Lab 4: Multiple Regression

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> "Statistics is the most important science in the whole world, for upon it depends the practical application of every other [science] and of every art."
>
> *Florence Nightengale*

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:

For students enrolled in the Clincal Sciences program of a South American Medical School, examine the effect of DREEM: Social Self Perception, DREEM: Academic Self Perception, Resilience, BDI, and Age on Med School Quality of Life (MS.QoL). After controlling for all the variables of interest, which is the strongest predictor of Med School Quality of Life - and what is the nature of that relationship?

Step1: Reflect on the Question:

Download the syntax and data files from Canvas.

We'll need to update the SDSRegressionR package because of an update. Then, let's load in our SDSRegressionR package so that we can use some of it's functions later:

```
#Load our class packags
library(SDSRegressionR)
library(tidyverse)
```

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.

```
clinical <- read_csv("data/Clinical.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. The first student with a Med School QoL of 10 is how old?

3. Of the first 10 participants, how many have a Med School QoL over 5?

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

```
View(clinical)
```

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 Medical School Quality of Life?
   - What type of variable is this?

2. Is there just one other predictor of the above variable?
   - List the variables that will be used to predict the above variable of interest.

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(clinical)
```

```
##  [1] "IDR"             "Female"          "Group"
##  [4] "QoL"             "MS.QoL"          "WHOQOL.PH"
##  [7] "WHOQOL.PSY"      "WHOQOL.SOC"      "WHOQOL.ENV"
## [10] "DREEM.L"         "DREEM.T"         "DREEM.A.SP"
## [13] "DREEM.At"        "DREEM.S.SP"      "DREEM.GS"
## [16] "Resilience"      "BDI"             "Age"
## [19] "School.location" "State.Anxiety"   "Trait.anxiety"
```

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 to answer this Lab question. Why?

2. We should examine our independent variables for something called tolerance. Why?

3. We should also examine our residual vs. fitted plot for the MLR. Why?

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) ####
#Lab4: Multiple Linear Regression
library(SDSRegressionR)
library(tidyverse)

#import data...
clinical <- read.csv("data/Clinical.csv", stringsAsFactors=FALSE)
names(clinical)

#Inital correlationss
vars <- c("MS.QoL", "DREEM.S.SP", "DREEM.A.SP", "Resilience", "BDI", "Age")
library(psych)
corr.test(select(clinical, one_of(vars)))

#First model
q_mod <- lm(MS.QoL ~ DREEM.S.SP + DREEM.A.SP + Resilience + BDI + Age, data=clinical)
summary(q_mod)
library(car)
vif(q_mod)

#Remove the problem independent:
q_mod2 <- lm(MS.QoL ~ DREEM.S.SP + Resilience + BDI + Age, data=clinical)
summary(q_mod2)
vif(q_mod2)

#Good model: Check assumptions
residFitted(q_mod2)

#Find the outliers...
cooksPlot(q_mod2, key.variable = "IDR", print.obs = TRUE, sort.obs = TRUE)
threeOuts(q_mod2, key.variable = "IDR")

#Remove the outlier(s)
good_clin <- clinical %>%
  filter(IDR %not in% c("IDR897"))

#Final Model
q_mod_f <- lm(MS.QoL ~ DREEM.S.SP + Resilience + BDI + Age, data=good_clin)
summary(q_mod_f)
residFitted(q_mod_f) #Just checking
confint(q_mod_f) #Confidence intervals for the slopes (for reporting)
lmBeta(q_mod_f) #Standardized Betas for our final model
pCorr(q_mod_f) #Partial and Part correlation coefficients

#Predict
library(emmeans)
ref_grid(q_mod_f)
```

```
lsmeans(q_mod_f, "DREEM.S.SP", at=list(DREEM.S.SP=10))

#Visualize...
#Get the graph....
s.sp_gr <- simpleScatter(good_clin, DREEM.S.SP, MS.QoL,
              title="Social Perception and Quality of Life",
              xlab="DREEM Social Self Perception", ylab="Quality of Life")

#New "mean" data and prediction for fit and confidence
library(psych)
describe(good_clin$DREEM.S.SP)
pgr <- summary(lsmeans(q_mod_f, "DREEM.S.SP", at=list(DREEM.S.SP=seq(8, 26, 0.5))))
pgr

#Add fit and confidence
s.sp_gr +
  geom_line(data=pgr, aes(x=DREEM.S.SP, y=lsmean), color="red") +
  geom_ribbon(data=pgr, aes(x=DREEM.S.SP, y=lsmean, ymin=lower.CL, ymax=upper.CL), alpha=0.3)
```

Question 1

Our initial Multiple Regression model ("q_mod") shows a multiple R^2 of _____ . This overall model was significant with an F(_____, _____) = _____.

Question 2

Investigation of the initial model's multicollinearity showed a problem. The variable "DREEM.A.SP" had a Variance Inflation Factor (VIF) value of _____, well above the acceptable cut-off of _____ . The corresponding tolerance value for this variable was _____ (this answer is **not** in the above syntax…but you know it).

Quesiton 3

After removing the offending independent variable, the new model showed a Multiple R^2 of _____, with an Adjusted R^2 of _____ (F(4, 167) = _____, $p < 0.05$).

Question 4

We found and removed the single outlier by using the following code:

```
cooksPlot(q_mod2, key.variable = "IDR")
```

The participant "IDR897" was removed because of a high _____ value of _____.

Question 5

After removing the outlier, the final model shows that the most impactful predictor of Med School Quality of Life was _____, with a slope of _____ (t(_____) = _____, $p < 0.05$). 95% confidence interval for the slope of the coefficient was between _____ and _____. This variable *uniquely* contributed to the overall variance of Med School Quality of Life by _____ %.

Question 6

The final model allows us to say that we are 95% confident that students with Social Self Perception score of 10 will have an *average* Med School Quality of Life between _____ 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).

> We investigated the effects of multiple variables on the the prediction of Med School Quality of Life. These variables included DREEM: Social Self Perception, DREEM: Academic Self Perception, Resilience, BDI, and Age. An initial model showed that DREEM: Academic Self Perception was highly multicollinear with a Variance Inflation Factor of _____. This independent was subsequently removed from the model. Further investigation showed the presence of a single outlier with a large _____. The final model met our assumptions of _____, normality of residuals, and _____.
>
> The final model had an R^2 value of _____, with an Adjusted R^2 of _____ ($F(4, 166) =$ _____, $p < 0.05$). The best predictor of Med School Quality of Life was $b =$ _____ with a Standardized beta $\beta =$ _____, (t(_____) = _____, $p < 0.05$) (See Table 1 for a list of all coefficients.)

```
library(stargazer)
stargazer(q_mod, q_mod2, q_mod_f, title="Final Model of Med School Quality of Life",
          column.labels   = c("First Model", "Second Model", "Final Model"),
          model.numbers = FALSE, single.row=TRUE, header=FALSE,
          omit.stat="ser")
```

Table 1: Final Model of Med School Quality of Life

	Dependent variable:		
	MS.QoL		
	First Model	Second Model	Final Model
DREEM.S.SP	0.167 (0.191)	0.214*** (0.029)	0.204*** (0.027)
DREEM.A.SP	0.048 (0.194)		
Resilience	0.003 (0.011)	0.003 (0.011)	0.006 (0.011)
BDI	0.004 (0.024)	0.001 (0.020)	0.007 (0.019)
Age	−0.057 (0.037)	−0.059* (0.035)	−0.070** (0.034)
Constant	4.161** (1.716)	4.438*** (1.302)	4.546*** (1.244)
Observations	172	172	171
R^2	0.353	0.353	0.355
Adjusted R^2	0.333	0.337	0.340
F Statistic	18.094*** (df = 5; 166)	22.730*** (df = 4; 167)	22.863*** (df = 4; 166)
Note:			*p<0.1; **p<0.05; ***p<0.01

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!