KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

STAT 513: Statistical Modeling

Term 231, Major Exam I

Monday October 09, 2023, 05:45 PM

Name: ________________________________ ID #: ____________________

Question No	Full Marks	Marks Obtained
1	**38**	
2	**18**	
3	**24**	
Total	**80**	

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.

[Blank page]

Q1: (2 x 18 = 38 pts.) Multiple choice or fill in the blank questions. Any MCQ with more than one option circled will be considered wrong.

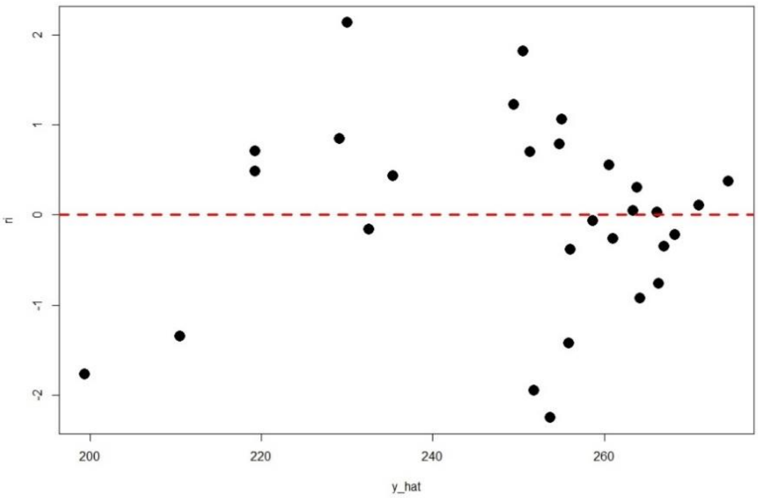
(i) Which one of the following diagrams would be useful in depicting the median value in the data?

(A) Pie Chart
(B) Bar Chart
(C) Scatter Plot
(D) Box Plot
(E) Stem and Leaf Diagram

(ii) Which one of the following diagrams would be useful in visually examining the relationship between two quantitative variables?

(A) Scatter Plot
(B) Box Plot
(C) Pie Chart
(D) Bar Chart
(E) Stem and Leaf Diagram

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



Considering the given 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.

(iv) Wilcoxon signed rank test is used for testing hypothesis related to

(A) Population variance
(B) Population median
(C) Population standard deviation
(D) Population percentiles
(E) None of the others

(v) Which one of the following statements is true?

(A) Z test is used for testing hypothesis about the population variance.
(B) T test is used for testing hypothesis about the population variance.
(C) T test is more flexible than the Z test.
(D) T test cannot be applied to a sample of size more than 30
(E) Z test cannot be applied to a sample of size more than 30

(vi) Which one of the following is **not** a linear regression model?

(A) $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 e^{X_{3i}} + \epsilon_i$
(B) $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}^3 + \beta_3 X_{3i}^2 + \epsilon_i$
(C) $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \ln X_{3i} + \epsilon_i$
(D) $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e^{\beta_3 X_{3i}} + \epsilon_i$
(E) $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} X_{2i} + \beta_3 X_{3i} + \epsilon_i$

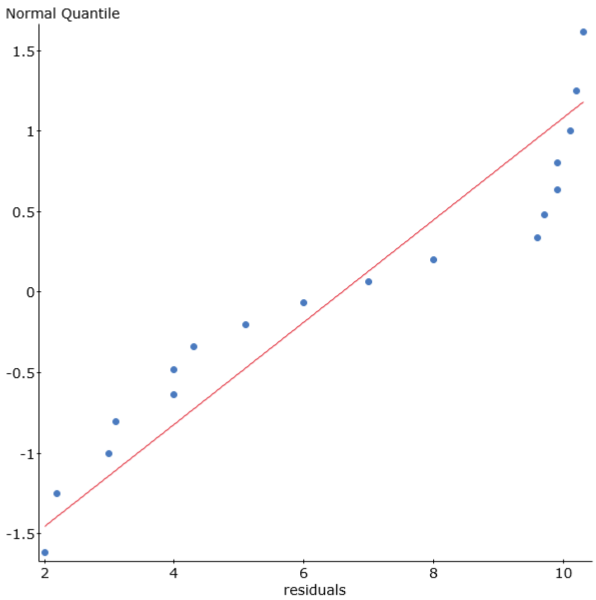
(vii) Which one of the following is true about the prediction interval computed from a multiple linear regression model?

(A) The width of prediction interval increases with an increase in significance level α.
(B) It is wider than the confidence interval for mean response.
(C) The predicted value can fall outside the prediction interval.
(D) A prediction interval cannot be computed when there is only one predictor in the model.
(E) The actual value of response always falls inside the prediction interval.

(viii) 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) Due to a unit increase in X_1, y increases on average by 1.67 units, keeping X_2 & X_3 fixed.
(B) Due to a unit increase in X_2, y decreases on average by 0.33 units, keeping X_1 & X_3 fixed.
(C) Due to a unit increase in X_3, y decreases on average by 1.92 units, keeping X_1 & X_2 fixed.
(D) Due to a unit increase in X_1, y decreases on average by 1.67 units, keeping X_2 & X_3 fixed.
(E) None of the others

(ix) Suppose we use a least squares linear regression model on a set of data points (x,y). We find the coefficient of correlation is -0.7, the regression line is given by y = -51x + 32 and the QQ plot of residuals is given as:



Which one of the following is true for this model?

(A) The model is good because the correlation is negative.
(B) The model is good because the slope coefficient is negative.
(C) The model is not good because the slope coefficient is negative.
(D) The model is good because QQ plot shows that the residuals are normally distributed.
(E) The model is not good because QQ plot shows that the residuals are not normally distributed.

(x) If the correlation coefficient between the two variables X and y is close to +1, what does that mean?

(A) X is causing the change in y.
(B) y is causing the change in X.
(C) When X increases y also increases, and vice versa.
(D) When X increases y decreases, and vice versa.
(E) X and y both are causing the change in each other.

(xi) In regression analysis, the difference between actual value of response variable and fitted value is called

(A) independent variable
(B) variance inflation factor
(C) analysis of variance
(D) residual
(E) outlier

(xii) Which one of the following is **not** true for regression analysis?

(A) SSE ≤ 0
(B) SSR ≥ 0
(C) SSR $\leq$ SST
(D) SSE $\leq$ SST
(E) SST ≥ 0

(xiii) In a linear regression model, we performed a Breusch-Pagan test and found the test statistic $BP = 8.46$ with the p-value = 0.21. What do we conclude from this output assuming $\alpha = 0.05$?

(A) None of the predictors is significant.
(B) The equal variance assumption has not failed.
(C) The normality assumption has failed.
(D) The linearity assumption has not failed.
(E) None of the others.

(xiv) 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 observed response
(B) squares of the fitted values
(C) difference between observed and predicted response
(D) absolute of the fitted values
(E) None of the others

(xv) For testing the significance of a predictor X_1 in simple linear regression, we can define Z-test based on $Z = \frac{\hat{\beta}_1}{\sqrt{\sigma^2/S_{XX}}}$. Why this test is impractical for regression analysis?

(A) S_{XX} is never known.
(B) σ^2 is never known.
(C) $\hat{\beta}_1$ is never known.
(D) Normal distribution PDF cannot be integrated.
(E) CDF of Normal distribution is not available in closed form.

(xvi) In a linear regression model, we performed Lilliefors test and found the test statistic $D = 0.19$ with the p-value = 0.035. What do we conclude from this output assuming $\alpha = 0.05$?

(A) All predictors are insignificant.
(B) The linearity assumption has failed.
(C) The normality assumption has failed.
(D) The equal variance assumption has failed.
(E) None of the others.

(xvii) What is the difference between mathematical and statistical relationships?

- (A) The error term.
- (B) Nothing, they are both the same.
- (C) Statistical relationships are exact while mathematical are approximate.
- (D) The intercept.
- (E) The slope.

(xviii) In multiple linear regression analysis, a partial F test is used for

- (A) Testing the normality assumption.
- (B) Testing the independence assumption.
- (C) Testing the assumption of equal variance.
- (D) Testing the significance of some predictors.
- (E) None of the others.

(xix) 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: $H_0: \frac{\beta_3}{10} - 2\beta_2 = -0.5$ against $H_1: \frac{\beta_3}{10} - 2\beta_2 \neq -0.5$.
Write down the $\boldsymbol{T}$ matrix and $\boldsymbol{c}$ vector for testing the above hypotheses.

$\boldsymbol{T} = [\qquad\qquad\qquad\qquad\qquad\qquad]$, $\boldsymbol{c} = [\qquad\qquad]$

Q2: (3+5+5+3+2 = 18 pts.) Data on the thrust of a jet turbine engine and four predictors are available with $n = 32$. Several models are applied to the given dataset and the resulting R outputs are given below:

model1: $\boldsymbol{y = B_0 + B_2X_2 + \varepsilon}$

```
Call:
lm(formula = y ~ x2, data = jet_turbine_engine)

Residuals:
    Min      1Q  Median      3Q     Max 
-172.91  -95.74  -35.49   51.55  489.55 

Coefficients:
                Estimate   Std. Error t value Pr(>|t|)    
(Intercept) -27070.92141   2051.38282  -13.20 5.01e-14 ***
x2               1.04584      0.06937   15.08 1.53e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 175.6 on 30 degrees of freedom
Multiple R-squared:  0.8834,	Adjusted R-squared:  0.8795 
F-statistic: 227.3 on 1 and 30 DF,  p-value: 1.533e-15

> nortest::ad.test(x=rstandard(model1))

	Anderson-Darling test

data:  rstandard(model1)
A = 2.1329, p-value = 0.00001502

> lmtest::bptest(model1)

	studentized Breusch-Pagan test

data:  model1
BP = 0.0046336, df = 1, p-value = 0.9457

> car::durbinWatsonTest(model1)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1901218      2.346469   0.324
 Alternative hypothesis: rho != 0
```

model2: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = jet_turbine_engine)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.595 -18.056   4.516  17.017  44.965 

Coefficients:
              Estimate Std. Error t value     Pr(>|t|)    
(Intercept) -3900.2496  2651.1738  -1.471       0.1528    
x1             1.4549     0.1634   8.906 0.0000000016 ***
x2             0.1882     0.1196   1.574       0.1272    
x3             0.7653     0.4219   1.814       0.0808 .  
x4           -17.0861     2.7906  -6.123 0.0000015324 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.48 on 27 degrees of freedom
Multiple R-squared:  0.997,	Adjusted R-squared:  0.9966 
F-statistic:  2276 on 4 and 27 DF,  p-value: < 2.2e-16

> anova(model1,model2)
Analysis of Variance Table

Model 1: y ~ x2
Model 2: y ~ x1 + x2 + x3 + x4
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     30 925016                                  
2     27  23460  3    901556 345.86 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Report at least 4 decimal points of your numerical answers.

(i) It can be seen from the above outputs that predictor **x2** was significant in model1, but it became insignificant in model2. What is/are the possible reason(s)? [3 pts.]

(ii) For model1, if we are interested in testing the assumption of homoscedasticity (equal variance), then fill the following blanks: [5 pts.]

H_0: ____

H_1: ____

p-value of the test = ____________

Assuming $\alpha = 0.05$, we reject H_0 if ______________________________

Conclusion ____

____

(iii) With reference to model2, if we are interested in testing the significance of **x3**, then fill the following blanks: [5 pts.]

H_0: ____

H_1: ____

Test statistic = ____________

p-value of the test = ____________

Conclusion ____

____

(iv) In the presence of **x2**, are **x1**, **x3** and **x4** contributing significantly? i.e.
H_0: $\beta_1 = \beta_3 = \beta_4 = 0$ against H_1: At least one $\beta_j \neq 0$ for $j = 1, 3$ or 4. [3 pts.]

Test statistic = ____________

p-value of the test = ____________

Conclusion ____

____

(v) What percent of the variation in thrust of a jet turbine engine is explained by the four predictors?

____________ [2 pts.]

Name: ________________________ ID #: _______________ Data Code: ____

Report at least 4 decimal points of your numerical answers.

Q3: (2+3+3+5+2+3+3+3 = 24 pts.) Data on the thrust of a jet turbine engine and four predictors are available with $n = 32$. Fit a multiple linear regression model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$.

(i) The fitted regression equation is [2 pts.]

$\hat{y} =$ ____

(ii) Predict the thrust of a jet turbine engine when $x_1 = 2080$, $x_2 = 30200$, $x_3 = 1710$ and $x_4 = 105$. [3 pts.]

The predicted value is equal to ____________________.

(iii) A 99% prediction interval for the thrust of a jet turbine engine when $x_1 = 2080$, $x_2 = 30200$, $x_3 = 1710$ and $x_4 = 105$ is given as: [3 pts.]

[__________________ , __________________]

(iv) Is the prediction done in part (ii) interpolation or extrapolation? Provide all the details of your solution before writing the final answer. [5 pts.]

(v) Construct a 99% confidence interval estimate for β_3 i.e. the average change in y due to a unit change in X_3, holding the other predictors. [2 pts.]

[__________________ , __________________]

(vi) Test the following constraints: $H_0: \frac{\beta_3}{10} - 2\beta_2 = -0.5$ against $H_1: \frac{\beta_3}{10} - 2\beta_2 \neq -0.5$. The $\boldsymbol{T}$ matrix and $\boldsymbol{c}$ vector for testing the above hypotheses is given as: [3 pts.]

$\boldsymbol{T} = [0 \quad 0 \quad -2 \quad 0.1 \quad 0]$, $\boldsymbol{c} = [-0.5]$

For testing the above hypothesis, the p-value is given by ________________________.

(vii) For testing the normality assumption, perform the Lilliefors test on studentized residuals (ri) and report your findings. [3 pts.]

p-value = ________________________

Conclusion:

(viii) For testing the equal variance assumption, perform the Breusch-Pagan test and report your findings. [3 pts.]

p-value = ________________________

Conclusion:

Good Luck