Lab 2: Simple Linear Regression

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> "The best thing about being a statistician is that you get to play in everyone's backyard."
>
> *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:

You have data from 51 undergraduate students enrolled at an international university (University of Daugavpils, Latvia). Participants were measured for blood cortisol levels and percentage body fat. Participants were also rated on their perceived level of "agreeableness" from a single photograph. The primary research question is: Does blood cortisol level, significantly predict perceived agreeableness? If so, what is the nature of that relationship? Additional questions to follow…

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 packages
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.

```
agree <- read_csv("data/CortisolAgree.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 age of the first student with a cortisol level under 300?

3. Of the first 10 participants, how many had a negative agreeableness score?

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

```
View(agree)
```

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 agreeableness score?
   - What type of variable is this?

2. Which variable tells us the about the blood cortisol levels?
   - What type of variable is this?

Again, these can be answered by looking at the dataframe, and with the help of the *names()* function. Remember, R is case-sensitive.

```
names(agree)
```

```
## [1] "UniqueID"      "Age"           "Agreeableness" "AntiHBS"
## [5] "PlasmaCort"    "PerFat"
```

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 a Pearson correlation to begin answering this Lab question. Why?
2. We will use simple regression to answer this Lab question. Why?

3. We should generate a scatterplot of these two variables before we continue our analysis. Why?

4. We should look for outliers before we complete our analysis. 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) ####
##Lab 2: Simple Regression

#Read in the data (or use the Import Dataset button option)
agree <- read_csv("data/CortisolAgree.csv")

#Run a correlation matrix, and get t and p values
library(psych)
corr.test(select(agree, Agreeableness, AntiHBS, PlasmaCort, PerFat, Age))
corr.test(select(agree, Agreeableness, AntiHBS, PlasmaCort, PerFat, Age))$t
corr.test(select(agree, Agreeableness, AntiHBS, PlasmaCort, PerFat, Age))$p

#Check for linearity
simpleScatter(agree, PlasmaCort, Agreeableness, line=TRUE)

#Run a Simple Linear Regression model
c_mod <- lm(Agreeableness ~ PlasmaCort, agree)
c_mod

#Check for Homoscedastity
residFitted(c_mod)

#Check for outliers
studResidPlot(c_mod, key.variable = "UniqueID")
levPlot(c_mod, key.variable = "UniqueID")
cooksPlot(c_mod, key.variable = "UniqueID")
threeOuts(c_mod, key.variable = "UniqueID")

#Remove "outliers"
agree_no_out <- agree %>%
  filter(UniqueID %not in% c(61, 44))

#Let's look at the outliers too...
agree %>%
  filter(UniqueID %in% c(61, 44))

#Re-run the correlations to see a change
library(psych)
corr.test(select(agree_no_out, Agreeableness, AntiHBS, PlasmaCort, PerFat, Age))
corr.test(select(agree_no_out, Agreeableness, AntiHBS, PlasmaCort, PerFat, Age))$t
corr.test(select(agree_no_out, Agreeableness, AntiHBS, PlasmaCort, PerFat, Age))$p

#Re-run the Simple Linear Regression Model
simpleScatter(agree_no_out, PlasmaCort, Agreeableness, line=TRUE)
c_mod2 <- lm(Agreeableness ~ PlasmaCort, agree_no_out)
c_mod2
```

Question 1

Our initial correlation matrix shows the underlying relationship of cortisol and agreeableness to be _____ . This relationship was _____, (t(_____) = _____, p = _____.

The initial scatter plot shows us that the relationship between of cortisol and agreeableness is _____, _____, and _____ .

Question 2

After running our initial model, we see that cortisol can account for _____ % of the variance in perceived agreeableness.

Quesiton 3

The initial model shows us that the as cortisol levels increase increases by _____ (in micromols), the average score of subject agreeableness will changed by _____ units.

Question 4

We found and removed two outliers by using the following code:

```
c_mod <- lm(Agreeableness ~ PlasmaCort, agree)
threeOuts(c_mod, key.variable = "UniqueID")
```

```
agree_no_out <- agree %>%
  filter(UniqueID %not in% c(61, 44))
```

The participant with UniqueID 61 had a cortisol level of _____, with an agreeableness score of _____. The participant with UniqueID had a cortisol level of _____, with an agreeableness score of _____.

Question 5

After removing the outliers, the correlation between cortisol and agreeableness _____, to a value of _____.

Question 6

Once the outliers were removed, the new Simple Linear Regression model slope was _____, with an Intercept of _____. The new model accounts for _____% of the variance in agreeableness from cortisol levels.

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).

> An intial investigation into the relationship between blood cortisol levels and perceived agreeableness showed a significant Pearson correlation, r = _____, t(_____) = , p = _____. The initial model shows cortisol levels acocunting for _____ % of the variance in percieved agreebleness. Two outliers with a Cook's distance of _____ and _____, were found and removed. After outlier removal, the model could account for _____ %, of the variance, r = _____, t(_____) = , p = _____. Finally, the new model showed an increase in slope from _____ to _____.

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!