King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
Department of Mathematics

STAT 513: Statistical Modelling

Term 212, Second Major Exam, Tuesday April 19, 2022, 09:00 PM

Name: ________________________________ ID #: ____________________

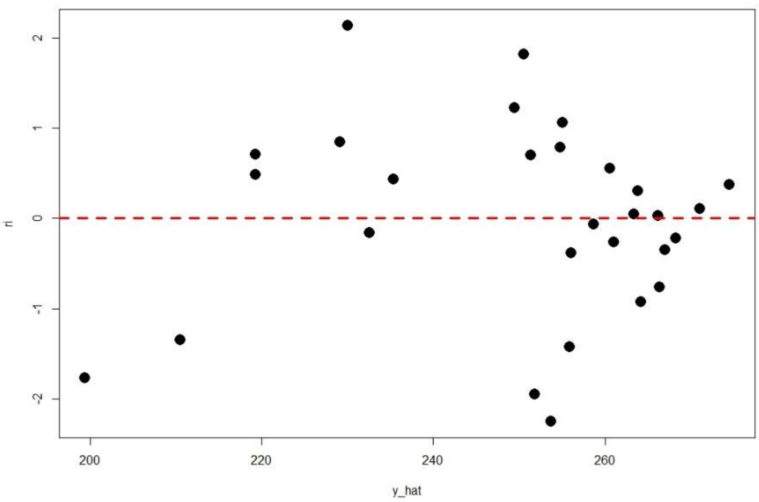
Please mark the correct answer to each of the questions by completely darkening the circle of your choice with a dark pen or pencil.

MULTIPLE CHOICE:	A	B	C	D	E
Q.No.1: -	O	O	O	O	O
Q.No.2: -	O	O	O	O	O
Q.No.3: -	O	O	O	O	O
Q.No.4: -	O	O	O	O	O
Q.No.5: -	O	O	O	O	O
Q.No.6: -	O	O	O	O	O
Q.No.7: -	O	O	O	O	O
Q.No.8: -	O	O	O	O	O
Q.No.9: -	O	O	O	O	O
Q.No.10: -	O	O	O	O	O

Code: 00

Score: $\frac{\quad}{10}$

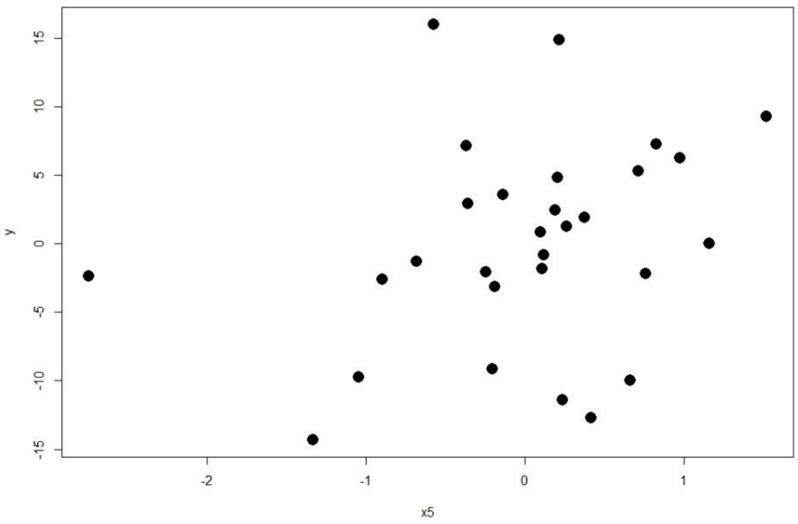
Q1: Consider the simple linear regression model fit to the solar energy data. The plot of residuals vs fitted response is given as follows:



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

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

Q2: A multiple linear regression model is fitted with 5 predictors i.e.
$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i \qquad \forall\, i = 1,2,\dots,n$
The partial regression plot for predictor X_5 is given below:



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

A. In presence of X_1, X_2, X_3 and X_4 in the model, X_5 is not contributing significantly.
B. The coefficient of X_5 in the fitted regression equation is negative.
C. The partial relationship of X_5 with Y is quadratic.
D. There is strong positive correlation between Y and X_5.
E. There is evidence of severe multicollinearity.

Q3: 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. $\boldsymbol{y' = \ln\sqrt{y}}$ **and** $\boldsymbol{x' = e^X}$

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 $x' = e^{\frac{1}{X}}$

Q4: 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 $COVRATIO_{22} = 1.81$. In light of the given information, which of the following statements is true?

A. 22nd observation is improving the precision of model.

B. 22nd observation is degrading the precision of model.

C. The variance of 22nd observation is 1.81.

D. The covariance of 22nd observation with all the other observations is 1.81.

E. 22nd observation has no effect on $Var - Cov(\hat{\beta})$.

Q5: 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 $DFFIT_9 = -0.0219$. 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 $\hat{\beta}_0$.

Q6: 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.

Q7: 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 → Major 2 score

X_1 → 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. $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_2^2 + \epsilon$

Q8: In a linear regression model, there are 2 continuous predictors and 2 categorical predictors. The notations are given as follows:

y → response variable

X_1 → Continuous Predictor 1

X_2 → Continuous Predictor 2

X_3 → Categorical Predictor 1 with 2 categories

X_4 → Categorical Predictor 2 with 4 categories

We introduce the following 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. 6
B. 4
C. 5
D. 3
E. 7

Q9: 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 $H_0\colon \beta_3 = \beta_5 = 0$ which statistical test should be used?

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

Q10: Box-Cox transformation is used for finding the optimal power transformation on

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

King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
Department of Mathematics

STAT 513: Statistical Modelling

Term 212, Second Major Exam, Tuesday April 19, 2022, 09:00 PM

Name: ________________________________ ID #: ____________________

Q1: 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 90% 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.

To do prediction in R, you can follow these steps:

Suppose you called the fitted linear model as model1 where the response is denoted by y1 and the predictor is denoted by x1. Write down the new data for which you want to do prediction i.e. `new.data <- data.frame(x1=x10)` where x10 is the value you want to plug in to the model for prediction. Then use `predict(model1 , newdata=new.data , interval = "prediction", level = 1-α)` for obtaining the prediction and prediction interval. Finally, de-transform everything.

Q2: Code_____ **Report at least 4 decimal points.**

Write code # before starting.

(6 pts.) 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

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

You can use the following R code for creating the new predictor:

```
S1 <- ifelse(X-15<0,0,X-15)
```

Final fitted model:

$\hat{y} =$ ________ + ___________(X) + ____________(S_1) + ____________(S_2)

Predicted y when $X = 35$: ______________________

Q3: 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 x1→ Compression ratio x2→ Rear axle ratio

x3→ Overall length (in.) x4→ Weight (lb)

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:

(1.5 pts.) Cook's D_9 = __________________ Comment:

(1.5 pts.) $DFBETAS_{2,13}$ = __________________ Comment:

(1.5 pts.) $DFFITS_{17}$ = ____________________ Comment:

(1.5 pts.) $COVRATIO_{21}$ = ___________________ Comment:

You can use the following R codes for influential analysis using `car` library:
`lm.influence(model1)` for hat matrix, `options(scipen = 2)` for displaying all decimals, `cooks.distance(model1)` for Cook's distance, `dffits(model1)` for DFFITS, `dfbetas(model1)` for DFBETAS, `covratio(model1)` for covariance ratio