1. When pesticides are applied to vegetable crops intended for human consumption, solvents may be used to extract the pesticides before the vegetables are sent to market. In a study designed to test the effectiveness of a partiular solvent, three different concentrations of solvent were used, each being applied to 8 batches of radishes. The amount of pesticide left in the radishes after extraction was recorded for each of the 24 batches, giving the following data.

Solvent concentration i	Low ($i = 1$)	Medium ($i = 2$)	High ($i = 3$)
	25.1	20.9	19.7
	23.7	22.5	18.8
	21.0	19.5	18.1
	22.1	20.7	17.3
	27.7	20.2	18.1
	23.8	20.8	19.3
	23.4	20.4	18.9
	24.6	23.3	18.7
Mean $\overline{y_{i\cdot}}$	23.925	21.038	18.612
Standard deviation s_i	2.011	1.249	0.759

Specify a model that describes these data and can be used to test for differences in effectiveness of the three levels of solvent concentration.

[2 marks]

What assumptions must be made about the data, and how could these assumptions be checked?

[4 marks]

Carry out an analysis of variance to test for differences in effectiveness.

[10 marks]

Comment on your results.

[4 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 σ^2.

(i) Show that the least squares estimator $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$;

(ii) Show that $\text{Var}\left[\widehat{\boldsymbol{\beta}}\right] = \sigma^2 \left(X^T X\right)^{-1}$;

(iii) Derive the distribution of $\widehat{\boldsymbol{\beta}}$.

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

[10 marks]

(b) Consider now the case of simple linear regression, where

$$Y_i \;=\; \beta_1 + \beta_2 x_i + \epsilon_i \qquad\qquad (i = 1, 2, \ldots, n).$$

Write down the matrix X in this case, find $\left(X^T X\right)^{-1}$, and hence derive formulae for the least squares estimates $\widehat{\beta}_1$, $\widehat{\beta}_2$ in terms of the observed data.

[10 marks]

3. In a study of the use of drugs in the treatment of leprosy, two drugs were compared. Ten patients were treated, five patients receiving Drug 1, the remaining five patients receiving Drug 2. For each patient the numbers of leprosy bacilli present were measured before and after treatment, resulting in the following data.

Patient number (i)	Drug (j)	Bacilli before (x_i)	Bacilli after (Y_i)
1	1	10	6
2	1	7	0
3	1	4	2
4	1	13	8
5	1	16	11
6	2	7	2
7	2	9	4
8	2	8	2
9	2	20	14
10	2	6	1

It is proposed to model these data by the relationship

$$Y_i \;=\; \alpha_j + \beta\,(x_i - \bar{x}) + \epsilon_i \qquad\qquad (\Omega)$$

for $i = 1, 2, \ldots, 10$.

(i) Interpret the model (Ω) in non-statistical terms.

[2 marks]

(ii) The model (Ω) may be expressed in the form $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\alpha_1, \alpha_2, \beta)^T$. Calculate the value of $\bar{x}$, and hence write down the matrix X.

[3 marks]

Question 3 continued overleaf

(iii) Compute least squares parameter estimates $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\beta}$.

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

[6 marks]

(iv) Consider the simplified model

$$Y_i = \alpha + \beta\,(x_i - \bar{x}) + \epsilon_i \qquad (\omega)$$

for $i = 1, 2, \ldots, 10$.

Fitting model (Ω) to the data using SAS gives residual Sum of Squares $SS_\Omega = 12.58$. In the case of model (ω), the residual Sum of Squares is found to be $SS_\omega = 14.18$. Perform a hypothesis test to determine which of these two models should be preferred for these data.

[6 marks]

Explain in non-statistical terms the meaning of your result.

[1 mark]

(v) Consider now the model

$$Y_i = \alpha_j + \epsilon_i \qquad (\omega')$$

for $i = 1, 2, \ldots, 10$.

Without performing any further calculations, write down the values of the least squares parameter estimates $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ for model (ω'), explaining the reasoning behind your answer.

[2 marks]

4. (a) Explain what is meant by (i) a balanced factorial experiment; (ii) a Latin square design.

[6 marks]

Outline briefly the relative advantages and disadvantages of a Latin square design versus a factorial experiment.

[4 marks]

(b) An experiment was carried out to compare the growth rates of seeds from three different suppliers. The seeds were grown in three special cabinets in each of which there were three locations (left, centre and right). The resulting data were analysed using the following SAS program.

```
data seeds;
input cabinet location supplier yield;
cards;
1 1 1 35
1 2 3 46
1 3 2 33
2 1 3 40
2 2 2 42
2 3 1 41
3 1 2 37
3 2 1 41
3 3 3 39
;
proc anova data=seeds;
class cabinet location supplier;
model yield = cabinet location supplier;
run;
quit;
```

Name the type of design used for this experiment.

[2 marks]

Question 4 continued overleaf

Some of the output from the above program is given below.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	?	103.33333333	?	?	?
Error	?	18.66666667	?		
Corrected Total	?	122.00000000			

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CABINET	?	14.00000000	?	?	?
LOCATION	?	60.66666667	?	?	?
SUPPLIER	?	28.66666667	?	?	?

Test the hypothesis that there are no differences between the three suppliers.

[8 marks]

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

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

If l denotes the log of the likelihood function, let

$$u = \frac{dl}{d\theta} = l'.$$

Given that $\mathbf{E}\left[u(Y)\right] = 0$, show that

$$\mathbf{E}\left[Y\right] = -\frac{c'(\theta)}{b'(\theta)}. \qquad (*)$$

[2 marks]

Show that if $Y \sim \text{Bin}\,(n,\theta)$ then Y belongs to the exponential family in canonical form, and hence use equation $(*)$ above to confirm that $\mathbf{E}[Y] = n\theta$.

[4 marks]

(b) Explain what is meant by the term *generalised linear model*, including the definition of the *link function* in your explanation. Define also the term *maximal model* in this context.

[8 marks]

For a generalised linear model with a binary response variable, give four possible link functions, outlining their relative advantages and disadvantages.

[6 marks]

6. A popular 'old wives' tale' is that if it rains on Saint Swithin's Day (July 15th), then it will rain for the next 40 days and nights. For each of 12 years ($i = 1, 2, \ldots, 12$) the following data were recorded.

$$x_i = \begin{cases} 0 & \text{if Saint Swithin's Day was rainy,} \\ 1 & \text{if Saint Swithin's Day was dry,} \end{cases}$$

Y_i = time (days) before the first completely dry day after Saint Swithin's Day.

Treating 'time' as a continuous random variable, it is proposed to model these data using a generalised linear model with exponential response distribution and log link, so that for $i = 1, 2, \ldots, 12$ the density of Y_i is given by

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

where the parameters μ_i satisfy

$$\ln(\mu_i) = \alpha + \beta x_i. \qquad (*)$$

Write down an expression for the log-likelihood $l(\alpha, \beta)$.

[3 marks]

The data observed are given below, Y_i values being recorded to the nearest day.

x_i	0	0	0	0	0	0	0	1	1	1	1	1
Y_i	2	12	5	18	7	8	1	6	1	1	9	8

By substituting these values into your expression for $l(\alpha, \beta)$, show that

$$l(\alpha, \beta) = -12\alpha - 5\beta - 53e^{-\alpha} - 25e^{-\alpha}e^{-\beta}$$

[2 marks]

Hence derive two equations satisfied by the maximum likelihood estimates $\widehat{\alpha}$, $\widehat{\beta}$ and solve to find the values of these maximum likelihood estimates.

[8 marks]

Question 6 continued overleaf

The deviance for model $(*)$ can be calculated to be $D = 8.892$. Test whether model $(*)$ provides a good fit to the data.

[3 marks]

To see whether weather conditions on Saint Swithin's Day affect the response, the model

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

was also fitted to the data, and found to have a deviance $D = 9.379$. Which of the models $(*)$ or $(**)$ would be preferred?

[3 marks]

In non-statistical terms, what is your conclusion?

[1 mark]

7. (a) What is a *contingency table*?

[2 marks]

If the data in a contingency table are to be analysed as a generalised linear model, then which probability distribution and which link function should be used?

[2 marks]

In testing for an interaction between the two factors in a contingency table, two different test statistics may be used. Write down the formulae for these two statistics, defining clearly all terms used. How are the values of the two statistics related to one another?

[6 marks]

(b) It is desired to classify defects found on furniture produced in a certain factory according to (1) type of defect and (2) time of day produced. Furniture defects are classified as one of four types A, B, C, D and 'time of day' is classified into three periods 1, 2, 3. A total of 309 furniture defects were recorded, resulting in the following data.

Type of defect / Time period	A	B	C	D	Total
1	15	21	45	13	94
2	26	31	34	5	96
3	33	17	49	20	119
Total	74	69	128	38	309

Calculate the deviance value for these data, and hence test whether there is evidence of any relationship between type of defect and time of production.

[8 marks]

What further analysis is appropriate when there is evidence of a relationship?

[2 marks]