THE OPEN UNIVERSITY OF SRI LANKA

B.Sc. /B.Ed. Degree Programme, Continuing Education Programme

APPLIED MATHEMATICS-LEVEL 04

APU2141/APE4141- REGRESSION ANALYSIS I

FINAL EXAMINATION 2016/2017

Duration: Two Hours.

Date: 05.01.2018	**Time: 9.30am-11.30am**

Instructions:

- **This question paper consists of 06 questions. Answer only four questions.**
- **Statistical Tables are provided. When reading values, you may use the closest degrees of freedom given in the table.**
- **Where appropriate, consider that the regression models are fitted using the method of least squares.**
- **In all tests, use the significance level as 0.05.**
- **Non-programmable calculators are permitted.**

1. i) State whether each of the following statements is true or false.
   a) Pearson correlation coefficient close to 1 implies that the population regression relation is linear.
   b) Pearson correlation coefficient close to zero indicates that there is no relationship between the response and the predictor variable.
   c) Coefficient of determination is always less than the Pearson correlation coefficient.

   ii) The following summary statistics were computed from the systolic blood pressure (SBP), y and the age measured on a sample of 25 individuals. Ages of the persons in the sample had varied from 20 years to 65 years.

   $$\sum x_i = 1104; \sum y_i = 3223; \sum x_i^2 = 53272.0; \sum y_i^2 = 416469.54; \sum x_i y_i = 144102.35$$

   a) Compute the Pearson correlation coefficient between the SBP and age.
   b) Compute the coefficient of determination and explain what it measures in relation to this study.

2. In a study on modeling the association between the dried weight (milligrams) of a medicinal plant, y and the amount of fertilizer added (milligrams), x, a researcher fitted the model $y = \beta_0 + \beta_1 x + \varepsilon$ using the method of least squares. The regression sum of squares and the total sum of squares obtained were 5714.494 and 6029.454 respectively. The values of the predictor variable used for the data collection were 0, 1, 2, 3 and 4 with four replicates at each level.

   i) Compute $\sum (x_i - \bar{x})^2$.

   ii) Estimate the slope parameter.

   iii) If the total dried weight of the plants used for data collection was 909 milligrams, find the equation of the fitted model.

   iv) Based on the fitted model, estimate:

      a) The expected dried weight of a plant that had received 3 milligrams of the fertilizer.

      b) The likely age of a plant that had a dried weight of 20 milligrams

3. The following summary statistics were obtained from the heights (cm) of 25 children at ages in the range of 10 to 15 years. The researcher wants to fit the model $y = \beta_0 + \beta_1 x + \varepsilon$, where y denotes the height and x denotes the age (years).

$\sum x_i = 323; \sum y_i = 2539; \sum x_i^2 = 4245; \sum y_i^2 = 259091; \sum x_i y_i = 33049$

   i) Estimate the slope parameter.

   ii) Estimate the standard error of the estimate obtained in part (i).

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

   iv) Using the results obtained in part (iii) or otherwise, test the validity of the hypothesis that a one year change in age is associated with a 3cm change in the height.

4. A researcher measured the yield (grams) of a chemical reaction, y after adding known amounts (mg) of a catalyst, x to each sample. The following summary statistics were computed from the data on 24 samples.

Variable	Sample mean	Standard deviation
yield	19.94	5.057
catalyst	7.08	4.791

Pearson correlation between yield and catalyst = 0.781.

i) Construct an analysis of variance (ANOVA) table for fitting the model $y = \beta_0 + \beta_1 x + \varepsilon$.

ii) Using the ANOVA table constructed in part (i), test whether β_1 is zero or not. Clearly state the conclusions you make in relation to this study.

iii) Based on the ANOVA table, estimate the random variation in the yield of the chemical reaction.

5. State whether each of the following statements is true or false for fitting a linear regression model using least squares. In each case, give reasons for your answer.

a) Plot of residuals against the predictor variable and plot of residuals against the fitted values have the same pattern.

b) If there are extreme observations in the data, there should be large residuals.

c) Zero is not included in the 95% confidence interval for the slope parameter indicates that the predictor variable is not related with the response variable.

d) Sum of residuals can be used to test whether the random errors have mean zero.

e) If a plot of residuals against the fitted values is linear, constant variance assumption is violated.

f) The joint confidence level of the individual 95% confidence intervals for the slope and intercept parameters is less than or equal to 95%.

6. i) State whether each of the following statements is true or false for fitting two simple linear regression models for the same response variable with different predictor variables:

   a) The model with the larger value for the slope parameter indicates stronger relationship between the response variable and the corresponding predictor variable.

   b) The model with the smaller residual sum of squares gives better fits to the observed data.

ii) The percentage of defectives produced by a machine, y is suspected to be linearly associated with the work experience (months), X_1 and the length of training (months), X_2. The following information was obtained by fitting simple linear regression models to y.

Model1: $y = \beta_0 + \beta_1 X_1 + \varepsilon$

Estimate for the slope parameter = 3.053

regression sum of squares = 5773.342

residual sum of squares = 687.831

Model2: $y = \gamma_0 + \gamma_1 X_2 + \varepsilon$

Estimate for the slope parameter = 0.887

a) Construct an analysis of variance (ANOVA) table for fitting Model2.

b) Compute adjusted R^2 for fitting Model2 and explain what it measures in relation to this study.