# M.Sc. (Honours) Examination, 2025
# Semester-II
# Statistics
# MSC-21-Inference-II
# Time: 3 hrs Full Marks:40

Answer any **four** questions of the following.

1. (a) Let $X_1, X_2, \cdots, X_n$ constitute a random sample from $U(0, \theta)$. For a MP test $H_0 : \theta = 1$ against $H_1 : \theta = 2$ with $\alpha = 0.01$ calculate the power of it.

   (b) Let $X$ be a random observation having discrete distribution so that under $H_0$ the distribution is discrete uniform on $\{0, 1, 2, 3\}$ and under $H_1$, $P(X = 0) = \frac{1}{5}$, $P(X = 1) = \frac{1}{5}$, $P(X = 2) = \frac{1}{5}$, $P(X = 3) = \frac{2}{5}$. Can you construct an MP test for $\alpha = \frac{2}{5}$?

   (c) Define a $\alpha$ similar test.

   4+4+2

2. (a) Suppose out of $n$ sample observations, $R_i$ be the rank of $i$-th ordered observation and $R_j$ be the rank of $j$-th ordered observation. Then find out covariance between $R_i$ and $R_j$.

   (b) Define a test having Neyman structure with respect to a sufficient statistic $T$.

   (c) Suppose $X$ and $Y$ be the independent Poisson random variables with parameters $\lambda$ and $\mu$ respectively. Propose a UMP test for $H_0 : \mu \leq \lambda$ against $H_1 : \mu > \lambda$.

   3+3+4

3. (a) Let $X_1$, $X_2$ and $X_3$ be collected from $U(\theta, \theta + 1)$. Does this family have Monotone likelihood ratio property? Also construct a UMP test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

   (b) Establish that Kolmogorov-Smirnov one sample test statistic is distribution free.

   5+5

4. (a) Let $X_1, X_2, \cdots, X_n$ be a random sample from $f(x, \lambda) = 2\lambda e^{-\lambda x^2} x$; $X > 0$. Construct a UMP test for testing $H_0 : \lambda = 1.5$ against $H_1 : \lambda > 1.5$.

   (b) For an exponential family with single parameter $\theta$ and sufficient statistic $T(x)$, to construct a UMPU test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, show that $E_{\theta_0}[T\phi(T)] = \alpha E_{\theta_0}(T)$ , $\alpha$ being the level of significance.

   5+5

5. (a) Define one sample Wilcoxon signed rank test statistic, $T^+$. Show that it is a linear rank statistic.

   (b) For $n = 4$, what is the range of $T^+$? Hence prepare a probability distribution table of $T^+$.

   (c) Find its mean for $n = 4$.

   4+4+2

6. (a) Suppose $F_X()$ and $F_Y()$ be the probability distribution functions of $X$ and $Y$ respectively. Discuss a nonparametric test in order to check $X$ is stochastically larger than $Y$, clearly stating the null and alternative hypothesis.

   (b) State and prove Neyman and Pearson fundamental lemma.

5+5

**MSc in Statistics Sem-II Examination 2025**
**Subject: Statistics Paper: MSC-22**
**(Applied Multivariate Analysis)**
**Full Marks: 40 Time: 3 hours**

Answer any **four** from the following seven questions of equal marks.
Notations are of their usual meanings.

1. (a) Discuss the role of eigenvectors and eigenvalues in Principal Component Analysis (PCA) and how they help in dimensionality reduction.

   (b) Explain how you decide on the number of PCs to be retained. Also discuss the scree plot and how it is used to decide the number of components to retain.

   5+5

2. (a) Write down the orthogonal factor model with assumptions and establish how it explains the covariance matrix.

   (b) Explain the concept of communality in factor analysis. Can you use factor model always? Answer with reasons.

   5+5

3. (a) Define Cluster Analysis. What are the main objectives and types of clustering techniques?

   (b) Describe the Hierarchical Agglomerative Clustering Algorithm. Explain the role of linkage criteria and compare single, complete, and average linkage methods.

   3+7

4. (a) Define the concept of the Expected Cost of Misclassification (ECM). Derive the decision rule that minimizes the ECM for a two-population case.

   (b) Explain how to assess the classification accuracy of a discriminant function using a confusion matrix.

   6+4

5. (a) Find out the discriminant rule for two multivariate normal populations and interpret.

   (b) Compare Canonical Correlation Analysis with Principal Component Analysis.

   5+5

6. (a) Define the concept of distance measure in clustering and explain with examples.

   (b) Let the covariance matrix be defined as:

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

   where, $-1 \leq \rho \leq 1$. Find PCs of the covariance matrix and interpret.

5+5

7. (a) Explain how MANOVA helps when dependent variables are correlated. Why is it preferred over multiple separate ANOVAs?

   (b) Derive the hypothesis structure in one-way MANOVA. Clearly define the null and alternative hypotheses.

6+4

**M.Sc. Examination, 2025**
**Semester-II**
**Statistics**
**Course: MSC-23**
**(Regression Techniques)**
**Time: 3 Hours** **Full Marks: 40**

Questions are of value as indicated in the margin.
Notations have their usual meanings

**Answer *any four* questions.**

**1.** **(a)** Consider multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Illustrate the significance of hat matrix in multiple regression. Prove that the matrices $H$ and $I - H$ are idempotent. Show that in the multiple linear regression model $Var(\hat{Y}) = \sigma^2 H$.

**(b)** Assume $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{W}^{-1}$ with a known positive–definite weight matrix $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$. Show that the weighted least–squares estimator is $\widehat{\boldsymbol{\beta}}_{\text{WLS}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$, $\text{Var}\big(\widehat{\boldsymbol{\beta}}_{\text{WLS}}\big) = \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$. Discuss why, when the weights are correctly specified, WLS is more efficient than OLS.

5 + 5

**2.** (a) Suppose we wish to find the least-square estimator of $\beta$ in the multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ subject to a set of equality constraints on $\beta$, say $\mathbf{T}\beta = \mathbf{c}$. Show that the estimator is $\tilde{\beta} = \hat{\beta} + (\mathbf{X'X})^{-1}\mathbf{T'}[\mathbf{T}(\mathbf{X'X})^{-1}\mathbf{T}]^{-1}(\mathbf{c} - \mathbf{T}\hat{\beta})$, where $\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'y}$.

(b) Explain 'piecewise polynomial fitting' and 'spline regression.' Suppose we fit a degree-$k$ polynomial regression model using $n$ observations with distinct $x_i$ values. Show that the model matrix $\mathbf{X}$ has full column rank.

5 + 5

**3.** (a) What is the studentized residual and when is it used? Show that, the studentized residuals $(r_i)$ can be expressed as $r_i = \dfrac{e_i}{\sqrt{MS_{Res}\left[1 - \left(\dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}}$, for $i = 1, 2, \cdots, n$. (notations have their usual meaning).

(b) Write short note on PRESS Residual and PRESS statistic. **(OR)** Write a short note on M-estimators as Robust Regression for Linear Model.

5 + 5

**4.** **(a)** Define the *Variance Inflation Factor* (VIF) for the $j$-th predictor in a multiple regression model and prove

$$\text{VIF}_j = \frac{1}{1 - R_j^2},$$

where $R_j^2$ is the coefficient of determination when $x_j$ is regressed on the remaining predictors. Describe how large VIF values inflate $\text{Var}(\widehat{\beta}_j)$.

**(b)** Suppose that there are only two regressor variables, $x_1$ and $x_2$ . Let $r_{jy}$ is the simple correlation between $x_j$ and $y$ , $j = 1, 2$. The model, assuming that $x_1, x_2$ , and $y$ are scaled to unit length, is $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Show that, strong multicollinearity between $x_1$ and $x_2$ results in large variances and covariances for the least - squares estimators of the regression coefficients.

5 + 5

CONTINUED

**5.** (a) What is a Generalized Linear Model? When do we use logistic regression? Briefly describe the principle of Logistic Regression and Probit Regression.

(b) Write short note on: leverage point and influential point.

$5 + 5$

**6.** (a) Explain how the following problematic scenario (s) can be handled in light of Ridge regression.

(i) The Least Squared (LS) estimate depends upon $(X'X)^{-1}$ , we would have problems in computing $\beta_{LS}$ if $X'X$ were singular or nearly singular.

(ii) In above case(s), a small changes to the elements of $X$ lead to large changes in $(X'X)^{-1}$, and the least squared estimator $\beta_{LS}$ may provide a good fit to the training data, but it may not fit sufficiently well to the test data.

(b) Define the Ridge estimator $\hat{\beta}_{\mathbf{R}}$. Obtain the mean squared error of the Ridge estimator. Justify the following statement:

Ridge estimate will not necessarily provide the best " fit " to the data.

$(2\frac{1}{2} + 2\frac{1}{2}) + 5$

---

**M.Sc. Examination 2025**
**Semester-II**
**Statistics**
**Course: MSC-24 (Design of Experiments)**
**Full Marks: 40** **Time: 3 Hours**

**(Answer any four questions.)**

1. (a) State and prove a necessary and sufficient condition for the estimability of a linear parametric function $\boldsymbol{p'\tau}$ under a general block design.

   (b) Prove that for a block design with $v$ treatments and $b$ blocks, $rank(C)+b = rank(D)+v$, where the notations have their usual meanings. 6+4

2. (a) When is a block design called structurally connected?

   (b) In a connected block design, prove that the average variance of all elementary treatment contrasts is given by $\frac{2\sigma^2}{H}$, where $H$ is the harmonic mean of all non-zero eigenvalues of the $C$ matrix and $\sigma^2$ is the error variance.

   (c) Originally, a randomized block design with $v$ treatments and $b$ blocks was planned. However, due to a mistake, treatment 1 was applied twice and treatment 2 was not applied, in the first block. Is the resulting design connected and orthogonal? 2+4+4

3. In the context of a proper block design with incidence matrix $N$, if the treatment effects ($\boldsymbol{\tau}$) are considered to be fixed and the block effects ($\boldsymbol{\beta}$) are considered to be random, then prove that

   (a) $E(\boldsymbol{Q}) = C\boldsymbol{\tau}$.

   (b) $E(\boldsymbol{Q^*}) = C^*\boldsymbol{\tau}$

   (c) $$Disp\begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{Q^*} \\ G \end{pmatrix} = \begin{pmatrix} C\sigma^2 & O & \mathbf{0} \\ O & \frac{C^*}{w_2} & \mathbf{0} \\ \mathbf{0'} & \mathbf{0'} & \frac{n}{w_2} \end{pmatrix}$$

where $w_2$ is the reciprocal of the inter-block variance. $C, G$ & $\boldsymbol{Q}$ denote respectively the C-matrix, the vector of the adjusted treatment totals and the grand total. Further, $C^* = \frac{NN'}{k} - \frac{\boldsymbol{rr'}}{n}$ & $\boldsymbol{Q^*} = \boldsymbol{T} - \boldsymbol{Q} - \frac{G\boldsymbol{r}}{n}$. Other symbols have their usual meanings. 10

4. (a) Define a Hadamard matrix and state it's connection with a BIBD.

   (b) Construct BIBD s with the following parameters. You should properly state the results you use in each case. (Give the control block only.) 2+(4+4)

   i. $v = 13, b = 26, r = 6, k = 3, \lambda = 1$

   ii. $v = 9, b = 12, r = 4, k = 3, \lambda = 1$

5. (a) For a BIBD with parameters $(v, b, r, k, \lambda)$, prove that $b \geq v$.

   (b) Prove that the complementary design obtained from a BIBD with parameters $(v, b, r, k, \lambda)$, is itself a BIBD with parameters $v_1 = v, b_1 = b, r_1 = b - r, k_1 = v - k, \lambda_1 = b - 2r + \lambda$

   (c) Distinguish between a residual design and a derived design with examples. 3+4+3

6. (a) Construct the layout of a $3^3$ experiment conducted in blocks of $3^2$ plots such that the effect $ABC^2$ is confounded.

   (b) One of the blocks of a $3^3$ experiment conducted in $3^2$ blocks is $(AB^2, A^2C, BC^2)$. Identify the confounded effects.

   (c) Mention the disadvantages of a split-plot design. 4+4+2

---

# MSc Semester-II Examination 2025

Subject: Statistics
Paper: MSC-25
(Practical on Applied Multivariate and Inference II)

Answer all the following questions.
Nonations are of usual meanings.

Full Marks: 40 Time: 4 hours

1. The variance–covariance matrix of a bivariate dataset is:

$$S = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$$

   (a) Find the eigenvalues and eigenvectors of $S$.

   (b) Compute the proportion of total variance explained by the first principal component.

   (c) State which component should be retained if the criterion is at least 70% variance explained.

6

2. The following table gives the Euclidean distances between four objects:

| | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| $A$ | 0 | 4 | 6 | 8 |
| $B$ | 4 | 0 | 5 | 7 |
| $C$ | 6 | 5 | 0 | 3 |
| $D$ | 8 | 7 | 3 | 0 |

   (a) Using single linkage and average linkage hierarchical clustering, show the sequence of cluster merges.

   (b) Draw the dendrograms (scale can be approximate).

6

3. Two groups ($G_1$ and $G_2$) have the following sample statistics:

$$G_1 : \ \bar{x}_1 = 10, \ s_1^2 = 4, \ n_1 = 5$$

$$G_2 : \ \bar{x}_2 = 14, \ s_2^2 = 5, \ n_2 = 5$$

   Assume equal population variance.

(a) Compute the pooled variance.

(b) Find the discriminant function $D(x) = ax + b$.

(c) Classify an observation $x = 13$ assuming equal prior probabilities.

6

4. Let $X_i, (i = 1(1)5)$ be a random sample of size 5 from a Poisson distribution with mean $\lambda$. Construct a test of size 0.1 for testing $H_0 : \lambda = 1$ against $H_1 : \lambda > 1$. Justify if it is UMP.

4

5. Let $X$ and $Y$ be independent Poisson($\lambda$) and Poisson ($\mu$).Construct a UMP test for $H_0 : \mu \leq \lambda$ against $H_1 : \mu > \lambda$ at level of significance 0.15.

4

6. Ten students take a test and their scores (out of 100) are as follows: 95, 80, 40, 52, 60, 82, 82, 59, 65, 50. Test the null hypothesis that the cumulative distribution of the proportion of right answers a student gets on the test is $F_0(x) = x^2(3 - 2x)$, $0 \leq x \leq 1$.

4

7. The following data shows the age of diagnosis of type II diabetes in young adults.

| Gender | Female | Female | Female | Female | Female | Male | Male | Male | Male |
|---|---|---|---|---|---|---|---|---|---|
| Age | 19 | 22 | 16 | 24 | 29 | 20 | 11 | 17 | 15 |

(a) Do you believe age of diagnosis in males is lower than that of females? Write null and alternative hypothesis.

(b) Define U statistic you used for testing it?

(c) Report your p-value and hence conclude.

5

8. Viva-voce and Practical Notebook.

5

**M.Sc. Examination 2025**
**Semester-II**
**Statistics (Practical)**
**Course: MSC-26**

**Full Marks: 40** **Time: 4 Hours**

Notations have their usual meanings.
R programming can be used for computation purposes.

1. The determination of usual activity at four different distances (A, B, C, D) as treatments was the subject of a recent experiment. Four different subjects (as blocks) chosen at random from a large group were used for this purpose. Assuming a mixed effect model,

   (a) Find unbiased estimates of the variance components.

   (b) Test whether the treatment effects differ significantly or not.

   The data are given below:

| **Subject** | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| 1 | – | 16 | 30 | 27 |
| 2 | 5 | 10 | – | 18 |
| 3 | 7 | 28 | 35 | – |
| 4 | 10 | – | 51 | 26 |

5+5

2. The following table gives the plants and the yields (in suitable units) of a manurial experiment involving two factors $N$ and $P$ each at 3 levels.

   Analyse the data partitioning the treatment SS into 8 orthogonal components.

| **Replicate-1** | | **Replicate-2** | | **Replicate-3** | |
|---|---|---|---|---|---|
| 12 | 223 | 01 | 269 | 02 | 191 |
| 00 | 236 | 20 | 233 | 11 | 300 |
| 10 | 240 | 12 | 266 | 01 | 278 |
| 21 | 300 | 00 | 213 | 21 | 209 |
| 22 | 189 | 22 | 226 | 22 | 226 |
| 01 | 160 | 11 | 240 | 10 | 233 |
| 11 | 284 | 02 | 282 | 12 | 182 |
| 20 | 271 | 21 | 209 | 00 | 270 |
| 02 | 259 | 10 | 293 | 20 | 258 |

8

3. A study was conducted attempting to relate home ownership to family income. Twenty households were selected and family income was estimated, along with information concerning home ownership ($y = 1$ indicates *Yes* and $y = 0$ indicates *No*). (See Table 1; You may obtain the dataset by the R command: `library(MPV); attach(p13.2);`):

   (a) Fit a logistic regression model to the response variable $y$. Use a simple linear regression model as the structure for the linear predictor. Does the model deviance indicate that the logistic regression model is adequate?

   (b) Expand the linear predictor to include a quadratic term in income. Is there any evidence that this quadratic term is required in the model? 3+4

4. The concentration of $NbOCl_3$ in a tube-flow reactor as a function of several controllable variables is given in Table 2 (Table b.6; you may obtain table.b6 in R package MPV).

   (a) Fit a multiple regression model relating the concentration of $NbOCl_3$ ($y$) to the concentration of $COCl_2$ ($x_1$) and mole fraction ($x_4$). Test for the significance of the regression.

   (b) Determine the contribution of $x_1$ and $x_4$ to the model. Are both regressors $x_1$ and $x_4$ necessary? Test for multicollinearity. 3+4

5. A statistician claims the following ridge regression code (R code) addresses multicollinearity, but there is a fundamental flaw. Fix it and explain. 3

```
library(MASS); set.seed(1);
x1 <- rnorm(10); x2 <- x1 + rnorm(10,0,0.01); y <- 2 + 3*x1 - x2 + rnorm(10);
lm.ridge(y ~ x1 + x2, data=data.frame(y,x1,x2), lambda=0.5, center=FALSE, scale=FALSE)
```

6. Practical Notebook and viva-voce 5

| Household | Income | Home Ownership Status | Household | Income | Home Ownership Status |
|---|---|---|---|---|---|
| 1 | 38000 | 0 | 11 | 38700 | 1 |
| 2 | 51200 | 1 | 12 | 40100 | 0 |
| 3 | 39600 | 0 | 13 | 49500 | 1 |
| 4 | 43400 | 1 | 14 | 38000 | 0 |
| 5 | 47700 | 0 | 15 | 42000 | 1 |
| 6 | 5000 | 0 | 16 | 54000 | 1 |
| 7 | 4500 | 1 | 17 | 51700 | 1 |
| 8 | 4800 | 0 | 18 | 39400 | 0 |
| 9 | 45400 | 1 | 19 | 40900 | 0 |
| 10 | 52400 | 1 | 20 | 52800 | 1 |

Table 1: Home ownership to family income (p13.2 from MPV package in R)

| | y | x1 | x2 | x3 | x4 |
|---|---|---|---|---|---|
| 1 | 0.000 | 0.010 | 90.900 | 0.016 | 0.018 |
| 2 | 0.000 | 0.011 | 84.600 | 0.016 | 0.017 |
| 3 | 0.000 | 0.011 | 88.900 | 0.016 | 0.016 |
| 4 | 0.001 | 0.012 | 488.700 | 0.019 | 0.008 |
| 5 | 0.000 | 0.012 | 454.400 | 0.019 | 0.007 |
| 6 | 0.000 | 0.012 | 439.200 | 0.019 | 0.006 |
| 7 | 0.000 | 0.012 | 447.100 | 0.019 | 0.007 |
| 8 | 0.000 | 0.012 | 451.600 | 0.019 | 0.006 |
| 9 | 0.001 | 0.012 | 487.800 | 0.019 | 0.015 |
| 10 | 0.001 | 0.012 | 467.600 | 0.019 | 0.013 |
| 11 | 0.001 | 0.009 | 95.400 | 0.016 | 0.035 |
| 12 | 0.001 | 0.010 | 87.100 | 0.016 | 0.034 |
| 13 | 0.001 | 0.010 | 82.700 | 0.016 | 0.032 |
| 14 | 0.001 | 0.010 | 87.000 | 0.016 | 0.034 |
| 15 | 0.001 | 0.011 | 516.400 | 0.019 | 0.016 |
| 16 | 0.001 | 0.012 | 488.400 | 0.019 | 0.015 |
| 17 | 0.001 | 0.011 | 534.500 | 0.019 | 0.016 |
| 18 | 0.002 | 0.010 | 542.300 | 0.019 | 0.016 |
| 19 | 0.002 | 0.007 | 98.800 | 0.016 | 0.038 |
| 20 | 0.002 | 0.007 | 84.800 | 0.016 | 0.036 |
| 21 | 0.003 | 0.004 | 69.600 | 0.016 | 0.033 |
| 22 | 0.003 | 0.007 | 436.900 | 0.019 | 0.026 |
| 23 | 0.003 | 0.008 | 406.300 | 0.019 | 0.020 |
| 24 | 0.003 | 0.007 | 447.900 | 0.019 | 0.020 |
| 25 | 0.002 | 0.009 | 58.500 | 0.016 | 0.033 |
| 26 | 0.003 | 0.008 | 394.300 | 0.018 | 0.067 |
| 27 | 0.003 | 0.007 | 461.000 | 0.017 | 0.077 |
| 28 | 0.003 | 0.006 | 469.200 | 0.017 | 0.078 |

Table 2: (table.b6 from R package MPV)

# M.Sc. Examination, 2024
## Semester-II
## Statistics
## MSC-21-Inference-II
## Time: 3 hrs Full Marks:40

Answer any **four** questions of the following.

1. (a) Let $X$ be a random variable having density function $f \in \{f_0, f_1\}$ when $f_0(x) = 1$; $0 < x < 1$ and $f_1(x) = \frac{1}{3}$; $0 < x < 3$. For testing $H_0 : f = f_0$ against $H_1 : f = f_1$, based on a single observation find out the power of most powerful test when $\alpha$ (size)= .05.

   (b) Propose a level $\alpha$ MP test for $H_0 : X \sim N(0, 1/2)$ against $H_1 : X \sim Cauchy(0, 1)$ based on a single observation.

   (c) Define a $\alpha$ similar test.

   4+4+2

2. (a) For a nonparametric test of median ($\mu$) being zero under $H_0$, let $X_\alpha, \alpha = 1(1)n$ be a continuous random variable for $\alpha = 1(1)n$. $R_\alpha^+$ is the rank of $|X_\alpha|$. Further define an indicator variable $Z_\alpha$ such that

$$Z_\alpha = \begin{cases} 1 \; if \; X_\alpha > \mu \\ 0 \; if \; X_\alpha < \mu \end{cases}$$

   Show that under $H_0$, $R_\alpha^+$ and $Z_\alpha$ are independently distributed.

   (b) Define a test having Neyman structure with respect to a sufficient statistic $T$.

   (c) Suppose $X$ and $Y$ be the independent Poisson random variables with parameters $\lambda$ and $\mu$ respectively. Propose a UMP test for $H_0 : \mu \leq \lambda$ against $H_1 : \mu > \lambda$.

   3+3+4

3. (a) Let $X_1$, $X_2$ and $X_3$ be collected from $U(\theta, \theta + 2)$. Does this family have Monotone likelihood ratio property? Also construct a UMP test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

   (b) Establish that Kolmogorov-Smirnov one sample test statistic is distribution free.

   5+5

4. (a) Let $X_1, X_2, \cdots, X_n$ be a random sample from $f(x, \lambda) = 2\lambda e^{-\lambda x^2} x$; $x > 0$. Construct a UMP test for testing $H_0 : \lambda = 1$ against $H_1 : \lambda > 1$.

   (b) For an exponential family with single parameter $\theta$ and sufficient statistic $T(x)$, to construct a UMPU test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, show that $E_{\theta_0}[T\phi(T)] = \alpha E_{\theta_0}(T)$ , $\alpha$ being the level of significance.

   5+5

5. (a) Define one sample linear rank statistic associated with one sample location test.

(b) Find its mean and variance.

(c) Show that one sample Wilcoxon signed rank test statistic is a particular case of it.

5+5

6. (a) Suppose $F_X()$ and $F_Y()$ be the probability distribution functions of $X$ and $Y$ respectively. Discuss a nonparametric test in order to check $X$ is stochastically larger than $Y$, clearly stating the null and alternative hypothesis.

(b) Let $P_0$,$P_1$, $P_2$ be the probability distributions assigning to the integers 1, 2, 3, 4, 5 the following probabilities:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $P_0$ | 0.03 | 0.02 | 0.02 | 0.01 | 0.92 |
| $P_1$ | 0.06 | 0.05 | 0.08 | 0.02 | 0.79 |
| $P_2$ | 0.09 | 0.05 | 0.12 | 0 | 0.74 |

Determine whether there exists a UMP test for $H_0 : P = P_0$ against $H_1 : P \neq P_0$ at $\alpha$(size)= .05.

5+5

# MSc in Statistics Sem-II Examination 2024

Subject: Statistics
Paper: MSC-22
(Applied Multivariate Analysis))

Answer any **four** from the following seven questions of equal marks.
Nonations are of usual meanings.

Full Marks: 40 Time: 3 hours

1. (a) Explain the Principal Component Analysis (PCA) concept and its primary objectives.
   (b) Explain which one you prefer for carrying out PCA, out of the covariance and correlation matrix.

   5+5

2. (a) Write down the orthogonal factor model and show how it explains the variance-covariance matrix.
   (b) Establish in PCA how the total system variance is expressed in terms of the sum of eigenvalues of the variance-covariance matrix.

   5+5

3. (a) Why do we perform factor rotation? Explain how the rotation of factor axes does not change the original factor model.
   (b) Give some real examples where PCA is used.

   7+3

4. (a) What is the primary goal of Discriminant Analysis? Give two examples.
   (b) Explain Fisher's Discriminant Function and how it is derived.

   5+5

5. (a) What is the primary purpose of Canonical Correlation Analysis (CCA)? Derive the pair of Canonical variables along with their Canonical Correlation from a general set-up.
   (b) Give two real examples of the application of cluster analysis.

8+2

6. (a) Let you be given some observations on a group of people on their heights, eye colour, body weight, and gender. Explain how you will find a similarity matrix to cluster them using a hierarchical clustering algorithm.

   (b) Describe how you will determine the number of clusters in k-means clustering method.

6+4

7. (a) What is the primary objective of Multivariate Analysis of Variance (MANOVA)? Write down the assumptions required for conducting MANOVA.

   (b) Discuss the interpretation of Wilk's Lambda $\Lambda$ in the context of MANOVA.

5+5

**M.Sc. Examination, 2024**
**Semester-II**
**Statistics**
**Course: MSC-23**
**(Regression Techniques)**
**Time: 3 Hours** **Full Marks: 40**

Questions are of value as indicated in the margin.
Notations have their usual meanings

Answer any four questions.

**1.** (a) Consider the simple linear regression model, $y = \beta_0 + \beta_1 x + \epsilon$, with $E(\epsilon) = 0$, $var(\epsilon) = \sigma^2$, and $\epsilon$ uncorrelated. Show that $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2/S_{xx}$. Also show that $\text{Cov}(\bar{y}, \beta_1) = 0$.

(b) Consider the multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Illustrate the significance of the hat matrix in multiple regression. Prove that the matrices $H$ and $I - H$ are idempotent. Show that in the multiple linear regression model $\text{Var}(\hat{Y}) = \sigma^2 H$.

$2 + 3 + 3 + 2$

**2.** (a) Consider the multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Find the expression for the least-squares estimator $\hat{\beta}$ of $\beta$. Show that the least-squares estimator can be written as $\hat{\beta} = \beta + \mathbf{R}\varepsilon$, where $\mathbf{R} = (\mathbf{X^T X})^{-1}\mathbf{X^T}$.

(b) Consider a correctly specified regression model with $p$ terms, including the intercept. Make the usual assumptions about $\epsilon$. Prove that $\sum_{i=1}^{n} Var(\hat{y}_i) = p\sigma^2$.

$5 + 5$

**3.** (a) What is the studentized residual and when is it used? Show that, the studentized residuals $(r_i)$ can be expressed as $r_i = \dfrac{e_i}{\sqrt{MS_{Res}\left[1 - \left(\dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}}$, for $i = 1, 2, \cdots, n$ and argue why we can write $r_i = \dfrac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$ (notations have their usual meaning).

(b) Write short note on PRESS Residual and PRESS statistic.

$5 + 5$

**4.** (a) Briefly describe the principle of Logistic Regression and Probit Regression. Consider a binary logistic regression model where the response variable $Y$ takes values in $\{0, 1\}$. The model is defined as:

$$\text{logit}(P(Y = 1|\mathbf{X})) = \mathbf{X}^T\boldsymbol{\beta},$$

where $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)^T$ is the vector of predictors including an intercept term, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients. Derive the log-likelihood function for the logistic regression model, Compute the gradient of the log-likelihood function with respect to the regression coefficients $\boldsymbol{\beta}$ and show that the Hessian matrix of the log-likelihood function is negative semi-definite.

(b) Among two logit models, how do you determine which model is better? Justify.

$5 + 5$

**5.** (a) Describe a formal test for Lack of Fit under the suitable assumption(s).

(b) Write a short note on influential points and leverage points.

$5 + 5$

CONTINUED

**6.** (a) Define 'piecewise polynomial fitting' and 'spline regression.' Under what circumstances is ridge regression used? Can the correlation matrix provide any indication of multicollinearity? Explain your answer.

(b) Suppose we wish to find the least-square estimator of $\beta$ in the multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ subject to a set of equality constraints on $\beta$, say $\mathbf{T}\beta = \mathbf{c}$. Show that the estimator is

$$\tilde{\beta} = \hat{\beta} + (\mathbf{X'X})^{-1}\mathbf{T'}[\mathbf{T}(\mathbf{X'X})^{-1}\mathbf{T}]^{-1}(\mathbf{c} - \mathbf{T}\hat{\beta}),$$

where $\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'y}$. Discuss situations in which this constrained estimator might be appropriate.

$5 + 5$

**7.** (a) Consider the polynomial regression model of degree $k$ given by:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i,$$

where $i = 1, 2, \ldots, n$, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are independent and identically distributed (i.i.d.) random errors. Prove that the model matrix $\mathbf{X}$ for the polynomial regression model has full column rank if the $x_i$ values are distinct.

(b) Illustrate how the linearization can be done by a Taylor series expansion of the nonlinear Regression function, followed by the iteration method of parameter estimation
(or)
Write a short note on M-estimators as Robust Regression for Linear Model.

$5 + 5$

---

**M.Sc. Examination 2024**
**Semester-II**
**Statistics**
**Course: MSC-24 (Design of Experiments)**
**Full Marks: 40** **Time: 3 Hours**

**(Answer any four questions.)**

1. (a) Provide the detailed inter-block analysis of an incomplete block design with $b$ blocks and $v$ treatments, starting from the appropriate model and assumptions.
   (b) Prove that the inter and intra block estimates of the treatment contrast $\boldsymbol{p'\tau}$ are independent of each other. 7+3

2. (a) For a BIBD with parameters $(b, v, r, k, \lambda)$, if $b$ is divisible by $r$, then prove that $b \geq v + r - 1$.
   (b) Let $N$ be the incidence matrix of a symmetric BIBD with parameters $(v, r, \lambda)$. If $v$ is even, show that $(r - \lambda)$ is a perfect square.
   (c) For a symmetric BIBD with parameters $(v, k, \lambda)$, show that any two blocks have exactly $\lambda$ treatments in common. 3+3+4

3. (a) What is a split-plot design? Write down the underlying model, hypotheses and the complete ANOVA table of the design.
   (b) State the advantages and disadvantages of the design.
   (c) Find the efficiency of split-plot design with respect to a randomized block design. 4+3+3

4. What do you mean by the connectedness of a block design? Give the rank definition and the structural definition of connectedness. Show that these definitions are equivalent. 1+2+7

5. Construct BIBD s with the following parameters. You should properly state the results you use in each case. (Give the control block only.) 3+4+3
   (a) $v = 13, b = 26, r = 6, k = 3, \lambda = 1$
   (b) $v = 12, b = 44, r = 11, k = 3, \lambda = 2$
   (c) $v = 9, b = 12, r = 4, k = 3, \lambda = 1$

6. (a) Consider the $3^3$ experiment conducted in 2 replications in blocks of $3^2$ plots. The following information is given below.

   **Replicate 1:** 100, 112, 202, 211, 220, 121, 010, 021, 002
   **Replicate 2:** 001, 102, 012, 110, 200, 121, 222, 211, 020

   Identify the confounded effects.
   (b) In the context of the combined inter-intra block analysis of a block design, find unbiased estimators of the variance components involved in the model. 3+7

---

# MSc Semester-II Examination 2024

Subject: Statistics
Paper: MSC-25
(Practical on Applied Multivariate and Inference II)

Answer all the following questions.
Nonations are of usual meanings.

Full Marks: 40 Time: 4 hours

1. Two groups have the following means and pooled covariance matrix:

   Group 1 mean: $\underline{\mu_1} = (2, 3)'$ Group 2 mean: $\underline{\mu_2} = (4, 5)'$
   Pooled covariance matrix:

$$S = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

   (a) Using the above data derive a linear discriminant function to classify a new observation into anyone of the two populations.
   (b) Classify the observation $X = (3, 4)$ using the derived discriminant function.

   6

2. Given the following set of data for two sets of variables $X_1, X_2$ and $Y_1, Y_2$, where the covariance matrix between the variables is:

$$\text{Cov}(X_1, X_2, Y_1, Y_2) = \begin{pmatrix} 4 & 2 & 3 & 1 \\ 2 & 5 & 1 & 2 \\ 3 & 1 & 6 & 3 \\ 1 & 2 & 3 & 7 \end{pmatrix}$$

   Compute the canonical correlations between the two sets of variables $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$.

   6

3. Consider a dataset with three variables $X_1$, $X_2$, and $X_3$ and the following covariance matrix:

$$\Sigma = \begin{pmatrix} 2 & 1 & 0.5 \\ 1 & 3 & 1 \\ 0.5 & 1 & 2 \end{pmatrix}$$

Perform Principal Component Analysis (PCA) using the covariance matrix and interpret the results.

5

4. Let the distribution of $X$ be given by

| $X$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P_\theta$ | $\theta$ | $2\theta$ | $0.9 - 2\theta$ | $0.1 - \theta$ |

where $0 < \theta < 0.1$ for testing $H_0 : \theta = .04$ against $H_1 : \theta > 0.04$ at $\alpha = 0.05$. Determine which of the following test is UMPU.

(a) $\phi(0) = 1$, $\phi(1) = \phi(2) = \phi(3) = 0$
(b) $\phi(1) = 0.5$, $\phi(x) = 0$ if $x \neq 1$
(c) $\phi(3) = 1$, $\phi(x) = 0$ if $x \neq 3$

4

5. An urn contains 7 marbles of which $m$ are white and (7-m) are black. To test $H_0 : m = 5$ against $H_1 : m > 5$ one draws 4 marbles without replacement. The null hypothesis is rejected if the sample contains 2 or 3 while balls otherwise accepted. Construct a test and its size.

4

6. Use an appropriate nonparametric test to judge whether the square roots of the following numbers are coming from a $U(0, 1)$ distribution. The data set is as follows.

$0.0123, 0.1039, 0.1954, 0.2621, 0.2802, 0.3217, 0.3645, 0.3919, 0.4240, 0.4814$
$0.5139, 0.5846, 0.6275, 0.6541, 0.6889, 0.7621, 0.8320, 0.8871, 0.9249, 0.9634.$

5

7. Let there be a random sample of 10 observations, viz., 2,-1.2, 5, 7,-1.2,4, 2.9, 3.9, 2.5,3 are selected from $N(\mu, \sigma^2)$. Construct a test for $H_0 : \sigma^2 > 2 \, or \, \sigma^2 < 1$ against $H_1 : 1 < \mu < 2$ for a level of significance 0.5.

5

8. Viva-voce and Practical Notebook.

5

**M.Sc. Examination 2024**
**Semester-II**
**Statistics (Practical)**
**Course: MSC-26**

**Full Marks: 40** **Time: 4 Hours**

Notations have their usual meanings.
R programming can be used for computation purposes.

1. The effects of developer strength (A) and development time (B) on the density of photographic plate film are being studied. Three strengths and three times are used, and four replicates of a $3^2$ factorial experiment are run. The data from this experiment is as follows.

| | Development Time (Minutes) | | |
|---|---|---|---|
| Developer strength | 10 | 12 | 14 |
| 1 | 0, 2, 5, 4 | 1, 3, 4, 2 | 2, 5, 4, 6 |
| 2 | 4, 6 7, 5 | 6, 8, 7, 7 | 9, 10, 8, 5 |
| 3 | 7, 10, 8, 7 | 10, 10 8, 7 | 12, 10, 9, 8 |

Analyze the data to test whether these two factors and their interaction have significant effect on the density. 8

2. The determination of visual acuity at four different distances (say A, B, C, and D, as treatments) was the subject of a recent experiment. Four different subjects (as blocks) chosen at random from a large group were used for this purpose. The data recorded were as follows:

| **Distance** | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| **Subject** | | | | |
| 1 | - | 16 | 30 | 27 |
| 2 | 5 | 10 | - | 18 |
| 3 | 7 | 28 | 35 | - |
| 4 | 10 | - | 51 | 26 |

Assuming a mixed effect model, find

(a) the estimates of error variance and the variance of the block effects,

(b) the BLUE of the treatment effects. 5+4

3. The compressive strength of an alloy fastener used in aircraft construction is being studied. Ten loads were selected over the range 2500 – 4300 psi, and a number of fasteners were tested at those loads. Attach the data given in p13.3 of r package MPV. The numbers of fasteners failing at each load were recorded.

   (a) Fit a generalized linear model to the data using a simple linear regression model as the structure for the linear predictor. You may make suitable assumptions, transform the data if necessary, and fit a logistic regression.

   (b) Evaluate the adequacy of the generalized linear model based on model deviance.

   (c) Expand the linear predictor to include a quadratic term to assess whether this addition is necessary.

   (d) Calculate Wald statistics for each parameter of the quadratic model to evaluate their significance.

   (e) Determine approximate 95% confidence intervals for the model parameters of the quadratic logistic regression model.

   5

4. The concentration of $NbOCl_3$ in a tube-flow reactor as a function of several controllable variables is given in Table B.6 (attach table.b6 of R package MPV).

   (a) Fit a multiple regression model relating the concentration of $NbOCl_3$ ($y$) to the concentration of $COCl_2$ ($x_1$) and mole fraction ($x_4$).

   (b) Test for the significance of the regression.

   (c) Calculate $R^2$ and $R^2_{\text{Adj}}$ for this model.

   (d) Using t-tests, determine the contribution of $x_1$ and $x_4$ to the model. Are both regressors $x_1$ and $x_4$ necessary?

   (e) Is multicollinearity a potential concern in this model?

   5

5. The table (data p13.1 of R package MPV) presents the test-firing results for 25 surface-to-air antiaircraft missiles at targets of varying speed. The result of each test is either a hit ($y = 1$) or a miss ($y = 0$).

   (a) Fit a logistic regression model to the response variable $y$. Use a simple linear regression model as the structure for the linear predictor.

   (b) Does the model deviance indicate that the logistic regression model from part (a) is adequate?

   (c) Provide an interpretation of the parameter $\beta_1$ in this model.

   (d) Expand the linear predictor to include a quadratic term in target speed.

   (e) Is there any evidence that this quadratic term is required in the model?

   5

6. Analyse the chemical process data in Table B.5 of R package MPV (attach table.b5) for evidence of multicollinearity. Use the variance inflation factors and the condition number of $X^T X$. 3

7. Practical Notebook and viva-voce 5

Table 1: p13.1

| | x | y |
|---|---|---|
| 1 | 400.00 | 0.00 |
| 2 | 220.00 | 1.00 |
| 3 | 490.00 | 0.00 |
| 4 | 210.00 | 1.00 |
| 5 | 500.00 | 0.00 |
| 6 | 270.00 | 0.00 |
| 7 | 200.00 | 1.00 |
| 8 | 470.00 | 0.00 |
| 9 | 480.00 | 0.00 |
| 10 | 310.00 | 1.00 |
| 11 | 240.00 | 1.00 |
| 12 | 490.00 | 0.00 |
| 13 | 420.00 | 0.00 |
| 14 | 330.00 | 1.00 |
| 15 | 280.00 | 1.00 |
| 16 | 210.00 | 1.00 |
| 17 | 300.00 | 1.00 |
| 18 | 470.00 | 1.00 |
| 19 | 230.00 | 0.00 |
| 20 | 430.00 | 0.00 |
| 21 | 460.00 | 0.00 |
| 22 | 220.00 | 1.00 |
| 23 | 250.00 | 1.00 |
| 24 | 200.00 | 1.00 |
| 25 | 390.00 | 0.00 |

Table 2: table.b6

| | y | x1 | x2 | x3 | x4 |
|---|---|---|---|---|---|
| 1 | 0.00 | 0.01 | 90.90 | 0.02 | 0.02 |
| 2 | 0.00 | 0.01 | 84.60 | 0.02 | 0.02 |
| 3 | 0.00 | 0.01 | 88.90 | 0.02 | 0.02 |
| 4 | 0.00 | 0.01 | 488.70 | 0.02 | 0.01 |
| 5 | 0.00 | 0.01 | 454.40 | 0.02 | 0.01 |
| 6 | 0.00 | 0.01 | 439.20 | 0.02 | 0.01 |
| 7 | 0.00 | 0.01 | 447.10 | 0.02 | 0.01 |
| 8 | 0.00 | 0.01 | 451.60 | 0.02 | 0.01 |
| 9 | 0.00 | 0.01 | 487.80 | 0.02 | 0.02 |
| 10 | 0.00 | 0.01 | 467.60 | 0.02 | 0.01 |
| 11 | 0.00 | 0.01 | 95.40 | 0.02 | 0.04 |
| 12 | 0.00 | 0.01 | 87.10 | 0.02 | 0.03 |
| 13 | 0.00 | 0.01 | 82.70 | 0.02 | 0.03 |
| 14 | 0.00 | 0.01 | 87.00 | 0.02 | 0.03 |
| 15 | 0.00 | 0.01 | 516.40 | 0.02 | 0.02 |
| 16 | 0.00 | 0.01 | 488.40 | 0.02 | 0.01 |
| 17 | 0.00 | 0.01 | 534.50 | 0.02 | 0.02 |
| 18 | 0.00 | 0.01 | 542.30 | 0.02 | 0.02 |
| 19 | 0.00 | 0.01 | 98.80 | 0.02 | 0.04 |
| 20 | 0.00 | 0.01 | 84.80 | 0.02 | 0.04 |
| 21 | 0.00 | 0.00 | 69.60 | 0.02 | 0.03 |
| 22 | 0.00 | 0.01 | 436.90 | 0.02 | 0.03 |
| 23 | 0.00 | 0.01 | 406.30 | 0.02 | 0.02 |
| 24 | 0.00 | 0.01 | 447.90 | 0.02 | 0.02 |
| 25 | 0.00 | 0.01 | 58.50 | 0.02 | 0.03 |
| 26 | 0.00 | 0.01 | 394.30 | 0.02 | 0.07 |
| 27 | 0.00 | 0.01 | 461.00 | 0.02 | 0.08 |
| 28 | 0.00 | 0.01 | 469.20 | 0.02 | 0.08 |

Table 3: table.b5

| | y | x1 | x2 | x3 | x4 | x5 | x6 | x7 |
|---|---|---|---|---|---|---|---|---|
| 1 | 36.98 | 5.10 | 400.00 | 51.37 | 4.24 | 1484.83 | 2227.25 | 2.06 |
| 2 | 13.74 | 26.40 | 400.00 | 72.33 | 30.87 | 289.94 | 434.90 | 1.33 |
| 3 | 10.08 | 23.80 | 400.00 | 71.44 | 33.01 | 320.79 | 481.19 | 0.97 |
| 4 | 8.53 | 46.40 | 400.00 | 79.15 | 44.61 | 164.76 | 247.14 | 0.62 |
| 5 | 36.42 | 7.00 | 450.00 | 80.47 | 33.84 | 1097.26 | 1645.89 | 0.22 |
| 6 | 26.59 | 12.60 | 450.00 | 89.90 | 41.26 | 605.06 | 907.59 | 0.76 |
| 7 | 19.07 | 18.90 | 450.00 | 91.48 | 41.88 | 405.37 | 608.05 | 1.71 |
| 8 | 5.96 | 30.20 | 450.00 | 98.60 | 70.79 | 253.70 | 380.55 | 3.93 |
| 9 | 15.52 | 53.80 | 450.00 | 98.05 | 66.82 | 142.27 | 213.40 | 1.97 |
| 10 | 56.61 | 5.60 | 400.00 | 55.69 | 8.92 | 1362.24 | 2043.36 | 5.08 |
| 11 | 26.72 | 15.10 | 400.00 | 66.29 | 17.98 | 507.65 | 761.48 | 0.60 |
| 12 | 20.80 | 20.30 | 400.00 | 58.94 | 17.79 | 377.60 | 566.40 | 0.90 |
| 13 | 6.99 | 48.40 | 400.00 | 74.74 | 33.94 | 158.05 | 237.08 | 0.63 |
| 14 | 45.93 | 5.80 | 425.00 | 63.71 | 11.95 | 130.66 | 1961.49 | 2.04 |
| 15 | 43.09 | 11.20 | 425.00 | 67.14 | 14.73 | 682.59 | 1023.89 | 1.57 |
| 16 | 15.79 | 27.90 | 425.00 | 77.65 | 34.49 | 274.20 | 411.30 | 2.38 |
| 17 | 21.60 | 5.10 | 450.00 | 67.22 | 14.48 | 1496.51 | 2244.77 | 0.32 |
| 18 | 35.19 | 11.70 | 450.00 | 81.48 | 29.69 | 652.43 | 978.64 | 0.44 |
| 19 | 26.14 | 16.70 | 450.00 | 83.88 | 26.33 | 458.42 | 687.62 | 8.82 |
| 20 | 8.60 | 24.80 | 450.00 | 89.38 | 37.98 | 312.25 | 468.38 | 0.02 |
| 21 | 11.63 | 24.90 | 450.00 | 79.77 | 25.66 | 307.08 | 460.62 | 1.72 |
| 22 | 9.59 | 39.50 | 450.00 | 87.93 | 22.36 | 193.61 | 290.42 | 1.88 |
| 23 | 4.42 | 29.00 | 450.00 | 79.50 | 31.52 | 155.96 | 233.95 | 1.43 |
| 24 | 38.89 | 5.50 | 460.00 | 72.73 | 17.86 | 1392.08 | 2088.12 | 1.35 |
| 25 | 11.19 | 11.50 | 450.00 | 77.88 | 25.20 | 663.09 | 994.63 | 1.61 |
| 26 | 75.62 | 5.20 | 470.00 | 75.50 | 8.66 | 1464.11 | 2196.17 | 4.78 |
| 27 | 36.03 | 10.60 | 470.00 | 83.15 | 22.39 | 720.07 | 1080.11 | 5.88 |

Table 4: p13.3

| | x | n | r |
|---|---|---|---|
| 1 | 2500.00 | 50.00 | 10.00 |
| 2 | 2700.00 | 70.00 | 17.00 |
| 3 | 2900.00 | 100.00 | 30.00 |
| 4 | 3100.00 | 60.00 | 21.00 |
| 5 | 3300.00 | 40.00 | 18.00 |
| 6 | 3500.00 | 85.00 | 43.00 |
| 7 | 3700.00 | 90.00 | 54.00 |
| 8 | 3900.00 | 50.00 | 33.00 |
| 9 | 4100.00 | 80.00 | 60.00 |
| 10 | 4300.00 | 65.00 | 51.00 |

# M.Sc. (Honours) Examination, 2023
# Semester-II
# Statistics
# MSC-21-Inference-II
# Time: 3 hrs Full Marks:40

Answer any **four** questions of the following.

1. (a) Let $X$ be a random variable having density function $f \in \{f_0, f_1\}$ when $f_0(x) = 1$; $0 < x < 1$ and $f_1(x) = \frac{1}{3}$; $0 < x < 3$. For testing $H_0 : f = f_0$ against $H_1 : f = f_1$, based on a single observation find out the power of most powerful test when $\alpha$ (size)= .05.

   (b) Propose a level $\alpha$ MP test for $H_0 : X \sim N(0, 1/2)$ against $H_1 : X \sim Cauchy(0, 1)$ based on a single observation.

   (c) Define a $\alpha$ similar test.

   4+4+2

2. (a) For a nonparametric test of median ($\mu$) being zero under $H_0$, let $X_\alpha, \alpha = 1(1)n$ be a continuous random variable for $\alpha = 1(1)n$. $R_\alpha^+$ is the rank of $|X_\alpha|$. Further define an indicator variable $Z_\alpha$ such that

$$Z_\alpha = \begin{cases} 1 \; if \; X_\alpha > \mu \\ 0 \; if \; X_\alpha < \mu \end{cases}$$

   Show that under $H_0$, $R_\alpha^+$ and $Z_\alpha$ are independently distributed.

   (b) Define a test having Neyman structure with respect to a sufficient statistic $T$.

   (c) Suppose $X$ and $Y$ be the independent Poisson random variables with parameters $\lambda$ and $\mu$ respectively. Propose a UMP test for $H_0 : \mu \leq \lambda$ against $H_1 : \mu > \lambda$.

   3+3+4

3. (a) Let $X_1$, $X_2$ and $X_3$ be collected from $U(\theta, \theta + 2)$. Does this family have Monotone likelihood ratio property? Also construct a UMP test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

   (b) Establish that Kolmogorov-Smirnov one sample test statistic is distribution free.

   5+5

4. (a) Let $X_1, X_2, \cdots, X_n$ be a random sample from $f(x, \lambda) = 2\lambda e^{-\lambda x^2} x$; $x > 0$. Construct a UMP test for testing $H_0 : \lambda = 1$ against $H_1 : \lambda > 1$.

   (b) For an exponential family with single parameter $\theta$ and sufficient statistic $T(x)$, to construct a UMPU test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, show that $E_{\theta_0}[T\phi(T)] = \alpha E_{\theta_0}(T)$ , $\alpha$ being the level of significance.

   5+5

5. (a) Define one sample linear rank statistic associated with one sample location test.

(b) Find its mean and variance.

(c) Show that one sample Wilcoxon signed rank test statistic is a particular case of it.

5+5

6. (a) Suppose $F_X()$ and $F_Y()$ be the probability distribution functions of $X$ and $Y$ respectively. Discuss a nonparametric test in order to check $X$ is stochastically larger than $Y$, clearly stating the null and alternative hypothesis.

(b) Let $P_0$,$P_1$, $P_2$ be the probability distributions assigning to the integers 1, 2, 3, 4, 5 the following probabilities:

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $P_0$ | 0.03 | 0.02 | 0.02 | 0.01 | 0.92 |
| $P_1$ | 0.06 | 0.05 | 0.08 | 0.02 | 0.79 |
| $P_2$ | 0.09 | 0.05 | 0.12 | 0 | 0.74 |

Determine whether there exists a UMP test for $H_0 : P = P_0$ against $H_1 : P \neq P_0$ at $\alpha$(size)= .05.

5+5

**M.Sc. Semester II Examination 2023**
**Subject: Statistics**
Paper: MSC 22
Applied Multivariate Analysis

Full Marks: 40 Time: 3 hours

Answer any four of the following six questions of equal marks.
(Notations carry usual meanings)

1. (a) Derive Principal Components of a $p$- variate random vector. Cite two real-life applications of principal component analysis.
   (b) Prove or disprove- "Factor Model solution always exists".

   6+4

2. (a) Derive an optimum rule for discriminating two populations.
   (b) Let $f_1(x) = (1 - |x|)$ for $|x| \leq 1$, & $f_1(x) = 0$ for other values of $x$ and $f_2(x) = (1 - |x - 0.5|)$ for $-0.5 \leq x \leq 0.5$ & $f_2(x) = 0$ for other values of $x$. Sketch the two densities. Also identify the classification regions to discriminate the two populations (Assume prior probabilities and cost of misclassifications are requal.)

   5+5

3. (a) Write down hierarchical clustering algorithm. Differentiate among single, complete and average linkage clustering methods.
   (b) Compare and contrast canonical correlation analysis (CCA) with multiple regression analysis. Under what circumstances would one choose CCA over multiple regression, and vice versa?

   5+5

4. (a) Find the principal components and proportion of total system variance explained by each when the covariance matrix is given by
   $$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ 0 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$
   where $-\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$.
   (b) What is factor rotation and why is that performed? Explain how factor rotation does not change the factor model representation.

   5+5

5. (a) How does MANOVA differ from the univariate ANOVA (Analysis of Variance) technique? Explain the key advantages of using MANOVA when dealing with multiple dependent variables.
   (b) Interpretation of MANOVA results is essential for understanding the relationships between the independent and dependent variables. Explain how to interpret significant MANOVA findings, including the significance of individual dependent variables.

   4+6

6. (a) Explain how wouldyou obtain solution of a factor model with reasons.
   (b) Write a short note on K-means clustering method.

   5+5

**M.Sc. Examination, 2023**
**Semester-II**
**Statistics**
**Course: MSC-23**
**(Regression Techniques)**
**Time: 3 Hours** **Full Marks: 40**

Questions are of value as indicated in the margin.
Notations have their usual meanings

Answer any four questions.

**1.** Consider the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $E(\epsilon_i) = 0$, $var(\epsilon_i) = \sigma^2$, and and $\epsilon_i$'s are uncorrelated, $i \in \{1, 2, \cdots, n\}$.

(a) Show that $SSR = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}$.

(b) Show that $E(SSR) = \sigma^2 + \beta_1^2 S_{xx}$ and $E(MSE) = \sigma^2$.

(c) Consider the maximum-likelihood estimator $\tilde{\sigma}^2$ of $\sigma^2$. Find the bias in $\tilde{\sigma}^2$.

(d) Prove that the maximum value of $R^2$ is less than 1 if the data contain repeated (different) observations on $y$ at the same value of $x$.

$2 + 3 + 3 + 2$

**2.** (a) Consider the multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Find the expression for the least-squares estimator $\hat{\beta}$ of $\beta$. Show that the least-squares estimator can be written as $\hat{\beta} = \beta + \mathbf{R}\varepsilon$, where $\mathbf{R} = (\mathbf{X^T X})^{-1}\mathbf{X^T}$.

(b) Consider a correctly specified regression model with $p$ terms, including the intercept. Make the usual assumptions about $\epsilon$. Prove that $\sum_{i=1}^{n} Var(\hat{y}_i) = p\sigma^2$.

$5 + 5$

**3.** (a) Write a short note on PRESS Residual and PRESS statistic.

(b) Diagnose if the following statement is True/False with suitable explanation(s).

> A studentized residual ($r_i$) is just a deleted residual $d_i$ divided by its estimated standard deviation $s(d_i)$ (first formula). This turns out to be equivalent to the ordinary residual divided by a factor that includes the mean squared error based on the estimated model with the $i^{th}$ observation deleted, $MSE(i)$, and the leverage, $h_{ii}$ (second formula). In other words, $r_i = \dfrac{d_i}{s(d_i)} = \dfrac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$.

Hence (or otherwise), explain that the studentized residual for a given data point depends not only on the ordinary residual but also on the size of the mean squared error (MSE) and the leverage.

$5 + 5$

**4.** (a) (i) Briefly describe the principle of Logistic Regression and Probit Regression.

(ii) Among two logit models, how do you determine which model is better? Justify.

(b) (i) Describe a formal test for Lack of Fit under the suitable assumption(s).

(ii) Write a short note on influential points and leverage points.

$(2\frac{1}{2} + 2\frac{1}{2}) + (2\frac{1}{2} + 2\frac{1}{2})$

CONTINUED

**5.** (a) Explain how the following problematic scenario (s) can be handled in light of Ridge regression.

(i) The Least Squared (LS) estimate depends upon $(X'X)^{-1}$ , we would have problems in computing $\beta_{LS}$ if $X'X$ were singular or nearly singular.

(ii) In above case(s), a small changes to the elements of $X$ lead to large changes in $(X'X)^{-1}$, and the least squared estimator $\beta_{LS}$ may provide a good fit to the training data, but it may not fit sufficiently well to the test data.

(b) Define the Ridge estimator $\hat{\beta}_{\mathbf{R}}$. Obtain the mean squared error of the Ridge estimator. Justify the following statement:

Ridge estimate will not necessarily provide the best " fit " to the data.

$(2\frac{1}{2} + 2\frac{1}{2}) + 5$

**6.** (a) Compare between Least-Squared estimation in linear regression vs. Least-Squared estimation in nonlinear regression. Illustrate how the linearization can be done by a Taylor series expansion of the nonlinear Regression function, followed by the iteration method of parameter estimation.

(b) Write a short note on R-estimators as Robust Regression for Linear Model.
**(OR)**
Write a short note on M-estimators as Robust Regression for Linear Model.

$5 + 5$

---

**M.Sc. Examination 2023**
**Semester-II**
**Statistics**
**Course: MSC-24 (Design of Experiments)**
**Full Marks: 40** **Time: 3 Hours**

**(Answer any four questions.)**

1. (a) What is a split-plot design? Write down the underlying model, hypotheses and the detailed analysis procedure of the design.
   (b) Find the efficiency of split-plot design with respect to a randomized block design.

   7+3

2. Consider a randomized block design with $v$ blocks and $v$ treatments. Augment treatment $i$ with block $i(i = 1, 2, \ldots v)$.
   (a) Is the resultant design connected?
   (b) Is it orthogonal?
   (c) Is $\tau_2 - 2\tau_3 + \tau_4$ estimable? If so, find the expression of its BLUE and its standard error. You should simplify your answer as much as possible.

   3+3+4

3. Construct BIBD s with the following parameters. You should properly state the results you use in each case.
   (a) $v = 13, b = 26, r = 6, k = 3, \lambda = 1$
   (b) $v = 15, b = 15, r = 7, k = 7, \lambda = 2$
   (c) $v = 9, b = 12, r = 4, k = 3, \lambda = 1$

   3+4+3

4. (a) For a symmetric BIBD with parameters $(v, k, \lambda)$, show that any two blocks have exactly $\lambda$ treatments in common.
   (b) In addition, if $v$ is even, then prove that $(k - \lambda)$ is a perfect square.
   (c) Define the efficiency factor of a BIBD. Prove that it is less than 1.

   4+3+3

5. Derive the inter-block estimate of the contrast $\boldsymbol{p}'\boldsymbol{\tau}$ of the treatment effects and the standard error of the estimate. Show that the estimator is uncorrelated with the intra-block estimator.

   7+3

6. (a) Construct the layout of a $(3^3, 3^2)$ experiment confounding $ABC^2, BC^2$.

(b) Consider the $3^3$ experiment conducted in 2 replications in blocks of $3^2$ plots. The following information is given below.

**Replicate 1:** 100, 112, 202, 211, 220, 121, 010, 021, 002
**Replicate 2:** 001, 102, 012, 110, 200, 121, 222, 211, 020

Identify the confounded effects.

(c) Write down the ANOVA table of a $3^2$ factorial experiment.

4+3+3

---

**M.Sc. Semester II Examination 2023**
**Subject: Statistics**
Paper: MSC 25
(Practical on Applied Multivariate and Inference II )
Full Marks: 40 Time: 4 hours

Answer all the following questions.
(Notations carry usual meanings)

1. Let $X$ and $Y$ be independent distributed Poisson with $\lambda$ and Poisson with $\mu$ respectively. Construct a UMP test for testing $H_0 : \mu \leq \lambda$ against $H_1 : \mu \geq \lambda$ against the alternative $\lambda = .4$ and $\mu = .5$ at level of significance 0.1.

6

2. An urn contains 8 marbles of which $m$ are white and (10-m) are black. To test $H_0 : m = 5$ against $H_1 : m > 5$ one draws 4 marbles without replacement. The null hypothesis is rejected if the sample contains 2 or 3 while balls otherwise accepted. Construct a test and its size.

4

3. Let a random sample of 10 observations, viz., 2,-1.2, 5, 7,-1.2,4, 2.9, 3.9, 2.5,3 are selected from $N(\mu, 16)$. Construct a test for $H_0 : \mu > 3 or \mu < 2$ against $H_0 : 2 < \mu < 3$ for a level of significance .5.

5

4. Consider the following two features of few items.

| *Item* | Feature 1 | Feature 2 |
|---|---|---|
| $A$ | 1.5 | 1.0 |
| $B$ | 2.0 | 1.5 |
| $C$ | 3.0 | 5.0 |
| $D$ | 4.0 | 4.5 |
| $E$ | 3.5 | 4.0 |
| $F$ | 9.0 | 8.5 |
| $G$ | 8.5 | 9.0 |
| $H$ | 8.0 | 8.0 |
| $I$ | 1.0 | 2.0 |

Construct either of hierarchical (show dendrogram) or K-means (maximum 3) clusters.

8

5. The effectiveness of advertising for two rival products (Brand X and Brand Y) was compared. market research at a local shopping center was carried out with the participants being shown adverts for two rival brands of coffee, which they then rated on the overall likelihood of them buying the product (out of 10, with 10 being "'definitely going to buy the product"'). Below is the chart of rating.

| Brand X | | Brand Y | |
|---|---|---|---|
| Participant | Rating | Participant | Rating |
| 1 | 3 | 1 | 9 |
| 2 | 4 | 2 | 7 |
| 3 | 2 | 3 | 5 |
| 4 | 6 | 4 | 10 |
| 5 | 2 | 5 | 6 |
| 6 | 5 | 6 | 8 |

On the basis of this rating do you think these two brands of coffee are the same?

4

6. Suppose you are conducting a study to compare two different teaching methods, Method A and Method B, in terms of their effects on students' score on two subjects: Math and Science. Let you have the following data for a sample of students.

Method A
Math Scores:[85, 92, 78, 88, 76]
Science Scores:[78, 86, 82, 75, 80]
Method B
Math Scores:[92, 88, 75, 90, 85]
Science Scores:[85, 80, 88, 82, 78]

Assuming equal sample sizes for both methods, perform a one-way MANOVA to determine if there are any significant differences between the teaching methods in terms of the combined Math and Science scores.

8

7. Viva-voce and Practical Notebook

5

**M.Sc. Examination, 2023**
**Semester-II**
**Statistics**
**Course: MSC-26 (Practical)**
**Time: 4 Hours** **Full Marks: 40**

**1.** Consider the following incomplete block design (table 1). The yields are given as the entries of the table 1. Perform a complete intra-block analysis. Also find an estimate of the error variance. You may use R software for the calculations. 12

| Treatment Block | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | ··· | 16 | 30 | 25 | ··· |
| 2 | 5 | 10 | 18 | ··· | ··· |
| 3 | 7 | 28 | ··· | ··· | 35 |
| 4 | 10 | ··· | ··· | 20 | 52 |
| 5 | ··· | ··· | 24 | 11 | 40 |
| 6 | 12 | 24 | 29 | ··· | ··· |

Table 1: incomplete block design (The yields are given as the entries of the table)

**2.** The following table (table 2) gives the plants and yields (in suitable units) of a manurial experiment involving two factors N and P each at 3 levels.
Test for the significance of main effects and 2-factor interaction effects. 6

Replicate 1:

| Treatment | 12 | 00 | 10 | 21 | 22 | 01 | 11 | 20 | 02 |
|---|---|---|---|---|---|---|---|---|---|
| Yield | 223 | 236 | 240 | 300 | 189 | 160 | 284 | 271 | 259 |

Replicate 2:

| Treatment | 01 | 20 | 12 | 00 | 22 | 11 | 02 | 21 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| Yield | 269 | 233 | 266 | 213 | 226 | 240 | 282 | 209 | 293 |

Replicate 3:

| Treatment | 02 | 11 | 01 | 21 | 22 | 10 | 12 | 00 | 20 |
|---|---|---|---|---|---|---|---|---|---|
| Yield | 191 | 300 | 278 | 209 | 226 | 233 | 182 | 270 | 258 |

Table 2: plants and yields (in suitable units)

CONTINUED

**3.** The kinematic viscosity of a certain solvent system depends on the ratio of the two solvents and the temperature (Data table $b.10$ of Table 3). You may attach the data using the following R code:

```
library("MPV")
data(table.b10)
```

(a) Fit a multiple linear regression model relating the viscosity to the two regressors.
(b) Test for significance of the regression. What conclusions can you draw?
(c) Use t-tests to assess the contribution of each regressor to the model. Discuss your findings.
(d) Calculate $R^2$ and $R^2_{Adj}$ for this model. Compare these values to the $R^2$ and $R^2_{Adj}$ for the simple linear regression model relating the viscosity to temperature only.
(e) Find a 99% C.I for the regression coefficient for temperature for both models in part d. Discuss any differences.

5

**4.** A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine their oldest vehicle's age and total family income. A follow-up survey was conducted 6 months later to determine if they had actually purchased a new vehicle during that time period ( $y = 1$ indicates yes, and $y = 0$ indicates no). You may attach the data in Table 3 using the following R code:

```
library("MPV")
data(p13.5)
```

(a) Fit a logistic regression model to the data and interpret the model coefficients $\beta_1$ and $\beta_2$.
(b) Does the model deviance indicate that the logistic regression model is adequate?
(c) What is the estimated probability that a family with an income of \$45,000 and a car that is 5 years old will purchase a new vehicle in the next 6 months?
(d) Expand the linear predictor to include an interaction term. Is there any evidence that this term is required in the model?
(e) Find approximate 95% confidence intervals on the model parameters for the logistic regression model.

5

**5.** Hald cement data: The response variable $y$ is the heat evolved in a cement mix. The four explanatory variables are ingredients of the mix, i.e., $x_1$: tricalcium aluminate, $x_2$: tricalcium silicate, $x_3$: tetracalcium alumino ferrite, $x_4$: dicalcium silicate. You may attach the data by using the following R code (or in Table 3):

```
library("BAS");
 data(Hald);
```

(a) Using the Hald cement data, find the eigenvector associated with the smallest eigenvalue of $X^TX$. Interpret the elements of this vector.
(b) What can you say about the source of multicollinearity in these data?

2+2

**6.** Analyse the chemical process data in Table b.5 (to be found in the "MPV" package of R, or in Table 3) for evidence of multicollinearity. Use the variance inflation factors and the condition number of $X^TX$.
You may attach the data by using the following R code:

```
library("MPV")
data(table.b5)
```

3

**7.** Practical Notebook & Viva voce 5

ATTACHMENT FOLLOWS

ATTACHMENT

| | y | x1 | x2 | x3 | x4 |
|---|---|---|---|---|---|
| 1 | 78.50 | 7.00 | 26.00 | 6.00 | 60.00 |
| 2 | 74.30 | 1.00 | 29.00 | 15.00 | 52.00 |
| 3 | 104.30 | 11.00 | 56.00 | 8.00 | 20.00 |
| 4 | 87.60 | 11.00 | 31.00 | 8.00 | 47.00 |
| 5 | 95.90 | 7.00 | 52.00 | 6.00 | 33.00 |
| 6 | 109.20 | 11.00 | 55.00 | 9.00 | 22.00 |
| 7 | 102.70 | 3.00 | 71.00 | 17.00 | 6.00 |
| 8 | 72.50 | 1.00 | 31.00 | 22.00 | 44.00 |
| 9 | 93.10 | 2.00 | 54.00 | 18.00 | 22.00 |
| 10 | 115.90 | 21.00 | 47.00 | 4.00 | 26.00 |
| 11 | 83.80 | 1.00 | 40.00 | 23.00 | 34.00 |
| 12 | 113.30 | 11.00 | 66.00 | 9.00 | 12.00 |
| 13 | 109.40 | 10.00 | 68.00 | 8.00 | 12.00 |

(a) Hald Dataset

| | x1 | x2 | y |
|---|---|---|---|
| 1 | 0.92 | -10.00 | 3.13 |
| 2 | 0.92 | 0.00 | 2.43 |
| 3 | 0.92 | 10.00 | 1.94 |
| 4 | 0.92 | 20.00 | 1.59 |
| 5 | 0.92 | 30.00 | 1.32 |
| 6 | 0.92 | 40.00 | 1.13 |
| 7 | 0.92 | 50.00 | 0.97 |
| 8 | 0.92 | 60.00 | 0.85 |
| 9 | 0.92 | 70.00 | 0.75 |
| 10 | 0.92 | 80.00 | 0.67 |
| 11 | 0.75 | -10.00 | 2.27 |
| 12 | 0.75 | 0.00 | 1.82 |
| 13 | 0.75 | 10.00 | 1.49 |
| 14 | 0.75 | 20.00 | 1.25 |
| 15 | 0.75 | 30.00 | 1.06 |
| 16 | 0.75 | 40.00 | 0.92 |
| 17 | 0.75 | 50.00 | 0.80 |
| 18 | 0.75 | 60.00 | 0.71 |
| 19 | 0.75 | 70.00 | 0.63 |
| 20 | 0.75 | 80.00 | 0.57 |
| 21 | 0.57 | -10.00 | 1.59 |
| 22 | 0.57 | 0.00 | 1.32 |
| 23 | 0.57 | 10.00 | 1.12 |
| 24 | 0.57 | 20.00 | 0.96 |
| 25 | 0.57 | 30.00 | 0.83 |
| 26 | 0.57 | 40.00 | 0.73 |
| 27 | 0.57 | 50.00 | 0.65 |
| 28 | 0.57 | 60.00 | 0.58 |
| 29 | 0.57 | 70.00 | 0.52 |
| 30 | 0.57 | 80.00 | 0.47 |
| 31 | 0.36 | -10.00 | 1.16 |
| 32 | 0.36 | 0.00 | 0.99 |
| 33 | 0.36 | 10.00 | 0.86 |
| 34 | 0.36 | 20.00 | 0.75 |
| 35 | 0.36 | 30.00 | 0.67 |
| 36 | 0.36 | 40.00 | 0.59 |
| 37 | 0.36 | 50.00 | 0.53 |
| 38 | 0.36 | 60.00 | 0.48 |
| 39 | 0.36 | 70.00 | 0.44 |
| 40 | 0.36 | 80.00 | 0.40 |

(b) Table b.10

| | x1 | x2 | y |
|---|---|---|---|
| 1 | 45000.00 | 2.00 | 0.00 |
| 2 | 40000.00 | 4.00 | 0.00 |
| 3 | 60000.00 | 3.00 | 1.00 |
| 4 | 50000.00 | 2.00 | 1.00 |
| 5 | 55000.00 | 2.00 | 0.00 |
| 6 | 37000.00 | 5.00 | 1.00 |
| 7 | 31000.00 | 7.00 | 1.00 |
| 8 | 40000.00 | 4.00 | 1.00 |
| 9 | 75000.00 | 2.00 | 0.00 |
| 10 | 43000.00 | 9.00 | 1.00 |
| 11 | 50000.00 | 5.00 | 1.00 |
| 12 | 35000.00 | 7.00 | 1.00 |
| 13 | 65000.00 | 2.00 | 1.00 |
| 14 | 53000.00 | 2.00 | 0.00 |
| 15 | 48000.00 | 1.00 | 0.00 |
| 16 | 49000.00 | 2.00 | 0.00 |
| 17 | 37500.00 | 4.00 | 1.00 |
| 18 | 71000.00 | 1.00 | 0.00 |
| 19 | 34000.00 | 5.00 | 0.00 |
| 20 | 27000.00 | 6.00 | 0.00 |

(c) p13.5

| | y | x1 | x2 | x3 | x4 | x5 | x6 | x7 |
|---|---|---|---|---|---|---|---|---|
| 1 | 36.98 | 5.10 | 400.00 | 51.37 | 4.24 | 1484.83 | 2227.25 | 2.06 |
| 2 | 13.74 | 26.40 | 400.00 | 72.33 | 30.87 | 289.94 | 434.90 | 1.33 |
| 3 | 10.08 | 23.80 | 400.00 | 71.44 | 33.01 | 320.79 | 481.19 | 0.97 |
| 4 | 8.53 | 46.40 | 400.00 | 79.15 | 44.61 | 164.76 | 247.14 | 0.62 |
| 5 | 36.42 | 7.00 | 450.00 | 80.47 | 33.84 | 1097.26 | 1645.89 | 0.22 |
| 6 | 26.59 | 12.60 | 450.00 | 89.90 | 41.26 | 605.06 | 907.59 | 0.76 |
| 7 | 19.07 | 18.90 | 450.00 | 91.48 | 41.88 | 405.37 | 608.05 | 1.71 |
| 8 | 5.96 | 30.20 | 450.00 | 98.60 | 70.79 | 253.70 | 380.55 | 3.93 |
| 9 | 15.52 | 53.80 | 450.00 | 98.05 | 66.82 | 142.27 | 213.40 | 1.97 |
| 10 | 56.61 | 5.60 | 400.00 | 55.69 | 8.92 | 1362.24 | 2043.36 | 5.08 |
| 11 | 26.72 | 15.10 | 400.00 | 66.29 | 17.98 | 507.65 | 761.48 | 0.60 |
| 12 | 20.80 | 20.30 | 400.00 | 58.94 | 17.79 | 377.60 | 566.40 | 0.90 |
| 13 | 6.99 | 48.40 | 400.00 | 74.74 | 33.94 | 158.05 | 237.08 | 0.63 |
| 14 | 45.93 | 5.80 | 425.00 | 63.71 | 11.95 | 130.66 | 1961.49 | 2.04 |
| 15 | 43.09 | 11.20 | 425.00 | 67.14 | 14.73 | 682.59 | 1023.89 | 1.57 |
| 16 | 15.79 | 27.90 | 425.00 | 77.65 | 34.49 | 274.20 | 411.30 | 2.38 |
| 17 | 21.60 | 5.10 | 450.00 | 67.22 | 14.48 | 1496.51 | 2244.77 | 0.32 |
| 18 | 35.19 | 11.70 | 450.00 | 81.48 | 29.69 | 652.43 | 978.64 | 0.44 |
| 19 | 26.14 | 16.70 | 450.00 | 83.88 | 26.33 | 458.42 | 687.62 | 8.82 |
| 20 | 8.60 | 24.80 | 450.00 | 89.38 | 37.98 | 312.25 | 468.38 | 0.02 |
| 21 | 11.63 | 24.90 | 450.00 | 79.77 | 25.66 | 307.08 | 460.62 | 1.72 |
| 22 | 9.59 | 39.50 | 450.00 | 87.93 | 22.36 | 193.61 | 290.42 | 1.88 |
| 23 | 4.42 | 29.00 | 450.00 | 79.50 | 31.52 | 155.96 | 233.95 | 1.43 |
| 24 | 38.89 | 5.50 | 460.00 | 72.73 | 17.86 | 1392.08 | 2088.12 | 1.35 |
| 25 | 11.19 | 11.50 | 450.00 | 77.88 | 25.20 | 663.09 | 994.63 | 1.61 |
| 26 | 75.62 | 5.20 | 470.00 | 75.50 | 8.66 | 1464.11 | 2196.17 | 4.78 |
| 27 | 36.03 | 10.60 | 470.00 | 83.15 | 22.39 | 720.07 | 1080.11 | 5.88 |

(d) Table b.5

Table 3: The datasets: (only for reference and are **NOT recommended to be typed manually** in R). Students may use the pre-installed R packages and R codes mentioned in the question to retrieve the datasets quickly and save time.

CONTINUED