King Fahd University of Petroleum & Minerals
Department of Mathematics

STAT 510: Regression Analysis

Term 241, Midterm Exam

Saturday November 09, 2024

Time allowed: 150 minutes

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	50	
2	24	
3	26	
Total	100	

Important instructions:

- All types of mobile phones or smart watches are not allowed during the examination.
- 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.

[Blank Page]

Q1. (2 x 25 = 50 pts.) Multiple choice questions.

(i) In a multiple linear regression model $y = X\beta + \epsilon$, where X is the design matrix, β is the coefficient vector, and ϵ is the error vector, the hat matrix H is defined as $H = X(X^TX)^{-1}X^T$. How do we differentiate between interpolation and extrapolation in this model?

(a) If $\max(h_{ii}) > h_{00}$, then it is extrapolation.

(b) If $h_{00} = 0$, then it is extrapolation.

(c) If the new data point has a small value of h_{00}, then it is extrapolation.

(d) If $\max(h_{ii}) < h_{00}$, then it is extrapolation.

(e) If h_{00} for the new data point is significantly smaller than the other diagonal elements, it is extrapolation.

Here, h_{ii} are the diagonal elements of H, $h_{00} = x_0^T(X^TX)^{-1}x_0$, where x_0 is the new data point.

(ii) Fit a multiple linear regression model $y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4X_{4i} + \beta_5X_{5i} + \epsilon_i$, we write down $C = \left[\begin{smallmatrix} 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{smallmatrix}\right]$ and $d = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$. What hypotheses are we testing?

(a) $H_0 : \beta_1 = \beta_3$ against $H_1 : \beta_1 \neq \beta_3$

(b) $H_0 : \beta_1 = \beta_3$ and $\beta_4 = 0$ against H_1 : At least one constraint in H_0 is not true

(c) $H_0 : \beta_1 - \beta_3 = -1$ against $H_1 : \beta_1 - \beta_3 \neq -1$

(d) $H_0 : \beta_1 + \beta_3 = 0$ against $H_1 : \beta_1 + \beta_3 \neq 0$

(e) $H_0 : \beta_1 = \beta_3 = 0$ against H_1 : At least one $\beta_j \neq 0$ for $j = 1, 3$

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

(a) σ^2 is never known.

(b) S_{XX} is never known.

(c) $\hat{\beta}_1$ is never known.

(d) The normal distribution does not have a PDF.

(e) The variance of the normal distribution is undefined.

(iv) 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

(v) If the correlation coefficient between the two variables X and y is 0.93, 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 decreases, and vice versa.

(d) X and y both are causing the change in each other.

(e) When X decreases, y also decreases, and vice versa.

(vi) Which of the following statements is true regarding the formal test for lack-of-fit in regression analysis?

(a) It requires replication of the response variable at least at one level of the predictor variable.

(b) It is used to test if the linear model is a good fit without considering the assumptions of normality and independence.

(c) It tests for the adequacy of a non-linear regression model.

(d) It can be used without needing any assumptions about the variance of the residuals.

(e) It is not possible to compute if the regression model is linear.

(vii) In multiple linear regression, multicollinearity refers to:

(a) A situation where the predictor variables are highly correlated with each other.

(b) A case where the response variable is correlated with multiple predictor variables.

(c) The violation of the linearity assumption in the regression model.

(d) The condition where residuals are heteroscedastic.

(e) The presence of outliers in the dataset.

(viii) In a multiple linear regression model $y = X\beta + \epsilon$, where X is the design matrix, β is the coefficient vector, and ϵ is the error vector, the hat matrix H is defined as $H = X(X^TX)^{-1}X^T$. What is the expected value value of $\epsilon^T H\epsilon$ i.e. $\mathbb{E}[\epsilon^T H\epsilon]$?

(a) 0

(b) $\sigma^2 \cdot n$

(c) $\sigma^2 \cdot \text{tr}(H)$

(d) $\sigma^2 \cdot \text{tr}(X^TX)$

(e) $\sigma^2 \cdot (n - k - 1)$

(ix) For a simple linear regression model, let $k_i = \frac{x_i - \bar{x}}{S_{xx}}$, where x_i are the data points, $\bar{x}$ is the mean of the x_i's, and S_{xx} is the sum of squared deviations of the x_i's from their mean, i.e., $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Which of the following is the value of $\sum_{i=1}^{n} k_i x_i$?

(a) $\bar{x}$

(b) 1

(c) $n\bar{x}$

(d) $\sum_{i=1}^{n} x_i$

(e) 0

(x) Which of the following statements is true for the least squares method in linear regression?

(a) It maximizes the correlation between the response and predictor variables.

(b) It minimizes the sum of squared residuals between the observed and predicted values.

(c) It minimizes the sum of the squares of the response variable.

(d) It uses the variance of the residuals to determine the best-fitting line.

(e) It maximizes the goodness of fit by adjusting the slope.

(xi) In a regression analysis, how is autocorrelation in the residuals typically detected?

(a) Using the Durbin-Watson test.

(b) By calculating the Variance Inflation Factor.

(c) By inspecting the residual vs. fitted plot.

(d) By comparing the adjusted R^2 values across different models.

(e) Using a Shapiro-Wilk test on the residuals.

(xii) In a multiple linear regression, the adjusted $\mathbf{R^2}$ is preferred over the regular $\mathbf{R^2}$ because:

(a) It always increases when a new predictor is added to the model.

(b) It adjusts for the number of predictors in the model and sample size.

(c) It is unaffected by the number of predictors in the model.

(d) It measures only prediction accuracy of the model.

(e) It is directly interpretable in terms of model fit.

(xiii) What does the term "studentized residual," denoted by r_i, specifically address when compared to "standardized residuals," denoted by d_i?

(a) It divides the residual by the exact standard deviation of the residual, improving the scaling.

(b) It uses the residuals from a model fitted using a subset of predictors.

(c) It normalizes the residuals to have a mean of 1 and variance of 0.

(d) It is calculated by multiplying the residuals by their standard errors.

(e) It incorporates the correlation between errors in the dataset.

(xiv) In the context of a multiple linear regression model $\mathbf{y} = X\beta + \epsilon$, where:

- $\mathbf{y}$ is the vector of observed responses,
- X is the matrix of predictors (design matrix),
- β is the vector of regression coefficients, and
- ϵ is the vector of errors,

the derivative of the expression $\beta^T X^T X \beta$ with respect to β is:

(a) $X^T X \beta$

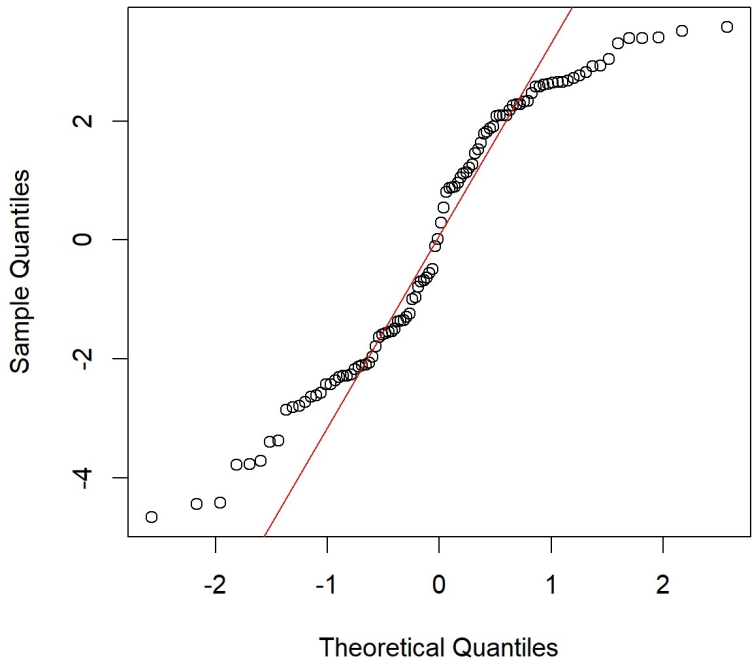
(b) $2X^T X$

(c) $X^T \beta$

(d) 2β

(e) $2X^T X \beta$

(xv) 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 Q-Q 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 positive.

(c) The model is not good because the slope coefficient is negative.

(d) The model is not good because the Q-Q plot shows that the residuals are not normally distributed.

(e) The model is good because the Q-Q plot shows that the residuals are normally distributed.

(xvi) Which of the following statements is true regarding the assumption of normality in the residuals of a regression model?

(a) Normality of residuals is only needed for model estimation, not inference.

(b) Normality of residuals is required to ensure that the residuals are independent.

(c) Normality of residuals is important for making valid inferences about the model parameters.

(d) Violating normality assumptions will cause multicollinearity in the model.

(e) Non-normal residuals can be easily corrected by increasing the sample size.

(xvii) In a linear regression model, we performed a Breusch-Pagan test and found the test statistic BP = 3.94 with the p-value = 0.027. What do we conclude from this output?

(a) None of the predictors is significant.

(b) The equal variance assumption has not failed.

(c) The linearity assumption has not failed.

(d) The normality assumption has not failed.

(e) The equal variance assumption has failed.

(xviii) What does heteroscedasticity in a regression model mean?

(a) The variance of the residuals is not constant across the range of predictor(s).

(b) The residuals are correlated with one another.

(c) The residuals are normally distributed with a mean of zero.

(d) The regression model includes non-linear terms.

(e) The errors are autocorrelated.

(xix) In the context of partial regression plots, what does a curvilinear pattern in the plot indicate?

(a) The predictor variable should be treated as response.

(b) There is a high multicollinearity between the predictor and other variables in the model.

(c) The response variable has non-constant variance.

(d) The model is overfitted and may need to be simplified.

(e) The relationship between the response and the predictor may not be linear.

(xx) Which of the following is the primary purpose of a QQ-plot of residuals?

(a) To check if the residuals follow a normal distribution.

(b) To assess if the residuals have constant variance across all levels of the predictors.

(c) To determine the presence of outliers in the dataset.

(d) To identify if there are any influential points that could affect the regression model.

(e) To assess the goodness of fit for the regression model.

(xxi) In a multiple linear regression model, the matrix X^TX, where X is the design matrix, is equal to:

(a)
$$\begin{bmatrix} n & 0 & 0 & \dots & 0 \\ 0 & \sum X_{1i}^2 & 0 & \dots & 0 \\ 0 & 0 & \sum X_{2i}^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sum X_{ki}^2 \end{bmatrix}$$

(b)
$$\begin{bmatrix} X_{1i} & X_{2i} & \dots & X_{ki} \\ X_{1i} & X_{2i} & \dots & X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1i} & X_{2i} & \dots & X_{ki} \end{bmatrix}$$

(c)
$$\begin{bmatrix} n & \sum X_{1i} & \sum X_{2i} & \dots & \sum X_{ki} \\ 1 & \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \dots & \sum X_{1i}X_{ki} \\ 1 & \sum X_{1i}X_{2i} & \sum X_{2i}^2 & \dots & \sum X_{2i}X_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \dots & \sum X_{ki}^2 \end{bmatrix}$$

(d)
$$\begin{bmatrix} n & \sum X_{1i} & \sum X_{2i} & \dots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \dots & \sum X_{1i}X_{ki} \\ \sum X_{2i} & \sum X_{1i}X_{2i} & \sum X_{2i}^2 & \dots & \sum X_{2i}X_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{ki} & \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \dots & \sum X_{ki}^2 \end{bmatrix}$$

(e)
$$\begin{bmatrix} n & 1 & 1 & \dots & 1 \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \dots & \sum X_{1i}X_{ki} \\ \sum X_{2i} & \sum X_{1i}X_{2i} & \sum X_{2i}^2 & \dots & \sum X_{2i}X_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{ki} & \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \dots & \sum X_{ki}^2 \end{bmatrix}$$

(xxii) In a linear regression model, the coefficient of determination indicates which of the following?

(a) The proportion of the variance in the response variable that is explained by the predictor(s).

(b) The mean squared error of the residuals.

(c) The correlation between the response and predictor variables.

(d) The significance of revised model.

(e) The exact change in the response variable for each unit change in the predictor variable(s).

(xxiii) What is the primary purpose of using the F-test in multiple linear regression?

(a) To determine if individual predictors are linearly related.

(b) To check for normality of the residuals.

(c) To detect multicollinearity between the predictor variables.

(d) To assess the overall significance of the regression model.

(e) To construct a prediction interval.

(xxiv) In the context of linear regression, the residuals represent:

(a) The variance explained by the predictor variable(s).

(b) The proportion of variance in the response variable explained by the model.

(c) The slope of the regression line.

(d) The difference between the observed and predicted values of the response variable.

(e) The correlation between the response and predictor variables.

(xxv) Which of the following techniques can be used to detect multicollinearity in multiple regression models?

(a) Anderson-Darling test

(b) Durbin-Watson test

(c) Variance Inflation Factor

(d) R-squared

(e) F-statistic

Q2. 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.
- x_1 (Operational Expenses): Total annual operational expenses in millions of dollars.
- x_2 (Production Output): The annual production output, measured in thousands of units.
- x_3 (Advertising and Marketing Spend): Annual budget allocated to advertising and marketing, in millions of dollars.
- x_4 (Market Penetration Index): A composite index representing market penetration of the company.

The R outputs for some fitted regression models are provided below:

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
```

(2 pts.) Based on model 1, interpret the estimated coefficient for x_2. Based on the estimated coefficient for x_2, can we comment on the relationship between production output (x_2) and revenue (y)?

(3 pts.) In model 2, we are testing the significance of the predictor x_3, in the presence of other predictors. Complete the following statements:

- H_0:
- H_1:
- p-value of the test:
- Decision:
- Conclusion:

(3 pts.) 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:

(2 pts.) In model 1, the predictor x_2 (Production Output) was statistically significant, but in model 2 it became insignificant. What could be the possible reasons for this change in significance?

(2 pts.) 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:

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

- H_0:
- H_1:
- p-value of the test:
- Decision:
- Conclusion:

(2 pts.) In the presence of x_2, do x_1, x_3, and x_4 contribute significantly to model 2? Specifically, test $H_0 : \beta_1 = \beta_3 = \beta_4 = 0$ against H_1: At least one $\beta_j \neq 0$ for $j = 1, 3, 4$.

- p-value of the test:
- Decision:
- Conclusion:

(3 pts.) To assess the assumption of homoscedasticity (equal variance) in model 1, complete the following statements:

- H_0:
- H_1:
- p-value of the test:
- Decision:
- Conclusion:

(1 pt.) What percentage of the variation in revenue is explained by the four predictors in model 2?

(3 pts.) For model 1, check the assumption of no autocorrelation (independence) in the residuals. Fill in the blanks below to complete the hypothesis testing framework:

- H_0:
- H_1:
- p-value of the test:
- Decision:
- Conclusion:

Name: _____ ID #: _____ Version: _____

Q3. (3+5+3+1+3+4+5+2 = 26 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.

Fit a multiple linear regression model expressed as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

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

[_____, _____]

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

The prediction is an example of _____.

Detailed Explanation and Results:

(iii) Construct a 90% confidence interval for the coefficient of `proximity`, representing the average change in housing price for a one kilometer increase in the proximity while holding other predictors constant.

The 90% confidence interval for the coefficient of `proximity` is:

[_____, _____]

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

The adjusted R-squared value is: _____

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

Multicollinearity test results:

Implication:

(vi) Evaluate the assumption of no autocorrelation (first order) for the fitted model using Breusch-Pagan test. Based on the test, discuss whether autocorrelation in residuals appears to be an issue or not.

H_0:

H_1:

p-value:

Decision:

Conclusion:

(vii) 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:

(viii) Construct a 99% confidence interval for the correlation between `price` and `proximity`.

The 99% confidence interval is:

[_____, _____]