KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

STAT 513: Statistical Modeling

Term 231, Major Exam II

Monday December 04, 2023, 07:00 PM

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**24**	
2	**09**	
3	**17**	
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.

Q1: (2 x 12 = 24 pts.) Multiple choice or fill in the blank questions. Any MCQ with more than one option circled will be considered wrong.

i. In multiple linear regression, which one of the following in an indication of the presence of multicollinearity?

A. large differences between the characteristic roots of $X'X$
B. high coefficient of determination of the model
C. large differences between the diagonal elements of hat matrix
D. strong correlation between the response and predictor variables
E. a value of Cook's distance more than one

ii. Why is the number of indicator variables to be entered into the regression model always equal to the number of categories (c) minus 1?

A. To control for other variables in the model
B. To increase the R-squared value
C. To fix the violation of constant variance assumption
D. To avoid the situation of perfect multicollinearity
E. To reduce the effect of influential observation

iii. What is/are the problem(s) associated with cubic spline model: $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \gamma_2 K^2 + \gamma_3 K^3 + \epsilon$ where $K = \begin{cases} 0 & \text{if } x \le k \\ x - k & \text{if } x > k \end{cases}$ and the knot is at point k.

A. The line is not continuous.
B. The first derivative of the line is not continuous.
C. The second derivative of the line is not continuous.
D. All of above.
E. None of above

iv. In a linear regression model $y = \beta_0 + \beta_1 X + \gamma_1 I + \epsilon$ with X as a continuous predictor and I as an indicator variable, how do we interpret the value of γ_1?

A. the difference between the two R-square values
B. the difference between the slopes
C. average change in y due to a unit change in X
D. the difference between the means of response variable y
E. difference between the variances of response variable y

v. 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. the assumptions of linearity and normality have failed.
E. only the assumption of equal variance has failed.

vi. 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. does not affect the regression equation significantly.
E. significantly affects the estimated coefficients.

vii. The Box-Cox method is applied when

A. the error term needs a power transformation.
B. the response variable needs a power transformation.
C. the predictor needs any transformation.
D. the response variable needs any 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. using the generalized least squares method
D. using the weighted least squares method
E. downweighing the influential observation

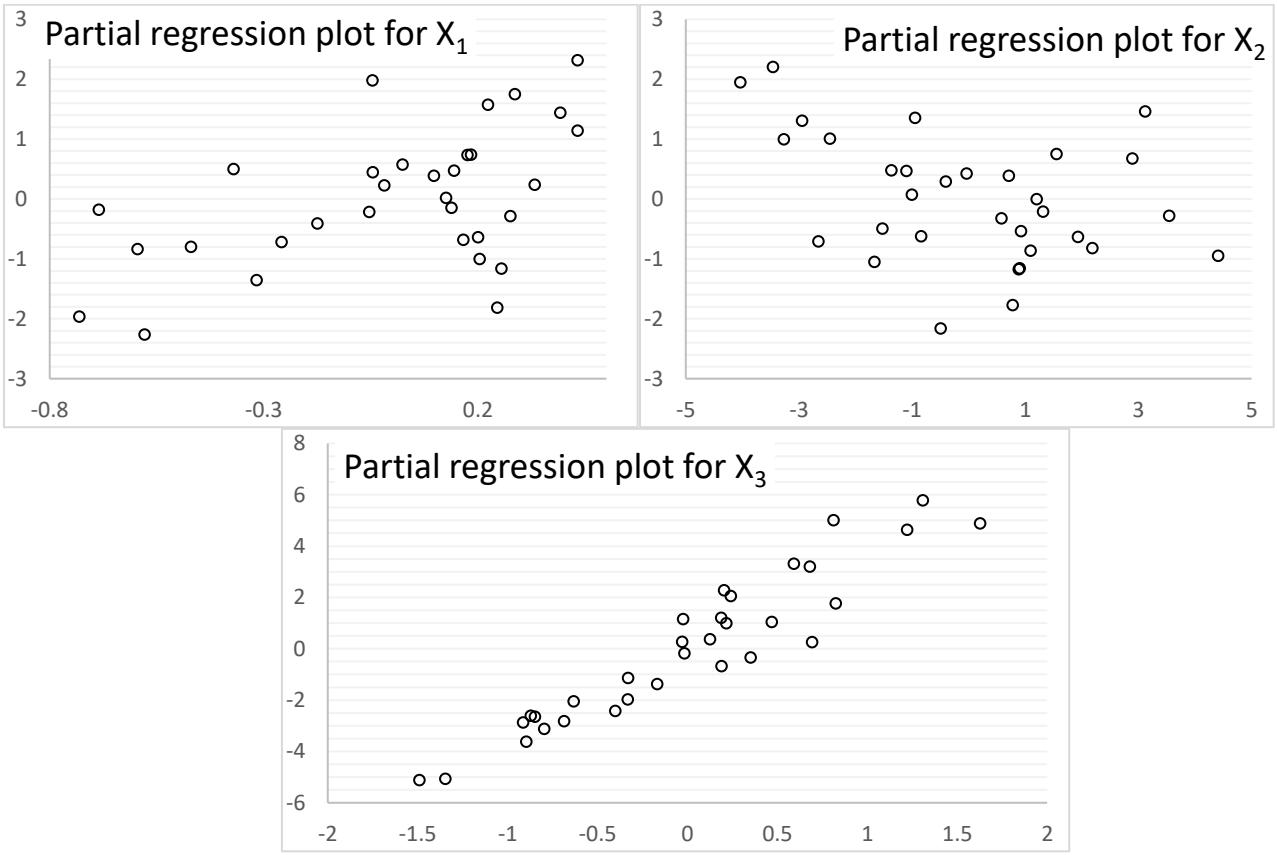
ix. A dataset is considered with response variable y and one predictor X. Suppose that the relationship between y and X is intrinsically linear and is given as $y = 2^{\left[\frac{\frac{1}{\sqrt{X}}}{\beta_1 + \frac{\beta_0}{\sqrt{X}}}\right]}$. Transform the variables such that the relationship becomes 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. $y_1 = \frac{1}{\ln y}$ and $X_1 = \sqrt{X}$
D. $y_1 = \sqrt{y}$ and $X_1 = e^X$
E. $y_1 = \frac{1}{\sqrt{y}}$ and $X_1 = e^{\frac{1}{X}}$

x. In a simple linear regression model, if the coefficient of determination is positive, then

A. the regression equation must have a positive slope
B. the regression equation must have a negative slope
C. the regression equation must have a positive intercept
D. the regression equation must fit all observations without any error
E. none of the others

xi. 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

xii. Which one of the following tests can be used to test the assumption of correct model specification?

A. Breusch-Pagan test
B. Shapiro-Wilk test
C. Lack-of-Fit test
D. Durbin-Watson test
E. Breusch–Godfrey test

Name: _____ ID #: _____

Q2: (9 pts.) Attached file contains the data on response variable y and predictor X. Suppose that the relationship between y and X is intrinsically linear and is given as $y = 2^{\left[\frac{\frac{1}{\sqrt{X}}}{\beta_1+\frac{\beta_0}{\sqrt{X}}}\right]}$.
Draw a scatter plot of y against X and you will notice that the relationship is not linear.
Now, transform the variables such that the relationship p becomes linear i.e. the new variables are $y_1 = \frac{1}{\ln y}$ and $X_1 = \sqrt{X}$. Draw the scatter plot of y_1 against X_1 and you will notice that the relationship is linear.

Report your answer correct up to 6 decimal points. Dataset Q2-code_____

Fit a linear regression model on the transformed variables y_1 and X_1. The fitted model is given as:

$$\hat{y}_1 = \_\_\_\_\_ + \_\_\_\_\_X_1$$

Predict the original response y when $X = 0.0081$. Also construct a 99% prediction interval for original y when $X = 0.0081$.

$\hat{y}_{X=0.0081}$ =_____

Lower Prediction Limit = _____ Upper Prediction Limit = _____

Paste the RStudio code of your solution on Blackboard.

Q3: (5+3+4+2+3 = 17 pts.) A Student performance survey is designed to examine the factors influencing academic student performance. The dataset consists of 3748 student records, with each record containing information about various features and a performance index.

Features:
Hours Studied: The total number of hours spent studying by each student.
Previous Scores: The scores obtained by students in previous tests.
Extracurricular Activities: Whether the student participates in extracurricular activities (Yes or No).
Sleep Hours: The average number of hours of sleep the student had per day.
Sample Question Papers Practiced: The number of sample question papers the student practiced.

Target Variable:
Performance Index: A measure of the overall performance of each student. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

Report your answer correct up to 6 decimal points. Dataset Q3-code_____

a) Construct an indicator variable against Extracurricular Activities for “Yes” category, keeping “No” as default. Fit a multiple linear model for predicting Performance Index of student based on Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours and Sample Question Papers Practiced. Compute the $R^2_{Prediction}$ for the fitted model, write computational details and interpret your answer.

Paste the RStudio code of your solution on Blackboard.

b) Compute the COVRATIO for all the students. How many students have COVRATIO outside the interval $1 \pm 6\left(\frac{k+1}{n}\right)$? Write the computational details briefly.

Paste the RStudio code of your solution on Blackboard.

c) Delete all the observations from data having COVRATIO outside the interval $1 \pm 6\left(\frac{k+1}{n}\right)$. Refit the model and recompute COVRATIO for all the remaining n_r students. Is/are there any value(s) outside the interval $1 \pm 6\left(\frac{k+1}{n_r}\right)$? If yes, what is/are the roll number(s) of corresponding student(s)? Write the computational details briefly.

Paste RStudio code of your solution on Blackboard.

d) Fit robust regression model to the complete dataset (with $n = 3748$) using Bisquare function. What was the weight assigned to 75th observation (student with roll number 75) in the data?

Fitted Model:

Weight assigned to 75th observation:

e) Obtain a 90% interval estimate for the performance index of a student who studied for 5 hours, scored 70 in previous tests, participates in extracurricular activities, sleeps 6 hours per day and practiced 5 sample question papers. Write the fitted model, computed interval, and its interpretation.

Paste RStudio code of your solution on Blackboard.