immediately to the left of age 65 and the bandwidth immediately to the right of 65 is 0.1199 (or 11.9%-age points), but it is not stat. sig.  With both wider and narrower band-widths, the size of the jump is larger and smaller, respectively

```
Estimates:
           Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
LATE       3.493      173           0.11995   0.1800      0.6665   0.5051
Half-BW    1.746       69           0.14425   0.1415      1.0195   0.3079
Double-BW  6.985      315           0.04279   0.1151      0.3718   0.7100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Look at that break!

Notice lots of smoothing of the lines happening here; this is our preferred bandwidth



# How did I do that?

```
# or use RDplot function in QMSS package
?RDplot
RDplot(rd.godoc, col = c("blue", "green"), pts = T, xlab = "Age")
```

# How to model Fuzzy RD?

- It really looks like an IV approach, because we need to control for treatment "take over"
- In the first stage, we instrument the treatment variable (D) using the assignment variable (X)
- In the second stage, we use the instrumented treatment variable (predicted D) as an independent variable explaining our dependent variable (Y)

# Another great package is here

## RDDtools: an R package for Regression Discontinuity Design

**RDDtools** is a new R package under development, designed to offer a set of tools to run all the steps required for a Regression Discontinuity Design (RDD) Analysis, from primary data visualisation to discontinuity estimation, sensitivity and placebo testing.

### Installing RDDtools

This github website hosts the source code. One of the easiest ways to install the package from github is by using the R package **devtools**:

```
library(devtools)
install_github(repo = "RDDtools", username = "MatthieuStigler", subdir = "RDDtools")
```

Note however the latest version of RDDtools only works with R 3.0, and that you might need to install Rtools if on Windows.

### Documentation

The (preliminary) documentation is available in the help files directly, as well as in the *vignette*. The vignette can be accessed from R with vignette("RDDtools"), or by accessing the pdf stored on this github.

# Another great resource is here


*

# A few more examples

Legewie, Joscha. "Racial profiling and use of force in police stops: How local events trigger periods of increased discrimination." *American journal of sociology* 122.2 (2016): 379-424. Link

Hoekstra, Mark. 2009. “The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach.” *Review of Economics and Statistics* 91 (4): 717–24. Link