KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

STAT 510: Regression Analysis

Term 231, Major Exam II

Saturday December 02, 2023, 07:00 PM

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**10**	
2	**13**	
3	**06**	
4	**08**	
5	**07**	
6	**06**	
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.

Q1: (1 x 10 = 10 pts.) All multiple choice questions.

i. SSE of a regression model can never be

A. larger than SST
B. larger than 1
C. smaller than 1
D. equal to SSR
E. smaller than SSR

ii. Weighted least squares is a method of model estimation when

A. only the assumption of independence has failed.
B. only the assumption of linearity has failed.
C. the assumptions of independence and equal variance have failed.
D. only the assumption of equal variance has failed.
E. the assumptions of linearity and normality have failed.

iii. Which one of the following is true for the estimated regression equation $\hat{y} = 2.3 - 1.67X_1 + 0.33X_2 + 1.92X_3$?

A. A unit increase in X_1 causes y to increase by 1.67 units.
B. A unit increase in X_1 causes y to decrease by 1.67 units.
C. A unit increase in X_2 causes y to decrease by 0.33 units.
D. A unit increase in X_3 causes y to decrease by 1.92 units.
E. None of the others.

iv. Which one of the following is not a required assumption for a regression model?

A. the variance of error term is same for all levels of X.
B. the values of error term are independent.
C. the predictors are not strongly correlated.
D. the error terms are normally distributed.
E. the response is linearly related to the predictor(s).

v. In a linear regression model, a pure leverage point

A. produces a large difference between e_i and $e_{(i)}$.
B. produces a large, scaled residual r_i.
C. reduces the practicality of the model.
D. significantly affects the estimated coefficients.
E. does not affect the regression equation significantly.

vi. A regression analysis is inappropriate when

A. you want to make prediction for one variable.
B. variance of the error term is different for different levels of predictor(s).
C. the pattern of data forms a reasonably straight line.
D. response variable is strongly correlated with the predictor(s).
E. the sample size is less than 30.

vii. The Box-Cox method is applied when

A. the error term needs a power transformation.
B. the predictor needs any transformation.
C. the response variable needs any transformation.
D. the response variable needs a power transformation.
E. no transformation is needed.

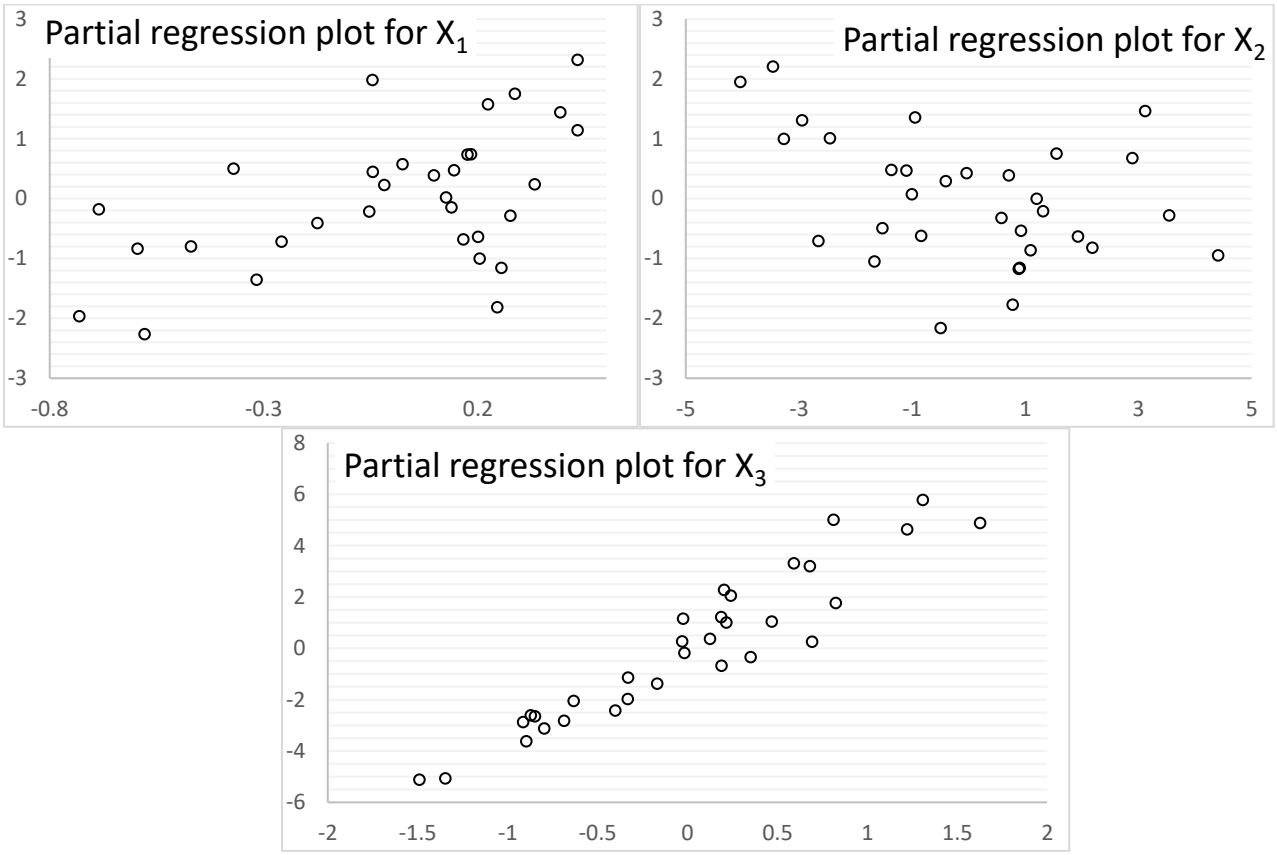
viii. If the influential observation is not an error value and a valid observation from the intended population, then which of the following is the best treatment of the influential observation?

A. deleting the influential observation
B. taking another sample from the intended population
C. downweighing the influential observation
D. using the generalized least squares method
E. using the weighted least squares method

ix. A dataset is considered with response variable y and one predictor X. Suppose that the relationship between y and X is intrinsically linear and is given as $y = 2^{\left[\frac{1/\sqrt{X}}{\beta_1 + \beta_0/\sqrt{X}}\right]}$. Transform the variables such that the relationship becomes linear. The transformed variables are

A. $y_1 = 2^y$ and $X_1 = \frac{1}{\sqrt{X}}$
B. $y_1 = \ln y$ and $X_1 = \ln X$
C. $y_1 = \sqrt{y}$ and $X_1 = e^X$
D. $y_1 = \frac{1}{\ln y}$ and $X_1 = \sqrt{X}$
E. $y_1 = \frac{1}{\sqrt{y}}$ and $X_1 = e^{\frac{1}{X}}$

x. Suppose data are available on the response variable y and three predictors X_1, X_2, X_3 and we fitted the multiple regression model. After fitting the model, the partial regression plots are created for all predictors given as follows:



The fitted model is given as $\hat{y}_i = 5.3 + 1.8X_{1i} - 0.19X_{2i} + \hat{\beta}_3 X_{3i}$. Which one of the following is closest to the correct value of $\hat{\beta}_3$?

A. 3.6
B. 0.1
C. -2.4
D. 16.9
E. 0

Q2: (2+4+4+3 = 13 pts.) The director of admissions of a small college administered a newly designed entrance test to 20 students selected at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year can be predicted from the entrance test score. The results of the study can be found in the attached Excel file named "data1_code____".

(a) Fit a simple linear regression model to the data and plot the estimated regression function along with the data in R. Does the line appear to fit well? Paste your R program on Blackboard and elaborate the results below:

(b) Fit the robust regression models to the data using Huber's and Bisquare functions. Plot all the 3 fitted models on the same scatter diagram. Does this improve the fit? Which model do you think best represents the data? Paste your R program on Blackboard and elaborate the results below:

(c) Obtain an approximate 99% interval estimate for the mean freshman GPA of students with entrance test score is 4.0. Which model do you prefer to use? Why? Interpret the recommended interval estimate. Paste your R program on Blackboard and elaborate the results below:

d) What is the weight assigned to 20^{th} observation (student with GPA of 3.7 and entrance test score of 3.2) in the dataset by the three models? Paste your RStudio code on Blackboard.

Simple Linear Regression Model _____

Robust Regression Model by Huber's Function _____

Robust Regression Model by Bi-square Function _____

Q3: (6 pts.) For a multiple linear regression model $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$, mathematically derive and simplify an expression covariance between two residuals e_i and e_j i.e. $\mathrm{Cov}(e_i, e_j)$ for $i \neq j$.

Q3 continues…

Q4: (8 pts.) For a multiple linear regression model $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$, the Cook's distance is defined as $D_i = \frac{[\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)}]^T(\boldsymbol{X}^T\boldsymbol{X})[\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)}]}{(k+1)MSE}$ where $\widehat{\boldsymbol{\beta}}$ is the vector of estimated coefficients using all n observations and $\widehat{\boldsymbol{\beta}}_{(i)}$ is the vector of estimated coefficients using all observations except i^{th}. Mathematically express D_i as a function of fitted values $\hat{y}_i$.

Note: Derive all the results that you use in your solution.

Q4 continues…

Q5: (7 pts.) For a multiple linear regression model $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$, suppose that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $V(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{P}$, where $\boldsymbol{P}$ (not equal to $\boldsymbol{I}_{n\times n}$) is a known $n \times n$ matrix. Under the said situation, the ordinary least square estimates $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is not appropriate. Derive an appropriate vector of estimates for $\boldsymbol{\beta}$ under this situation, say $\widetilde{\boldsymbol{\beta}}$. Also, derive the expected value of your estimator vector i.e. $E(\widetilde{\boldsymbol{\beta}})$.

Q5 continues…

Q6: (6 pts.) For a multiple linear regression model $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$, prove that the residuals $\boldsymbol{e}$ are independent of fitted values $\widehat{\boldsymbol{y}}$.

Q6 continues…