

Introduction to Linear Models and Regression, 2026

Full Marks: 50 Time : 3 hrs

Group A is mandatory. From Group B, answer as many as you want but maximum you can score in this group is 25

[Bold-faced letters are used to denote vectors]

Group A

1. In a feeding experiment, there are 4 types of cattle denoted by A, B, C, D. These are divided into 3 groups and fed on 3 different types of rations, viz. I, II, III. These ration types are chosen at random from a set of available choices of rations. The following results were obtained, the numbers in the table being the gains in weights (in kg). There are two cattle in each group. Analyse the data. In particular, specify the model, the hypothesis, assumptions, test statistic and its distribution, conclusions and any additional comments.

	Ration I	Ration II	Ration III
Cattle Type			
A	3.0, 3.6	6.1, 4.9	4.5, 5.6
B	7.0, 8.2	7.0, 6.1	8.5, 7.1
C	4.5, 6.0	7.0, 5.0	4.2, 4.0
D	9.0, 11.0	14.0, 13.0	9.0, 10.0

$F_{0.05, 3, 6} = 4.76$, $F_{0.05, 2, 6} = 5.14$, $F_{0.05, 2, 12} = 3.89$, $F_{0.05, 6, 12} = 3.00$ (10)

2. Consider the following data on the hours of study (x) and test score (y).

x	1	2	3	4	5	6	7
y	21.3	38.7	61.5	78.9	86.2	88.4	96.1

It is observed that, for the first 4 hours of study, scores increase rapidly. After 4 hours, scores still increase, but at a slower rate. Fit appropriate model (assume linearity wherever appropriate). (7)

3. For a study on prolonged use of an over-the counter medicine, the following data was collected on years of usage (x) and proportion of users having a particular side-effect (y). For each x, the proportion was computed based on the information from 100 volunteers.

Two different types of regression models were used, for which the R output is given below.

- Identify the models.
- Which one do you prefer and why.
- Explain the R-output of your preferred model. (2+2+4)

Regression Model 1

```

Residuals:
    Min       1Q   Median       3Q      Max
-0.030571 -0.021971  0.002247  0.012683  0.052120

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.099611    0.019755  -5.042 0.000999 ***
x             0.122909    0.003881  31.668 1.08e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02648 on 8 degrees of freedom
Multiple R-squared:  0.9921,    Adjusted R-squared:  0.9911
F-statistic: 1003 on 1 and 8 DF,  p-value: 1.076e-09

```

Regression Model 2

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.99144    0.20648  -14.49  <2e-16 ***
x             0.60889    0.04036   15.09  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 313.2973  on 9  degrees of freedom
Residual deviance:   1.4782  on 8  degrees of freedom
AIC: 52.129

Number of Fisher Scoring iterations: 3

```

Group B

4. Consider a multiple linear regression set-up with x_1 as the dependent variable and $x_2, x_3, \dots, x_p$ as the explanatory variables. Show that among all linear functions of $x_2, x_3, \dots, x_p$, multiple correlation gives the maximum correlation between x_1 and the linear function of $x_2, x_3, \dots, x_p$ as the predictor of x_1 . (5)
5. Suppose $X \sim N_p(0, \sigma^2 I_p)$. If P_1 is a $m \times p$ orthogonal matrix.
 - i. Find the distribution of $U = \frac{1}{\sigma^2} (X'X - Z'Z)$ where, $Z = P_1 X$.
 - ii. Show that U and Z are independently distributed. (5+2)
6. During a statistical analysis of advertising expense and its impact on sales, with both the variables of continuous type, a linear regression model was fit. Following is the summarized R output for the regression model along with the residual plots.

```

Call:
lm(formula = Income ~ adv_exp)

Coefficients:
(Intercept)      adv_exp
 49443.384         8.048

> summary(model)

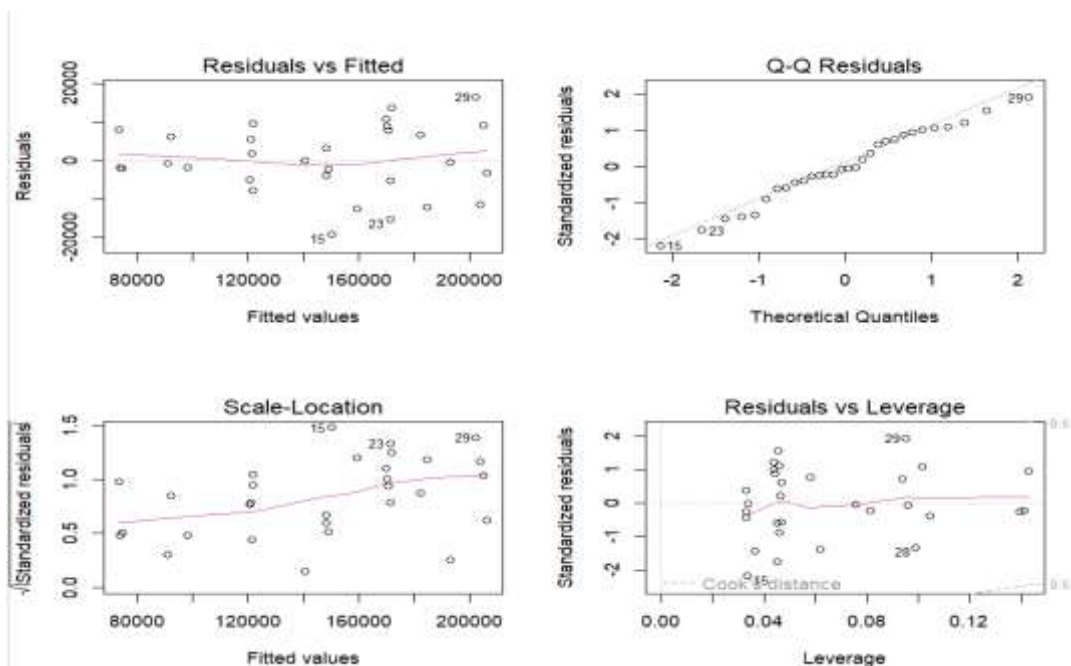
Call:
lm(formula = Income ~ adv_exp)

Residuals:
    Min       1Q   Median       3Q      Max
-19287.1  -4827.3   -638.3    7363.0  16351.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.944e+04  4.289e+03  11.53 3.81e-12 ***
adv_exp      8.048e+00  3.265e-01   24.65 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8999 on 28 degrees of freedom
Multiple R-squared:  0.9559,    Adjusted R-squared:  0.9544
F-statistic: 607.5 on 1 and 28 DF,  p-value: < 2.2e-16

```



- i. Make the analysis for regression diagnostics
- ii. If the model is doing well, justify your answer. If it is not, suggest possible remedy. (5+2)

7. Under the usual notations of general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, show that, if every BLUE is expressed as $\mathbf{a}'\mathbf{Y}$, then the coefficient vector $\mathbf{a}$ is a linear combination of the columns of $\mathbf{X}$ and vice-versa. (5)
8. Let X_i follows Gamma $(\alpha_i, 1)$, for $i = 1(1)k$ and X_i 's are mutually independent. Then find the joint pdf of $Y_1, Y_2, \dots, Y_{k-1}$, where, $Y_i = \frac{X_i}{X_1 + X_2 + \dots + X_k}$, for $i = 1, 2, \dots, k-1$. (6)