STAT 300 Written Assignment 2

Note 1: Please show steps with proper justifications in your solutions. Partial credits are given to intermediate steps and reasoning. Also define any notation that you use in your solutions. Provide R codes you used

Note 2: Discussion of ideas learned in class is encouraged (with other students, TAs or the instructor). This helps the leaning process. But individual work turned in by each student should be your own work. Do not copy or paraphrase solutions from other students or from other sources. DO NOT provide your solutions to another student. Failure to comply with these rules will result in an automatic 0 for your work, and additional academic penalties.

Question 1 (*10 marks*)

A researcher is studying the effect of different training methods on exam scores. There are three training methods: *Self-Study*, *Online Course*, and *In-Person Class*. A sample of 600 students was randomly assigned to one of the three methods, and their final exam scores were recorded. In addition to training method, the researcher collected data on each student's *study hours per week* and *previous GPA*.
consider the data file ”students_data.csv”

(a) Create an appropriate visualization to examine the distribution of exam scores across different training methods. Additionally, visualize the relationship between study hours, previous GPA, and exam scores. Comment on any patterns observed. *(2 marks)*

(b) Write down the regression equation that includes all given predictors. Interpret the coefficients. *(2 marks)*

(c) A student completed the *Online Course* by studying 8 hours per week and previouse GPA of 3.8. Predict their expected exam score based on your model. *(1 mark)*

(d) The researcher wants to test whether the training method significantly affects exam scores. Formulate the null and alternative hypotheses for this test and describe how you would test them using an ANOVA framework. *(2 marks)*

(e) Assess whether the assumptions of the linear regression model are met by examining diagnostic plots. Discuss any violations observed and their potential impact. *(2 marks)*

(f) Suppose the researcher suspects that the effect of study hours on exam scores may depend on previous GPA. Modify the regression equation to include an interaction term for these variables and explain how this changes the interpretation of the model. *(1 marks)*

Question 2 (*10 marks*)

An environmental scientist is studying the relationship between the concentration of a pollutant (*Concentration*) in water with temperature (*Temp*) and pH level (*pH*). The data contains measurements for 50 samples.
Consider the data file "data.csv"

(a) Create a scatterplot matrix to visualize the relationships between *Concentration*, *Temp*, and *pH.* Include histograms along the diagonal. Comment on any apparent relationships or patterns in the data, particularly noting any signs of nonlinearity or interaction effects. *(2 marks)*

(b) Fit a multiple linear regression model with *Concentration* as the response variable and *Temp* and *pH* as predictors. Write down the model equation and interpret the coefficients. *(2 marks)*

(c) Based on the visualization in part (a), discuss whether the linear model in part (b) seems adequate. *(1 mark)*

(d) If the linear model in part (b) does not seem adequate, suggest and fit an extended model that could address the observed shortcomings. Justify your choice of the extended model, write down the new model equation, and interpret the updated coefficients. Make sure to report p-value and your conclusion at the 5% significance level. *(2 marks)*

(e) Compare the R^2 and adjusted R^2 values for the models in (b) and (d). What do these values indicate about the added value of the extended model? *(1 mark)*

(f) Create a residual plot for both models. Do any of the plots indicate violations of regression assumptions? *(2 marks)*

Question 3 (*10 marks*)

A research team is investigating the effects of habitat type (Coastal vs. Offshore) and time of day (Morning vs. Evening) on the foraging success of Larus glaucescens (Glaucous-winged Gull). They conduct a study where they record the number of successful foraging attempts per hour for 20 randomly selected gulls in each condition. The data are summarized below (mean number of successful foraging attempts per hour):

Table 1: Mean number of successful foraging attempts per hour for *Larus glaucescens*

Habitat Type	**Morning**	**Evening**
Coastal	5.2	7.1
Offshore	6.5	4.8

(a) Create an interaction plot for the data, with habitat type on the x-axis and separate lines for morning and evening foraging success. Describe the general pattern of the means in the plot.*(2 marks)*

(b) State the null and alternative hypotheses for the main effects and the interaction effect. *(2 marks)*

(c) Compute the interaction term. *(2 marks)*

(d) Suppose the sum of squares for habitat type is 10.3, for time of day is 8.7, for the interaction term is 12.5, and sum of square for total is 46.7. Complete the ANOVA table below by calculating the missing Mean Squares (MS) and F-statistics. *(2 marks)*

Source	SS	df	MS	F
Habitat Type	10.3	1	?	?
Time of Day	8.7	1	?	?
Interaction	12.5	1	?	?
Error	?	36	?	
Total	46.7	39		

(e) Based on the calculated F-values, determine which effects are significant at the significance level of 0.05. What ecological conclusions can be drawn from the results regarding glaucous-winged gulls foraging behaviour? *(2 marks)*

Question 4 (*10 marks*)

A researcher is investigating the variability of caffeine content in specialty coffee. The caffeine content (in mg) of six different 16 oz. coffees from a specific coffee shop is recorded. The values are:

$$564.4, 498.2, 259.2, 303.3, 299.5, 307.2$$

(a) Generate 1000 bootstrap samples from the given data. Compute the sample mean for each bootstrap sample and visualize the empirical bootstrap distribution. *(2 marks)*

(b) Compute the bootstrap estimate of the standard deviation of the sample mean and interpret its significance in the context of variability. *(1 mark)*

(c) Construct a 95% bootstrap confidence interval for the mean using the percentile method. *(1 marks)*

(d) Suppose the coffee shop claims that the mean caffeine content is 300 mg. Use a bootstrap hypothesis test to assess this claim at a 5% significance level. Formulate the null and alternative hypotheses and determine the bootstrap p-value. *(3 marks)*

(e) Compare the bootstrap hypothesis test results to the classical one-sample t-test. Discuss any differences observed and their implications. *(2 marks)*

(f) Suppose the sample size were increased to $n = 50$. How would you expect the bootstrap and classical t-test results to change? Justify your answer. *(1 mark)*