Lab 8: Quant by Quant Interaction

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> "Doing statistics is like doing crosswords except that one cannot know for sure whether one has found the solution."
>
> *John Tukey*

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:

Examine the effect of Age, D125 metabolite, BMI, and parathyroid hormone (PTH) on Vitamin D blood serum levels (D25). Does Age moderate the relationship of BMI on D25 or PTH on D25? If an interaction exists, investigate and report.

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:

```
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.

```
vit <- read_csv("data/VitaminD.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 for the model?

2. What was weight for the first male?

3. Of the first 10 participants, how many had a standardized bmi for age score less than -2.00?

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

```
View(vit)
```

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 of Vitamin D serum levels?
   - What type of variable is this?

2. What is the "variable(s) of interest" for this model?
   - What is the moderator for this model?

3. List *all* the variables that will go into this model:

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(vit)
```

```
##  [1] "PartID"   "sex"      "ageyears" "height"   "weight"   "bmi"
##  [7] "zbfa"     "wasted"   "zhfa"     "stunted"  "period"   "school"
## [13] "pth"      "D25"      "D125"
```

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 Multiple Linear Regression with an interaction to answer this Lab question. Why?

2. We'll need to mean center or variable(s) of interest and moderator. Why?

3. We'll use Simple Slopes at -1SD and +1SD. Why?

4. We will also need to look at the "Regions of Significance." Why? What do they tell us?

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) ####
#Lab8: Quantitative Interaction

library(SDSRegressionR)

#Load Data
vit <- read_csv("data/VitaminD.csv")
names(vit)

#Any categorical variables?

#Run first model for diagnostics FIRST
#First run
v_mod <- lm(D25 ~ D125 + ageyears + bmi + pth + bmi*ageyears + pth*ageyears, data=vit)
summary(v_mod)

#Diagnostics
library(car)
vif(v_mod)
residFitted(v_mod)
cooksPlot(v_mod, key.variable = "PartID", print.obs=TRUE, sort.obs=TRUE)

#Remove bad outlier(s)
g_vit <- vit %>%
  filter(PartID %not in% c("ID178"))

#Re-run
v_mod2 <- lm(D25 ~ D125 + ageyears + bmi + pth + bmi*ageyears + pth*ageyears, data=g_vit)
summary(v_mod2)

#Find the Simple Slope locations (Pick-a-Point)
m_data <- modelData(v_mod2)
mean(m_data$ageyears, na.rm=TRUE)
mean(m_data$ageyears, na.rm=TRUE) - sd(m_data$ageyears, na.rm=TRUE)
mean(m_data$ageyears, na.rm=TRUE) + sd(m_data$ageyears, na.rm=TRUE)

#lsmeans for simple slopes
library(emmeans)
simple_mns <- emmeans(v_mod2, "bmi",
                      at=list(bmi=c(0,1), ageyears=c(8.18, 9.84, 11.50)),
                      by="ageyears")
simple_mns
pairs(simple_mns, reverse=TRUE)

# And Find the Regions of Significance
lmROS(v_mod2, interest = "bmi", moderator = "ageyears")
```

```
#Better graphs
#ROS
ROS +
  labs(x="Age in Years", title="ROS of Age", subtitle="BMI predicting D25 Serum")

#Simple Slopes
mns <- summary(emmeans(v_mod2, "bmi",
                       at=list(bmi=seq(10, 30, 2), ageyears=c(8.18, 9.84, 11.50)),
                       by="ageyears"))
simpleScatter(g_vit, x=bmi, y=D25, ptalpha = 0,
              title="BMI on D25 Moderated by Age") +
  geom_line(data=mns, aes(x=bmi, y=emmean, linetype=factor(ageyears))) +
  scale_linetype_manual(name = "Age Values",
                        values = c("dashed", "dotdash", "dotted"),
                        labels = c("One SD Below","Mean of Age", "One SD Above"))

#More fun from the R community!
library(jtools)
sim_slopes(v_mod2, pred = "bmi", modx = "ageyears")
johnson_neyman(v_mod2, pred = "bmi", modx = "ageyears")
interact_plot(v_mod2, pred = "bmi", modx = "ageyears")
```

Question 1

After removing the outliers, which of the two interactions investigated were significant?

Question 2

The overall model was _____, with an F(_____, _____) = _____, p _____ 0.05.

Question 3

The interaction between BMI and Age was significant: t(_____) = _____, p = _____.

Question 4

Just by looking at the model output, the Simple Slope for the effect of BMI on D25 blood serum is 5.118, for what value of **Age**?

Question 5

As the value of _____ increases by one unit, the slope between BMI and D25 blood serum changes by _____.

Question 6

The Simple Slope for 1SD below the mean of age is _____, (t(_____) = _____, p = _____), while the Simple Slope for 1SD above the mean of age is _____ (t(_____) = _____, p = _____).

Question 7

The Region(s) of Significance on Age that show a change in significance for the slope of BMI on D25 are _____ and _____.

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
Examine the effect of Age, D125 metabolite, BMI, and parathyroid hormone (PTH) on Vitamin D blood serum levels (D25). Does Age moderate the relationship of BMI on D25 or PTH on D25? If an interaction exists, investigate and report.

> After establishing a model free from influential outliers, we investigated the possibility of Age moderating the relationships of BMI and Parathyroid Hormone (PTH) on Vitamin D25 blood serum levels. All quantitative variables in the interactions were mean centered prior to inclusion in the model. The overall model showed good prediction of D25 (F(_____, _____) = _____, $p < 0.05$), with an R^2 of _____. The interaction between Age and BMI was significant (b = _____, t(_____) = _____, $p < 0.05$) while the interaction between Age and PTH was not (b = _____, t(_____) = _____, p = _____).
>
> Decomposing the interaction of Age and BMI showed that as age increases, the slope of BMI on D25 becomes negative. Using the Pick-a-Point approach recommended by Aiken & West (1991), at 1 Standard Deviation (1.68 years) below the mean of Age the Simple Slope of BMI on D25 was _____ (t(_____) = _____, p = _____), while at 1 Standard Deviation above the mean of Age the Simple Slope was _____ (t(_____) = _____, $p < 0.05$). Using the Johnson-Neyman technique (Bauer & Curran, 2005), the Regions of Significance for the impact of BMI on D25 were determined to be at Age values of _____ and _____. Below Age values of 10.08, the impact of BMI on D25 becomes _____ and _____. Children older than _____ will have show a significant _____ relationship between BMI and D25 blood serum levels.

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!