KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

MATH 587: Advanced Applied Regression

Term 222, Midterm Exam

Saturday March 18, 2023, 06:00 PM

Name: ________________________________ ID #: ____________________

Question No	Full Marks	Marks Obtained
1	**33**	
2	**19**	
3	**18**	
Total	**70**	

Instructions:

1. 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.

2. Show all the calculation steps. There are points for the steps so if your miss them, you lose points. For multiple choice type questions, showing calculation steps is not required.

3. Report **at least 3 decimal points** of your numerical answers.

Q1: (1.5 x 22 = 33 pts.) Multiple choice questions.

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

(A) Nothing, they are both same.

(B) Statistical relationships are exact while mathematical are approximate.

(C) The error term.

(D) The intercept.

(E) The slope.

(002) 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.

(003) Assuming the normality of error term, what is the probability that the fitted regression line coincides with the population regression line i.e. $\hat{\beta}_0$ is exactly equal to β_0 and $\hat{\beta}_1$ is exactly equal to β_1?

(A) Approximately zero.

(B) Approximately one.

(C) Approximately half.

(D) Approximately 0.25.

(E) Approximately 0.75.

(004) For testing the significance of a predictor X_1, 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) σ^2 is never known.

(B) S_{XX} 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.

(005) In regression analysis, which one of the following is not a required assumption?

(A) The expected value of error term is one.

(B) The errors are Normally distributed.

(C) The values of error term are independent.

(D) The variance of the error term is same for all levels of X.

(E) The relationship between the predictor(s) and response is linear.

(006) If the total variation in our response variable (i.e. SST) is small, what does that mean?

(A) Coefficient of determination will be high.

(B) Coefficient of determination will be low.

(C) Coefficient of correlation will be positive.

(D) Coefficient of correlation will be negative.

(E) None of the other.

(007) Which one of the following is not true?

(A) SSR $\geq$ 0

(B) SSR $\leq$ SST

(C) SSE $\leq$ 0

(D) SSE $\leq$ SST

(E) SST $\geq$ 0

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

(A) Testing the significance of some predictors.

(B) Testing the normality assumption.

(C) Testing the independence assumption.

(D) Testing the assumption of equal variance.

(E) None of the others.

(009) In linear regression analysis, variance inflation factor is used for

(A) Testing the significance of some predictors.

(B) Testing the normality assumption.

(C) Testing the independence assumption.

(D) Testing the assumption of equal variance.

(E) None of the other.

(010) What is the main objective of scaling the residuals?

(A) To identify the presence of unusual observations.

(B) To decrease the magnitude of error values.

(C) To increase the precision of model.

(D) To identify the presence of multicollinearity.

(E) None of the others.

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

(A) All predictors are insignificant.

(B) The normality assumption has failed.

(C) The linearity assumption has failed.

(D) The equal variance assumption has failed.

(E) None of the others.

(012) 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?

(A) None of the predictors is significant.

(B) The normality assumption has failed.

(C) The linearity assumption has failed.

(D) The equal variance assumption has failed.

(E) None of the other.

(013) F test for lack of fit can only be performed if

(A) replicates on y for at least some levels of X are available.

(B) sample size $n > 30$.

(C) no. of predictors $k > 1$.

(D) the equal variance assumption has failed.

(E) correlation between y and X is positive.

(014) If the coefficient of determination is equal to 1, then the correlation coefficient

(A) must also be equal to 1

(B) can be either -1 or +1

(C) can be any value between -1 to +1

(D) must be < 0

(E) must be > 0

(015) Regression analysis was applied between \$ sales (y) and \$ advertising (X) across all the branches of a major international corporation. The following regression function was obtained after fitting the model: $\hat{y} = 5000 + 2.5X$. If the advertising budgets of two branches of the corporation differ by \$30,000, then what will be the predicted difference in their sales?

(A) \$70,000

(B) \$35,000

(C) \$5000

(D) \$2.5

(E) \$30,000

(016) Suppose you use regression to predict the height of graduate students by using their father's height as the explanatory variable. Height was measured in feet from a sample of 100 graduates. Now, suppose that the height of both the students and their fathers are converted to centimeters. The impact of this conversion is:

(A) the sign of the slope will change.

(B) the magnitude of the slope will change.

(C) the slope will remain the same.

(D) both (A) and (B) are correct.

(E) all (A), (B) and (C) are correct.

(017) In simple linear regression, least square method calculates the best-fitting line for the observed data by minimizing the sum of the

(A) absolute of the residuals

(B) squares of the horizontal deviations

(C) squares of the fitted values

(D) difference between observed and predicted response

(E) absolute of the fitted values

(018) If the coefficient of determination for a simple linear regression model is equal to 1, then which one of the following is not true?

(A) $\hat{y}_i = y_i \ \forall \ i = 1,2,\dots,n$

(B) SST = SSR

(C) SSE > 0

(D) sum of square of the errors is zero.

(E) sum of absolute of the errors is zero.

(019) In regression analysis, the variable that is being predicted is

(A) usually denoted by y

(B) called independent variable

(C) called indicator variable

(D) usually denoted by X

(E) called influential variable

(020) 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) None of the others

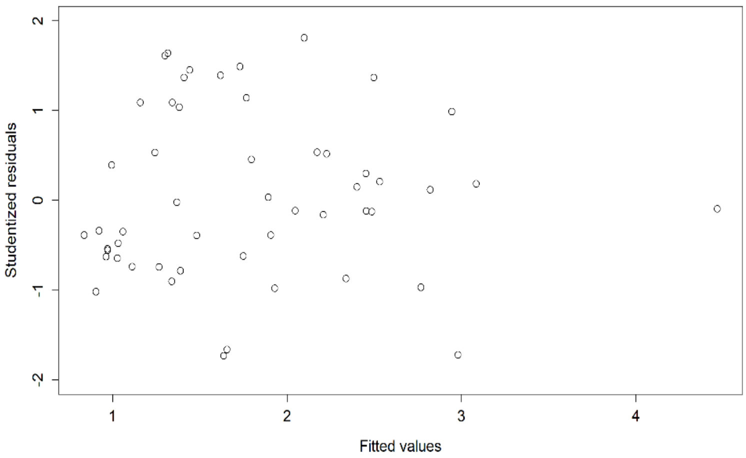
(B) A unit increase in X_1 causes y to increase by 1.67 units, keeping X_2 & X_3 fixed.

(C) A unit increase in X_1 causes y to decrease by 1.67 units, keeping X_2 & X_3 fixed.

(D) A unit increase in X_2 causes y to decrease by 0.33 units, keeping X_2 & X_3 fixed.

(E) A unit increase in X_3 causes y to decrease by 1.92 units, keeping X_2 & X_3 fixed.

(021) A research engineer is investigating the use of windmill to generate electricity. He has collected data on the DC output (volts) from his windmill and the corresponding wind velocity (miles per hour).



We fitted a simple linear regression line for predicting the DC output. The plot of the residuals against predicted DC outputs is given. This plot indicates

(A) Homoscedasticity assumption is violated.

(B) Normality assumption is violated.

(C) No linear relationship between y and X.

(D) Independence assumption is violated.

(E) None of the others.

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

(A) residual

(B) independent variable

(C) variance inflation factor

(D) analysis of variance

(E) outlier

Q2: (3+5+5+4+2 = 19 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: $\boldsymbol{y = \beta_0 + \beta_2 X_2 + \epsilon}$

```
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

> EnvStats::anovaPE(model1)
              Df  Sum Sq Mean Sq  F value    Pr(>F)    
x2             1 7007942 7007942 575.7192 0.0001587 ***
Lack of Fit   27  888498   32907   2.7034 0.2246818    
Pure Error     3   36517   12172                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> 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: $\boldsymbol{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon}$

```
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 3 decimal points of your numerical answers.

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

(002) For model1, if we are interested in testing the assumption of homoscedasticity (equal variance), then fill the following blanks:

H_0: ____

H_1: ____

p-value of the test = ____________________

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

Conclusion ____

____

(003) With reference to model2, if we are interested in testing the significance of **x3**, then fill the following blanks:

H_0: ____

H_1: ____

Test statistic = ___________________

p-value of the test = ___________________

Conclusion ____

____

(004) 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.

Test statistic = ___________________

p-value of the test = ___________________

Conclusion ____

____

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

____

Name: ________________________ ID #: _______________ Data Code: ____

Report at least 3 decimal points of your numerical answers.

Q3: (2+3+3+5+5 = 18 pts.) Data on the thrust of a jet turbine engine and four predictors are available with $n = 32$. Fit a multiple linear regression model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$.

(001) The fitted regression equation is

$\hat{y} =$ ____

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

The predicted value is equal to ____________________.

(003) 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:

[__________________ , __________________]

(004) Is the prediction done in part (002) interpolation or extrapolation? Provide all the details of your solution before writing the final answer.

(005) Is there evidence of the presence of multicollinearity? Provide all the details of your solution before writing the final answer.

Good Luck