KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

STAT 510: Regression Analysis

Term 231, Major Exam I

Saturday October 14, 2023, 07:30 PM

Name: ________________________________ ID #: ____________________

Question No	Full Marks	Marks Obtained
1	**15**	
2	**12**	
3	**08**	
4	**10**	
5	**05**	
Total	**50**	

Instructions:

1. Mobiles are not allowed in the exam. If you have your **mobile** with you, **turn it off** and put it **on the table/floor** so that it is visible to the proctor.

2. Show all the calculation steps. There are points for the steps so if you miss them, you lose points. For multiple choice type questions, showing calculation steps is not required.

3. Report **at least 4 decimal points** of your numerical answers.

[Blank page]

Q1: (2+2+3+4+4 = 15 pts.) Data on the thrust of a jet turbine engine and four predictors are available with $n = 32$. Fit a multiple linear regression model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$.

Download the dataset from Blackboard and write Code # ___________

(i) Predict the thrust of a jet turbine engine with $x_1 = 2080$, $x_2 = 30200$, $x_3 = 1710$ and $x_4 = 105$.

The predicted value is equal to _____________________.

(ii) A 99% interval estimate for the thrust of a jet turbine engine with $x_1 = 2080$, $x_2 = 30200$, $x_3 = 1710$ and $x_4 = 105$ is given as:

[___________________ , ___________________]

(iii) Is the prediction done in part (i) interpolation or extrapolation? Provide all the details of your solution before writing the final answer.

(iv) Estimate the correlation between $\hat{\beta}_1$ and $\hat{\beta}_4$. Provide all the details of your solution before writing the final answer.

(v) Test the hypothesis that X_2 and X_3 are contributing significantly to the model in presence of other predictors.

H_0: ____

H_1: ____

P-value = ______________________________

Decision and conclusion:

Q2: (1x12 = 12 pts.) Multiple choice or fill in the blank questions. Any MCQ with more than one option circled will be considered wrong.

(i) For a simple linear regression model, how many parameters are to be estimated?

(A) 0
(B) 1
(C) 2
(D) 3
(E) 4

(ii) A simple linear regression model is used to predict the monthly electricity bill based on mean daily temperature of that month. The data are given as:

y	125	110	95	90	110	130
X	30	40	50	60	70	80

The fitted model is given as $\hat{y}_i = 105 + 0.05X_i$. A customers decided to use this model to predict his monthly electricity bill for four different months with mean daily temperatures 36, 42, 54 and 61 using the formula $\hat{y}_0 \pm 2\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$. Which of the computed prediction intervals will be narrowest?

(A) The month with mean daily temperature of 36.
(B) The month with mean daily temperature of 42.
(C) The month with mean daily temperature of 54.
(D) The month with mean daily temperature of 61.
(E) All four of them will be of the same width.

(iii) For a simple linear regression model, if SSR is found to be zero then which one of the following is true?

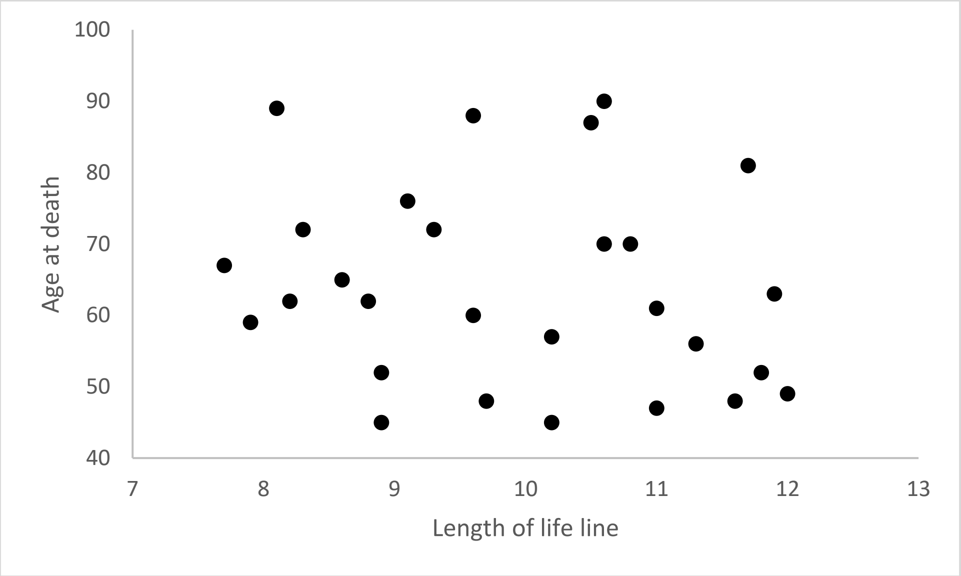
(A) The correlation between two variables is zero.
(B) p-value for the significance of full model will be zero.
(C) Intercept of the model will be zero.
(D) MSE of the model will be zero.
(E) Mean of the predictor variable is equal to zero.

(iv) In regression analysis, if the response variable is measured in kilograms, all the predictor variable

(A) must also be in kilograms.

(B) must be in some unit of weight.
(C) must be in the same units of weight.
(D) cannot be in kilograms.
(E) can be any unit.

(v) Palm readers claim to be able to tell how long your life will be by looking at a specific line on your hand. The following is a scatter plot of age of person at death (in years) vs length of lifeline on the right hand (in cm) for a sample of 28 (dead) people.



A simple linear regression model is fitted to these data and F test is conducted to test the significance of model. The p-value for the F test will be around

(A) 1
(B) 0
(C) 0.05
(D) 0.01
(E) 0.1

(vi) Which one of the following is true about the prediction interval computed from a multiple linear regression model?

(A) The width of prediction interval increases with an increase in significance level α.
(B) It is wider than the confidence interval for mean response.
(C) The predicted value can fall outside the prediction interval.
(D) A prediction interval cannot be computed when there is only one predictor in the model.
(E) The actual value of response always falls inside the prediction interval.

(vii) Which one of the following is **not** true for regression analysis?

(A) SSR ≥ 0

(B) SSR ≤ SST
(C) SSE ≤ SST
(D) SST ≥ 0
(E) SSE ≤ 0

(viii) Fit a multiple linear regression model $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i$, we write down the T matrix and c vector as follows:

$$T = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad , \quad c = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

What hypotheses are we testing?

(A) $H_0: \beta_1 = \beta_3 = 0$ against H_1: At least one $\beta_j \neq 0$ for $j = 1,3$
(B) $H_0: \beta_1 = \beta_3$ against $H_1: \beta_1 \neq \beta_3$
(C) $H_0: \beta_1 = \beta_3$ and $\beta_4 = 0$ against H_1: At least one constraint in $H_0$is not true
(D) $H_0: \beta_1 - \beta_3 = -1$ against $H_1: \beta_1 - \beta_3 \neq -1$
(E) None of the above

(ix) If the correlation coefficient between the two variables X and y is close to +1, what does that mean?

(A) X is causing the change in y.
(B) When X increases y decreases, and vice versa.
(C) X and y both are causing the change in each other.
(D) The mean of y is +1 when X=0.
(E) None of the above.

(x) In simple linear regression, least square method calculates the best-fitting line for the observed data by minimizing the sum of the

(A) squares of the observed response
(B) squares of the fitted values
(C) difference between observed and predicted response
(D) absolute difference between observed and predicted response
(E) None of the others

(xi) For testing the significance of a predictor X_1 in simple linear regression, we can define Z-test based on $Z = \frac{\widehat{\beta}_1}{\sqrt{\sigma^2 / S_{XX}}}$. Which one of the following is true for this test?

(A) S_{XX} is never known so the test cannot be conducted.

(B) The test cannot be used in real life.
(C) $\hat{\beta}_1$ is always positive.
(D) There are no critical values because normal distribution PDF cannot be integrated.
(E) The test is only applicable for samples of size 30 or more.

(xii) For a multiple linear regression model $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$, which one of the following is not true about the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$?

(A) It is symmetric i.e. $\boldsymbol{H}^{\boldsymbol{T}} = \boldsymbol{H}$.
(B) It is idempotent i.e. $\boldsymbol{HH} = \boldsymbol{H}$.
(C) $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\beta}$
(D) Its trace is equal to $k + 1$.
(E) $(\boldsymbol{I} - \boldsymbol{H})$ is also idempotent.

Q3: (8 pts.) For a multiple linear regression model $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$ or equivalently in matrix notation $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$, mathematically derive the maximum likelihood estimates for $\beta_0, \beta_1, \dots, \beta_k$ and σ^2 assuming $\epsilon_i \sim N(0, \sigma^2)$.

Q3 continues…

Q4: (10 points) Consider a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where β_0, β_1 are unknown parameters and x_i's are fixed. The OLS estimates of β_0 and β_1 are given as $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ where $S_{XY} = \sum xy - \frac{(\sum y)(\sum x)}{n}$ and $S_{XX} = \sum x^2 - \frac{(\sum x)^2}{n}$.
Mathematically derive the expression for $E(\hat{\beta}_0 \hat{\beta}_1)$.

Hint: The covariance between 2 variables X and Y is $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$
Note: You can use the unbiasedness of $\hat{\beta}_0$ and $\hat{\beta}_1$ without deriving it (if needed anywhere.)

Q4 continues…

Q5: (5 pts.) For a multiple linear regression model $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$, we have proved in the class that $MSE = \frac{SSE}{n-k-1}$ is unbiased for error variance σ^2 i.e. $E(MSE) = \sigma^2$. Derive an expression for $E(MSE^2)$.