43

**RAJARATA UNIVERSITY OF SRI LANKA**
**FACULTY OF APPLIED SCIENCES**

**Bachelor of Science in Applied Sciences**

**Third Year - Semester I Examination – July/August 2023**

**MAT 3203 – REGRESSION ANALYSIS**

**Time: Two (02) hours**

---

Answer **THREE** questions and **NO MORE**.

Calculators, Statistical tables, and graph papers will be provided.

---

1. a) Consider a data set where a simple linear regression (SLR) model is to be fitted with the equation $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\ ; i = 1,2,3,\dots,n$. Answer the following questions:

   i. What are the typical assumptions made about the term representing experimental error ($\varepsilon_i$) in the context of a simple linear regression (SLR) model?

   ii. By minimizing a suitable function, show that the least square estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ of $\beta_0$ and $\beta_1$ are given by

   $$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

   $$\widehat{\beta_1} = \frac{SS_{xy}}{SS_{xx}}$$

   Where $SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

   iii. Show that $\hat{\beta}_1$ is an unbiased estimator of $\beta_1$.

   b) A clinician recorded the age in years (x) and total cholesterol level of blood (y) for 20 patients suffering from a Cardiovascular disease. The summary statistics for the data are as follows:

$$\sum x_i = 809, \ \sum y_i = 68.3, \ SS_{xx} = 3630.95$$

$$SS_{xy} = 201.665, \ and \ SS_{yy} = 12.9455$$

i. Find the equation of the fitted simple linear regression model.

ii. Obtain the analysis of variance table and hence test the hypothesis that $\beta_1 = 0$.

iii. Calculate the coefficient of determination and interpret it in terms of this problem.

iv. Obtain 95% confidence interval for $\beta_1$. Hence, verify the results obtained in part (ii).

(100 marks)

2. a) The following information is extracted from the output obtained by fitting a linear regression model for the reaction time (minutes), y, of a chemical experiment using temperature as a predictor variable.

| Source of variation | Sum of squares | Degrees of freedom |
|---|---|---|
| Regression | 63075.666 | 2 |
| Residual | | |
| Total | 64221.872 | 42 |

i. What is the sample size used in this study?

ii. State whether the model fitted is a simple linear regression model or a multiple linear regression model. Give reasons for your answer.

iii. Construct the analysis of variance table and test whether the variables in the fitted model significantly help to predict the response variable. Clearly state the findings.

iv. Estimate the random variation in the reaction time based on the fitted model.

b) A researcher wants to fit the mean response function $E(y) = \beta_0 + \beta_1 x$ for the dried weight (mg) of a medical plant, with age (months) as the predictor variable $x$. The summary statistics computed from 26 observations are as follows:

$$\sum x = 106.0, \sum y = 64.6, \sum x^2 = 476.0, \ \sum y^2 = 224.56, \ \sum xy = 314.30$$

i. State any additional assumptions that are needed to obtain the least squares estimates for the regression parameters.

ii. Calculate the least square estimates for the slope parameter.

iii. Construct 95% confidence interval for the slope parameter.

iv. Using part (iii) or otherwise, test the hypothesis that an increase in the age by 2 months is associated with an increase in the dried weight by 2 milligrams and clearly state the findings.

(100 marks)

3. a) Show that,

i. $Var(\hat{y}) = \sigma^2 \boldsymbol{H}$

ii. The matrix $\boldsymbol{H}$ and $\boldsymbol{I} - \boldsymbol{H}$ are idempotent, that is, $\boldsymbol{HH} = \boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - \boldsymbol{H}$. (*Hint* $\boldsymbol{H} = X(X'X)^{-1}X'$)

b) Two explanatory variables are used to predict a dependent variable Price ($Y$), with independent variables length ($X_1$) and width ($X_2$). Write down a multiple linear regression model which can be used as a basis for the analysis, and define the meanings and properties of the terms in the model.

| Y | $X_1$ | $X_2$ |
|---|---|---|
| 14 | 120 | 60 |
| 20 | 120 | 80 |
| 37 | 120 | 100 |
| 36 | 120 | 120 |
| 31 | 150 | 75 |
| 42 | 150 | 100 |
| 54 | 150 | 125 |
| 64 | 150 | 150 |
| 38 | 180 | 90 |
| 66 | 180 | 120 |
| 64 | 180 | 150 |

| | | |
|---|---|---|
| 77 | 180 | 180 |
| 79 | 240 | 120 |
| 93 | 240 | 160 |
| 119 | 240 | 200 |
| 135 | 240 | 240 |

i. Plot scatter diagrams of price against each length and width. What do these graphs show?

ii. Fit a multiple linear model and interpret the estimated regression coefficients. *(Use matrix approach)*

iii. Can $H_0: [\beta_1, \beta_2] = 0$ be rejected with $\alpha = 0.05$? Use the ANOVA approach.

iv. Compute the coefficient of determination ($R^2$) and comment on it.

(100 marks)

4. A study was performed to investigate the shear strength of soil ($y$) related to depth in feet ($x_1$) and moisture content (in %) ($x_2$). Ten observations were collected, and the following summary quantities were obtained.

$\sum_{i=1}^{10} x_{1i} = 223$, $\sum_{i=1}^{10} x_{2i} = 553$, $\sum_{i=1}^{10} y_i = 1916$, $\sum_{i=1}^{10} x_{1i}{}^2 = 5200.9$,

$\sum_{i=1}^{10} x_{2i}{}^2 = 31729$, $\sum_{i=1}^{10} x_{1i}x_{2i} = 12352$, $\sum_{i=1}^{10} x_{1i}y_i = 43550.8$, $\sum_{i=1}^{10} x_{1i}x_{2i} =$

$12352$, $\sum_{i=1}^{10} x_{2i}y_i = 104736.8$, $\sum_{i=1}^{10} y_i^2 = 371595.6$

By considering multiple linear regression model, $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i; i = 1,2,\ldots,10$ for the above data:

i. Estimate the parameters $\alpha, \beta_1$ and $\beta_2$ in the model.

ii. What is the predicted shear strength of soil when $x_1 = 18$ feet and $x_2 = 43\%$?

iii. Compute the standard errors of the parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$.

iv. Find the 95% confidence interval for the parameters $\beta_1$ and $\beta_2$.

(100 marks)

**---END---**