THE OPEN UNIVERSITY OF SRI LANKA

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

APPLIED MATHEMATICS-LEVEL 05/04

ADU5301/APU2141- REGRESSION ANALYSIS 1

FINAL EXAMINATION 2019/2020

Duration: Two Hours.

Date: 05.02.2021 **Time: 9.30am – 11.30am**

Instructions:

- **This question paper consists of 06 questions. Answer only four questions.**
- **Statistical Tables are attached. 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. A researcher interested in quantifying the linear association between the experience of machine operator (measured in months) and the amount of wastage of raw material during operation (measured in grams) collected the data presented in the table.

Experience (months)	0	0	4	4	6	6	8	8	12	12
Wastage (grams)	13	11	11	8	7	5	6	4	3	2

i) If you were to fit a simple linear regression model to the data, which variable would you choose as the response variable? Give reasons for your choice.

ii) A student stated that unless there is a strong linear association between the two variables, a simple linear regression model cannot provide a good fit to the data. Do you agree with this statement? Give reasons for your answer.

iii) Suggest a suitable measure that can be calculated from the data to assess the strength of the linear association between the given variables.

iv) Calculate the value of the measure proposed in part (iii) for this data and clearly describe what you conclude about the association between the variables based on this value.

v) Let k denote the value of the measure obtained in part (iv). State whether each of the following statement is true or false. In each case, give reasons for your answer.

a) If the data collected on experience were converted to weeks and the measure proposed in part (iii) was recalculated after this conversion, the new value will be equal to k,

b) It was later realized that, due to an error in the scale used to measure the wastage, each measurement of wastage must be corrected by adding 0.02. If the measure proposed in part (iii) was re-calculated after correcting for this error, the new value will be equal to $k + 0.02$.

2. A researcher fitted a simple linear regression model to the time (minutes) required to dissolve a chemical compound in an acidic medium, using temperature (^0C) of the sample as the predictor variable. During data collection, temperatures of the samples were controlled at 10^0C, 20^0C, 30^0C, 40^0C and 50^0C and three replicates were collected at each of these temperature levels. Accidentally, the researcher had lost the raw data, but the following information obtained from the least squares fit of the model $y = \beta_0 + \beta_1 x + \in$ were available.

Regression sum of squares = 21385.8

Total sum of squares = 23192.3

i) What is the sample size, n used in this study?

i) Calculate $\sum(x_i - \bar{x})^2$, where x_i $(i = 1,2,\cdots,n)$ denote the temperature of the i^{th} sample.

ii) The researcher is aware that when the temperature is increased, the chemical compound dissolves fast. Based on the given information, estimate β_1 and explain what it measures in relation to this study.

iii) Calculate the residual sum of squares and the mean squared error.

iv) Construct a 95% confidence interval for β_1.

v) Using the results obtained in part (iv) or otherwise, test the validity of the hypothesis that a 5^0C increase in temperature is associated with a 12 minutes reduction in the time required to dissolve the chemical compound. Clearly state your findings.

3. The following summary statistics were computed from the wing length (cm) and tail length (cm) of 28 adult birds of a particular species. The researcher wants to fit a simple linear regression model to the wing length, using tail length as the predictor variable. The tail length of birds in the sample had varied from 6.0cm to 7.9cm.

Variable	Descriptive Statistic	
	Mean	Standard deviation
Wing length (y)	10.83	0.59
Tail length (x)	6.72	0.55

Pearson correlation between wing length and tail length = 0.90

i) Find the least squares estimates for the slope and the intercept of the fitted line in the proposed simple linear regression model fit.

ii) Write down the equation of the fitted line.

iii) Find the fitted value for the wing length of a bird with a tail length of 7.1cm and explain what it measures, in relation to this study.

iv) Estimate the mean difference in the wing lengths of two birds with 0.5cm difference in the tail lengths.

v) From the least squares model fit, estimates for the standard errors of the slope and intercept parameters were found to be 0.093 and 0.629 respectively. Estimate the standard error of the estimate computed in part (iv).

4. A researcher recorded the increase in plant height four weeks after weekly applying different known amounts of bio char to 25 medicinal plants with similar initial height. The following summary statistics were computed from the data.

$\sum x_i = 37.5,\ \sum y_i = 234.85,\ \sum x_i^2 = 76.6,\ \sum y_i^2 = 2522.95,\ \sum x_i y_i = 421.53.$

The researcher wants to test whether the bio char has a significant effect on the increase in height of the medicinal plant based on the findings from fitting the model $y = \beta_0 + \beta_1 x + \in$; weekly amount of bio char applied is taken as the predictor variable.

i) Write down the null and the alternative hypotheses you would test to address the researcher's objectives. Clearly describe the notation you use.

ii) Compute the least squares estimate for the slope parameter.

iii) Construct an analysis of variance (ANOVA) table that can be used to address the researcher's objectives.

iv) Using the ANOVA table constructed in part (iii), test the hypothesis stated in part (i) at a 0.05 significance level. Clearly state the findings.

5. In a study on the effect of protein on the weight gain of rats, a researcher measured the weight gain (mg) of 34 rats, three months after giving a diet with known amounts (mg) of protein, daily. The protein contents in the daily diets were in the range from 0 mg to 5mg. The model $y = \beta_0 + \beta_1 x + \in$ was fitted to the data using the method of least squares with weight gain as the response variable and the protein content in the diet as the predictor variable. The following table gives some parts of the output from the model fit.

Parameter	Estimate	Standard error
β_0	1.52	0.14
β_1	0.31	0.05

Residual sum of squares = 6.05

i) In relation to this study, clearly state the assumptions that the researcher must make in choosing the model and obtaining valid estimates for the parameters given in the table.

ii) Estimate the random variation in the weight gains of rats.

iii) A randomly chosen rat that had received a weekly diet containing 3mg of protein content had a weight gain of 4mg. Calculate the residual corresponding to this observation and clearly explain what it measures in relation to this study.

iv) Clearly explain what is measured by the standard error of the estimator for the slope parameter with estimated value 0.05, in the given table.

v) State whether each of the following statements is true or false for a least squares fit of the model $y = \beta_0 + \beta_1 x + \in$. In each case, give reasons for your answer.

a) There is no random error in an observation that has residual equal to zero.

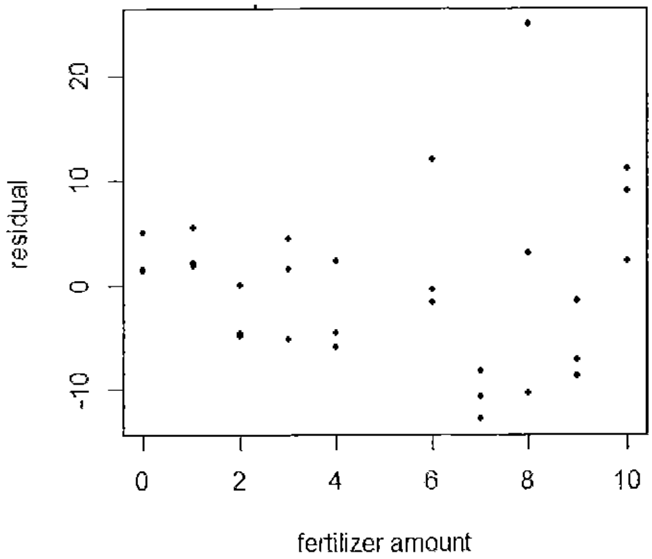
b) Sum of the residuals obtained from fitting the model to a set of data will always be equal to zero.

6. i) State whether fitting the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$ using the method of least squares for the response variable y with x as the predictor variable will be appropriate to meet the researcher's objective in each study given below. Give reasons for your answer.

   a) A researcher wants to estimate the age at which males stop increase in height, based on the heights measured on males of ages 5 to 30 years, choosing age as the predictor variable.

   b) A researcher wants to predict the final examination mark (y) of a student with a midterm mark of 10, based on the fitted model to the final examination marks of 40 students with midterm marks above 30.

   ii) The following figure illustrates a plot of residuals against the predictor variable obtained from fitting the model $y = \beta_0 + \beta_1 x + \epsilon$ using the method of least squares, to the data collected on the yield (y) of a tomato species with the amount of fertilizer added (x) as the predictor variable.



State whether each of the following statements is true or false according to the above plot. In each case, give reasons for your answer.

a) For small values of the fertilizer amount, the fitted model overestimates the yield.

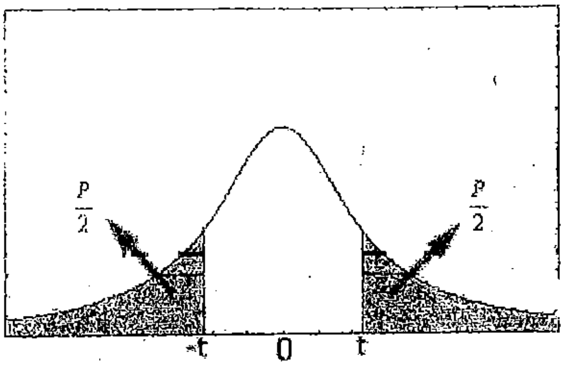
b) A plot of residuals against the fitted values will have a similar pattern as in the given plot.

c) Relying on the validity of the fitted model, we can conclude that the random errors do not have constant variance.

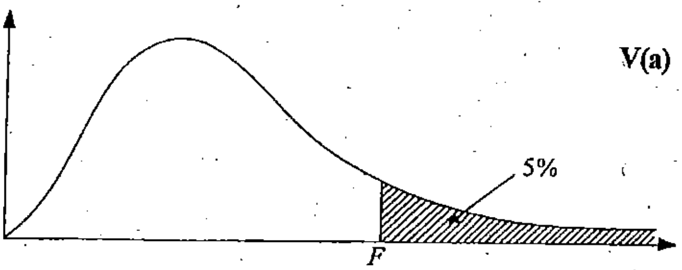
d) Since there are similar number of positive and negative residuals, the random errors will have a normal distribution.

e) A plot of fitted values against the predictor variable will lie along a straight line.

Table A2: Student's t - Distribution



P	50	20	10	5	2	1	0.2	0.1
Degrees of freedom								
1	1.00	3.08	6.31	12.7	31.8	63.7	318	637
2	0.82	1.89	2.92	4.30	6.96	9.92	22.3	31.6
3	0.76	1.64	2.35	3.18	4.54	5.84	10.2	12.9
4	0.74	1.53	2.13	2.78	3.75	4.60	7.17	8.61
5	0.73	1.48	2.02	2.57	3.36	4.03	5.89	6.87
6	0.72	1.44	1.94	2.45	3.14	3.71	5.21	5.96
7	0.71	1.42	1.89	2.36	3.00	3.50	4.79	5.41
8	0.71	1.40	1.86	2.31	2.90	3.36	4.50	5.04
9	0.70	1.38	1.83	2.26	2.82	3.25	4.30	4.78
10	0.70	1.37	1.81	2.23	2.76	3.17	4.14	4.59
12	0.70	1.36	1.78	2.18	2.68	3.05	3.93	4.32
15	0.69	1.34	1.75	2.13	2.60	2.95	3.73	4.07
20	0.69	1.32	1.72	2.09	2.53	2.85	3.55	3.85
24	0.68	1.32	1.71	2.06	2.49	2.80	3.47	3.75
30	0.68	1.31	1.70	2.04	2.46	2.75	3.39	3.65
40	0.68	1.30	1.68	2.02	2.42	2.70	3.31	3.55
60	0.68	1.30	1.67	2.00	2.39	2.66	3.23	3.46
∞	0.67	1.28	1.64	1.96	2.33	2.58	3.09	3.29



V(a)

5 ප්‍රතිශතයට අනුරූප F ව්‍යාප්තිය
F பரம்பலின் 5 சதவீத புள்ளிகள்
5 percent points of the F distribution

	$n_1 =$ 1	2	3	4	5	6	7	8	10	12	24
$n_2 =$ 2	18.5	19.0	19.20	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.64
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.77
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.53
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.84
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.41
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.12
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	2.90
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.74
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.51
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.29
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.08
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	1.98
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	1.89
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.08	2.00	1.79
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.70