University of Kelaniya – Sri Lanka
Centre for Distance and Continuing Education

Bachelor of Science (External) Third Year Second Semester - 2019
2025 – March
Faculty of Science

Statistics
STAT 37543 - Statistical Models

No. of Questions: Four (04) **No. of Pages: Four (04)** **Duration: Two & Half ($2\frac{1}{2}$)**

Answer all the questions

(Non-programmable calculators are allowed)

Note: Keep the final answer in three decimal places. Consider $\alpha = 0.05$ if it is not specified.

1. A linear model for pairs of observations $\{(x_i\, y_i): i = 1,2,\ldots,n)\}$ is given by $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1,2,\ldots,n$. Here β_0, β_1 and ϵ_i are in usual notations. ϵ_i are independent and $\epsilon_i \sim N(0, \sigma^2)$.

   (a) Derive the least square estimators of β_1, and β_0.

   (b) Show that $Var(\widehat{\beta_1}) = \frac{\sigma^2}{S_{XX}}$.

   (c) Write down the formula for 95% confidence interval for β_1.

   (d) The administration of a certain university aims to examine the relationship between students' entrance test scores and their GPA at the end of their freshman year. The table below presents the entrance test scores and corresponding GPA for a random selection of eight students.

Table 1:Test scores and corresponding GPA of students

Student No	Test Score	GPA
1	55	3.0
2	63	3.2
3	37	1.9
4	65	3.6
5	72	3.8
6	59	3.1
7	48	2.6
8	41	2.2

   (i). Define the independent and dependent variables.

   (ii). Draw a scatter plot and examine whether there is a relationship between entrance test score and GPA.

   (iii). Estimate the parameters and fit a simple linear regression model to estimate the GPA at the end of the freshmen year.

   (iv). Find 95% confidence intervals for β_1.

2. *a)* The table below displays the prices of houses (Y) in thousands of dollars, household assets (X_1) in thousands of dollars, and floor areas (X_2) in square feet for a sample of houses currently listed on the market.

Table 2: House Price, Household Assets, and Floor Area Data for market listings

Y	8.6	10.3	9.4	13	10.6	8.1
X_1	74	102	87	142	110	85
X_2	100	110	90	130	90	80

(i). Write down the formula in matrix notation to obtain the $\beta's$.

(ii). Find the missing entries in the following matrices.

$$X^TX = \begin{bmatrix} \_\_\_\_ & \overline{62938} & \overline{61610} \\ \_\_\_\_ & \_\_\_\_ & 61600 \end{bmatrix} \qquad X^TY = \begin{bmatrix} \overline{6205.3} \\ 6131 \end{bmatrix}$$

(iii). Fit a multiple linear regression model using the data from Table 2 to estimate the relationship between house prices (dependent variable) and the independent variables: household assets and floor area.

(b) Using the fitted model using R software is shown below.

(i). Comment on overall significance of the fitted model.

(ii). Identify the significant variables of the above model.

(iii). Write the updated model using only the significant variables.

```
Residuals:
       1        2        3        4        5        6
 0.04962 -0.06923  0.38253 -0.11485  0.30017 -0.54824

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)          1.09196    1.692   0.1893
X1                   0.01187    4.696   0.0183
X2                   0.01609    1.602   0.2075

Residual standard error: 0.4311 on 3 degrees of
freedom Multiple R-squared:  0.9638,    Adjusted R-
squared:  0.9396
F-statistic: 39.89 on 2 and 3 DF,  p-value: 0.0069
```

(c) Fill in the missing standardized residuals and leverage values (h_i), then utilize the provided theoretical values and your understanding of residual analysis to create a QQ plot for the previously mentioned fitted model. Finally, provide an interpretation of the plot. (MSE=0.2585, $\sum (X_i - \bar{X})^2 = 2938$)

Table 3: Data for QQ plot construction

Y_i	8.6	10.3	9.4	13	10.6	8.1
h_i	0.3967	0.2241	0.2241	____	0.2007	0.2432
Standardized Residuals	1.0554	____	0.6886	0.2654	-0.2172	-1.9258

Theoretical Quantiles	-1.0675	-0.5659	-0.1800	0.1799	0.5659	1.0675

(i) Complete the missing entries of the Table 3 for the fitted model for house prices (Y).

(ii) Using the given theoretical values and applying the knowledge of residual analysis, construct a QQ plot for the fitted model and interpret the resulting plot.

3.

(a) Define the following terms used in experimental design:

(i). Replication,

(ii). Randomization.

(b) What are the assumptions used in the completely randomized design (CRD)?

(c) An economist aims to examine whether the mean housing prices differ across three areas with varying levels of air pollution. To conduct the analysis a completely design is used, and a random sample of house prices from each of the three pollution levels is provided in Table 4.

Table 4: A random sample of house prices

Observation	Pollution Levels		
	Low	Moderate	High
1	120	61	40
2	68	59	55
3	40	110	73
4	95	75	45
5	83	80	64

(i). Write the appropriate linear model for this problem.

(ii). Complete the ANOVA table for this experiment and state your conclusion at 5% level of significance.

(iii). Based on the ANOVA table, what can you say about the variability between the pollution levels and variability within the pollution levels.

4.

(a) State two advantages and two disadvantages of Randomized Completely Block Design (RCBD)

(b) State the mathematical model for the randomized completely Block Design (RCBD) with usual notations.

(c) An agronomist aims to evaluate the effects of four different fertilizers on the growth of a specific type of plant. To account for potential variability among plants, the agronomist selects five fields of plants and applies all four fertilizers in a randomized order within each field. The observed plant growth results are provided in the Table 5.

Table 5: The observed plant growth results

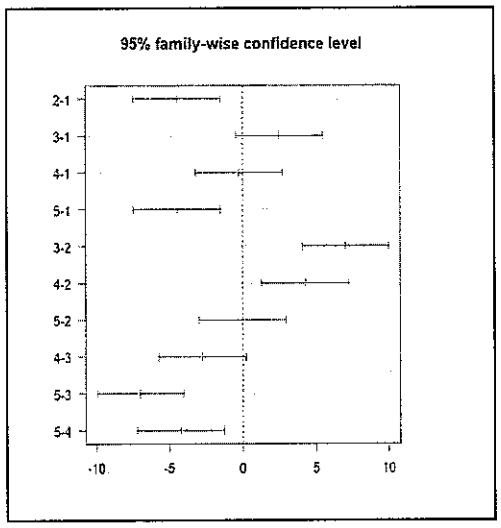
Fertilizer	Field of Plants						
	1	2	3	4	5	Total ($y_{i\cdot}$)	Mean ($\overline{y_{i\cdot}}$)
1	73	68	74	71	67	353	70.6
2	73	67	75	72	70	357	89.25
3	73	68	78	73	68	360	90
4	73	71	75	75	69	363	72.6
Total ($y_{\cdot j}$)	292	274	302	291	274	$y_{\cdot\cdot} = 1433$	
Mean ($\overline{y_{\cdot j}}$)	73	68.5	75.5	72.75	68.5		

(i). State the hypothesis to be tested to achieve the objectives of the agronomist.

(ii). Analyze the data and give an appropriate conclusion.

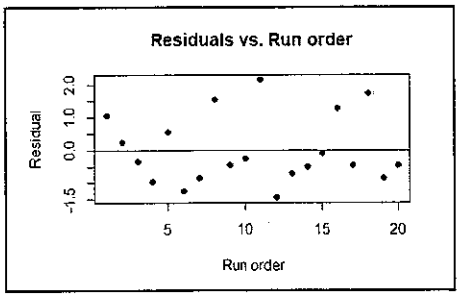
(iii). Which block comparisons (fields) show significant differences according to the Tukey's HSD test, shown in figure 1? (Provide proper reasons for your selections.

Figure 1: Differences in the mean level of groups



(iv). Based on the following plot in figure 2, describe the potential ability to conduct the experiment again.

Figure 2: Residuals vs run order



_____END_____