THE OPEN UNIVERSITY OF SRI LANKA

B. Sc. Degree Programme, Continuing Education Programme

APPLIED MATHEMATICS-LEVEL05

ADU5301- REGRESSION ANALYSIS I

FINAL EXAMINATION 2024/2025

**Duration: Two Hours.**

| **Date: 18.05.2025** | **Time: 9.30 a.m- 11.30 a.m** |
|---|---|

**Answer FOUR questions only.**

**Instructions:**

- **This question paper consists of 06 questions. Answer only four questions.**
- **Statistical Tables are attached at the end of the question paper. When reading values, you may use the closest degrees of freedom given in the table.**
- **Where appropriate, method of least squares will be used to fit the regression model.**
- **Non-programmable calculators are permitted.**

1. An election analyst is interested in examining the relationship between the change in the number of campaign events ($x$) by the candidate and the percentage change in votes received by the candidate ($y$), by comparing two general elections (referred to as current and previous). The percentage change in votes is measured to one decimal place. Changes in both variables may be positive or negative. Following data summaries were computed from the data collected on 30 candidates for whom records from both the current and the previous election were available.

$$n = 30,\ \sum x_i = 60,\ \sum y_i = 184.0,\ \sum x_i^2 = 180.0,\ \sum y_i^2 = 1382.56,\ \sum x_i y_i = 533.85$$

i) Apply the method of least squares to fit a simple linear regression model to the percentage change in votes, using the change in campaign events as the predictor variable. Write down the equation of the fitted regression line.

ii) Mr. P had the same number of campaign events as in the previous election but received a 1% increase in votes in the current election. Calculate the deviation of the percentage change in votes for Mr. P in the current election from the value predicted by the fitted model.

iii) Explain what is meant by the random error in the observable percentage in votes of Mr. $P$ and estimate it. In relation to this study, what does the random error represent?

iv) Clearly state all the assumptions made on the random error when computing least squares estimates for a simple linear regression model for the percentage change in votes.

2. i) Briefly describe why it is important to examine a scatter plot of response against the predictor variable prior to interpreting the value of Pearson correlation coefficient.

ii) Pearson correlation coefficient based on observations collected on daily hours spent on social media ($X$) and score received for a certain vocabulary test ($Y$), on a sample of 30 students in the age group of 10 to 12 years is -0.02. Furthermore, the standard deviations of the vocabulary test scores and the daily hours spent on social media were 4 and 14 hours respectively. State whether you agree or disagree with each of the following conclusions made by a student based on the given values. In each case, give reasons for your answer.

a) Data indicates that the performance on the vocabulary test of students in the said group is not related to the time spent on social media.

b) Daily hours spent on social media is not a good predictor of the said vocabulary test score.

c) Around 4% of the variation in the vocabulary test score can be explained by a simple linear regression model with daily hours spent on social media as the predictor variable.

d) As more time is spent on social media, the performance of the vocabulary test will drop.

e) Provided that the assumptions needed for a least squares fit of a simple linear regression model with time spent on the social media as the predictor variable predicts that associated with one hour of increase in the daily time spent on the social media, the vocabulary test score will be dropped by 0.006 units.

3. In a study to examine whether daily sugar intake is a significant predictor of body mass index (BMI), the average daily sugar intake (to the nearest gram) and BMI ($kg/m^2$) to the nearest single decimal point, were recorded on a random sample of 40 individuals. Following summary statistics were calculated from the sugar intake of persons in the sample ($x$):

minimum = 58; maximum = 125; $\sum x_i = 3660.0$; $\sum x_i^2 = 350656$

Assume that the data satisfy the required assumptions for a least squares fit of a simple linear regression model. The following information was extracted from the least squares fit of a simple linear regression model to the BMI, using daily sugar intake as the predictor variable.

Intercept = -38.532; Slope = 1.116

Mean squared error = 89.5522

i) Clearly explain what the slope parameter measures, in relation to this study.

ii) The estimated intercept from the fitted model is negative. How would you explain this result to a student, based on the data collected in this study?

iii) Clearly explain what the mean squared error measures, in relation to this study.

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

v) Using a 5% significance level, test whether the daily sugar intake is a significant predictor of body mass index or not. Clearly state your findings.

4. In a study on how the price of a used car depends on its age and mileage, a researcher collected the selling price (in rupees), age of the car (in years), and mileage (in kilometers) of a random sample of 40 used cars of the same make and model, from a used car seller.

a) If the researcher decided to choose a univariate multiple linear regression model, advise the researcher to choose the response and explanatory variables. Give reasons for your choice of variables.

b) If the researcher decided to choose only the data collected on the age and the selling price and to fit a linear regression function, selecting the variables appropriately, write down the model equation. You need to clearly describe the notation you use.

c) In relation to this study, describe what the slope parameter in your regression function in the model described in part (b) measures.

d) Give one advantage and one disadvantage of using the model described in part (a) over to selecting the model described in part (b).

e) If both models are fitted using least squares, which of the two models have smaller residual sum of squares. Give reasons for your answer.

f) Suppose the researcher computed the ordinary residuals based on the least squares fit of the model described in part (b). Describe all the residual plots that you would advise the researcher to construct, and clearly describe the use of each plot.

5. To study the relationship between daily water consumption (in liters), $x$, and hours spent under the sun (in hours), $y$, among athletes, a researcher recorded the number of hours spent in the playground and the water intake from 26 athletes on a given day. The following summary statistics were computed from the data collected.

$$n = 26, \sum x_i = 123.0, \sum x_i^2 = 697.0, \sum y_i = 90.0; \sum y_i^2 = 336.56; \sum x_i y_i = 478.9$$

A simple linear regression model is to be fitted to the water intake, using the method of least squares, taking hours spent in the playground as the predictor variable.

Part of the Analysis of variance (ANOVA) table obtained by fitting a simple linear regression model for $y$ using $x$ as the predictor variable is given below.

| Source of variation | Sum of squares | Degrees of freedom |
|---|---|---|
| Regression | (a) | (d) |
| Residual | (b) | (e) |
| Total | (c) | (f) |

i) Compute the missing values indicated by (a), (b), (c), (d), (e) and (f) in the given table.

ii) Calculate the proportion of variation in the water intake that can be explained using the hours spent in the playground.

iii) Using a 5% significance level, test whether the predictor variable significantly contributes to predict the variation in the response variable. Clearly state your findings.

6. In a calibration study, a researcher added different weights ranging from 2 grams to 5 grams and measured the reading of the weight of a spring balance. For each weight, 4 replicate readings were collected. The researcher only measured these two variables.

i) Define each of the following terms and describe each of them, in relation to this study:

a) Mean response

b) Residual

c) Regression function

ii) A plot of ordinary least squares residuals by fitting a simple linear regression model with additive errors to the readings from the spring balance with weight added as the predictor variable is given below. A student made each of the following conclusions based on this plot. State whether you agree with this student or not in each case, giving reasons for your answer.

a) The observations are correlated.

b) The plot indicates that the model fits better for observations taken by adding low weights to the spring balance.

c) The random errors in the response have zero mean.

d) A quadratic term in the predictor variable needs to be added to the chosen regression function.

e) The model gives equal predicted values for several observations.

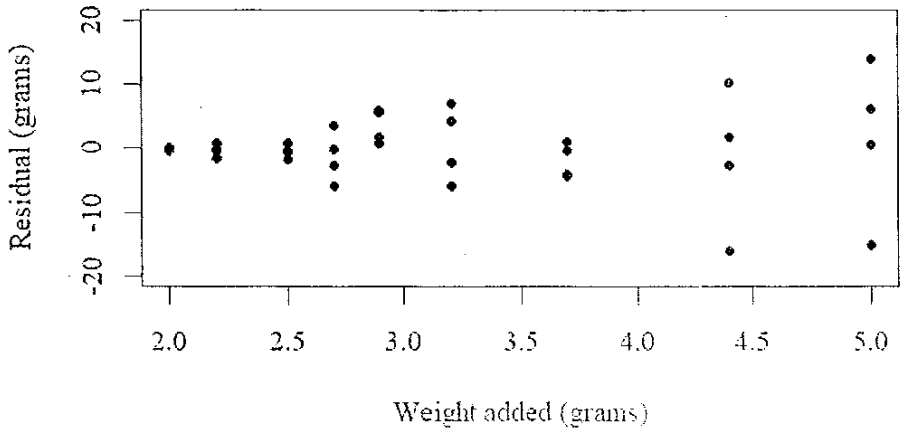
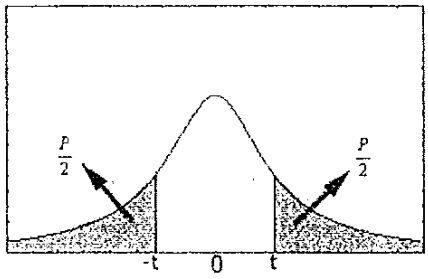


Table A2: Student's t - Distribution




| 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 |

7.

# *F* ව්‍යාප්තිය
# *F* பரம்பல்
# The *F* Distribution

*F* ව්‍යාප්තියේ යම් *P* ප්‍රතිශතයක් *F* ට පිටතින් පිහිටීමට අවශ්‍ය *F* හි අගය මෙම වගුවෙන් ලබාගත හැක.

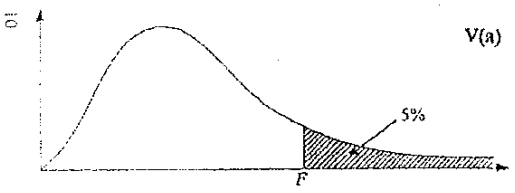
විවලතාවක් සඳහා වූ ස්වායත්ත ඇස්තමේන්තු දෙකක අනුපාතයෙන් *F* ව්‍යාප්තිය ලැබේ. මෙම එක් එක් ඇස්තමේන්තුවකට අනුරූප සුවිලත අංක ඇති බැවින්, භාවිතා කළ යුත්තේ කුමන *F* ව්‍යාප්තිය දැයි දන ගැනීමට, ලවයේ සුවිලත අංකය වන $n_1$ ද, හරයේ සුවිලත අංකය වන $n_2$ ද දිය යුතුය.

இவ் அட்டவணையானது, *F* பரம்பலின் ஒரு தரப்பட்ட நூற்றுவீதத்திலும் கூடிய, *F* பெறுமதிகளைத் தருகின்றது.

*F* பரம்பலானது, மாறுந்திறன் ஒன்றின் இரு சார மதிப்பீடுகளை ஒன்றை மற்றதால் பிரிக்க உண்டாகும். இச் சார மதிப்பீடுகள் ஒவ்வொன்றிற்கும் சுயாதீனப்படி இருப்பதால் எந்த *F* பரம்பலானது கருதப்படல் வேண்டுமெனக் குறிப்பிட, தொகுதியின் சுயாதீனப்படி $n_1$, பகுதியின் சுயாதீனப்படி $n_2$, இரண்டும் தரப்படல் வேண்டும்.

These tables give the values of *F* for which a given percentage of the *F*-distribution is greater than *F*.

The *F*-distribution arises when two independent estimates of a variance are divided one by the other. Each of these estimates has its degrees of freedom associated with it, thus to specify which particular *F*-distribution is to be considered, the degrees of freedom of both the numerator $n_1$, and the denominator $n_2$, must be given.




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 |