KING FAHD UNIVERSITY OF PETROLEUM & MINERALS, DHAHRAN, SAUDI ARABIA
DEPARTMENT OF MATHEMATICS

STAT 310: Regression Analysis

Term 221, Second Major Exam
Saturday November 19, 2022, 06:00 PM

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**07**	
2	**08**	
3	**05**	
4	**05**	
Total	**25**	

Instructions:

1. Formula sheet will be provided to you in exam. You are not allowed to bring, with you, formula sheet or any other printed/written paper.
2. Mobiles are not allowed in exam. If you have your **mobile** with you, **turn it off** and put it **on the table** so that it is visible to proctor.
3. Show all the calculation steps. There are points for the steps so if your miss them, you lose points. For multiple choice type questions, showing calculation steps is not required.
4. Derive every result that you use in your solution, unless mentioned otherwise.
5. Anything bold in a question indicates that it is a vector or matrix.

Q1: (3+2+2 = 7 pts.) For a multiple linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, the OLS estimates of $\boldsymbol{\beta}$ vector are given as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ with the variance covariance matrix given as $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, the fitted values as $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, the residuals as $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ and the hat matrix as $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

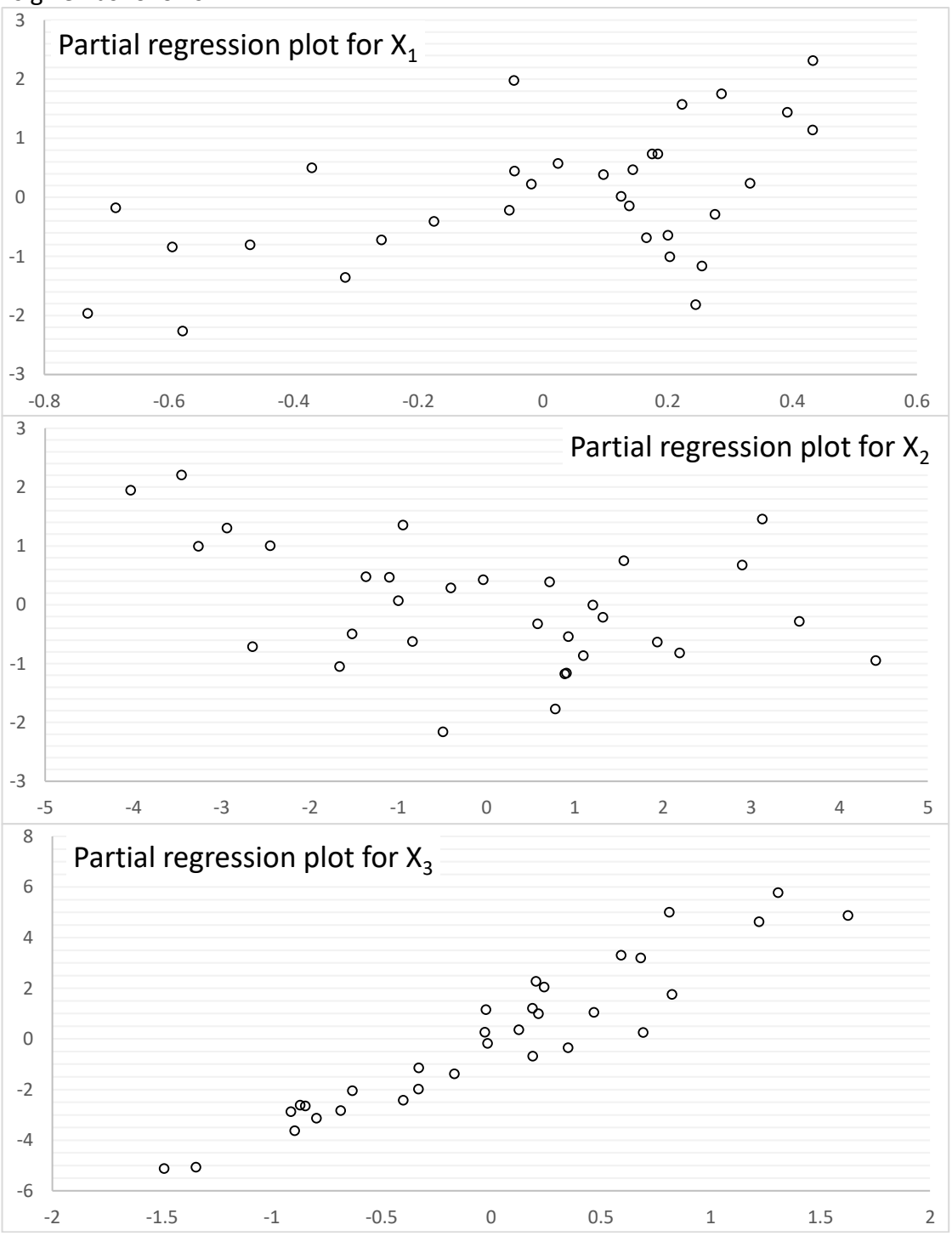
(a) Mathematically prove that $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$.

(b) Mathematically prove that $\mathrm{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$.

(c) Derive an expression for the correlation between e_i and e_j, in terms of elements of $\boldsymbol{H}$ matrix.

Q2: (1 x 8 = 8 pts.)

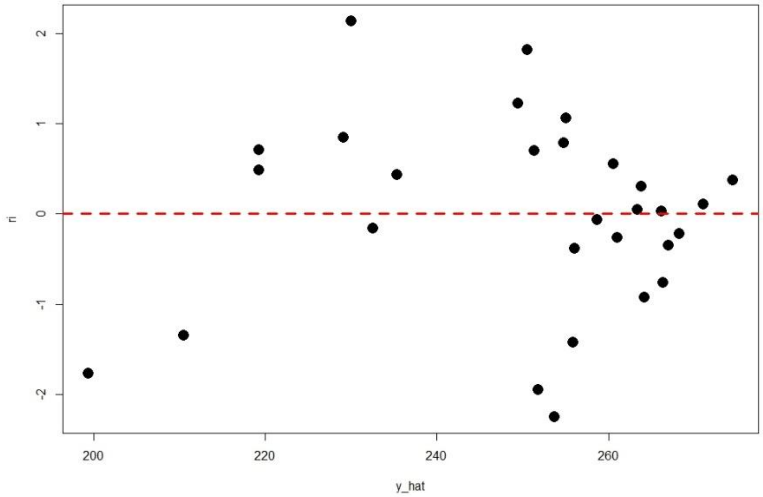
(2.1) Suppose data are available on the response variable y and three predictors X_1, X_2, X_3 and we fitted the multiple regression model. The fitted model is given as $\hat{y}_i = 5.3 + 1.8X_{1i} - 0.19X_{2i} + \hat{\beta}_3 X_{3i}$. After fitting the model, the partial regression plots are created for all predictors given as follows:



Which one of the following is the correct value of $\hat{\beta}_3$ in fitted model with three predictors?

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

(2.2) Consider the simple linear regression model fit to the solar energy data. The plot of residuals vs fitted response is given as follows:



In light of the above plot, which of the following statements is true?

A. The assumption of normality is violated.
B. The assumption of independence is violated.
C. The assumption of constant variance is violated.
D. The assumptions of independence and normality are violated.
E. The assumptions of normality and linearity are violated.

(2.3) A multiple linear regression model is fitted with 3 predictors i.e.

$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \qquad \forall\, i = 1,2,\dots,25$

A thorough influential analysis is performed to the model and it is found that $COVRATIO_{22} = \frac{\left|\left(X'_{(22)}X_{(22)}\right)^{-1} s^2_{(22)}\right|}{|(X'X)^{-1}MSE|} = 2.81$. Which of the following statements is the correct interpretation of given information?

A. 22nd observation is significantly degrading the precision of model.
B. 22nd observation is significantly improving the precision of model.
C. The variance of 22nd observation is 2.81.
D. The covariance of 22nd observation with all the other observations is 2.81.
E. 22nd observation has no effect on $Var - Cov(\widehat{\boldsymbol{\beta}})$.

(2.4) Box-Cox transformation is used for finding the optimal power transformation on

A. response
B. predictor
C. error
D. MSE
E. R^2

(2.5) A multiple linear regression model is fitted with 3 predictors i.e.

$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \qquad \forall\, i = 1,2,\ldots,25$

A thorough influential analysis is performed to the model and it is found that $DFFIT_9 = \frac{\hat{y}_9 - \hat{y}_{(9)}}{\sqrt{S^2_{(9)} \times h_{99}}} = -0.00219$. Which of the following statements is the correct interpretation of given information?

A. 9th observation is not significantly influencing the prediction.
B. 9th observation is not significantly influencing the coefficients in $\hat{\beta}$.
C. 9th observation is not significantly influencing $Var - Cov(\hat{\beta})$.
D. 9th observation is significantly influencing the prediction.
E. 9th observation is significantly influencing the intercept $\hat{\beta}_0$.

(2.6) Failure of constant variance assumption can be corrected by

A. transforming any one of the predictors
B. transforming all predictors
C. transforming the error terms
D. transforming the response
E. transforming the most significant predictors

(2.7) 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 $X_1 = e^{\frac{1}{X}}$

(2.8) A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 50 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. Shown in the data file are the age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y).

Fit a multiple linear regression model $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i$ and test the following constraints: H_0: $2\beta_1 + \beta_2 = 0$ and $\frac{\beta_3}{10} - 2\beta_2 = -0.5$ against H_1: At least one of the constraints in H_0 is not true.

Write down the $\boldsymbol{T}$ matrix and $\boldsymbol{c}$ vector for testing the above hypotheses.

$$\boldsymbol{T} = \begin{bmatrix} & \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} \end{bmatrix}$$

Q3: Code _____
Download the dataset from Blackboard and write down the code number in above blank.

(5 pts.) Data are available on the response variable y and a predictor X. The scatter and the residual plots are showing that the relationship between y and X is not linear. Suppose, it is known that y is linearly related to X^{α}.

(a) Use Box-Tidwell method to obtain an optimal value of α for the transformation. Start with $\alpha_0 = 1$ and report the following results after performing 2 iterations.

$\hat{\alpha}_1 =$ _____ , $\hat{\alpha}_2 =$ _____

(b) Note: Do not use the answer of part(a) in this question.
Suppose that the final answer of part (a) is $\hat{\alpha}_2 = -2$. Apply the required transformation on X and fit the model. Using the fitted model, predict the response when $X = 0.48$. Also construct a 99% prediction interval.

$\hat{y}_{X=0.48} =$ _____ , Critical value = _____

Lower Prediction Limit = _____ , Upper Prediction Limit = _____

Q4: Code _____
Download the dataset from Blackboard and write down the code number in above blank.

(5 pts.) A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 50 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. Shown in the data file are the age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y).
Fit a multiple linear regression model $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i$ and test the following constraints: H_0: $2\beta_1 + \beta_2 = 0$ and $\frac{\beta_3}{10} - 2\beta_2 = -0.5$ against H_1: At least one of the constraints in H_0 is not true.

The $\boldsymbol{T}$ matrix and $\boldsymbol{c}$ vector for testing the above hypotheses are $T = \begin{bmatrix} 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 0.1 & 0 \end{bmatrix}$ and $c = \begin{bmatrix} 0 \\ -0.5 \end{bmatrix}$. Perform the F test for testing the given constraints and fill in the following blanks:

Calculated value of F = _____

p-value = _____

Decision: (a) Reject H_0
(b) Fail to reject H_0

Conclusion: