Lab 5: Sequential Multiple Regression

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> "Data do not give up their secrets easily. They must be tortured to confess."
>
> *Jeff Hopper, Bell Labs*

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:

After controlling for Age, Gender, Parental status, and Full-time school status, does sense of coherence and stress impact account for a significantly greater proportion of the variance in quality of life of Norwegian nursing students? Explain the final model. Use the mean scales when able.

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:

```
#Load our class package
devtools::install_github("MichaelJMahometa/SDSRegressionR") #Yes - a new update
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.

```
coh <- read_csv("data/coherance.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 mean coherence score for the first participant 40 or older?
3. Of the first 10 participants, how many had a quality of life mean score under 3.0?

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

```
View(coh)
```

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 quality of life of a participant?
   - What type of variable is this?
   - What scale is this variable on?
2. What are the nuisance variables for the research question?
3. What are the variables of interest for the research question?

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(coh)
```

```
##  [1] "SubID"          "age"            "female"         "full_ed"
##  [5] "child"          "marital"        "practice"       "coherance"
##  [9] "coherance_mean" "quallife"       "quallife_mean"  "impact"
## [13] "impact_mean"
```

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 Sequential linear regression to answer this Lab question. Why?
2. We should examine our final model first. Why?
3. What test will be used to answer our primary research question. 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) ####
#Lab5: Sequential Regression

library(SDSRegressionR)

#import data...
coh <- read_csv("data/coherance.csv")
names(coh)

#Determine and run the final model
full <- lm(quallife_mean ~ age + female + child + full_ed + coherance_mean +
             impact_mean, data=coh)

#Look for any issues:
library(car)
vif(full)
residFitted(full)
cooksPlot(full, key.variable = "SubID", print.obs = TRUE,
          sort.obs=TRUE, save.cutoff = TRUE)
cooksCutOff * 2
threeOuts(full, key.variable = "SubID")

#Clean up
good_coh <- coh %>%
  filter(SubID %not in% c(...)) #You should complete this part...

#Re-run the final model
fullg <- lm(quallife_mean ~ age + female + child + full_ed + coherance_mean +
              impact_mean, data=good_coh)

#Get the "model data" for the nesting
good_coh_m2 <- modelData(fullg)

#Now for the Sequential Regression:
#Model 1:
m1_seq <- lm(quallife_mean ~ age + female + child + full_ed, data=good_coh_m2)
summary(m1_seq)
summary(m1_seq)$r.squared
lmBeta(m1_seq)
pCorr(m1_seq)

#Model 2:
m2_seq <- lm(quallife_mean ~ age + female + child + full_ed + coherance_mean +
               impact_mean, data=good_coh_m2)
summary(m2_seq)
summary(m2_seq)$r.squared
lmBeta(m2_seq)
```

```
pCorr(m2_seq)

#Now the Sequential Results
anova(m1_seq, m2_seq)
```

Question 1

Our first full Multiple Regression model ("full") showed that our data wasn't quite OK for the model. Why? Observation(s) _____ was(were) considered outliers because _____.

Question 2

What was the purpose of running the modelData() function?

Quesiton 3

The first model of the Sequential Regression showed a model R^2 of _____. This model was _____, with an F(_____, _____) = _____. The best predictor of quality of life was _____.

Question 4

The second model of the Sequential Regression showed a model R^2 of _____. This model was _____, with an F(_____, _____) = _____. The best predictor of quality of life was _____.

Question 5

Our Sequential Regression shows that the additional predictors of coherence and stress impact to the initial model increased our R^2 by _____. This change in R^2 was _____, with an F(_____, _____) = _____, p _____ 0.05.

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).

> Our primary research question investigated the additive effects of stress impact and coherance over and above the nuisance variables of age, gender, parent status, and full time education status, in the prediction of Quality of Life. This question was answered with a Sequential Regression Model. Our intial model of control variables accounted for _____% of the variance in Quality of Life, F(_____, _____) = _____, p _____ 0.05.
>
> The second model accounted for an additional _____ % of variance in Quality of Life. This change in R^2 was significant F(_____, _____) = _____, p _____ 0.05. The second model with all predictors showed a significant overall model, F(_____, _____) = _____, p _____ 0.05. Of the two additional predictors, only coherence showed a significant impact (b = _____, t(_____) = _____, $p < 0.05$) and could uniquely account for _____% of the variance in the mean quality of life.

Optional use of the stargazer package for pretty results: See next page:

```
library(stargazer)
library(knitr)
stargazer(m1_seq, m2_seq, title="Quality of Life Sequential Regression",
          column.labels   = c("First Model", "Second Model"),
          model.numbers = FALSE, single.row=TRUE, header=FALSE,
          omit.stat="ser")
```

Table 1: Quality of Life Sequential Regression

	Dependent variable:	
	quallife_mean	
	First Model	Second Model
age	−0.010 (0.009)	−0.001 (0.007)
female	−0.027 (0.131)	0.027 (0.102)
child	0.152 (0.166)	0.014 (0.128)
full_ed	0.066 (0.150)	0.195* (0.116)
coherance_mean		0.520*** (0.046)
impact_mean		−0.022 (0.038)
Constant	5.523*** (0.346)	2.713*** (0.365)
Observations	220	220
R^2	0.009	0.420
Adjusted R^2	−0.010	0.404
F Statistic	0.473 (df = 4; 215)	25.691*** (df = 6; 213)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
kable(anova(m1_seq, m2_seq), caption="Quality of Life Sequential
      Regression Rsults")
```

Table 2: Quality of Life Sequential Regression Rsults

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
215	90.55355	NA	NA	NA	NA
213	52.99640	2	37.55715	75.47375	0

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!