STAT448-25S1 Assignment 1

15% of your final grade

Instructions

Please adhere to the following guidelines for submitting your assignment:

1. Your assignment should be submitted on the Learn platform in one of the following formats: PDF or HTML R-markdown. This format should include written answers, R code, and results, all properly commented.
2. You have the option to complete the assignment in pairs. In such cases, please ensure that both names and student IDs are included on the assignment; both students must submit the assignment.
3. Please carefully read and follow the provided instructions for the assignment. Make sure all questions are answered.
4. Reminder: Ensure that your submission includes R code, output, and comments for the assigned tasks. These should be presented in a markdown HTML or PDF file.

Question 1

(15 **marks**) Three observations for a random response variable **Y** are {3, 9, 15}; the corresponding values observed for the explanatory variable **X** are {6, 7, 8}. Assume a linear model: $Y = \beta_0 + \beta_1 X + \epsilon$

a). Compute ordinary least squares estimates of the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ using linear algebra calculations by hand (either write by hand, use R-markdown or a Quarto document, either way, show full working) and with explanatory comments. (5 marks).

b). Calculate by hand the estimates of the residuals $\hat{\epsilon}$. (4 marks).

c). Perform the same matrix algebra calculations of part (a) and (b) using R. (4 marks).

d). Estimate the coefficients using the function **lm** in R. (2 marks).

Question 2

(15 marks) In the context of question 1, consider the case where the values observed for the explanatory variable $\mathbf{X}$ are $\{5, 5, 5\}$.

a). What happens to the coefficient estimates? (5 marks).

b). Using appropriate terminology give a statistical explanation of this situation. (5 marks).

c). Using appropriate terminology give a geometric explanation of this situation. (5 marks).

Question 3

(30 marks) Using the data in the provided CSV file (**Student__Scores__Dataset.csv**), generate a simple linear regression model to describe the relationship between student score (Response) and hours of study (Explanatory variable). Then answer the questions below. Note: Student scores are in the range $0 - 100$ and hours of study are in the range $0 - 10$.

a). Using the model summary, state the regression equation for student score (3 decimal places for coefficients is sufficient). (4 marks).

b). Using the β_1 coefficient from the model equation, provide an interpretation of this coefficient in relation to the response variable. (4 marks).

c). Using the model summary, do hours of study have a significant effect on student scores? Justify your answer. (5 marks).

d). Again using the model summary, does your model provide a good fit for the observed data? Justify your answer. (5 marks).

e). Validate your regression model using appropriate residual plots. What do you observe. Is the fit adequate? Do the residuals suggest a better fit is possible? (5 marks).

f). Provide a well labelled plot of the observations, also plot the regression line and include the regression equation in the plot title or sub-title. (4 marks).

g). Using your regression equation, make student score predictions for the following hours of study, $4.36, 6.86$, and 8.84 (to 2 decimal places). Why might it not be valid to make predictions outside the hours of study range in your data? (3 marks).

Question 4

(40 **marks)** Simple linear regression models serve as crucial tools for exploring relationships between variables. These models offer insights into how the response variable (typically plotted on the y-axis) changes as the explanatory variable (usually on the x-axis) varies. Given that many biological variables are continuous and tend to follow a normal distribution, simple linear regression models play a foundational role in elucidating these associations between variables of interest.

In this context, we will employ a simple linear regression model to investigate the connections between fertility rate and age in female rhesus macaques from Cayo Santiago.

We will use reproductive data from Cayo Santiago rhesus macaque females, as documented in Luevano et al. (2022). Our goal is to determine whether female fertility is influenced by age through the application of simple linear regression analysis. This dataset represents authentic information collected via daily visual censuses conducted by the staff of the Caribbean Primate Research Center (CPRC) at the University of Puerto Rico-Medical Sciences Campus.

The original dataset contains a total of $14,401$ rows, each providing information on the reproductive performance of females at various stages of their lives. We have calculated the mean age-specific fertility rate, which is defined as the number of offspring produced at age 'x' divided by the total number of females of age 'x'. To achieve this, we have grouped the rows by age and computed the mean fertility rate for each age category. You can access the resulting dataset, **'macaque.csv'**, to assist in completing the tasks and addressing the questions posed in this analysis.

a). Plot **mean age specific fertility** versus **age**, making sure to provide a suitable title and axis labels. Describe any pattern in mean fertility rate as females get older. Is the association between mean fertility and age linear or nonlinear? (10 marks).

b). Build a simple linear regression model to describe the association between mean age-specific fertility rate and age. Provide the linear regression equation (3 decimal places for coefficients is sufficient). What is your interpretation of the equation? (10 marks).

c). If you plot the model onto the plot from $Q4(a)$ what do you observe? Is the fit adequate? Do the residuals suggest a better fit is possible? Would adding a quadratic term to the model result in a better fit? (17 marks).

d). Using your regression equation, make mean age-specific fertility predictions for the following age values, $6.50, 14.25$ and 18.75 (to 2 decimal places). (3 marks).

The End

Have you answered all the questions?

References

Luevano, L., C. Sutherland, S. Gonzalez, and R. Hernández Pacheco. 2022. "Rhesus Macaques Compensate for Reproductive Delay Following Ecological Adversity Early in Life." *Ecology and Evolution* 12. https://doi.org/10.1002/ece3.8456.