Lab 1: Correlation

SDS358: Applied Regression Analysis

Michael J. Mahometa, Ph.D.

.

> ”Statistics is the grammar of science”
>
> *Karl Pearson*

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 selected European countries, is there a linear relationship between the incidence of Prostate Cancer (IPC) and the use of oral contraceptives (OC). Out of the all the types of contraception, which has the highest relationship to IPC?

Step1: Reflect on the Question:

Download the syntax and data files from Canvas.

Let's load in our **SDSRegressionR** and the **tidyverse** packages so that we can use some of it's functions later:

```
#Load our class package
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.

```
world <- read_csv("data/whole_world_ocp.csv")
```

```
## Parsed with column specification:
## cols(
##   Countryorarea = col_character(),
##   gdp = col_integer(),
##   incidence = col_double(),
##   mortality = col_double(),
##   Pill = col_double(),
##   IUD = col_integer(),
##   Condom = col_integer(),
##   Vaginalbarrier = col_integer(),
##   europe = col_integer()
## )
```

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 is the GDP for Ethiopia?

3. Of the first 10 countries, how many have an oral contraceptive percentage of more than 10%?

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

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 incidence of Prostate cancer per county?
   - What type of variable is this?

2. Which variable tells us the percentage of women in a country using oral contraceptives?
   - 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(world)
```

```
## [1] "Countryorarea"  "gdp"            "incidence"      "mortality"
## [5] "Pill"           "IUD"            "Condom"         "Vaginalbarrier"
## [9] "europe"
```

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 correlation to answer this Lab question. Why?

2. We should generate a scatterplot of these two variables before we continue 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):

```r
#### Here is the R script you will use:  (remember that # indicates a comment) ####
library(SDSRegressionR)
library(tidyverse)

#Load the Data:
world <- read_csv("data/whole_world_ocp.csv")

#Subset for countries in Europe
new_world <- world %>%
  filter(europe == 1)

# Visualize and describe the first variable of interest
histogram(new_world$incidence)
fivenum(new_world$incidence)
mean(new_world$incidence)
sd(new_world$incidence)

# Visualize and describe the second variable of interest
histogram(new_world$Pill)
fivenum(new_world$Pill)
mean(new_world$Pill)
sd(new_world$Pill)

# Create a scatterplot
simpleScatter(new_world, Pill, incidence)

# Add line of best fit
simpleScatter(new_world, Pill, incidence, line = TRUE)

# Calculate the correlation coefficient
cor(select(new_world, Pill, incidence))

# Create a correlation matrix
cor(select(new_world, Pill, incidence, Condom, Vaginalbarrier),
    use="pairwise.complete.obs")

# Create a correlation matrix: p-values
library(psych)
corr.test(select(new_world, Pill, incidence, Condom, Vaginalbarrier),
          use="pairwise.complete.obs")$p
```

Question 1

What do the histogram and descriptive statistics tell us about the distribution of the our variables of interest?

On average, European countries have an Prostate Cancer Incidence of: ___________ .
Of these European countries, an average of ___________ women used oral contraception. (The median is

appropriate here.)

Question 2

What does the scatterplot show us?
The relationship looks __________, __________, and __________ .

Quesiton 3

The correlation, rounded to three decimal places, between the Prostate Cancer Incidences and oral contraception use is: r = __________ .
How many times does this value appear in the correlation matrix?
Is this the highest relationship to Prostate Cancer Incidence?

Question 4

On the scatterplot, we see a data point that doesn't really "follow" the others. This country has a low oral contraception usage (about 17%), but a high IPC.

Use this code to help identify this country:

```r
#Identify a specific record
filter(new_world, Pill > 17 & incidence < 35)

#Remove that record
noHungary <- new_world %>%
  filter(Countryorarea != "Hungary")
```

What happens to the correlation value with this country removed?

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).

> There is a moderately strong, __________, linear relationship between the use of oral contraception and incidence of Prostate Cancer (IPC) in european contries, r = __________, with a p-value of __________ . The average IPc was __________, while the average use of oral contraception was __________ . Of all the contraception choices (OC, IUD, Condom, and VB), __________ had the highest relationship to IPC (all other r's less than 0.010). Removal of a potential outlier of Hungary (with a low oral conception of 17% and a high IPC) showed a new Pearson Correlation of r = __________.

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!