KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

MATH 587: Advanced Applied Regression

Term 222, Final Exam

Wednesday May 17, 2023, 07:00 PM

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**10**	
2	**06**	
3	**04**	
4	**03**	
5	**04**	
6	**07**	
7	**06**	
8	**03**	
9	**02**	
Total		

Instructions:

1. Mobiles are not allowed in 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 mathematical/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.

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

(001) A multiple linear regression model is fitted with 5 continuous predictors i.e.
$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i \qquad \forall\, i = 1,2, \dots, n$
For testing the hypothesis H_0: $\beta_3 = \beta_5 = 0$ which statistical test should be used?

A. Partial F test
B. Full F test
C. T test
D. Z test
E. Durbin Watson test

(002) 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 $y = \left[e^{\beta_0 + \beta_1 e^X}\right]^2$. Transform the variables such that the relationship becomes linear. The transformed variables are

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

(003) A multiple linear regression model is fitted with 3 predictors i.e.
$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i \qquad \forall\, i = 1,2, \dots, 39$
A thorough influential analysis is performed to the model and it is found that $COVRATIO_{14} = 2.81$. In light of the given information, which of the following statements is true?

A. 14th observation is improving the precision of model.
B. 14th observation is degrading the precision of model.
C. The variance of 14th observation is 2.81.
D. The covariance of 14th observation with all the other observations is 2.81.
E. 14th observation has no effect on $Var - Cov(\hat{\beta})$.

(004) In a linear regression model, there are 2 continuous predictors and 2 categorical predictors. The notations are given as follows:
$y \rightarrow$ response variable
$X_1 \rightarrow$ Continuous Predictor 1
$X_2 \rightarrow$ Continuous Predictor 2
$X_3 \rightarrow$ Categorical Predictor 1 with 2 categories
$X_4 \rightarrow$ Categorical Predictor 2 with 4 categories

We introduce the following 4 dummy variables:

$I_3 = \begin{cases} 1 & \text{if } X_3 \text{ is equal to its category 1} \\ 0 & \text{otherwise} \end{cases}$

$I_4 = \begin{cases} 1 & \text{if } X_4 \text{ is equal to its category 1} \\ 0 & \text{otherwise} \end{cases}$

$I_5 = \begin{cases} 1 & \text{if } X_4 \text{ is equal to its category 2} \\ 0 & \text{otherwise} \end{cases}$

$I_6 = \begin{cases} 1 & \text{if } X_4 \text{ is equal to its category 3} \\ 0 & \text{otherwise} \end{cases}$

How many regression lines are accommodated in the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 I_3 + \beta_4 I_4 + \beta_5 I_5 + \beta_6 I_6 + \epsilon$?

A. 8
B. 7
C. 6
D. 5
E. 4

(005) Consider a response variable y and one predictor X. The values of X are shown below:
X = 1.00, 1.07, 1.025, 1.02, 1.045, 1.085, 1.06, 1.05, 1.095, 1.02.
Suppose that we wish to fit a second - order polynomial model using these levels for the regressor variable X. In light of the given information, which of the following statements is true?

A. There is potentially a problem of multicollinearity.
B. The variance of X is too low that can cause the problem of non-constant variance of ϵ.
C. The R^2 of the model will be too low.
D. The coefficient of X^2 in the regression will tend to ∞.
E. The regression equation cannot be estimated through the usual method of ordinary least squares.

(006) A multiple linear regression model is fitted with 3 predictors i.e.
$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i \qquad \forall\, i = 1,2,\ldots,39$
A thorough influential analysis is performed to the model and it is found that $DFFIT_9 = -0.00219$. In light of the given information, which of the following statements is true?

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 $\hat{\beta}_0$.

(007) Based on the performance of students in 1st Major exam, an instructor wants to predict the students' scores in 2nd Major exam. The data on the scores of students from two sections (i.e. Section 1 and Section 2) are available. Fit a regression equation to predict the students' score in 2nd Major exam based on the performance in 1st Major exam. The regression equation should have the capacity to accommodate change in intercept and change in slope of lines for both sections. The variables are defined as:

$y \rightarrow$ Major 2 score

$X_1 \rightarrow$ Major 1 score

$$X_2 = \begin{cases} 1 & \text{if student belongs to section 1} \\ 0 & \text{otherwise} \end{cases}$$

Which of the following models meets the requirements?

A. $\boldsymbol{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon}$
B. $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
C. $y = \beta_0 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon$
D. $y = \beta_0 + \beta_1 X_1 + \beta_{12} X_1 X_2 + \epsilon$
E. $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_2^2 + \epsilon$

(008) Ridge regression is preferred over ordinary least squares regression

A. in the presence of multicollinearity.
B. when the number of predictors is not enough.
C. if the variance of error term is high.
D. because of less bias.
E. it can handle the influential observations.

(009) 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 residuals
B. squares of the horizontal deviations
C. squares of the fitted values
D. difference between observed and predicted response
E. absolute of the fitted values

(010) The most common method of fixing the failure to constant variance assumption is

A. transforming the response variable.
B. transforming the predictor.
C. centering the predictors.
D. deleting the insignificant variables.
E. scaling the data using unit length scaling method.

Q2: Consider a Logistic regression model with 3 predictors and an intercept term.

(4 pts.) Using data consisting of 100 observations, regression analyst A fits the logit model. The estimated parameters using these data are: $\hat{\beta}_0 = 5$, $\hat{\beta}_1 = 1$, $\hat{\beta}_2 = 2$ and $\hat{\beta}_3 = -1$. Predict the original response (π_i) for an observation with $x_1 = 0.5$, $x_2 = 0.33$ and $x_3 = 0.25$. Write your simplified expression for the predicted π_i.

(2 pts.) Regression analyst B also fits the same model as Regression analyst A but without an intercept term. Comment on the difference of the deviances of the models fitted by regression analysts A and B.

Q3: Answer the following four questions related to multicollinearity.

(1 pt.) What is multicollinearity and explain if it is an assumption of model fitting or not.

(1 pt.) Name some methods of detecting the presence of multicollinearity.

(1 pt.) Explain what are the consequences of multicollinearity.

(1 pt.) Explain at least two possible ways of dealing with multicollinearity instead of removing the predictors from the model.

Q4: A researcher fits a cubic spline to a dataset containing the response variable y and a predictor X, with one knot at point k. The equation for the fitted model is given as:

$$E(Y_i) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + +\beta_{11}[(X-k)_+] + \beta_{12}[(X-k)_+]^2 + \beta_{13}[(X-k)_+]^3$$

(3 pts.) What are the drawbacks of fitted cubic spline and what adjustments (if any) do you recommend? Explain in detail.

Q5: A dataset on gasoline mileage performance for 25 automobiles is available where the description of variables is as follows:

y→ Miles/gallon x_1→ Compression ratio x_2→ Rear axle ratio
x_3→ Overall length (in.) x_4→ Weight (lbs.)

A model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$ is fitted and thorough influential analysis on the given data is performed. Interpret the following given values of influential diagnostics:

(1 pt.) Cook's D_9 = 1.3921

Comment:

(1 pt.) $DFBETAS_{2,13}$ = 0.19448

Comment:

(1 pt.) $DFFITS_{17}$ = -0.29114

Comment:

(1 pt.) $COVRATIO_{21}$ = 0.47891

Comment:

Q6: Code_____ **Report at least 4 decimal points.**

Write code # ↑ before starting.

A dataset is available for 63 samples of gold proximity where the response variable is
GOLD ----------->> presence/absence of gold, `1=Present, 0=absent (0.5km)`

The predictors are
AS ----------->> As level,
SB ----------->> Sb level,
LIN ----------->> presence/absence of lineament, `1=Present, 0 if absent (0.5km)`

(4 pts.) Fit a suitable model to predict the presence/absence of gold. Include the 2 main predictors in the model i.e. AS and SB. Also include a 2-way interaction of AS and LIN. Write down the fitted model and comment on the significance of AS, SB and 2-way interaction of AS and LIN.

(3 pts.) Use your model to predict the presence/absence of gold on a site where **As level** is 3.79, **Sb level** is 2.14 and lineament is present. Also provide an approximate 95% prediction interval. There is no credit for answers without detailed justification so provide all details.

Q7: Code_____ **Report at least 4 decimal points.**

Write code # ↑ before starting.

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 $y = \left[e^{\beta_0+\beta_1 e^X}\right]^2$.

Draw a scatter plot of y against X and you will notice that the relationship is not linear. You can use `plot(y,x)` command in R for that.

Now, transform the variables such that the relationship becomes linear i.e. the new variables are $y_1 = \ln\sqrt{y}$ and $X_1 = e^X$. Draw the scatter plot of y_1 against X_1 and you will notice that the relationship is linear.

Fit a linear regression model on the transformed variables y' and X'. The fitted model is given as:

(2 pts.) $\hat{y}_1 =$ [_____] + [_____]X_1

Predict the original response y when $X = 1.9$. Also construct a 92% prediction interval for original y when $X = 1.9$.

(2 pts.) $\hat{y}_{X=1.9} =$_____

(2 pts.) Lower Prediction Limit = _____ Upper Prediction Limit = _____

Hint: First, predict the transformed response ($\hat{y}_1$) using the fitted linear model. Also, find the prediction interval. Finally, de-transformed the predicted value and prediction interval.

Q8: Code_____ **Report at least 4 decimal points.**

Write code # ↑ before starting.

Download the Excel file for this question containing the data on two variables y and X. Fit a linear spline to these data to predict y using two knots i.e. $k_1 = 15$ and $k_2 = 29$. The line should be continuous at the knots. Also, predict y when $X = 35$.

Hint: Fit the model: $y = \beta_0 + \beta_1 X + \gamma_1 S_1 + \gamma_2 S_2 + \epsilon$ where

$$S_1 = \begin{cases} X - 15, & X > 15 \\ 0, & X \leq 15 \end{cases} \quad \text{and} \quad S_2 = \begin{cases} X - 29, & X > 29 \\ 0, & X \leq 29 \end{cases}$$

(2 pts.) Final fitted model:

$\hat{y} =$ _____ + _____(X) + _____(S_1) + _____(S_2)

(1 pts.) Predicted y when $X = 35$: _____

Q9: Code_____ **Report at least 6 decimal points.**

Write code # ↑ before starting.

Attached file contains the data on gasoline mileage performance for 25 automobiles where the description of variables is as follows:

y→ Miles/gallon x_1→ Compression ratio x_2→ Rear axle ratio
x_3→ Overall length (in.) x_4→ Weight (lbs.)

Fit the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$ and perform a thorough influential analysis on the given data and answer the following questions:

(0.5 pts.) Cook's D_9 = _____

(0.5 pts.) $DFBETAS_{2,13}$ = _____

(0.5 pts.) $DFFITS_{17}$ = _____

(0.5 pts.) $COVRATIO_{21}$ = _____

Good Luck