**Zachary Shin**

**Introduction to Linear Model: Hand Span and Height**

**As a result of completing this exercise you should be able to:**

* Calculate the correlation and covariance using cor and cov functions in R
* Use correlation coefficient to describe the relationship between two quantitative variables
* Create scatter plots using plot function
* Find the best fit line using a simple regression model with lm function
* Draw the best line on the scatter plot using abline function
* Get from the liner model
* Explain in your model

**Instructions**

1. We will **use the dataset collected from previous ISAT 251 students called “ISAT251\_HandSpanHeight.csv”.** The data were collected by each student, he/she 1) recorded his/her own hand span and height and 2) found other 4 people who are outside of our ISAT 251 class to measure their hand span and height. To measure the hand span: 1) open the right hand as widely as possible and 2) measure the distance from the thumb to pinky using *centimeter* as the unit

The dataset includes the following variables or features:

* student: a student observer’s jmu eid (the id on his/her duke email)
* participant.id: the id of the participant (1 for the first participant, 2 for the second participant, and so on)
* gender: the participant’s gender (M as male, F as female)
* right.hand.span: the span of the participant’s right hand (from thumb to pinky) in **centimeters**
* height: the height of the participant in **centimeters**.

The top row, header, of csv file will be labeled like this (with one word in each cell):

student participant.id gender age right.hand.span height

1. Now it’s time to **bring the data into R**.
2. Go to the **Canvas** and download the data file **“ISAT251\_HandSpanHeight.csv”**. Move the file to a directory on your local machine that you can access using R and RStudio.
3. Open the RStudio Statistical Software. Use file.choose function to navigate and get the full path to your file where your hand span and height data stored in it. Then, store the path to a variable call myfile.

> myfile <- file.choose()

> myfile

**The full path to *your* file is: /Users/zshindc/Downloads/ISAT251\_HandSpanHeight.csv**

1. Create a new variable called handspan\_height which will hold every observation for records everyone collected outside of the class. The argument header=TRUE tells R that the first row of my data set contains variable or column names.

> handspan\_height <- read.csv(myfile,header=TRUE)

Once you’ve loaded the CSV file using read.csv, you can check to make sure it’s there with the head command, which pulls out the first six observations only.

> head(handspan\_height)

**Snapshot and paste your R code and the first six lines of your data set here:**

**A picture containing bird

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

1. Before you jump into your analyses, please use the summary function to get an overview of your data as well s to check whether there are weird data values or missing values

> summary(handspan\_height)

**Snapshot and paste your R code and the summary of your data set here:**

**A picture containing bird

Description automatically generated**

**A screenshot of text

Description automatically generated**

**Now, we would like to explore the relationship between the two quantitative variables, right.hand.span and height. Specifically, the research question we’d like to answer is whether a taller person has a wider hand.**

1. Create a scatter plots with the two quantitative variables, right.hand.span and height, using plot function. (Please identify the x variable and y variable based on the research question we would like to answer.) **Snapshot and paste your R code and your scatter plot below and answer the following question. *According to the scatter plot, how would you describe the relationship between the two variables, in terms of direction, form, and strength?***

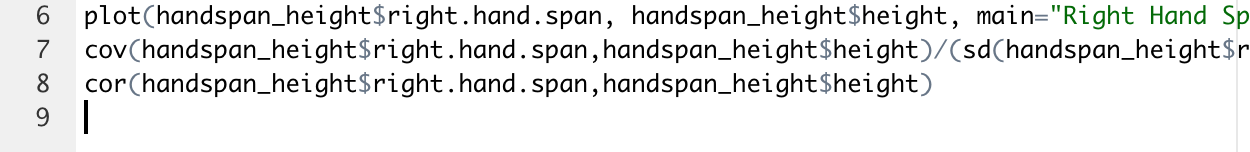
A screenshot of a social media post

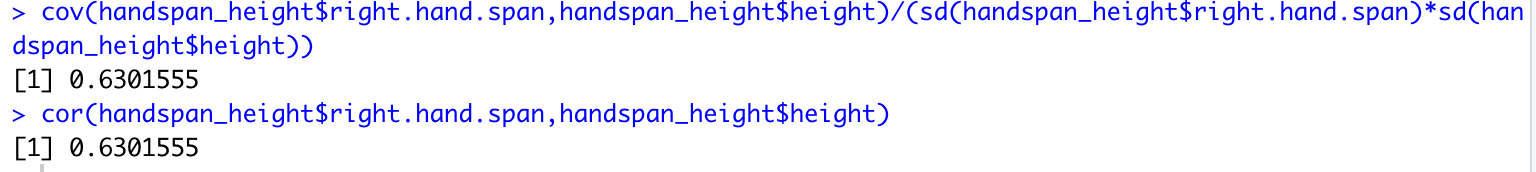
Description automatically generatedA close up of a piece of paper

Description automatically generated

There is a weak, positive, linear association between the two variables.

1. Identify the relationship between the two quantitative variables, right.hand.span and height. Calculate the covariance and correlation using cov, sd, cor functions.
   1. Run the below function to find and type your correlation coefficient out. **Snapshot and paste your R code and the output and answer the question. Do you get the same value from cov(x, y)/(sd(x)\*sd(y))equation and cor(x, y)function? (*Please replace x and y with your or proper variables*)**

****

****

**Sam values from both the equation and function.**

* 1. Answer the two questions: what does the correlation coefficient tell you? Does it align with what you find in step 2 from the scatter plot?

The correlation coefficient tells how strong a relationship is between two quantitative variables. Yes because the scatter plot is linearly positive.

1. Find the best fit line (simple linear regression model) using lm function. *Please replace x and y by your or proper variables.* Then, using summary function to show the output of lm function, specifically the **slope** and **intercept**.

> fit <- lm(y~x)

> summary(fit)

**Snapshot and paste your R code and the summary of your regression model below. In addition, use that information to type out the equation of your line model.**

A picture containing knife

Description automatically generated

A screenshot of text

Description automatically generated

Y = 2.5048x + 124.8027

1. Explain what tells you about the linear fit.

R^2 tells how close the data are to the fitted regression line.

1. Get the using the model you get from the linear regression model (the output of the lm function) in step 4. Show the value of and the value of the correlation coefficient (see step 3) squared (). Are they the same?

R^2 = 0.3971, 0.3944

Correlation coefficient^2 = 0.6301555^2 = 0.3971

R^2 and the correlation coefficient squared are the same

1. Create a scatter plots with the two quantitative variables, right.hand.span and height, using plot function, if you haven’t done so. Then, **add the best fit line on the scatter plot** using abline function. **Snapshot your codes and the plot and paste them below** and **answer the question, does the line you find out align with the correlation coefficient you found in step 3, and how?**

A close up of a piece of paper

Description automatically generated

