King Fahd University of Petroleum and Minerals
Mathematics Department

Dhahran, Saudi Arabia

STAT 310: Regression Analysis

T231

Exam 2

November 4th , 2023

Name:

ID:

Instructions:

- Save this file as “STAT 310 - Exam 2 - your name”
- Any questions that specify the use of R or Rstudio, you have to upload your codes that you used to solve these questions on the Blackboard.
- Duration of the test is 100 minutes.

Score:

Q1	/10	Q3	/30	Total	/100
Q2	/60				

Question 1

Answer with True or False: If you answer False then justify your answer:

1.1) All high leverage points and outliers must be removed from the model.

1.2) The coefficient of determination is sufficient to assess the goodness of the model

1.3) Assume we have the following regression model: $\log(\hat{y}) = 2.043 + 1.012\sqrt{x}$. Then, the correct interpretation of the estimate 1.012 is: The average increase in the log(y) is 1.012 for a given one unit increase in x.

1.4) The independence assumption is related to the independence of x from y.

1.5) Assume that when performing the Box-Cox transformation for $y = \beta_o + \beta_1 x + \epsilon$, the 95% confidence interval of λ was found to be [-0.1,0.8]. Then one proper transformation is $\sqrt{y} = \beta_o + \beta_1 x + \epsilon$.

Question 2

The data "Adrev " is uploaded on the blackboard, and it shows two variables:

- AdRevenue: The gross advertising revenue in $.
- AdPages: The number of pages used to advertise the product.

The data are for the top 70 US magazines ranked in terms of total gross advertising revenue in 2006. In particular we will develop regression models to predict gross advertising revenue per advertising page.
NOTE: It is an excel file.

2.1) Develop a simple linear regression model (i.e. $y = \beta_o + \beta_1 x + \epsilon$). And interpret these estimates

2.2) Construct a 95% confidence interval for the population slope and interpret this interval

2.3) Find a 95% prediction interval for the gross revenue when the number of pages is 300 and 150. And interpret these intervals

2.4) Plot a scatter plot between the number of pages and the revenue. Add the regression line to the plot. Comment on the plot

2.5) Check the normality assumption using the Q-Q plot and the histogram of the standardized errors . Comment on the plot.

2.6) Use the appropriate test to check for normality. What is your conclusion about the normality assumption?

2.7) Check the equality of variance assumption using the residuals plot. Comment on the plot.

2.8) Use the appropriate test to check for the equality of variance. What is your conclusion about the equality of variance assumption?

2.9) What is the percentage of variation in the Ad revenue that is explained by the variation in Ad pages?

2.10) What is the standard error of the estimate for this model?

2.11) Is a transformation needed? Justify your answer

2.12) Use the Box-Cox method to find the appropriate transformation. What value of λ is appropriate in this case? And write the transformation of the regression equation.

2.13) After the transformation. What is the new regression equation? And interpret the estimates

2.14) Check the following assumptions using the appropriate plots and test. Comment on these plots and tests?

(2.14.a) The normality of the errors

(2.14.b) The equality of the variance

2.15) Find a 95% prediction interval for the gross revenue when the number of pages is 300 and 150 using the new model. And interpret these intervals

2.16) Find the coefficient of determination and the standard error for the new model and compare it with the previous linear model.

2.17) Find the leverage points and outliers using the scatter plot

2.18) Remove these points and fit a new model (using the same transformation). Has the coefficient of determination improved significantly?

2.19) Remove these points and fit a new model (using the same transformation). Has the standard error of the estimate improved significantly?

Question 3

The data “startup” on the blackboard shows 5 variables:

- **RDSpend**: The total amount spent on the research and development projects in $.
- **Administration**: The total amount spent on administration in $.
- **MarketingSpend**: The total amount spent on marketing in $.
- **State**: The location of the startup “New York” or “California”.

Target Variable:

- **profit**: The total profit of each startup in $

The objective is to examine the factors influencing the profit of the startup. The dataset consists of 50 startups, with each record containing information about various predictors and the profit.

NOTE: it is a CSV file, import it as text.

Using this data answer the following:

Part (A): fit a linear regression model

$$y_{profit} = \beta_o + \beta_1 X_{RD\ spend} + \beta_2 X_{Admin} + \beta_3 X_{Marketing} + \beta_4 X_{State}$$

3.1) What is the estimated regression equation and interpret the estimated coefficients

3.2) Find the estimated correlation between all the numerical variables and the profit. Comment on these values

3.3) Plot a scatter for the following and comment on the relationship:

3.3.a) Profit Vs R&D spent
3.3.b) Profit Vs Administration
3.3.c) Profit Vs Marketing spent

3.4) Estimate the standard error for this model and find the coefficient of determination:

3.5) Plot a boxplot for the profit with respect to the location of the startup (the state). Comment on the plot

Part (B): fit a the following regression model

$$y_{profit} = \beta_o + \beta_1 X_{RD\ spend} + \beta_2 X_{Admin} + \beta_3 \sqrt{X_{Marketing}} + \beta_4 X_{State}$$

3.6) What variables are insignificant in this new model justify your answer. Use $\alpha = 0.1$

3.7) Remove all insignificant independent variable and fit a third model. What are the estimated values of the parameters

3.8) Estimate the standard error for this model and find the coefficient of determination:

3.9) Check the normality assumption using the Q-Q plot and the histogram of the standardized errors . Comment on the plot.

3.10) Use the appropriate test to check for normality. What is your conclusion about the normality assumption?

3.11) Check the equality of variance assumption using the residuals plot. Comment on the plot.

3.12) Use the appropriate test to check for the equality of variance. What is your conclusion about the equality of variance assumption?

3.13) Construct a 95% confidence interval for each slope in this model and interpret these slopes