**STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES**

**MASTER OF SCIENCE IN STATISTICAL SCIENCES**

**END OF SEMESTER EXAMINATION**

**STA 8103: LINEAR MODELS**

Date: 13th December 2024 Time: 5.00 pm – 7.30 pm

**INSTRUCTIONS:** ANSWER **QUESTION ONE** AND **ANY OTHER TWO** QUESTIONS

**Question 1 [20 Marks]**

(a) Outline any two approaches you would use to determine if interaction exists between two or more variables (4 marks).

(b) When modelling count data, how does one recognize over-dispersion and what are the appropriate measures to account for it? (5 marks)

(c) A computer solution for a regression model $Y=A+B_1X_1+B_2X_2+B_3X_3+D+E$ for a sample data set is given below:

| PREDICTOR | COEFF | STDEV | T-RATIO | P-VALUE |
|---|---|---|---|---|
| CONSTANT | 4.089 | 5.480 | 0.75 | 0.468 |
| X1 | 0.33157 | 0.03115 | 10.64 | 0.000 |
| X2 | -0.00775 | 0.01240 | -0.62 | 0.544 |
| D1 | 3.969 | 1.501 | 2.64 | 0.021 |
| D2 | -0.6788 | 0.9362 | -0.73 | 0.482 |

S = 1.727 R-SQ = 97.5% R-SQ(ADJ) = 96.7%

ANOVA

| SOURCE | DF | SS | MSS | F | P-VALUE |
|---|---|---|---|---|---|
| REGRESSION | 4 | 1402.10 | 350.53 | 117.57 | 0.000 |
| ERROR | 12 | 35.78 | 2.98 | | |
| TOTAL | 16 | 1437.88 | | | |

(I) Write down the estimated regression equation and the meaning of the values of the constant, the regression coefficients and the two R-SQ values.

(ii) Test at 2.5% level whether or not the coefficient of D2 is negative

(iii) Test the significance of the model at 1% level of significance (11 marks)

**Question 2 [20 Marks]**

(a) Briefly outline the following variable selection techniques for linear regression model.

(i) Akaike Information Criterion (AIC) (2 marks)
(ii) Bayesian Information Criterion (BIC) (2 marks)
(iii) Likelihood Ratio Test (LRT) (2 marks)
(iv) F-Test (2 marks)

(b) Discuss any 3 shrinkage methods of dealing with too many variables in a model. (6 marks)

(c) Suppose that an experimenter postulates a model $Y_i = \alpha + \beta x_{1i} + E_i \quad i = 1, 2, \ldots\ldots, n$ linearly relating $x_1$ to the response. Check if $b = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)Y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$ is an unbiased and consistent estimator of $\beta$ given that $E_i \sim N(0, \sigma^2)$ (6 marks)

**Question 3 [20 Marks]**

(a) What are full and less than full rank models? Discuss rank deficiency in relation t+o the set of data below

| Machine | Operator | Response |
|---|---|---|
| 1 | Me | 10 |
| 1 | Me | 8 |
| 1 | Me | 13 |
| 2 | You | 11 |
| 2 | You | 15 |
| 2 | You | 6 |

(5 marks)

(b) Using a binary logistic regression, explain in detail the three components of Generalized Linear Models and the assumptions of the models (8 marks)

(c) Discuss non linear regression models and use an example to explain Gauss-Newton and Levenberg-Marquardt algorithms as applied to these models. (7 marks)

**Question 4 [20 Marks]**

(a) Discuss the following statistical approaches:

(i) Reducing Confounding (2 marks)

(ii) Penalized Splines approximation (2 marks)

(ii Ridge regression (2 marks)

(b) What is mixed effects modelling and why does it matter. Explain the random effect part and the fixed effect part in a mixed model (6 marks)

(c) Residuals are useful in model diagnostics. Discuss any 4 model diagnostics that you could carry out using residuals and suggest any remedies should there be a violation of regression assumptions (8 marks)