117

THE OPEN UNIVERSITY OF SRI LANKA

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

APPLIED MATHEMATICS-LEVEL 05/04

ADU5301- REGRESSION ANALYSIS 1

FINAL EXAMINATION 2020/2021

Duration: Two Hours.

Date: 29.03.2022	**Time: 9.30am – 11.30am**

Instructions:

- **This question paper consists of 06 questions in 07 pages.**
- **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. The diesel consumption (in litres) of a certain machine, y, and the number of hours the machine was operated (x) in each of 10 days from 1st of March 2022 to 10th of March 2022 are given in the table.

x (hours)	4	4	4	8	8	8	8	12	12	12
y (litres)	2.5	2.8	2.4	3.1	3.2	3.4	3.1	5.4	5.2	5.6

Suppose that the simple linear regression model $y = \beta_0 + \beta_1 x + \in$ is appropriate for data collected on y throughout the year 2022.

i) Clearly describe what the term $\in$ represents and state the assumptions one must make to fit the model using the method of least squares.

ii) Obtain the least squares estimate for the slope parameter in the above model and explain what it measures in relation to this study.

iii) Obtain the equation of the fitted line.

iv) Suppose that 3.5 litres of diesel in stock is available for use in the machine on the 30th of March 2022. Assume that the machine can be operated only in shifts of length 2 hours. Estimate the maximum number of shifts that the machine can operate on the 30th of March 2022, consuming diesel from the allocated stock for that day.

v) If the machine is operated as the number of shifts calculated in part (iv), calculate the amount of diesel that will remain unused.

vi) A student stated that the amount of unused diesel calculated in part (v) represents the residual corresponding to the diesel usage on 30th of March 2022. Comment on the statement made by the student.

2) i) State whether the objectives of each study given below can be addressed by fitting a simple linear regression model using the method of least squares If the objectives can be addressed, state which variable you choose as the response variable, giving reasons for your choice. If you disagree, give reasons.

a) An experimenter recorded the photon counts emitted and the corresponding temperature while heating a sediment sample over a range of temperatures. The researcher is aware that a linear regression function is suitable to describe the relationship between the variables. The researcher has observed that at high temperatures photons emit fast, giving large errors in the photon counts compared to those recorded at low temperatures. The objective is to develop a model to predict the photon count at a given temperature.

b) An experimenter recorded the amounts of fertilizer applied and the dried weights at the age of six weeks, of 80 plants sampled from a large cultivation. The experimenter is aware that a regression function of the form $\beta_0 + \beta_1 x^3$is appropriate. The objective is to develop a model to predict the dried weight of the plant if a known amount of fertilizer is applied, assuming that the assumptions required are valid.

c) An experimenter wants to predict the height (inches) of his son at the age of 20 years based on the heights of his son measured at the end of each month during the period that the son is10 to 15 years old. The experimenter has found that a regression function of the form $\beta_0 + \beta_1 e^x$is suitable to describe the relationship between heights (y) and ages (x) and the assumptions required to fit the model are valid.

ii) The following summary statistics were computed from the times taken for a chemical reaction to complete (minutes), y and the amount of catalyst added (mg), x, to 30 chemical samples.

$$\sum x = 316, \sum y = 552, \sum x^2 = 4878, \sum y^2 = 11234, \sum xy = 5584.$$

Calculate the Pearson correlation coefficient and clearly explain what it measures in relation to this study.

Based on the value of the Pearson correlation coefficient, a student concluded that the catalyst added has no effect on the completion time of the chemical reaction. Do you agree with the conclusion made by the student? Give reasons for your answer.

3) A researcher fitted the model $y = \beta_0 + \beta_1 x + \in$, using the method of least squares, for the data collected on 24 adult birds from a population, where y and x denote the length (cm) and the neck circumference (cm) of the bird. Summary statistics calculated from the data collected on the two variables and part of the output obtained from the model fit are given below.

Variable	Sample mean	Standard deviation
Length	13.26	0.41
Neck circumference	7.01	0.18

Estimate for the slope parameter was 1.89 with a standard error of 0.27.

i) Write down the equation of the fitted model.

ii) Clearly explain what is measured by the value of the standard error indicated in the given output.

iii) Estimate the residual sum of squares.

iv) Estimate the random variation in the length of a randomly chosen bird from this population.

v) Based on the analysis, a student stated that the length of the bird and the neck circumference has a positive linear association. Determine the validity of the statement made by the student.

4) An experimenter measured the yields (grams), y from each of 30 experimental plots of a cultivation, four months after adding known amounts of a fertilizer, x. The fertilizer amounts used for the study were 0, 2, 4, 6, 8 and 10 grams per plot with 5 replicates for each level of fertilizer. Part of the analysis of variance table obtained from the least squares fit of the model $y = \beta_0 + \beta_1 x + \in$ is as follows:

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	1860.711	***	***
Residual	28	304.2552	***	
Total	29	2164.967		

i) Calculate the coefficient of determination for the model fit and clearly explain what it measures in relation to this study.

ii) Complete the ANOVA table by filling the missing values indicated with "***".

iii) Using critical values from the F-table (see Appendix) and a 5% significance level, test whether the predictor variable significantly helps to predict the variation in the response variable and clearly state your findings.

iv) State the assumptions needed to do the analysis carried out in part (ii).

v) Describe two plots that the researcher should examine to test the validity of assumptions mentioned in part (iii).

5) A researcher recorded the wastage of raw material per item produced (grams), y for different production rates (measured as number of items per hour), x. The following summary statistics were computed from the data collected on a sample of size 32.

$$\sum x = 378, \sum y = 182.5, \sum x^2 = 5376, \sum y^2 = 1202.89, \sum xy = 2496.4.$$

The researcher wants to test whether the production rate has a significant effect on the wastage of raw material (per item), based on the findings from least squares fit of the model $y = \beta_0 + \beta_1 x + \in$.

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.

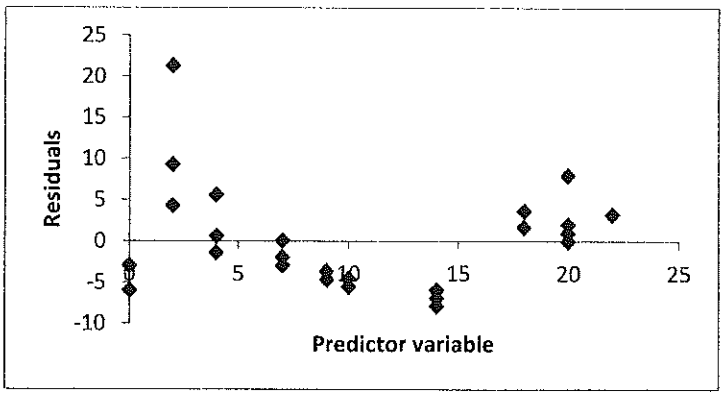
6). i) State whether each of the following statements is true or false for least squares fit of the simple linear regression model $y = \beta_0 + \beta_1 x + \in$ for the response variable y with x as the predictor variable. In each case, give reasons for your answer.

a) The residuals obtained from the model fit will always sum to zero.

b) Random error in an observation that falls on the fitted regression line is zero.

c) If the fitted values of two observations are equal, the said observations must be replicates.

ii) The following figure illustrates a plot of residuals against the predictor variable obtained from fitting the model $y = \beta_0 + \beta_1 x + \in$ 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.

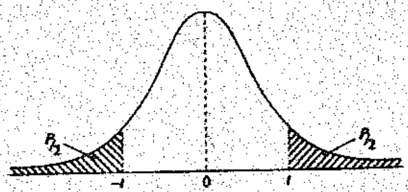
Appendix I

III

ස්ටුඩන්ට් t ව්‍යාප්තිය
නිදහස් අංශ වැඩිවන විට පටු ස්වරූපයක් ගනී.

மாணாக்கர்களின் t பரம்பல்
சுயாதீனப்படி அதிகரிக்கையில் ஒடுங்கும்

The Student's t Distribution
Narrows with increasing degrees of freedom ($d.f.$)



t ව්‍යාප්තියේ P නම් ප්‍රතිශතයක් $-t$ හා $+t$ පරාසයට පිටතින් පිහිටීමට අවශ්‍ය t හි අගයන් විවිධ නිදහස් අංශවලට සඳහා වගුගත කර ඇත.

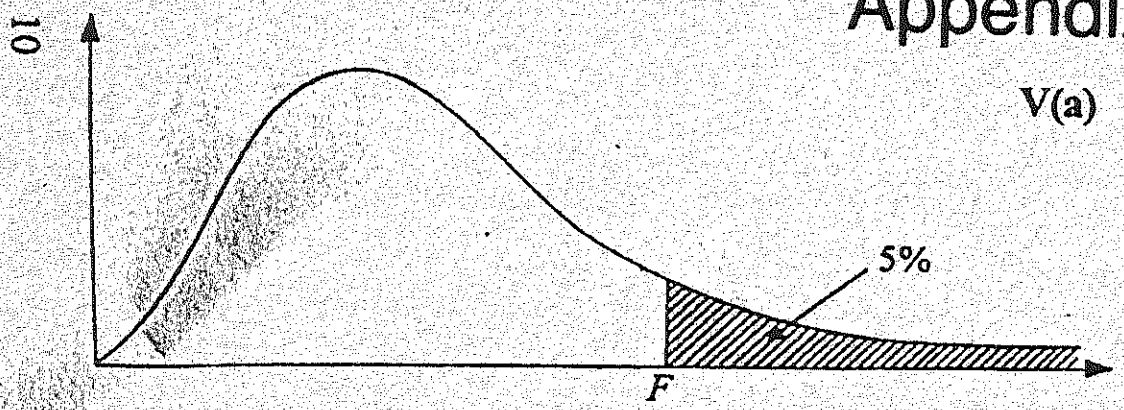
இவ் அட்டவணையானது மாணாக்கர்களின் பரம்பலின் வீச்சு $-t$ க்கும் $+t$ க்கும் வெளியே உள்ள, ஒரு குறிப்பிட்ட நூற்றுவீதம் P க்குரிய t யின் பெறுமதிகளைத் தருகின்றது. இவ் t பெறுமதிகள் வேறுபட்ட சுயாதீனப்படிகளுக்கு அட்டவணைப்படுத்தப்பட்டுள்ளன.

This table gives the values of t for which a particular percentage, P, of the Student's t distribution lies outside the range $-t$ to $+t$. These values of t are tabulated for various degrees of freedom.

P / නිදහස් අංශ / சுயாதீனப்படி / Degrees of freedom	50%	20%	10%	5%	2%	1%	0.2%	0.1%
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

Appendix II

V(a)



5 ප්‍රතිශතයට අනුරූප F ව්‍යාප්තිය

F பரம்பலின் 5 சதவீத புள்ளிகள்

5 percent points of the F distribution

n_2 \ n_1 =	1	2	3	4	5	6	7	8	10	12	24
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