KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DEPARTMENT OF MATHEMATICS

MATH 587: Advanced Applied Regression

Term 232, Midterm Exam
Saturday March 09, 2024, 3:30 PM

Time allowed **2 hours**

Name: ________________________________ ID #: ____________________

Question No	Full Marks	Marks Obtained
1	**27**	
2	**18**	
3	**15**	
Total	**60**	

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.

[Blank page]

Q1: (1.5 x 18 = 27 pts.) Multiple choice questions.

i. A plot of residuals against fitted values is primarily used to assess which assumption of the linear regression model?

(A) Outlier detection
(B) Homoscedasticity
(C) Normality of errors
(D) Independence of errors
(E) Multicollinearity

ii. Which one of the following statements can be **true** for a regression model $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i$?

(A) $SSE = -12.58$
(B) $SSR = 22.67$ and $SST = 13.56$
(C) $p - value = 1.092$ for testing significance of a predictor
(D) $VIF_3 = 0.57$ for X_3.
(E) $MSE = 1.982$

iii. If the correlation coefficient between two variables X and y is close to 0, how would you describe the relationship between two variables?

(A) There is perfect linear relationship between X and y.
(B) There is strong linear relationship between X and y.
(C) X and y are independent of each other.
(D) There is no significant linear relationship between X and y.
(E) The relationship between X and y is curvilinear.

iv. In multiple linear regression analysis, what does a large *p-value* signify for a predictor variable?

(A) The predictor has a strong influence on the model.
(B) The predictor follows a normal distribution.
(C) The predictor has no relationship with the response variable.
(D) The predictor may not be statistically significant.
(E) The predictor is strongly linearly related to the response variable.

v. What is likely to happen if an influential observation is retained when fitting a regression model?

(A) The regression coefficients will be biased, leading to inaccurate predictions.
(B) The regression coefficients will be unaffected by outliers.
(C) The model's performance will improve due to the presence of influential observations.
(D) The residuals will be evenly distributed, resulting in a better fit of the model.
(E) The model's interpretation will be simplified, facilitating easier communication of results.

vi. In a Normal Q-Q plot, if the observed data points deviate significantly from the straight line, what inference can be made?

(A) The normality assumption is not violated, suggesting the errors are normally distributed.
(B) The normality assumption is not affected by the plot's shape.
(C) The normality assumption is irrelevant for linear regression analysis.
(D) The normality assumption is inconclusive and requires further investigation.
(E) The normality assumption is violated, indicating non-normality of errors.

vii. How does scaling the residuals benefit multiple regression analysis?

(A) It standardizes the residuals, making it easier to identify outliers.
(B) It allows for easier interpretation of the regression coefficients.
(C) It reduces the number of predictor variables in the model.
(D) It increases the complexity of the regression model.
(E) It improves the goodness-of-fit of the regression model.

viii. What is the Hat matrix in multiple regression analysis and its primary use?

(A) It helps estimate the parameters of the regression model by understanding the relationship between predictor and response variables.
(B) It assists in pinpointing influential points in the dataset, aiding in identifying potential outliers or influential observations.
(C) It calculates standardized residuals, helping assess the model's goodness-of-fit and detect any unusual observations.
(D) It evaluates multicollinearity among predictor variables, which can affect the estimation and interpretation of regression coefficients.
(E) It helps us understand how each data point aligns with the predictor variables, showing their position in the dataset's overall structure.

ix. Why is detecting hidden extrapolation important in multiple regression?

(A) To increase the complexity of the regression model
(B) To decrease the computational time
(C) To eliminate outliers from the dataset
(D) To ensure the reliability of predictions
(E) To simplify the interpretation of regression coefficients

x. If the correlation coefficient between variables X and y is negative, which one of the following best explains this situation?

(A) X is causing y to decrease.
(B) When X increases, y decreases, and vice versa.
(C) An increase in X is causing y to decrease.
(D) Both X and y influence each other's change.
(E) When X decreases, y also decreases, and vice versa.

xi. Which one of the following is true for the estimated regression equation $\hat{y} = 4.2 - 2.14X_1 + 0.75X_2 - 1.63X_3$?

(A) Due to a unit increase in X_1, y increases on average by 2.14 units, holding X_2 & X_3.
(B) Due to a unit increase in X_2, y decreases on average by 4.2 units, holding X_1 & X_3.
(C) Due to a unit increase in X_3, y increases on average by 1.63 units, holding X_1 & X_2.
(D) Due to a unit increase in X_1, y decreases on average by 0.75 units, holding X_2 & X_3.
(E) None of the above.

xii. What characterizes multicollinearity in multiple regression analysis?

(A) When predictor variables in a regression model are highly correlated
(B) When predictor variables have no correlation
(C) When the dependent variable is highly correlated with the predictors
(D) When outliers are present in the dataset
(E) When the residuals are not normally distributed

xiii. In linear regression analysis, what is the primary objective of least squares estimation?

(A) Maximize the correlation coefficient.
(B) Optimize the range of the dataset.
(C) Minimize the sum of squared differences between observed and predicted values.
(D) Maximize the standard deviation of the residuals.
(E) Maximize the sum of absolute differences between observed and predicted values.

xiv. For a regression model $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \epsilon_i$, which one of the following T matrix and c vector are true for testing the hypothesis:
$H_0: \beta_1 = \beta_4 = 0$ against H_1: At least one $\beta_j \neq 0$ for $j = 1$ or 4.

(A) $T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$ and $c = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

(B) $T = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix}$ and $c = [0]$

(C) $T = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$ and $c = [0]$

(D) $T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$ and $c = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

(E) $T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$ and $c = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

xv. What is the primary purpose of the Breusch-Godfrey test in linear regression analysis?

(A) To assess the linearity of the relationship between predictor and response variables.
(B) To detect multicollinearity among predictor variables.
(C) To evaluate the normality assumption of the residuals.
(D) To test for the independence assumption.
(E) To examine the homoscedasticity of the residuals.

xvi. What conclusion can be drawn if the *p-value* of the Lilliefors test for normality is close to one in linear regression analysis?

(A) The normality assumption is violated, indicating non-normality of errors.
(B) The normality assumption is not relevant for linear regression analysis.
(C) The *p-value* indicates the presence of influential points in the dataset.
(D) The *p-value* suggests a perfect fit of the regression model to the data.
(E) The normality assumption is not violated, suggesting the errors are normally distributed.

xvii. Among the given equations, which one represents a non-linear regression model that cannot be transformed into a linear model?

(A) $E(y_i|X_i) = \beta_0 + \beta_1 e^{X_i}$
(B) $E(y_i|X_i) = \beta_0 + \beta_1 X_i^3$
(C) $E(y_i|X_i) = \frac{\beta_0 X_i}{\beta_1 + X_i}$
(D) $E(y_i|X_i) = e^{\beta_0 + \beta_1 X_i}$
(E) $E(y_i|X_i) = \beta_0 + \beta_1 \ln X_i$

xviii. What does the value of slope coefficient $\hat{\beta}_1 = 0.034$ indicate in a simple linear regression $y_i = \beta_0 + \beta_1 X_i + \epsilon_i$?

(A) For a one-unit increase in X, y increases by 0.034 units.
(B) We will fail to reject H_0: $\beta_1 = 0$ as $\hat{\beta}_1$ is close to zero
(C) The regression model is insignificant.
(D) The standard deviation of the residuals is 0.034.
(E) The correlation between variables X and y is very weak.

Q2: The dataset comprises information on various factors influencing individuals' salaries. The primary variable of interest is "*salary*" representing the annual income of individuals measured in dollars. Alongside salary, the dataset includes predictors such as "*age*, X_1" indicating the age of individuals in years, and "*education*, X_2" denoting their highest level of educational attainment, ranging from high school to advanced degrees. Additionally, "*experience*, X_3" reflects the number of years of general work experience, while "*relevant_experience*, X_4" specifies the years of experience directly related to the individual's job or industry. Lastly, "*last_promotion*, X_5" indicates the time elapsed since the individual's most recent career advancement.

```
model1 <- lm(salary ~ age + education + experience + relevant_experience + last_promotion, data = salary)

> summary(model1)

Coefficients:
                        Estimate    Std. Error  t value     Pr(>|t|)
(Intercept)             29845.625   4200.456    7.10        1.2e-10 ***
age                     522.535     55.123      9.48        2.1e-18 ***
education               706.211     85.456      8.27        3.2e-15 ***
experience              411.877     45.678      9.00        2.8e-17 ***
relevant_experience     684.736     65.789      10.41       1.2e-22 ***
last_promotion          158.561     95.123      1.67        0.098
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared: 0.847,  Adjusted R-squared: 0.835
F-statistic: 320.4 on 5 and 995 DF,  p-value: < 2.2e-16

> car::vif(model1)
age   education   experience  relevant_experience     last_promotion
2.3   3.1         2.7         1.5                     1.2

> nortest::lillie.test(rstandard(model1))

      Lilliefors (Kolmogorov-Smirnov) normality test

data:  rstandard(model1)
D = 0.036, p-value = 0.245

> lmtest::bptest(model1)

studentized Breusch-Pagan test

data:  model1
BP = 12.45, df = 5, p-value = 0.006

> lmtest::bgtest(model1, order=1)

      Breusch-Godfrey test for serial correlation of order up to 1
data:  model1
LM test = 0.513, df = 1, p-value = 0.392
```

```
model2 <- lm(salary ~ age + education + last_promotion, data = salary)

> summary(model2)

Coefficients:
                  Estimate    Std. Error  t value     Pr(>|t|)
(Intercept)       29561.345   4100.345    7.21        1.5e-13 ***
Age               484.369     54.234      8.93        2.1e-16 ***
Education         718.125     82.567      8.69        3.2e-15 ***
last_promotion    147.893     90.567      1.63        0.103
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared: 0.819,  Adjusted R-squared: 0.805
F-statistic: 260.8 on 3 and 997 DF,  p-value: < 2.2e-16

> anova(model1, model2)

Analysis of Variance Table

Model 1: salary ~ age + education + experience + relevant_experience + last_promotion
Model 2: salary ~ age + education + last_promotion
      Res.Df      RSS         Df     Sum of Sq    F           Pr(>F)
1     995         164578
2     997         198762      -2     -34184       12.524      0.0005 ***
```

Q2: Part I. (1.5 pts.) Which one of the following statements is true in context of model1?

(A) The change in "*salary*" for a one-unit increase in each predictor variable, holding all other predictors constant, is 522.535.

(B) The estimated average "*salary"* is 29845.625 when all predictor variables are zero.

(C) The estimated average "*salary*" is 4200.456 when all predictor variables are zero.

(D) The intercept of this model is close to zero and should be ignored in regression analysis.

(E) The estimated average "*salary*" is 320.4 when all predictor variables are at their mean values.

Q2: Part II. (1.5 pts.) What percentage of the variation in "*salary*" is explained by "*age*", "*education*", "*experience*", "*relevant_experience*", and "*last_promotion*"?

(A) 83.5%

(B) 81.9%

(C) 84.7%

(D) 80.5%

(E) 82.3%

Q2: Part III. (1.5 pts.) In context of model1, what can be inferred about multicollinearity among the predictors in the model?

(A) All predictors exhibit strong multicollinearity issues.
(B) Only the "*education*" predictor exhibits multicollinearity issues.
(C) There is no evidence of multicollinearity among the predictors in the model.
(D) The "*relevant_experience*" predictor is highly correlated with other predictors in the model.
(E) The "*last_promotion*" predictor has the highest degree of multicollinearity.

Q2: Part IV. (1.5 pts.) In context of model1, what does the *p-value* (`< 2.2e-16`) for the F-statistic in the model summary signify?

(A) All the predictors are significantly affecting the "*salary*".
(B) At least one predictor is significantly affecting the "*salary*".
(C) The predictors in the model are not statistically significant.
(D) There is insufficient evidence to determine the significance of the model.
(E) The model has perfect predictive power.

Q2: Part V. (3 pts.) In context of model1, test if "*last_promotion*" is significantly affecting "salary" in the presence of other predictors. Use 1% level of significance.

H_0: ____

H_1: ____

Test statistic = ___________________

p-value of the test = ___________________

Conclusion ____

____

Q2: Part VI. (3 pts.) What can be inferred about the assumption of normality in model1?

H_0: ____

H_1: ____

Test statistic = ___________________

p-value of the test = ___________________

Conclusion ____

____

Q2: Part VII. (3 pts.) What can be inferred about the assumption of homoscedasticity in model1?

H_0: ____

H_1: ____

Test statistic = ___________________

p-value of the test = ___________________

Conclusion ____

____

Q2: Part VIII. (3 pts.) What conclusion can be drawn about the joint significance of "*experience*" and "*relevant_experience*", in the presence of "*age*", "*education*" and "*last_promotion*"?

H_0: ____

H_1: ____

Test statistic = ___________________

p-value of the test = ___________________

Conclusion ____

____

Q3: The dataset comprises information on various factors influencing individuals' salaries. The primary variable of interest is "*salary*" representing the annual income of individuals measured in dollars. Alongside salary, the dataset includes predictors such as "*age*, X_1" indicating the age of individuals in years, and "*education*, X_2" denoting their highest level of educational attainment, ranging from high school to advanced degrees. Additionally, "*experience*, X_3" reflects the number of years of general work experience, while "*relevant_experience*, X_4" specifies the years of experience directly related to the individual's job or industry. Lastly, "*last_promotion*, X_5" indicates the time elapsed since the individual's most recent career advancement.

Download the data and RStudio codes sheet from Blackboard and write the data code below:

Midterm_code___

Fit a multiple linear regression model by regressing "*salary*" on all predictors i.e.

```
salary ~ age + education + experience + relevant_experience + last_promotion
```

Q3: Part I. (1 pt.) The fitted regression equation is

$\widehat{\text{salary}} =$____

Q3: Part II. (2 pts.) Based on fitted model, what would be the predicted salary of a new individual who is 35 years old, has 16 years of education, 8 years of experience, 6 years of relevant experience, and received their last promotion 2 years ago?

Q3: Part III. (3 pts.) Is the prediction done in Q3: Part II interpolation/extrapolation? Justify your answer and provide all the details.

Q3: Part IV. (2 pts.) Construct a 99% interval estimate for the average salary of all those individuals who are 35 years old, have 16 years of education, 8 years of experience, 6 years of relevant experience, and received their last promotion 2 years ago?

[___________________ , ___________________] Q3: Part V. (1 pt.) Do you suspect the presence of multicollinearity? Justify your answer and provide the details.

Q3: Part VI. (2 pts.) Conduct a Partial F-test for the joint significance of "*relevant_experience*" and "*last_promotion*", in the presence of "*age*", "*education*" and "*experience*"? Use 1% level of significance.

H_0: $\beta_4 = \beta_5 = 0$

H_1: At least one $\beta_j \neq 0$ for $j = 4$ or 5.

Test statistic = ___________________

p-value of the test = ___________________

Q3: Part VII. (2 pts.) For testing the normality assumption of model fitted in Q3: Part I, conduct a Lilliefors test.

H_0: The errors follow a Normal distribution.

H_1: The errors do not follow a Normal distribution.

Test statistic = ___________________

p-value of the test = ___________________

Q3: Part VIII. (2 pts.) For testing the homoscedasticity assumption of model fitted in Q3: Part I, conduct a Breusch-Pagan test.

H_0: The errors are homoscedastic.

H_1: The errors are heteroscedastic.

Test statistic = ___________________

p-value of the test = ___________________