KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

MATH 405: Learning from Data

Term 231, Major Exam II

Saturday December 02, 2023, 07:00 PM

Name: ________________________________ ID #: ___________________

Question No	Full Marks	Marks Obtained
1	**10**	
2	**05**	
3	**12**	
4	**13**	
Total	**40**	

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.

Q1: (5+5 = 10 pts.) For a multiple linear regression model $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(K+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$,

a) derive the estimates of $\boldsymbol{\beta}$ vector using the method of Maximum Likelihood Estimation (MLE) assuming $\epsilon \sim N(0, \sigma^2)$.

b) Also, show that for $k = 1$, MLE based estimated vector simplifies to $\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum(y-\bar{y})(x-\bar{x})}{\sum(x-\bar{x})^2} \end{bmatrix}$.

[Blank page]

Name: ________________________________ ID #: ___________________

Q2: (5 pts.) In a newly designed printer, the manager is interested in $X =$ time until the printer produces an error (in terms of number of days). Data is available on 31 printers and stored in the attached Excel file. Using the non-parametric Bootstrapping with ***10^5*** iterations, construct a 99% two-sided confidence interval for the median time until the printer produces an error. Paste your RStudio code on Blackboard and write your final answer along with the interpretation below:

My data file is Q2_code________

Q3: (2+4+3+3 = 12 pts.) Data on the thrust of a jet turbine engine (*y*) and four features are available with $n = 32$. Fit a linear model to these data using all the features. You can use the following formulas wherever needed:

$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T, \quad \boldsymbol{h_{00}} = \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_0}$

a) Using a global F-test, test the significance of full model. Report the value of *F-statistics*, the corresponding *p-value* and interpret your results.

b) The investigator of this study claims that the average change in *y* due to a unit change in *x2* keeping other features fixed, is the same as the average change in *y* due to a unit change in *x3* keeping other features fixed. Test the investigator's claim at 10% level of significance. Paste your RStudio code on Blackboard and, report your H0, H1 and results along with the interpretation below:

c) Find a 99% interval estimate for the thrust of a jet turbine engine when $x1 = 2080$, $x2 = 30200$, $x3 = 1710$ and $x4 = 105$. Paste your RStudio code on Blackboard and report your results along with the interpretation below:

d) Is the prediction done in part d) interpolation or extrapolation? Provide all the details of your solution before writing the final answer. Paste your RStudio code on Blackboard and report your results along with the interpretation below:

Q4: (3+2+2+3+3 = 13 pts.) Data are available from a medical study of patients with compensated liver disease. To see the impact of 4 different medicines, the patients are divided into 4 groups randomly and given the medication accordingly. After 6 months, the liver function score of patients is measured and provided in the Excel file named Q4-code_______.

a) Using one way analysis of variance, test the hypothesis that average liver function score of patients for all 4 groups is same and there is no significant difference among the four medications. Write your H0, H1 and p-value. Paste your RStudio code on Blackboard and report your results along with the interpretation below:

b) Test the assumption of homogeneity of variance. Write your H0, H1 and p-value. Paste your RStudio code on Blackboard and report your results along with the interpretation below:

c) Test the assumption of normality. Write your H0, H1 and p-value. Paste your RStudio code on Blackboard and report your results along with the interpretation below:

d) Assuming that the normality assumption failed in one way analysis of variance approach, apply a suitable non-parametric test for testing the hypothesis that average liver function score of patients for all 4 groups is same. Write the name of test, H0, H1 and p-value. Paste your RStudio code on Blackboard and report your results along with the interpretation below:

e) In continuation to part d), perform a post hoc analysis to find out which specific pair(s) of medicines have significantly different impact on the liver function score. Paste your RStudio code on Blackboard and report your results along with the interpretation below:

Good luck