1. Explain briefly the basic idea of one-way analysis of variance, and specify the model used. What assumptions must be made about the data?

   Show that the total sum of squares can be partitioned into a between-groups and a within-groups component.

   To investigate how the yield of a chemical experiment varies with the pressure of the reactor vessel, an experiment was carried out. Five independent trials were carried out at each of four equally spaced pressure values, and the yields recorded were as below.

Pressure	2000 mb	2500 mb	3000 mb	3500 mb
	36.5	34.7	46.3	43.4
	42.6	39.7	42.1	38.7
	38.3	43.4	39.4	41.3
	32.4	39.4	40.5	40.2
	37.4	41.3	42.7	39.6
Mean	37.44	39.70	41.98	40.20
Standard deviation	3.66	3.21	2.95	1.88
$\sum y^2$	7062.4	7921.8	8846.5	8094.3

   Carry out an analysis of variance to test whether the mean yield varies with pressure.

2. 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.

Show that the least squares estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} \;=\; \left(X^T X\right)^{-1} X^T \boldsymbol{Y}$$

and that $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

Suppose now that for a given $r \times p$ matrix A and a given $r \times 1$ vector $\boldsymbol{c}$ we wish to test the hypothesis that $A\boldsymbol{\beta} = \boldsymbol{c}$. Denoting by Ω the unconstrained model, and by ω the model with the constraint $A\boldsymbol{\beta} = \boldsymbol{c}$, define the residual Sums of Squares SS_Ω and SS_ω, and explain how they may be used to test the null hypothesis $\mathrm{H}_0 : A\boldsymbol{\beta} = \boldsymbol{c}$.

In an experiment to investigate how the yield of soya beans was affected by the distance between the rows, data were collected on the average yield of beans (Y) obtained using nine different row spacings (x). The data so obtained were as follows.

Row spacing (cm)	18	21	24	27	30	33	36	39	42
Yield	2.41	2.58	2.83	3.01	3.13	3.11	2.95	2.82	2.64

Fitting the cubic model

$$Y_i \;=\; \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i \qquad (i = 1, 2, \ldots, 9)$$

to these data, the least squares parameter estimates were found to be

$$\hat{\beta}_0 = -0.220, \quad \hat{\beta}_1 = 0.180, \quad \hat{\beta}_2 = -0.00147, \quad \hat{\beta}_3 = -2.87 \times 10^{-5}$$

with residual Sum of Squares equal to 0.01488.

The constraint $\beta_3 = 0$ was then imposed, and the resulting simplified (quadratic) model fitted to the same data, giving least squares parameter estimates

Question 2 continued overleaf

$$\hat{\beta}_0 = -0.900, \quad \hat{\beta}_1 = 0.254, \quad \hat{\beta}_2 = -0.00405$$

with residual Sum of Squares equal to 0.01573.

Use the above information to test the hypothesis that $\beta_3 = 0$.

For your preferred model, estimate the expected yield for a row spacing of 20 cm.

3. 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, each having mean zero and variance σ^2.

Discuss the meaning and implications of *orthogonality* for this model. Your discussion should include (but not be restricted to) a proof that if the columns of the design matrix X satisfy certain orthogonality conditions, then imposing the constraint $\beta_1 = \beta_2 = \cdots = \beta_q = 0$ for some $q < p$ does not affect the least squares estimates of the remaining parameters $\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p$.

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

Suppose we wish to estimate three parameters $\beta_1, \beta_2, \beta_3$ from four observations Y_1, Y_2, Y_3, Y_4 where

$$\begin{aligned} Y_1 &= \beta_1 + \beta_3 + \epsilon_1, \\ Y_2 &= \beta_1 - \beta_3 + \epsilon_2, \\ Y_3 &= \beta_2 + \beta_3 + \epsilon_3, \\ Y_4 &= \beta_2 + \epsilon_4, \end{aligned}$$

and $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ are independent random variables, each with mean zero, variance σ^2.

If the observed values are $y_1 = 2.3$, $y_2 = 5.4$, $y_3 = 1.0$, $y_4 = 6.2$, find the least squares estimates of $\beta_1, \beta_2, \beta_3$.

Find the least squares estimates of β_1, β_2 subject to the constraint $\beta_3 = 0$. Comment.

4. Explain what is meant by a factorial experiment. Outline briefly the advantages of a balanced factorial design.

   An experiment was performed to determine the effects of temperature (factor A), time of reaction (factor B), and 'alkali percent' (factor C) on the yield percent of pulp recovered from a cellulosic raw material. Three different temperatures, three different times of reaction, and four different 'alkali percent' levels were considered, and 2 observations taken at each of the $3 \times 3 \times 4$ possible factor combinations. Complete and interpret the following analysis of variance table for the observed data.

Source	SS	df	MS	F
A	2048.7			
B	371.9			
C	799.3			
A × B	80.6			
A × C	102.8			
B × C	11.3			
A × B × C	80.9			
Residual				
Total	3995.5	71		

   Without actually computing parameter estimates, write down your fitted model, including all constraints on the parameters.

   Estimate the error variance σ^2.

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

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

When is this of canonical form?

If l denotes the log of the likelihood function, let

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

Given that $\mathbf{E}[u]=0$ and $\mathbf{E}[u^2]=\mathbf{E}[-u']$, show that

$$\mathbf{E}[a(Y)] \;=\; -\frac{c'(\theta)}{b'(\theta)},$$

$$\text{and} \qquad \mathrm{Var}[a(Y)] \;=\; \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{(b'(\theta))^3}.$$

Show that the following distributions belong to the exponential family and can be written in canonical form, and hence determine the mean and variance of the distribution in each case.

(a) $f(y;\theta) \;=\; \theta \mathrm{e}^{-\theta y} \qquad (y \geq 0);$

(b) $f(y;\theta) \;=\; \binom{n}{y}\theta^y(1-\theta)^{n-y} \qquad (y=0,1,\ldots,n \text{ for some known } n);$

(c) $f(y;\theta) \;=\; \theta^2 y \mathrm{e}^{-\theta y} \qquad (y \geq 0).$

6. For a generalised linear model, define the term *deviance*. Explain briefly the role of deviance in the fitting of generalised linear models and in model selection.

   If responses $Y_1, Y_2, \ldots, Y_n$ are independent Poisson random variables with means $\mu_1, \mu_2, \ldots, \mu_n$, then show that in terms of observed values $y_1, y_2, \ldots, y_n$ the log-likelihood is given by

$$l(\boldsymbol{\mu}) = \sum_{i=1}^{n} y_i \ln(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \ln(y_i!).$$

   Hence show that the model with a common mean $\mu = \mu_1 = \mu_2 = \ldots = \mu_n$ has deviance

$$D = 2\sum_{i=1}^{n} y_i \ln(y_i/\bar{y})$$

   where $\bar{y} = (y_1 + y_2 + \cdots + y_n)/n$.

   Suppose the numbers of minor accidents on a particular North Sea oil rig in each of the last ten years were

$$1, 3, 7, 9, 2, 1, 1, 2, 0, 6.$$

   Use deviance to test whether a single Poisson distribution can be used to model the numbers of accidents over this ten year period.

   [Assume that when $x = 0$, then $x \ln x = 0$.]

7. In an experiment into the effects of a new drug on patients with severe asthma, the drug was tested at six different dose levels (from 190μg to 240μg) and the numbers of patients who could breathe adequately after inhaling the drug recorded. For $i = 1, 2, \ldots, 6$, the recorded data consisted of

$$\begin{aligned} x_i &= \text{Dose level,} \\ n_i &= \text{Number of patients receiving dose } x_i, \\ y_i &= \text{Number of patients breathing adequately after receiving dose } x_i. \end{aligned}$$

Thus y_i follows a Binomial distribution with parameters n_i and π_i, where π_i is a parameter to be estimated.

Write down the log-likelihood function $l(\boldsymbol{\pi})$ for a set of observations $y_1, y_2, \ldots, y_6$, and hence obtain expressions for the maximum likelihood estimates of $\pi_1, \pi_2, \ldots, \pi_6$ in terms of $y_1, y_2, \ldots, y_6$.

Suppose a reduced model is considered in which the parameters $\pi_1, \pi_2, \ldots, \pi_6$ are related by the logistic relationship

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i$$

for unknown parameters α, β, so that for this model,

$$\pi_i = \frac{\mathrm{e}^{\alpha+\beta x_i}}{1+\mathrm{e}^{\alpha+\beta x_i}}.$$

Show that the deviance for comparing the two models is given by

$$D = 2\sum_{i=1}^{6}\left\{y_i \ln\left(\frac{y_i}{n_i-y_i}\right) - n_i \ln\left(\frac{n_i}{n_i-y_i}\right) - y_i\left(\hat{\alpha}+\hat{\beta}x_i\right) + n_i \ln\left(1+\mathrm{e}^{\hat{\alpha}+\hat{\beta}x_i}\right)\right\}$$

where $\hat{\alpha}$, $\hat{\beta}$ are the maximum likelihood estimates of α, β.

Given that

$$\hat{\alpha} = -7.712,$$

Question 7 continued overleaf

$$\hat{\beta} = 0.0364,$$
$$D = 3.730,$$

then using the model with logistic link, estimate the proportion of patients who would be able to breathe adequately after receiving a dose of 215μg of the drug.

Using the analysis of deviance, which of the two models would be preferred for this data?