Lab 10: Segmented Regression/Regression Discontinuity

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> ”Scientists and artists have kind of the same job...It starts with a gut feeling, an idea, a flash of inspiration, and the rest is hard work.”
>
> *Tom Sachs*

Introduction

The basic idea of Lab is as follows: Answer a research question with the provided dataset. Each week, that research question (and data) will change depending on the topic we've covered the prior class days. Once we're done with Lab, you'll have a Lab Assignment, that will look a lot like the Lab: a research question you'll need to answer given some data. In Lab, you'll learn the procedure for answering the research question. For the Lab Assignment, you'll do that procedure for a grade (independently).

To help answer the research question, we'll follow some basic steps that we'll repeat throughout the semester:

- Reflect on the Question: Figure out the variables of interest, and the technique that's required.
- Analyze the Data: Perform the steps required for the technique.
- Draw Conclusions: Use the information that you got from the prior step to answer the research question in a concise, logical manner.

Let's get started:

Primary Research Question:

Does the underlying regression model predicting posttest math skills score from pretest math skills score change when under performing students are mandated to take additional tutoring sessions? Is there a significant improvement in student's math skills score on posttest, when flagged for inclusion into mandatory tutoring sessions? Cutoff for inclusion is 215 points or lower on the math skills pretest.

Step1: Reflect on the Question:

Download the syntax and data files from Canvas.

Let's load in our SDSRegressionR package so that we can use some of it's functions later:

Remember we need the *newest* version of the package for this lab.

```
#Load our class package
library(SDSRegressionR)
```

Next, we'll load in the data. Be sure to use the basic file structure we talked about the first Lab: Put your syntax in a folder specific to this Lab. Then, make a "data" folder in that same place - use lowercase. If you do that, then all of this syntax will work like a charm.

```
math <- read_csv("data/mathComp.csv")
```

Check the Data:

To make sure that we're working with the right data, and that we're all looking at it the same way, we'll answer some basic questions about the data before moving on:

1. How many observations are in the dataset?
2. What was the pretest score for the first male (gender == 1)?
3. Of the first 10 observations, how many students were younger than 12?

These questions can be answered simply by looking at the dataset once it's loaded in:

```
View(math)
```

Check the Variables of Interest

Let's find the variables that we need to answer the primary research question:

1. Which variable tells us the Outcome variable for this model?
    - What type of variable is this?

2. Which variable will be used on the x-axis?

3. At what value will this variable be "cut"?

Again, these can be answered by looking at the dataframe, and with the help of the *names()* function. Also, the codebook for the data frame is our friend. You can open this in R or Excel. Remember, R is case-sensitive.

```
names(math)
```

```
##  [1] "ID"       "gender"   "sped"     "frlunch"  "esol"     "black"
##  [7] "white"    "hispanic" "asian"    "age"      "pretest"  "posttest"
```

Reflect on the Method

The last part of Reflect on the Question asks about the method or technique we'll use.

1. We will use Segmented Regression/Discontinuity Design to answer this question. Why?

2. We'll need to run at least two regression models. Why?

3. Why would we *not* use bootstrapping to *find* a cut-point (for this Research Question)?

Step2: Analyze the Data

In this step, we'll run the provided syntax and answer some questions about the output to help us prepare for the final step.

Here's the syntax you'll need (from the .R syntax file):

```
#### Here is the R script you will use:  (remember that # indicates a comment) ####
#Lab10: Segmented Regression

library(SDSRegressionR)
#Bring in data
math <- read_csv("data/mathComp.csv")
names(math)

#Establish cut-off
cutoff <- 215

#Code the data for the cut-off
math2 <- math %>%
  mutate_at(vars(pretest), as.numeric) %>% #Initial catch all for numeric...
  mutate(pre1 = pretest, #Simplet replication
         pre2 = pretest - cutoff, #Start second segment counting...
         pre2 = case_when(pre1 <= cutoff ~ 0, #Make sure to start at zero BEFORE segment
                          TRUE ~ pre2),
         jump = case_when(pretest < cutoff ~ 0, #Define the segment status...
                          pretest >= cutoff ~ 1))
#Check
plyr::count(math2, c("pretest", "pre1", "pre2", "jump"))

#Inital model run and diagnostics
full <- lm(posttest ~ pre1 + jump + pre2, data=math2)
residFitted(full)
cooksPlot(full, key.variable="ID", print.obs=TRUE, sort.obs = TRUE)
threeOuts(full, key.variable="ID")

#Get good data...
#Let's remove ALL
cooks <- cooksPlot(full, key.variable="ID", print.obs=TRUE, sort.obs = TRUE)
g_math <- math2 %>%
  filter(ID %not in% cooks$ID)

#Initial look
simpleScatter(g_math, pretest, posttest, title="Raw Data")

#Look with means
mns_math <- g_math %>%
  group_by(pretest) %>%
  summarise(mean = mean(posttest, rm.rm=TRUE))

g_mns <- simpleScatter(mns_math, pretest, mean, title="Means Plot")
g_mns

#Run the model
```

```
seg <- lm(posttest ~ pre1 + jump + pre2, data=g_math)
summary(seg)

#Come up with prediction lines
library(skimr)
g_math %>%
  skim(pretest)

library(emmeans)
p1 <- summary(emmeans(seg, "pre1", at=list(pre1=c(180, cutoff), pre2=0, jump=0)))
p2 <- summary(emmeans(seg, "pre1", at=list(pre1=c(cutoff, 270), pre2=c(0, (270-cutoff)), jump=1),
                      by="pre2"))
p2 <- p2 %>% #Just the first and last row
  slice(c(1,4))

# p1 <- data.frame(pre1=c(min(g_math$pretest),215), pre2=0, jump=0)
# p1 <- data.frame(p1, predict(seg, p1))
# names(p1)[length(p1)] <- "pred"
# p2 <- data.frame(pre1=c(215,max(g_math$pre1)), pre2=c(0,max(g_math$pre2)), jump=1)
# p2 <- data.frame(p2, predict(seg, p2))
# names(p2)[length(p2)] <- "pred"

#Graph it!
g <- simpleScatter(g_math, pretest, posttest, title="Raw Data")
g +
  labs(subtitle="Segmented Regression") +
  geom_vline(xintercept = 215, linetype="dashed", color="green") +
  geom_line(data=p1, aes(x=pre1, y=emmean), color="red") +
  geom_line(data=p2, aes(x=pre1, y=emmean), color="red")

#Graph it! MEANS
g_mns +
  labs(subtitle="Segmented Regression") +
  geom_vline(xintercept = 215, linetype="dashed", color="green") +
  geom_line(data=p1, aes(x=pre1, y=emmean), color="red") +
  geom_line(data=p2, aes(x=pre1, y=emmean), color="red")

#Code for slope of zero
g_math <- g_math %>%
  mutate(pre1_is = case_when(pretest >= cutoff ~ cutoff,
                             TRUE ~ pretest))

#Re-run model
seg_is <- lm(posttest ~ pre1_is + jump + pre2, data=g_math)
summary(seg_is)
```

All quesitons are in reference to models run *after* outlier removal.

Question 1

Was there a significant "main effect" for the segmented regression? Report the meaningful statistics: b = ________, t(________) = ________, p = ________.

Question 2

How much did the mandated tutoring sessions "help" the students below the cutoff score?

Quesiton 3

What was effect of pretest score on posttest score for those **below** below the cutoff (and in the "treatment" group)?

Question 4

What was effect of pretest score on posttest score for those **not below** the cutoff (and **not in** the "treatment" group)?

Question 5

Was there a significant "interaction" for the segmented regression? Answer: ________ Report the meaningful statistics: b = ________, t(________) = ________, p = ________.

Question 6

If the interaction **had been** significant, what are some possible conclusions?

Step3: Draw Conclusions

The final step is for us to Draw Conclusions. We'll take the syntax we've been given from Analyze the Question, run it, then examine the output. The questions from the prior step help set us up for the Draw Conclusions part.

We'll "fill in the blanks" in a canned paragraph for the Lab. For the Lab Assignment, you'll need to come up with a similar paragraph all on your own (please don't steal mine).

Primary Research Question
Does the underlying regression model predicting posttest math skills score from pretest math skills score change when poor performing students are mandated to take additional tutoring sessions? Is there a significant improvement in student's math skills score on posttest, when flagged for inclusion into mandatory tutoring sessions? Cutoff for inclusion is 215 points or lower on the math skills pretest.

> We used a segmented regression (regression discontinuity design) to examine the effect of mandated tutoring sessions on math skills posttest performance. Students scoring 215 or below on the math skills pretest were included into the tutoring sessions. The overall model was significant (R^2 = _______, F(_______, _______) = _______, p < 0.05). Analysis showed a significant main effect of inclusion into the sessions, raising posttest scores, b = _______, t(_______) = _______, p < 0.05. The interaction effect (the difference in the relationship of pretest score to posttest score between the two segments) was also significant, b = _______, t(_______) = _______, p = _______. The effect of pretest on posttest for the mandated group was: b = _______, while the effect for the students above the cutoff criteria was: b = _______.

Lab Assignment

Now, with the tools at your disposal (the R syntax from Lab, and the logic of proceeding through the three steps of answering the research question), you'll have a Lab Assignment to complete (independently). For now, the Lab Assignment is to be completed in Canvas. It will follow the basic structure, and lead to the same place - answering the research question with a concise paragraph as in Draw Conclusions.

Good Luck!