



RAJARATA UNIVERSITY OF SRI LANKA
FACULTY OF APPLIED SCIENCES, MIHINTALE

B.Sc. (General Degree)
Third Year - Semester I Examination – September /October 2019
End Semester Examination

MAT 3203 – REGRESSION ANALYSIS

Time allowed: Two (02) hours

Answer all questions.

Calculators and statistical tables will be provided.

1. a) Consider the following regression model.

$$Y = \beta(X - \bar{X}) + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

Suppose that $(x_i, Y_i), i = 1, \dots, n$ is a data set from this model.

- i. Determine the least square estimator.

(15 Marks)

- ii. What is the distribution of least square estimator? (you have to give the mean and the variance of $\hat{\beta}$)

(30 Marks)

- b) In a study on the occurrence of sodium and chloride in a surface streams in Maldives, the following data for chloride concentration Y (in milligrams per liter) and roadway area in the watershed X (in percentage) were obtained.

| | | | | | | | | | | | | |
|-----|-----|-----|------|------|------|------|------|------|------|------|------|------|
| Y | 4.4 | 6.6 | 9.7 | 10.6 | 10.8 | 10.9 | 11.8 | 12.1 | 14.3 | 14.7 | 5.0 | 17.3 |
| X | .78 | .74 | 1.16 | 1.29 | 1.26 | 1.22 | 1.06 | 1.29 | 1.19 | 1.37 | 1.40 | 1.37 |

- i. Complete the following ANOVA table.

| Source | SS | DF | MSE | F- value |
|------------|---------|-------|-------|----------|
| Regression | 104.819 | | | |
| Error | | | | |
| Total | 143.737 | | | |

(35 Marks)

- ii. Write down the hypothesis that you would test using the ANOVA table and test this hypothesis at a significant level $\alpha = 0.05$ and write down your conclusion. **(15 Marks)**
- iii. Find the coefficient of determination and interpret your results. **(05 Marks)**

2. Given the observation.

| | | | | | | | | | | |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| x | 1.6 | 1.8 | 1.4 | 2.0 | 1.2 | 2.2 | 1.0 | 2.4 | 0.8 | 2.6 |
| y | 12 | 6 | 13 | 5 | 10 | 1 | 20 | 1 | 24 | 0 |

- a) Write down the model specification matrix X for the quadratic model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$. **(05 Marks)**

- b) If the matrix,

$$(X'X)^{-1} = \begin{pmatrix} 8.733 & -10.813 & 3.03 \\ -10.813 & 13.987 & -4.025 \\ 3.03 & -4.025 & 1.184 \end{pmatrix}$$

Find the estimates of β_0 , β_1 and β_2 . Write down the equation of the fitted regression model.

(15 Marks)

- c) Complete the ANOVA table

| Source | SS | df | MS | F |
|------------|--------|-------|-------|-------|
| Regression | 544.78 | | | |
| Residual | | | | |
| Total | 605.60 | | | |

(35 Marks)

- d) State the hypothesis that would test for the significance of the fitted regression model. Test the significance of the model at a significance level of $\alpha = 0.05$. **(15 Marks)**
- e) Write down the hypothesis that you would use to test whether the quadratic term is needed in the model and test this hypothesis at a significance level of $\alpha = 0.05$. **(20 Marks)**
- f) Use coefficient of determination to discuss the strength of the fitted model. **(10 Marks)**

3. a) Briefly explain the multiple linear regression.

(10 Marks)

- b) In an experiment to study factors influencing the wood specific gravity of slash pines, a sample of 20 mature wood samples were obtained, and measurements were taken on number of fibers/mm² in springwood (x_1), number of fibers/mm² in summerwood (x_2), % springwood (x_3), light absorption in springwood (x_4), and light absorption in summerwood (x_5).

Fitting the regression function $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_5x_5 + \epsilon$ resulted the $R^2 = 0.769$. (Use $\alpha = 0.01$ and $SST = 0.0196610$).

- i. Construct the ANOVA table.

(40 Marks)

- ii. State the hypothesis that you would test using the ANOVA table and test the hypothesis.

(10 Marks)

- iii. Write down your conclusions.

(05 Marks)

- iv. When x_1 , x_2 , and x_4 are all deleted, the resulting value of $R^2 = 0.654$. Does the data suggest that all of x_1 , x_2 , and x_4 have zero coefficient in the regression model? Test the hypothesis at $\alpha = 0.05$.

(35 Marks)

4. a) What is the difference between reduced model and full model?

(10 Marks)

- b) Write down the three procedures which are used in model selection and explain one of it.

(30 Marks)

- c) Briefly explain the uses of R^2 , R_{adj}^2 , MSE and C_p statistics in model selection procedure.

(20 Marks)

- d) For a multiple regression model with 5 independent variables, the following statistics for variable selection were obtained:

| Model | Variable | R^2 | R^2_{adj} | C_p | \sqrt{MSE} | 1 | 2 | 3 | 4 | 5 |
|-------|----------|-------|-------------|-------|--------------|---|---|---|---|---|
| A | 1 | 69.5 | 65.5 | 138.7 | 8.9639 | | | | X | |
| B | 1 | 68.6 | 63.6 | 142.5 | 9.0771 | | X | | | |
| C | 1 | 53.4 | 49.2 | 202.5 | 10.727 | X | | | | |
| D | 1 | 28.6 | 22.1 | 315.2 | 13.278 | | | X | | |
| E | 1 | 69.6 | 66.5 | 137.8 | 8.8832 | | | | | X |
| F | 2 | 97.9 | 97.4 | 2.7 | 5.4063 | X | X | | | |
| G | 2 | 98.2 | 97.7 | 2.5 | 2.7343 | X | | | X | |
| H | 2 | 93.5 | 92.2 | 22.4 | 4.1921 | | | X | X | |
| I | 2 | 84.7 | 81.6 | 62.4 | 6.4455 | | X | X | | |
| J | 2 | 68.0 | 61.1 | 138.2 | 9.3214 | | X | | X | |
| K | 2 | 78.98 | 85.87 | 3.6 | 5.945 | X | | | | X |
| L | 2 | 89.91 | 91.9 | 4.5 | 7.896 | | X | | | X |
| M | 3 | 98.2 | 97.6 | 3.0 | 2.3087 | X | X | | X | |
| N | 3 | 98.2 | 97.6 | 3.0 | 2.3121 | | X | X | X | |
| O | 3 | 98.1 | 97.5 | 3.5 | 2.3766 | X | | X | X | |
| P | 3 | 97.3 | 96.4 | 7.3 | 2.8638 | X | X | X | | |
| Q | 3 | 99.1 | 98.0 | 2.9 | 2.2456 | X | | X | | X |
| R | 3 | 94.7 | 92.1 | 5.3 | 3.345 | | X | | X | X |
| T | 4 | 98.2 | 97.4 | 5.0 | 2.4460 | X | X | X | X | |

- Select the best 1 variable, 2 variable and 3 variable models and justify your answer.
- Select also the best model among these models and justify your answer.

(40 Marks)

*** END ***