1. (a) Four specimens of each of five brands of a synthetic wood veneer material were subjected to a friction test. A measure of wear (in appropriate units) was determined for each specimen, resulting in the following data.

| Brand $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| | 2.3 | 2.2 | 2.2 | 2.4 | 2.3 |
| | 2.1 | 2.3 | 2.0 | 2.7 | 2.5 |
| | 2.4 | 2.4 | 1.9 | 2.6 | 2.3 |
| | 2.5 | 2.6 | 2.1 | 2.7 | 2.4 |
| Mean $\overline{y_{i\cdot}}$ | 2.325 | 2.375 | 2.050 | 2.600 | 2.375 |
| Standard deviation $s_i$ | 0.171 | 0.171 | 0.129 | 0.141 | 0.096 |

Carry out an analysis of variance to test for differences in mean wear for the five brands and comment on your results.

[12 marks]

(b) Consider the one-way ANOVA model for three groups, and suppose that for $i = 1, 2, 3$, group $i$ consists of $n_i$ observations. Writing the model in the form

$$\begin{aligned} Y_{1j} &= \mu + \alpha_1 + \epsilon_{1j} && (j = 1, 2, \ldots, n_1) \\ Y_{2j} &= \mu + \alpha_2 + \epsilon_{2j} && (j = 1, 2, \ldots, n_2) \\ Y_{3j} &= \mu - (\alpha_1 + \alpha_2) + \epsilon_{3j} && (j = 1, 2, \ldots, n_3) \end{aligned}$$

and considering this as a general linear model of the form $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, write down the components of the vector $\boldsymbol{\beta}$ and the design matrix $X$ in this case.

Find a condition under which the first column of $X$ is orthogonal to each of the other two columns of $X$. What implications does this have for parameter estimation? Write down the matrix $X^T X$ in this case.

[8 marks]

2. (a) Consider the general linear model, with

$$\boldsymbol{Y} \;=\; X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $X$ is a known $n \times p$ matrix of rank $p$, $\boldsymbol{\beta}$ is a vector of $p$ unknown parameters, and $\boldsymbol{\epsilon}$ is a random vector whose components are independent Normal random variables, each having mean zero and variance $\sigma^2$.

Show that the least squares estimator $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$, and derive an expression for the variance of $\hat{\boldsymbol{\beta}}$.

[You may assume without proof that for the general linear model, least squares estimates are given by $\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \boldsymbol{Y}$.]

[8 marks]

Write down an expression which may be used to estimate the unknown error variance $\sigma^2$.

[2 marks]

(b) A general linear model of the form $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is analysed in SAS using the following program.

```
data small;
input x1 x2 x3 y;
cards;
1 1 1 7
2 3 4 13
2 -2 2 5
4 3 2 21
;
proc glm data=small;
model y=x1 x2 x3 /noint;                        (*)
contrast 'Label1' x1 1 x2 -2 x3 -2;             (**)
run;
quit;
```

**Question 2 continued overleaf**

Write down the $X$ matrix for this model and the observed response values $\boldsymbol{y}$. Explain the meaning of the lines labelled (*) and (**) in the SAS program.

[6 marks]

Some of the output from this program is given below.

| Contrast | DF | Contrast SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Label1 | 1 | 0.36765036 | 0.36765036 | 0.22 | 0.7217 |

| Parameter | Estimate | T for H0: Parameter=0 | Pr > \|T\| | Std Error of Estimate |
|---|---|---|---|---|
| X1 | 4.059210526 | 8.25 | 0.0768 | 0.49232334 |
| X2 | 1.618421053 | 4.53 | 0.1382 | 0.35696474 |
| X1 | 0.059210526 | 0.12 | 0.9238 | 0.49232334 |

What conclusions can be drawn about the values of the parameters of the model?

[4 marks]

3. (a) Define the *rank* of a matrix. Describe briefly the importance of rank in the analysis of general linear models.

[4 marks]

(b) In a study of the relationship between the price of oranges and sales-per-customer, data were collected from 3 stores over 6 consecutive Saturdays. Sales-per-customer $Y_{ij}$ and price $x_{ij}$ (pence per lb) were recorded for each store $i$ on each of the days $j$ of the study, giving a total of 18 observations. The data may be modelled by the relationship

$$Y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij} \qquad (\Omega)$$

for $i = 1, 2, 3,\ j = 1, 2, 3, 4, 5, 6$.

Fitting model $(\Omega)$ to the data using SAS, parameter estimates were found to be $\hat{\alpha}_1 = 41.93$, $\hat{\alpha}_2 = 45.89$, $\hat{\alpha}_3 = 42.39$, $\hat{\beta} = -0.668$, with residual Sum of Squares $SS_\Omega = 271.43$.

(i) To test the hypothesis that differences between the 3 stores have no effect upon sales, the following reduced model was also fitted to the data

$$Y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij} \qquad (\omega_1)$$

Parameter estimates were $\hat{\alpha} = 38.16$, $\hat{\beta} = -0.563$, with residual Sum of Squares $SS_{\omega_1} = 319.60$.

Test the hypothesis that differences between stores have no effect upon sales.

[6 marks]

(ii) Next, to see whether price affects sales, the following model was fitted.

$$Y_{ij} = \alpha_i + \epsilon_{ij} \qquad (\omega_2)$$

Parameter estimates were $\hat{\alpha}_1 = 12.33$, $\hat{\alpha}_2 = 10.50$, $\hat{\alpha}_3 = 7.00$, with residual Sum of Squares $SS_{\omega_2} = 462.83$.

Test the hypothesis that price has no effect upon sales.

[6 marks]

(iii) Comment on your results. Which is the preferred model?

[3 marks]

For your preferred model, estimate the sales per customer of oranges in store 1 on a day when the price is 40 pence per lb.

[1 mark]

4. Three growth promoting methods $i = 1, 2, 3$ were applied to seeds from each of four varieties $j = 1, 2, 3, 4$ of turf grass, and 4 observations taken at each of the $3 \times 4$ possible factor combinations, giving 48 observations in total. Group mean yields $\overline{y_{ij\cdot}}$ are given below.

| Variety $j$ / Method $i$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 21.9 | 22.9 | 24.8 | 25.6 |
| 2 | 11.4 | 16.0 | 15.6 | 11.6 |
| 3 | 18.9 | 20.3 | 18.0 | 14.8 |

(i) Give a graphical representation of the group means which shows the relative importance of the two main effects and the interaction between them, and comment on your plot.

[6 marks]

(ii) Complete and interpret the analysis of variance table for these data presented below.

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Method | 825.7 | | | |
| Variety | 61.6 | | | |
| Interaction | 114.3 | | | |
| Residual | | | | |
| Total | 1666.3 | | | |

[12 marks]

(iii) What further data might it be worthwhile to collect?

[2 marks]

5. A random variable $Y$ with a single parameter $\theta$ belongs to the exponential family in canonical form if its probability mass function (or probability density function) can be written in the form

$$f_Y(y;\theta) \quad = \quad \exp\{yb(\theta)+c(\theta)+d(y)\}.$$

(a) If $Y$ is a Normal random variable with unknown mean $\mu$ and known standard deviation $\sigma > 0$, so that

$$f_Y(y;\mu) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} \qquad (-\infty < y < \infty)$$

then show that $Y$ belongs to the exponential family in canonical form.

[3 marks]

(b) Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent Normal random variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and common (known) standard deviation $\sigma$. Write down the log-likelihood of $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ in terms of the observed response values $y_1, y_2, \ldots, y_n$.

[2 marks]

Suppose now that the $\mu_i$ are given by

$$\mu_i \quad = \quad \boldsymbol{x}_i^T\boldsymbol{\beta}$$

for some unknown parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ and known vectors of explanatory variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ .
Write down the log-likelihood of $\boldsymbol{\beta}$.

[2 marks]

Denoting by $X$ the matrix with columns $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, and assuming that $X$ is of full rank $p$, show that the maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ are given in terms of observed response values $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ by

$$\widehat{\boldsymbol{\beta}} \quad = \quad \left(X^TX\right)^{-1}X^T\boldsymbol{y}$$

[9 marks]

(c) Suppose now that $Y$ is a Normal random variable with *known* mean $\mu$ and *unknown* standard deviation $\sigma$. Is it possible for the density $f_Y(y;\sigma)$ to be written as a member of the exponential family in canonical form? Explain your answer.

[4 marks]

6. In a five-year study of road traffic accidents, the number of accidents on a particular stretch of road was recorded for each consecutive six-month period. The data are given below.

| Time period $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of accidents $y_i$ | 2 | 5 | 7 | 12 | 10 | 10 | 8 | 9 | 12 | 12 |

Assume that the numbers of accidents $Y_1, Y_2, \ldots, Y_{10}$ are independent Poisson random variables with means $\mu_1, \mu_2, \ldots, \mu_{10}$.

(i) It is conjectured that the average number of accidents varies linearly over time, so that for $i = 1, 2, \ldots, 10$,

$$\mu_i \quad = \quad \alpha + \beta i \qquad\qquad (*)$$

for unknown parameters $\alpha$, $\beta$.

Fitting model $(*)$ in SAS produced maximum likelihood estimates $\hat{\alpha} = 3.3063$, $\hat{\beta} = 0.9807$, with deviance $D = 5.7244$. Does model $(*)$ provide a good fit to the data?

[4 marks]

(ii) To see whether there is any evidence of a change in mean accident levels over time, the model

$$\mu_i \quad = \quad \alpha \qquad\qquad (**)$$

was also fitted to the data, giving $\hat{\alpha} = 8.7000$ with deviance $D = 13.5293$.

Which of models $(*)$ and $(**)$ would be preferred for these data?

[4 marks]

(iii) As an alternative to linear variation, it is suggested that the mean number of accidents may vary exponentially with time. Thus we now consider the model

$$\mu_i \quad = \quad \exp\{\alpha + \beta i\} \qquad\qquad (***)$$

In this case, we find maximum likelihood estimates are $\hat{\alpha} = 1.5917$, $\hat{\beta} = 0.0969$, with deviance $D = 6.9409$.

**Question 6 continued overleaf**

Again, we wish to test for evidence of a change in mean accident levels over time by comparing model $(***)$ with model $(**)$. Which of these two models should be preferred?

[4 marks]

(iv) We also want to decide whether a linear or exponential trend provides a better description of the data, by comparing models $(*)$ and $(***)$. Of these two models, which should be preferred?

[4 marks]

(v) Using your preferred model (amongst the three available), estimate the number of accidents expected to occur during the year following the end of the study.

[4 marks]

7. Suppose $Y_1, Y_2, \ldots, Y_n$ are independent exponential random variables with means $\mu_1, \mu_2, \ldots, \mu_n > 0$, so that the density of $Y_i$ is given by

$$f(y_i; \mu_i) = \frac{\exp\{-y_i/\mu_i\}}{\mu_i} \qquad (y_i \geq 0)$$

(i) Write down an expression for the log-likelihood of $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ in terms of observed values $y_1, y_2, \ldots, y_n$, and derive expressions for the maximum likelihood estimates $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n$.

[6 marks]

(ii) Suppose now that $\mu_i = \alpha i$ $(i = 1, 2, \ldots, n)$ for some unknown parameter $\alpha$. Show that the log-likelihood of $\alpha$ is given by

$$l(\alpha) = -n\ln(\alpha) - \frac{1}{\alpha}\sum_{i=1}^{n}\left(\frac{y_i}{i}\right) - \sum_{i=1}^{n}\ln(i)$$

and hence find an expression for the maximum likelihood estimate $\widehat{\alpha}$.

[6 marks]

(iii) Find an expression for the deviance for comparing this model with the maximal model, simplifying your expression as far as possible.

[8 marks]



/End