Lab Challenge 10 – Correlation and Regression

Due Date: 11:59 pm, day before next class

Each challenge is graded out of 2 points:

- 0 points – no attempt or no progress to a solution
- 1 point – challenge not fully completed or completed with major errors
- 2 points – challenge fully completed with at most a small error

Deliverables

1. A single pdf document containing your solutions to the challenges you completed.
2. An RStudio file (.R extension) containing a *complete* script used to generate your results.

Challenges

Import the data "F2021_MATH_1350_Data.xlsx" from Learning Hub. Even though this data is only for MATH 1350 students, for the purpose of this lab we will consider this group of students to be a simple random sample of all BCIT students.

1. Consider the variables $X =$ Siblings and $Y =$ Income.Goal
   a. Create a scatter plot of Y against X. Add labels and a title.
   b. Calculate the linear correlation coefficient r and test whether it is statistically significant.
   c. Find the equation of the regression line $\hat{y} = a + bx$. Plot the regression line on top of the scatter plot (use `col = "red"`).
   d. Use the regression line to predict the income goal of a student who has $x = 4$ siblings.

2. Import the data set in the file `SOCR-HeightWeight.txt,` which contains the height and weight for a population of 25,000 people. Let $X =$ Height.Inches and $Y =$ Weight.Pounds.
   a. Create a scatter plot of Y against X. Add appropriate labels and a title.
   b. Calculate the population correlation coefficient ρ.
   c. Use R to *simulate* selecting 10^4 random samples (no replacement) of size $n = 30$ from this population. For each sample, find r and perform a significance test. Plot a histogram of the values of r obtained for the 10^4 samples. Label it appropriately.
   d. For what percentage of samples does the correlation indicate a *positive* linear correlation?
   e. For what percentage of random samples does the significance test correctly determine that there is a non-zero population correlation ρ? (This is called the *power* of the test.)

3. In this challenge, you will use the data for MATH 1350 students to find a multi-linear model for the variable Income.Goal.
   a. Find a multi-linear model for Income.Goal in terms of the other three numerical variables: $X_1 =$ Age, $X_2 =$ Height, and $X_3 =$ Siblings.
   b. If a student's Age increase by 1 year, what is the predicted change in their income goal?
   c. Predict Income.Goal for a student who is 23 years old, 147.5 cm tall, and has no siblings.