Homework 3 - MATH 141

Due Date: Thursday 09/23/2021, 11:59 PM

Instructions:

- Please provide complete answers/solutions for each question/problem.
- If it involves mathematical computations, please provide your reasoning and/or detailed solutions.
- There are two ways you can write your answers, a: by handwriting (either physically or digitally), or b: by typing on a template document with file type options, Word or RMarkdown, which can be downloaded from the course website.
- If you had handwritten your answers/solutions on a physical paper, make sure to label it properly and please scan your document using a scanner app for convenience. Suggestions: (1) "Tiny Scanner" for Android or (2) "Scanner App" for iOS.
- If you have questions or concerns, please feel free to ask the instructor.
- **Please save your work as pdf file(s), don't put your name in any part of the document, and submit it to the Gradescope page for this course. Your document upload will correspond to your name automatically in Gradescope.**

I. Simple Linear Regression (SLR)

I.A. Consider the scatterplot of x and y variables with a best fit linear model shown in Fig. 1 and the table of sample statistics in Table 1. For the following problems, you can use R as a calculator but write the formula you used to get to your answer.

1. The correlation coefficient of the two numerical variables x and y is $r = 0.925$. Compute the slope estimate of the linear fit using the least squares method.
2. Apply the point-slope equation using the means $(\bar{x}, \bar{y})$ and the slope to write the equation of the line.
3. Suppose that each observation is a book where x is the `complexity` in % and y is `read_time` in hours. Write the linear equation in this context and interpret the slope.
4. The SST of the model is 1.391×10^6 and the SSE of the model is 2.02×10^5. Compute the coefficient of determination (R^2). Interpret this value in the context of the model.
5. Consider the histogram of residuals in Fig. 2 (left) and the scatterplot of residuals in 2 (right). Is the linear regression model appropriate for the data it represents? Explain why.

Table 1: The number of observations, mean, and standard deviation of the x and y variables.

	n	mean	sd
x	500	49.7	29.5
y	500	88.5	52.8

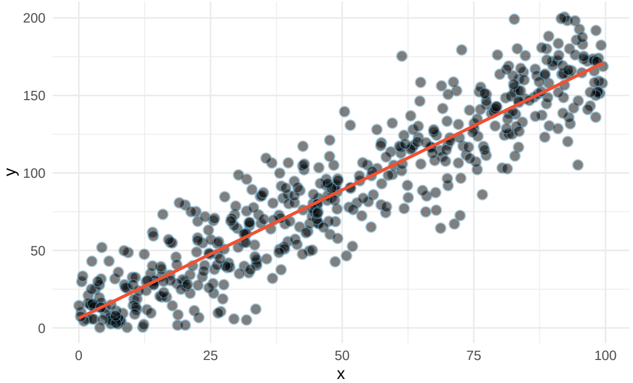

Figure 1: Scatterplot of variables x and y. The red line is the best fit linear model of the data.

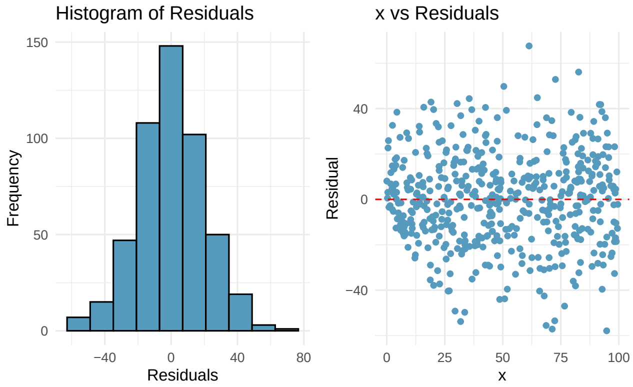

Figure 2: Histogram and scatterplot of the residuals.

I.B. Consider the scatterplot of x and y variables with a best fit linear model shown in Fig. 3 and the table of sample statistics in Table 2. For the following problems, you can use R as a calculator but write the formula you used to get to your answer.

1. State the assumptions of the least squares regression.
2. The correlation coefficient of the two numerical variables x and y is $r =$ -0.956. Compute the slope estimate of the linear fit using the least squares method.
3. Apply the point-slope equation using the means and the slope to write the equation of the line.
4. The SST of the model is 3.725×10^9 and the SSE of the model is 3.227×10^8. Compute the coefficient of determination (R^2). Interpret this value in the context of the model.
5. Consider the histogram of residuals in Fig. 2 (left) and the scatterplot of residuals in 2 (right). Even though the R^2 value is decent, explain why we might want to use other models for this type of data. What assumption is not met? What type of model can you suggest that might best fit this data?

Table 2: The number of observations, mean, and standard deviation of the x and y variables.

	n	mean	sd
x	100	50	29.3
y	100	3331	6134.2

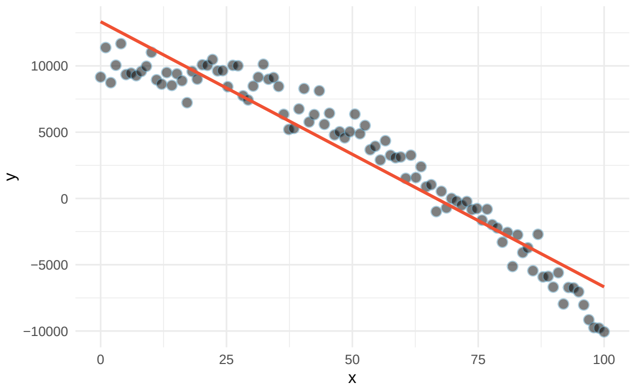

Figure 3: Scatterplot of variables x and y. The red line is the best fit linear model of the data.

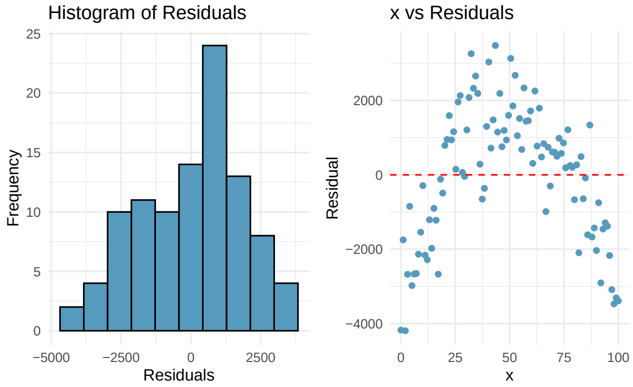

Figure 4: Histogram and scatterplot of the residuals.

II. Categorical Variables and MLR

II.A. Consider a SLR with one categorical predictor with two levels. We want to build a model that predicts a numerical response variable called "complexity" with a given predictor either group "A" or group "B".

1. Since the predictor is a categorical variable with text as levels, we can simplify the model by indicating that it is either group B or not group B (which means group A). Write the linear model equation in this context with the variable x_B with intercept b_0 and slope b_1. Explain the meaning of the variable x_B and its inputs.
2. Suppose that we fit this model into the data - which is shown as boxplots in Fig. 5, the least squares regression method yields the intercept $b_0 = 50.139$ and the slope $b_1 = 20.445$. If the value of $x_B = 0$ - which means that the input is group A, what is the estimated mean complexity? Does it fall close to the actual means?
3. If the value of $x_B = 1$ - which means that the input is group B, what is the estimated mean complexity?
4. Interpret the slope and intercept in this particular context.

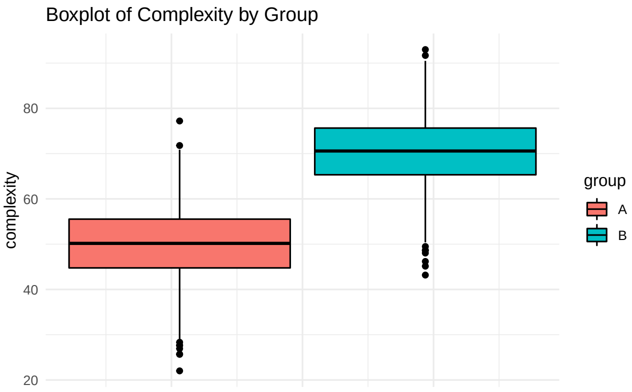

Figure 5: Histogram and scatterplot of the residuals.

II.B. The questions below are concepts regarding the Multiple Linear Regression (MLR).

1. Write the general form of the multiple linear regression model. What are the coefficients and what are the outcome and predictors - or the response and the explanatory variables.
2. What is a common issue of using Multiple Linear Regression (MLR) and how do we prevent such issue?
3. We learned that R^2 is the a measure of how well the linear regression fits the data. Explain the difference between R^2 and the adjusted R^2_{adj}? Include in your explanation in terms of the number of predictors in a linear regression model.

III. Textbook Exercises

Note: To view the selected exercises below, please refer to the textbook, *OpenIntro: Introduction to Modern Statistics (2021) by Mine Çetinkaya-Rundel and Johanna Hardin, First Edition.*

Section 7.5: Choose any 4 exercise problems below to answer.

6. Partners' ages and heights.

12. Crawling babies, correlation.

20. Starbucks, calories and carbs.

22. Body measurements, regression.

24. Cat weights.

30. Cherry trees.

Section 8.6: Choose any 4 exercise problems below to answer.

2. Dealing with categorical predictors.

4. Multiple regression fact checking.

6. Baby weights and mature moms.

8. Movie returns by genre.

10. Palmer penguins, predicting body mass.

12. Palmer penguins, backwards elimination.