KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

STAT 513: Statistical Modeling

Term 241, Midterm Exam
Monday October 28, 2024, 5:35 PM

Time allowed **150 minutes**

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**16**	
2	**14**	
3	**20**	
Total	**50**	

Instructions:

- 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.
- 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.
- Report **at least 4 decimal points** of your numerical answers.

Q1: (16 points) Answer the following 16 questions:

1. When should a partial regression plot be used?

   (A) To assess the relationship between a response variable and a predictor, controlling for other predictors
   (B) To determine the overall model fit
   (C) To visualize the distribution of residuals
   (D) To identify outliers in the dataset
   (E) To compare multiple models

2. Which technique is useful for visualizing the relationships among multiple numerical variables?

   (A) Box Plot
   (B) Histogram
   (C) Pie Chart
   (D) Heatmap
   (E) Bar Chart

3. What is the primary purpose of residual analysis in regression?

   (A) To determine the correlation between predictors
   (B) To calculate the R-squared value
   (C) To estimate the parameters of the model
   (D) To perform hypothesis testing
   (E) To assess the adequacy of the model

4. 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_1, y decreases on average by 1.67 units, keeping X_2 & X_3 fixed.
   (D) Due to a unit increase in X_3, y decreases on average by 1.92 units, keeping X_1 & X_2 fixed.
   (E) Due to a unit increase in X_3, y increases on average by 0.33 units, keeping X_1 & X_2 fixed.

5. 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 cannot be applied to a sample of size more than 30
   (D) Z test cannot be applied to a sample of size more than 30
   (E) T test is more flexible than the Z test.

6. 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

7. In linear regression models, the difference between actual value of response variable and fitted value is called

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

8. How can you visualize the proportions of different categories in a dataset?

   (A) Heatmap
   (B) Box Plot
   (C) Histogram
   (D) Pie Chart
   (E) Scatter Plot

9. 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.

10. Which of the following indicates that a linear regression model may not be appropriate?

    (A) A non-random pattern in the residuals vs. fitted plot
    (B) Residuals are symmetrically distributed
    (C) The R-squared value is close to 1
    (D) All residuals are small
    (E) F test for full model is significant.

11. How does a scatter plot matrix (created using pairs() in R) help in analyzing multiple numerical variables?

    (A) It displays the frequency distribution of each variable
    (B) It shows the pairwise relationships among all variables
    (C) It calculates the mean and median of each variable
    (D) It provides a summary of variance for each variable
    (E) It plots the cumulative distribution function for each variable

12. Which non-parametric test can be used as an alternative to one sample Z test when normality assumption is not met?

    (A) Anderson-Darling Test
    (B) Mann-Whitney U Test
    (C) Shapiro-Wilk Test
    (D) Wilcoxon Signed Rank Test
    (E) Durbin Watson Test

13. Which method is used to identify outliers in linear regression?

    (A) Standardized residuals
    (B) R-squared value
    (C) Coefficient of determination
    (D) Multicollinearity diagnostics
    (E) ANOVA test

14. What is the difference between mathematical and statistical relationships?

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

15. Which of the following is **not** an assumption of linear regression?

    (A) Predictors are uncorrelated i.e. no multicollinearity
    (B) Errors are normally distributed
    (C) Relationship between response and predictors is linear
    (D) Errors are uncorrelated
    (E) Constant variance of errors

16. 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: $H_0: \beta_1 = \beta_4 = 0$ against $H_1: \beta_j \neq 0$ for at least one $j = 1,4$.
    Write down the $\boldsymbol{C}$ matrix and $\boldsymbol{d}$ vector for testing the above hypotheses.

Q2: (3+3+3+2+3 = 14 pts.) This dataset contains financial and operational metrics for 32 mid-sized manufacturing companies. Each row represents a single company, and the variables are described as follows:

- `y` (Revenue): Annual revenue of each company in millions of dollars.
- `x1` (Operational Expenses): Total annual operational expenses in millions of dollars.
- `x2` (Production Output): The annual production output, measured in thousands of units.
- `x3` (Advertising and Marketing Spend): Annual budget allocated to advertising and marketing, in millions of dollars.
- `x4` (Market Penetration Index): A composite index representing the company's market penetration on a scale from 0 to 100.

model1: $y = B_0 + B_2X_2 + \varepsilon$

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

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: $y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \epsilon$

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

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


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) Based on Model 1, interpret the estimated coefficient for `x2`. Based on the estimated coefficient for `x2`, can we comment on the relationship between production output (`x2`) and revenue (`y`).

(ii) In Model 2, we are testing the significance of the predictor `x3`, in the presence of other predictors. Complete the following statements:

H_0:

H_1:

p-value of the test:

Decision:

Conclusion:

(iii) Considering the residuals in Model 1, evaluate whether the residuals meet the normality assumption. Use the provided output to fill in the blanks below:

H_0:

H_1:

p-value of the test:

Decision:

Conclusion:

(iv) Compare the residual standard errors of Model 1 and Model 2. What does this comparison suggest about the fit of the models?

Residuals Standard Error of Model 1: _____, Residuals Standard Error of Model 2: _____

Comment:

(v) Based on the F-statistic and p-value in Model 2, evaluate the overall significance of this model. Complete the following:

H0:

H1:

p-value of the test:

Decision:

Conclusion:

Name: _____ ID #: _____ Dataset Version ____

Q3: (2+2+4+2+1+2+4+3 = 20 pts.) The dataset includes information on n homes in the Eastern Region of Saudi Arabia, focused on predicting housing prices. This region encompasses both urban and suburban areas, including major cities like Dammam, Al Khobar, and Dhahran. Each observation represents one home, with key characteristics that influence the final sale price.

- `price`: The sale price of the home in Saudi Riyals (SAR).
- `area`: The size of the house lot in square meters, ranging from smaller urban plots to larger suburban or rural properties.
- `age`: The age of the house in years, values range from 0 (recently constructed) to about 100 years
- `rooms`: The total number of rooms in the house, excluding bathrooms.
- `garage`: The size of the garage in square meters, with values from 0 (no garage) up to 50 square meters for larger garages.
- `proximity`: The distance to the nearest city center (Dammam, Al Khobar, or Dhahran) in kilometers, with values ranging from 0.5 to 50 km.

(i) Determine the threshold age below which 95% of the houses in the sample fall, indicating that 5% of the houses have an age greater than this value.

Fit a multiple linear regression model for predicting the sale price of homes and answer the remaining questions.

(ii) Using the fitted regression model, calculate a 90% confidence interval for **average housing prices** of homes with an area of 900 square meters, an age of 10 years, 2 rooms, a garage size of 20 square meters, and a proximity of 43 kilometers to the nearest city center.

The 90% confidence interval is: [_____ , _____]

(iii) Assess whether the estimation of average housing prices done in (ii) represents interpolation or extrapolation. Provide a detailed explanation to support your answer before giving the final conclusion.

The prediction is an example of _____.

Reason:

(iv) Construct a 90% confidence interval for the coefficient of age, representing the average change in housing price for a one-year increase in the age of the house while holding other predictors constant.

The 90% confidence interval for the coefficient of age is: [_____ , _____]

(v) Calculate the adjusted R-squared value for the fitted model.

The adjusted R-squared value is: _____

(vi) Perform a diagnostic test to check for multicollinearity among predictors in the fitted model. What are the results?

Multicollinearity test results:

Implication:

(vii) Evaluate the assumption of no autocorrelation (independence) for the fitted model. Based on this value, discuss whether autocorrelation in residuals appears to be an issue and provide any recommendations.

Test name:

H_0:

H_1:

p-value:

Decision:

Conclusion:

(viii) Test whether the average change in house price due to one additional room, while holding other predictors constant, is equal to 1,000 SAR against the alternative that it is not equal to 1,000 SAR.

H_0: $\beta_3 = 1000$ against H_1: $\beta_3 \neq 1000$

Test name:

p-value:

Decision:

Conclusion:

[Best of luck]