Paper 4, Section I

5K Statistical Modelling

Consider the normal linear model where the n-vector of responses Y satisfies $Y = X\beta + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$ and X is an $n \times p$ design matrix with full column rank. Write down a $(1-\alpha)$-level confidence set for β.

Define the *Cook's distance* for the observation (Y_i, x_i) where x_i^T is the ith row of X, and give its interpretation in terms of confidence sets for β.

In the model above with $n = 100$ and $p = 4$, you observe that one observation has Cook's distance 3.1. Would you be concerned about the influence of this observation? Justify your answer.

[*Hint: You may find some of the following facts useful:*

1. *If* $Z \sim \chi^2_4$, *then* $\mathbb{P}(Z \leqslant 1.06) = 0.1$, $\mathbb{P}(Z \leqslant 7.78) = 0.9$.
2. *If* $Z \sim F_{4,96}$, *then* $\mathbb{P}(Z \leqslant 0.26) = 0.1$, $\mathbb{P}(Z \leqslant 2.00) = 0.9$.
3. *If* $Z \sim F_{96,4}$, *then* $\mathbb{P}(Z \leqslant 0.50) = 0.1$, $\mathbb{P}(Z \leqslant 3.78) = 0.9$.]

Paper 3, Section I

5K Statistical Modelling

In an experiment to study factors affecting the production of the plastic polyvinyl chloride (PVC), three experimenters each used eight devices to produce the PVC and measured the sizes of the particles produced. For each of the 24 combinations of device and experimenter, two size measurements were obtained.

The experimenters and devices used for each of the 48 measurements are stored in `R` as factors in the objects `experimenter` and `device` respectively, with the measurements themselves stored in the vector `psize`. The following analysis was performed in `R`.

```
> fit0 <- lm(psize ~ experimenter + device)
> fit <- lm(psize ~ experimenter + device + experimenter:device)
> anova(fit0, fit)
Analysis of Variance Table

Model 1: psize ~ experimenter + device
Model 2: psize ~ experimenter + device + experimenter:device
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     38 49.815
2     24 35.480 14    14.335 0.6926 0.7599
```

Let X and X_0 denote the design matrices obtained by `model.matrix(fit)` and `model.matrix(fit0)` respectively, and let Y denote the response `psize`. Let P and P_0 denote orthogonal projections onto the column spaces of X and X_0 respectively.

For each of the following quantities, write down their numerical values if they appear in the analysis of variance table above; otherwise write 'unknown'.

1. $\|(I-P)Y\|^2$
2. $\|X(X^TX)^{-1}X^TY\|^2$
3. $\|(I-P_0)Y\|^2 - \|(I-P)Y\|^2$
4. $\dfrac{\|(P-P_0)Y\|^2/14}{\|(I-P)Y\|^2/24}$
5. $\sum_{i=1}^{48} Y_i/48$

Out of the two models that have been fitted, which appears to be the more appropriate for the data according to the analysis performed, and why?

Paper 2, Section I

5K Statistical Modelling

Define the concept of an *exponential dispersion family*. Show that the family of scaled binomial distributions $\frac{1}{n}\text{Bin}(n,p)$, with $p \in (0,1)$ and $n \in \mathbb{N}$, is of exponential dispersion family form.

Deduce the mean of the scaled binomial distribution from the exponential dispersion family form.

What is the canonical link function in this case?

Paper 1, Section I

5K Statistical Modelling

Write down the model being fitted by the following R command, where $y \in \{0,1,2,\ldots\}^n$ and X is an $n \times p$ matrix with real-valued entries.

```
fit <- glm(y ~ X, family = poisson)
```

Write down the log-likelihood for the model. Explain why the command

```
sum(y) - sum(predict(fit, type = "response"))
```

gives the answer 0, by arguing based on the log-likelihood you have written down.
[*Hint: Recall that if $Z \sim Pois(\mu)$ then*

$$\mathbb{P}(Z=k) = \frac{\mu^k e^{-\mu}}{k!}$$

for $k \in \{0,1,2,\ldots\}$.]

Paper 4, Section II

13K Statistical Modelling

In a study on infant respiratory disease, data are collected on a sample of 2074 infants. The information collected includes whether or not each infant developed a respiratory disease in the first year of their life; the gender of each infant; and details on how they were fed as one of three categories (breast-fed, bottle-fed and supplement). The data are tabulated in R as follows:

```
  disease nondisease gender       food
1      77        381    Boy Bottle-fed
2      19        128    Boy Supplement
3      47        447    Boy Breast-fed
4      48        336   Girl Bottle-fed
5      16        111   Girl Supplement
6      31        433   Girl Breast-fed
```

Write down the model being fit by the R commands on the following page:

```
> total <- disease + nondisease
> fit <- glm(disease/total ~ gender + food, family = binomial,
+ weights = total)
```

The following (slightly abbreviated) output from R is obtained.

```
> summary(fit)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.6127     0.1124 -14.347  < 2e-16 ***
genderGirl      -0.3126     0.1410  -2.216   0.0267 *
foodBreast-fed  -0.6693     0.1530  -4.374 1.22e-05 ***
foodSupplement  -0.1725     0.2056  -0.839   0.4013
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
```

Briefly explain the justification for the standard errors presented in the output above.

Explain the relevance of the output of the following R code to the data being studied, justifying your answer:

```
 > exp(c(-0.6693 - 1.96*0.153, -0.6693 + 1.96*0.153))
[1] 0.3793940 0.6911351
```

[*Hint: It may help to recall that if* $Z \sim N(0,1)$ *then* $\mathbb{P}(Z \geqslant 1.96) = 0.025$.]

Let D_1 be the deviance of the model fitted by the following R command.

```
> fit1 <- glm(disease/total ~ gender + food + gender:food,
+ family = binomial, weights = total)
```

What is the numerical value of D_1? Which of the two models that have been fitted should you prefer, and why?

Paper 1, Section II

13K Statistical Modelling

Consider the normal linear model where the n-vector of responses Y satisfies $Y = X\beta + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$. Here X is an $n \times p$ matrix of predictors with full column rank where $n \geqslant p + 3$, and $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients. Let X_0 be the matrix formed from the first $p_0 < p$ columns of X, and partition β as $\beta = (\beta_0^T, \beta_1^T)^T$ where $\beta_0 \in \mathbb{R}^{p_0}$ and $\beta_1 \in \mathbb{R}^{p-p_0}$. Denote the orthogonal projections onto the column spaces of X and X_0 by P and P_0 respectively.

It is desired to test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$. Recall that the F-test for testing H_0 against H_1 rejects H_0 for large values of

$$F = \frac{\|(P - P_0)Y\|^2/(p - p_0)}{\|(I - P)Y\|^2/(n - p)}.$$

Show that $(I - P)(P - P_0) = 0$, and hence prove that the numerator and denominator of F are independent under either hypothesis.

Show that

$$\mathbb{E}_{\beta,\sigma^2}(F) = \frac{(n - p)(\tau^2 + 1)}{n - p - 2},$$

where $\tau^2 = \dfrac{\|(P - P_0)X\beta\|^2}{(p - p_0)\sigma^2}$.

[*In this question you may use the following facts without proof:* $P - P_0$ is an orthogonal projection with rank $p - p_0$; any $n \times n$ orthogonal projection matrix Π satisfies $\|\Pi\varepsilon\|^2 \sim \sigma^2\chi^2_\nu$, where $\nu = \text{rank}(\Pi)$; and if $Z \sim \chi^2_\nu$ then $\mathbb{E}(Z^{-1}) = (\nu - 2)^{-1}$ when $\nu > 2$.]

Paper 4, Section I

5J Statistical Modelling

The output X of a process depends on the levels of two adjustable variables: A, a factor with four levels, and B, a factor with two levels. For each combination of a level of A and a level of B, nine independent values of X are observed.

Explain and interpret the `R` commands and (abbreviated) output below. In particular, describe the model being fitted, and describe and comment on the hypothesis tests performed under the `summary` and `anova` commands.

```
> fit1 <- lm(x~a+b)

> summary(fit1)

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    2.5445      0.2449    10.39  6.66e-16  ***
a2            -5.6704      0.4859   -11.67   < 2e-16  ***
a3             4.3254      0.3480    12.43   < 2e-16  ***
a4            -0.5003      0.3734    -1.34    0.0923
b2            -3.5689      0.2275   -15.69   < 2e-16  ***

> anova(fit1)

Response:  x
           Df   Sum Sq  mean Sq  F value    Pr(>F)
a           3    71.51    23.84    17.79   1.34e-8  ***
b           1   105.11   105.11    78.44  6.91e-13  ***
Residuals  67    89.56     1.34
```

Paper 3, Section I

5J Statistical Modelling

Consider the linear model $Y = X\beta + \epsilon$ where $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$, with $\epsilon_1, \ldots, \epsilon_n$ independent $N(0, \sigma^2)$ random variables. The $(n \times p)$ matrix X is known and is of full rank $p < n$. Give expressions for the maximum likelihood estimators $\widehat{\beta}$ and $\widehat{\sigma}^2$ of β and σ^2 respectively, and state their joint distribution. Show that $\widehat{\beta}$ is unbiased whereas $\widehat{\sigma}^2$ is biased.

Suppose that a new variable Y^* is to be observed, satisfying the relationship

$$Y^* = x^{*\mathrm{T}}\beta + \epsilon^* ,$$

where x^* $(p \times 1)$ is known, and $\epsilon^* \sim N(0, \sigma^2)$ independently of ϵ. We propose to predict Y^* by $\widetilde{Y} = x^{*\mathrm{T}}\widehat{\beta}$. Identify the distribution of

$$\frac{Y^* - \widetilde{Y}}{\tau\,\widetilde{\sigma}},$$

where

$$\begin{aligned} \widetilde{\sigma}^2 &= \frac{n}{n-p}\widehat{\sigma}^2 , \\ \tau^2 &= x^{*\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x^* + 1 . \end{aligned}$$

Paper 2, Section I

5J Statistical Modelling

Consider a linear model $Y = X\beta + \epsilon$, where Y and ϵ are $(n \times 1)$ with $\epsilon \sim N_n(0, \sigma^2 I)$, β is $(p \times 1)$, and X is $(n \times p)$ of full rank $p < n$. Let γ and δ be sub-vectors of β. What is meant by *orthogonality* between γ and δ?

Now suppose

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 P_3(x_i) + \epsilon_i \quad (i = 1, \ldots, n) ,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables, $x_1, \ldots, x_n$ are real-valued known explanatory variables, and $P_3(x)$ is a cubic polynomial chosen so that β_3 is orthogonal to $(\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$ and β_1 is orthogonal to $(\beta_0, \beta_2)^{\mathrm{T}}$.

Let $\widetilde{\beta} = (\beta_0, \beta_2, \beta_1, \beta_3)^{\mathrm{T}}$. Describe the matrix $\widetilde{X}$ such that $Y = \widetilde{X}\widetilde{\beta} + \epsilon$. Show that $\widetilde{X}^{\mathrm{T}}\widetilde{X}$ is block diagonal. Assuming further that this matrix is non-singular, show that the least-squares estimators of β_1 and β_3 are, respectively,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \qquad \text{and} \qquad \widehat{\beta}_3 = \frac{\sum_{i=1}^n P_3(x_i) Y_i}{\sum_{i=1}^n P_3(x_i)^2} .$$

Paper 1, Section I

5J Statistical Modelling

Variables $Y_1, \ldots, Y_n$ are independent, with Y_i having a density $p(y \mid \mu_i)$ governed by an unknown parameter μ_i. Define the *deviance* for a model M that imposes relationships between the (μ_i).

From this point on, suppose $Y_i \sim \text{Poisson}(\mu_i)$. Write down the log-likelihood of data $y_1, \ldots, y_n$ as a function of $\mu_1, \ldots, \mu_n$.

Let $\widehat{\mu}_i$ be the maximum likelihood estimate of μ_i under model M. Show that the deviance for this model is given by

$$2\sum_{i=1}^{n}\left\{y_i \log \frac{y_i}{\widehat{\mu}_i} - (y_i - \widehat{\mu}_i)\right\}.$$

Now suppose that, under M, $\log \mu_i = \beta^{\mathrm{T}} x_i$, $i = 1, \ldots, n$, where $x_1, \ldots, x_n$ are known p-dimensional explanatory variables and β is an unknown p-dimensional parameter. Show that $\widehat{\mu} := (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)^{\mathrm{T}}$ satisfies $X^{\mathrm{T}} y = X^{\mathrm{T}} \widehat{\mu}$, where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and X is the $(n \times p)$ matrix with rows $x_1^{\mathrm{T}}, \ldots, x_n^{\mathrm{T}}$, and express this as an equation for the maximum likelihood estimate $\widehat{\beta}$ of β. [You are not required to solve this equation.]

Paper 4, Section II

13J Statistical Modelling

Let f_0 be a probability density function, with cumulant generating function K. Define what it means for a random variable Y to have a model function of exponential dispersion family form, generated by f_0.

A random variable Y is said to have an *inverse Gaussian distribution*, with parameters ϕ and λ (both positive), if its density function is

$$f(y; \phi, \lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi y^3}}\, e^{\sqrt{\lambda\phi}} \exp\left\{-\frac{1}{2}\left(\frac{\lambda}{y} + \phi y\right)\right\} \qquad (y > 0).$$

Show that the family of all inverse Gaussian distributions for Y is of exponential dispersion family form. Deduce directly the corresponding expressions for $E(Y)$ and $\mathrm{Var}(Y)$ in terms of ϕ and λ. What are the corresponding canonical link function and variance function?

Consider a generalized linear model, M, for independent variables Y_i $(i = 1, \ldots, n)$, whose random component is defined by the inverse Gaussian distribution with link function $g(\mu) = \log(\mu)$: thus $g(\mu_i) = x_i^{\mathrm{T}}\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of unknown regression coefficients and $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is the vector of known values of the explanatory variables for the i^{th} observation. The vectors x_i $(i = 1, \ldots, n)$ are linearly independent. Assuming that the dispersion parameter is known, obtain expressions for the score function and Fisher information matrix for β. Explain how these can be used to compute the maximum likelihood estimate $\widehat{\beta}$ of β.

Paper 1, Section II

13J Statistical Modelling

A cricket ball manufacturing company conducts the following experiment. Every day, a bowling machine is set to one of three levels, "Medium", "Fast" or "Spin", and then bowls 100 balls towards the stumps. The number of times the ball hits the stumps and the average wind speed (in kilometres per hour) during the experiment are recorded, yielding the following data (abbreviated):

```
Day   Wind   Level    Stumps
1     10     Medium   26
2     8      Medium   37
⋮     ⋮      ⋮        ⋮
50    12     Medium   32
51    7      Fast     31
⋮     ⋮      ⋮        ⋮
120   3      Fast     28
121   5      Spin     35
⋮     ⋮      ⋮        ⋮
150   6      Spin     31
```

Write down a reasonable model for $Y_1, \ldots, Y_{150}$, where Y_i is the number of times the ball hits the stumps on the i^{th} day. Explain briefly why we might want to include interactions between the variables. Write `R` code to fit your model.

The company's statistician fitted her own generalized linear model using `R`, and obtained the following summary (abbreviated):

```
>summary(ball)
Coefficients:
                 Estimate  Std. Error  z value  Pr(>|z|)
 (Intercept)     -0.37258     0.05388   -6.916  4.66e-12  ***
 Wind             0.09055     0.01595    5.676  1.38e-08  ***
 LevelFast       -0.10005     0.08044   -1.244  0.213570
 LevelSpin        0.29881     0.08268    3.614  0.000301  ***
 Wind:LevelFast   0.03666     0.02364    1.551  0.120933
 Wind:LevelSpin  -0.07697     0.02845   -2.705  0.006825  **
```

Why are `LevelMedium` and `Wind:LevelMedium` not listed?

Suppose that, on another day, the bowling machine is set to "Spin", and the wind speed is 5 kilometres per hour. What linear function of the parameters should the statistician use in constructing a predictor of the number of times the ball hits the stumps that day?

Based on the above output, how might you improve the model? How could you fit your new model in `R`?

Paper 4, Section I

5K Statistical Modelling

Define the concepts of an *exponential dispersion family* and the corresponding *variance function*. Show that the family of Poisson distributions with parameter $\lambda > 0$ is an exponential dispersion family. Find the corresponding variance function and deduce from it expressions for $E(Y)$ and $\mathrm{Var}(Y)$ when $Y \sim \mathrm{Pois}(\lambda)$. What is the canonical link function in this case?

Paper 3, Section I

5K Statistical Modelling

Consider the linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

for $i = 1, 2, \ldots, n$, where the ε_i are independent and identically distributed with $N(0, \sigma^2)$ distribution. What does it mean for the pair β_1 and β_2 to be *orthogonal*? What does it mean for all the three parameters β_0, β_1 and β_2 to be *mutually orthogonal*? Give necessary and sufficient conditions on $(x_{i1})_{i=1}^n, (x_{i2})_{i=1}^n$ so that β_0, β_1 and β_2 are mutually orthogonal. If $\beta_0, \beta_1, \beta_2$ are mutually orthogonal, find the joint distribution of the corresponding maximum likelihood estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$.

Paper 2, Section I

5K Statistical Modelling

The purpose of the following study is to investigate differences among certain treatments on the lifespan of male fruit flies, after allowing for the effect of the variable 'thorax length' (`thorax`) which is known to be positively correlated with lifespan. Data was collected on the following variables:

`longevity` lifespan in days

`thorax` (body) length in mm

`treat` a five level factor representing the treatment groups. The levels were labelled as follows: "00", "10", "80", "11", "81".

No interactions were found between thorax length and the treatment factor. A linear model with `thorax` as the covariate, `treat` as a factor (having the above 5 levels) and `longevity` as the response was fitted and the following output was obtained. There were 25 males in each of the five groups, which were treated identically in the provision of fresh food.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -49.98      10.61   -4.71  6.7e-06
treat10         2.65       2.98    0.89     0.37
treat11        -7.02       2.97   -2.36     0.02
treat80         3.93       3.00    1.31     0.19
treat81       -19.95       3.01   -6.64  1.0e-09
thorax        135.82      12.44   10.92   <2e-16

Residual standard error: 10.5 on 119 degrees of freedom
Multiple R-Squared: 0.656,  Adjusted R-squared: 0.642
F-statistics: 45.5 on 5 and 119 degrees of freedom, p-value: 0
```

(a) Assuming the same treatment, how much longer would you expect a fly with a thorax length 0.1mm greater than another to live?

(b) What is the predicted difference in longevity between a male fly receiving treatment `treat10` and `treat81` assuming they have the same thorax length?

(c) Because the flies were randomly assigned to the five groups, the distribution of thorax lengths in the five groups are essentially equal. What disadvantage would the investigators have incurred by ignoring the thorax length in their analysis (i.e., had they done a one-way ANOVA instead)?

(d) The residual-fitted plot is shown in the left panel of Figure 1 overleaf. Is it possible to determine if the regular residuals or the studentized residuals have been used to construct this plot? Explain.

(e) The Box–Cox procedure was used to determine a good transformation for this data. The plot of the log-likelihood for λ is shown in the right panel of Figure 1. What transformation should be used to improve the fit and yet retain some interpretability?

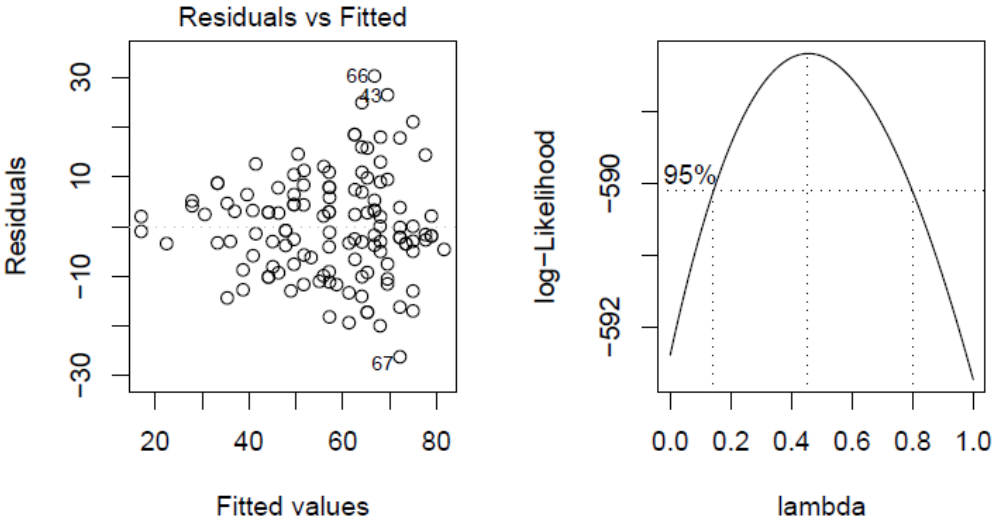

Figure 1: Residual-Fitted plot on the left and Box-Cox plot on the right

Paper 1, Section I

5K Statistical Modelling

Let $Y_1, \ldots, Y_n$ be independent with $Y_i \sim \frac{1}{n_i}\mathrm{Bin}(n_i, \mu_i)$, $i = 1, \ldots, n$, and

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^\top \beta\,, \tag{1}$$

where x_i is a $p \times 1$ vector of regressors and β is a $p \times 1$ vector of parameters. Write down the likelihood of the data $Y_1, \ldots, Y_n$ as a function of $\mu = (\mu_1, \ldots, \mu_n)$. Find the unrestricted maximum likelihood estimator of μ, and the form of the maximum likelihood estimator $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$ under the logistic model (1).

Show that the *deviance* for a comparison of the full (saturated) model to the generalised linear model with canonical link (1) using the maximum likelihood estimator $\hat{\beta}$ can be simplified to

$$D(y; \hat{\mu}) = -2 \sum_{i=1}^{n} \left[n_i y_i x_i^\top \hat{\beta} - n_i \log(1 - \hat{\mu}_i) \right].$$

Finally, obtain an expression for the deviance residual in this generalised linear model.

Paper 4, Section II

13K Statistical Modelling

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be jointly independent and identically distributed with $X_i \sim N(0,1)$ and conditional on $X_i = x$, $Y_i \sim N(x\theta, 1)$, $i = 1, 2, \ldots, n$.

(a) Write down the likelihood of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and find the maximum likelihood estimate $\hat{\theta}$ of θ. [You may use properties of conditional probability/expectation without providing a proof.]

(b) Find the Fisher information $I(\theta)$ for a single observation, (X_1, Y_1).

(c) Determine the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta)$. [You may use the result on the asymptotic distribution of maximum likelihood estimators, without providing a proof.]

(d) Give an asymptotic confidence interval for θ with coverage $(1-\alpha)$ using your answers to (b) and (c).

(e) Define the observed Fisher information. Compare the confidence interval in part (d) with an asymptotic confidence interval with coverage $(1-\alpha)$ based on the observed Fisher information.

(f) Determine the exact distribution of $\left(\sum_{i=1}^{n} X_i^2\right)^{1/2}(\hat{\theta} - \theta)$ and find the true coverage probability for the interval in part (e). [*Hint. Condition on* $X_1, X_2, \ldots, X_n$ *and use the following property of conditional expectation: for* U, V *random vectors, any suitable function* g*, and* $x \in \mathbb{R}$*,*

$$P\{g(U,V) \leqslant x\} = E[P\{g(U,V) \leqslant x | V\}].]$$

Paper 1, Section II

13K Statistical Modelling

The treatment for a patient diagnosed with cancer of the prostate depends on whether the cancer has spread to the surrounding lymph nodes. It is common to operate on the patient to obtain samples from the nodes which can then be analysed under a microscope. However it would be preferable if an accurate assessment of nodal involvement could be made without surgery. For a sample of 53 prostate cancer patients, a number of possible predictor variables were measured before surgery. The patients then had surgery to determine nodal involvement. We want to see if nodal involvement can be accurately predicted from the available variables and determine which ones are most important. The variables take the values 0 or 1.

`r` An indicator 0=no/1=yes of nodal involvement.

`aged` The patient's age, split into less than 60 (=0) and 60 or over (=1).

`stage` A measurement of the size and position of the tumour observed by palpation with the fingers. A serious case is coded as 1 and a less serious case as 0.

`grade` Another indicator of the seriousness of the cancer which is determined by a pathology reading of a biopsy taken by needle before surgery. A value of 1 indicates a more serious case of cancer.

`xray` Another measure of the seriousness of the cancer taken from an X-ray reading. A value of 1 indicates a more serious case of cancer.

`acid` The level of acid phosphatase in the blood serum where 1=high and 0=low.

A binomial generalised linear model with a logit link was fitted to the data to predict nodal involvement and the following output obtained:

```
Deviance Residuals:
     Min       1Q   Median     3Q     Max
  -2.332   -0.665   -0.300  0.639   2.150

Coefficients:
             Estimate  Std. Error  t value  Pr(>|z|)
(Intercept)    -3.079       0.987    -3.12    0.0018
aged           -0.292       0.754    -0.39    0.6988
grade           0.872       0.816     1.07    0.2850
stage           1.373       0.784     1.75    0.0799
xray            1.801       0.810     2.22    0.0263
acid            1.684       0.791     2.13    0.0334

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 70.252 on 52 degrees of freedom
Residual deviance: 47.611 on 47 degrees of freedom
AIC: 59.61

Number of Fisher Scoring iterations: 5
```

(a) Give an interpretation of the coefficient of `xray`.

(b) Give the numerical value of the sum of the squared deviance residuals.

(c) Suppose that the predictors, `stage`, `grade` and `xray` are positively correlated. Describe the effect that this correlation is likely to have on our ability to determine the strength of these predictors in explaining the response.

(d) The probability of observing a value of 70.252 under a Chi-squared distribution with 52 degrees of freedom is 0.047. What does this information tell us about the null model for this data? Justify your answer.

(e) What is the lowest predicted probability of the nodal involvement for any future patient?

(f) The first plot in Figure 1 shows the (Pearson) residuals and the fitted values. Explain why the points lie on two curves.

(g) The second plot in Figure 1 shows the value of $\hat{\beta} - \hat{\beta}_{(i)}$ where (i) indicates that patient i was dropped in computing the fit. The values for each predictor, including the intercept, are shown. Could a single case change our opinion of which predictors are important in predicting the response?

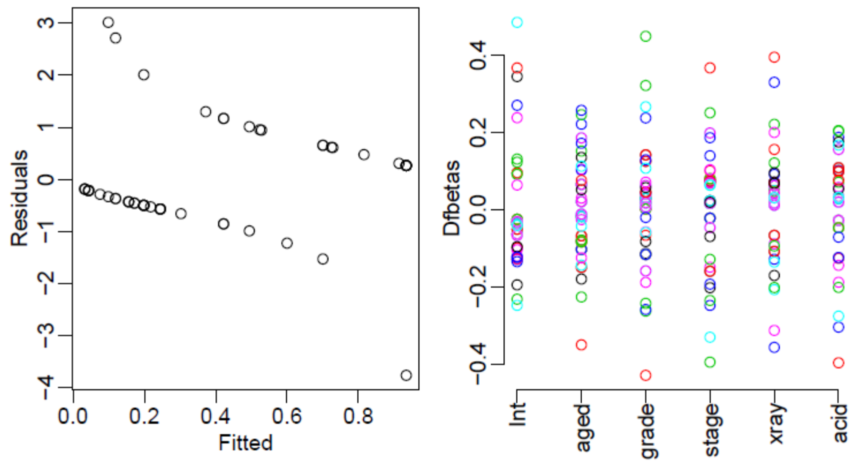

Figure 1: The plot on the left shows the Pearson residuals and the fitted values. The plot on the right shows the changes in the regression coefficients when a single point is omitted for each predictor.

Paper 1, Section I

5J Statistical Modelling

Let $Y_1, \ldots, Y_n$ be independent identically distributed random variables with model function $f(y, \theta)$, $y \in \mathcal{Y}$, $\theta \in \Theta \subseteq \mathbb{R}$, and denote by E_θ and Var_θ expectation and variance under $f(y, \theta)$, respectively. Define $U_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log f(Y_i, \theta)$. Prove that $E_\theta U_n(\theta) = 0$. Show moreover that if $T = T(Y_1, \ldots, Y_n)$ is any unbiased estimator of θ, then its variance satisfies $\mathrm{Var}_\theta(T) \geqslant (n\mathrm{Var}_\theta(U_1(\theta)))^{-1}$. [You may use the Cauchy–Schwarz inequality without proof, and you may interchange differentiation and integration without justification if necessary.]

Paper 2, Section I

5J Statistical Modelling

Let f_0 be a probability density function, with cumulant generating function K. Define what it means for a random variable Y to have a model function of exponential dispersion family form, generated by f_0. Compute the cumulant generating function K_Y of Y and deduce expressions for the mean and variance of Y that depend only on first and second derivatives of K.

Paper 3, Section I

5J Statistical Modelling

Define a generalised linear model for a sample $Y_1, \ldots, Y_n$ of independent random variables. Define further the concept of the link function. Define the binomial regression model with logistic and probit link functions. Which of these is the canonical link function?

Paper 4, Section I

5J Statistical Modelling

The numbers of ear infections observed among beach and non-beach (mostly pool) swimmers were recorded, along with explanatory variables: frequency, location, age, and sex. The data are aggregated by group, with a total of 24 groups defined by the explanatory variables.

`freq`	F = frequent, NF = infrequent
`loc`	NB = non-beach, B = beach
`age`	15-19, 20-24, 24-29
`sex`	F = female, M = male
`count`	the number of infections reported over a fixed time period
`n`	the total number of swimmers

The data look like this:

```
   count  n freq loc sex   age
1     68 31    F  NB   M 15-19
2     14  4    F  NB   F 15-19
3     35 12    F  NB   M 20-24
4     16 11    F  NB   F 20-24
[...]
23     5 15   NF   B   M 25-29
24     6  6   NF   B   F 25-29
```

Let μ_j denote the expected number of ear infections of a person in group j. Explain why it is reasonable to model $\texttt{count}_j$ as Poisson with mean $n_j\mu_j$.

We fit the following Poisson model:

$$\log(\mathbb{E}(\texttt{count}_j)) = \log(n_j\mu_j) = \log(n_j) + \mathbf{x}_j\beta,$$

where $\log(n_j)$ is an offset, i.e. an explanatory variable with known coefficient 1.

R produces the following (abbreviated) summary for the main effects model:

```
Call:
glm(formula = count ~ freq + loc + age + sex, family = poisson, offset = log(n))
[...]
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.48887    0.12271   3.984 6.78e-05 ***
freqNF      -0.61149    0.10500  -5.823 5.76e-09 ***
locNB        0.53454    0.10668   5.011 5.43e-07 ***
age20-24    -0.37442    0.12836  -2.917  0.00354 **
age25-29    -0.18973    0.13009  -1.458  0.14473
sexM        -0.08985    0.11231  -0.800  0.42371
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
[...]
```

Why are expressions freqF, locB, age15-19, and sexF not listed?

Suppose that we plan to observe a group of 20 female, non-frequent, beach swimmers, aged 20-24. Give an expression (using the coefficient estimates from the model fitted above) for the expected number of ear infections in this group.

Now, suppose that we allow for interaction between variables `age` and `sex`. Give the R command for fitting this model. We test for the effect of this interaction by producing the following (abbreviated) ANOVA table:

```
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        18     51.714
2        16     44.319  2   7.3948   0.02479 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Briefly explain what test is performed, and what you would conclude from it. Does either of these models fit the data well?

Paper 1, Section II

13J Statistical Modelling

The data consist of the record times in 1984 for 35 Scottish hill races. The columns list the record time in minutes, the distance in miles, and the total height gained during the route. The data are displayed in R as follows (abbreviated):

```
> hills
                 dist climb    time
Greenmantle       2.5   650  16.083
Carnethy          6.0  2500  48.350
Craig Dunain      6.0   900  33.650
Ben Rha           7.5   800  45.600
Ben Lomond        8.0  3070  62.267
[...]
Cockleroi         4.5   850  28.100
Moffat Chase     20.0  5000 159.833
```

Consider a simple linear regression of `time` on `dist` and `climb`. Write down this model mathematically, and explain any assumptions that you make. How would you instruct R to fit this model and assign it to a variable `hills.lm1`?

First, we test the hypothesis of no linear relationship to the variables `dist` and `climb` against the full model. R provides the following ANOVA summary:

```
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     34 85138
2     32  6892  2     78247 181.66 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Using the information in this table, explain carefully how you would test this hypothesis. What do you conclude?

The R command

```
summary(hills.lm1)
```

provides the following (slightly abbreviated) summary:

```
[...]
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.992039   4.302734  -2.090   0.0447 *
dist         6.217956   0.601148  10.343 9.86e-12 ***
climb        0.011048   0.002051   5.387 6.45e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
[...]
```

Carefully explain the information that appears in each column of the table. What are your conclusions? In particular, how would you test for the significance of the variable `climb` in this model?

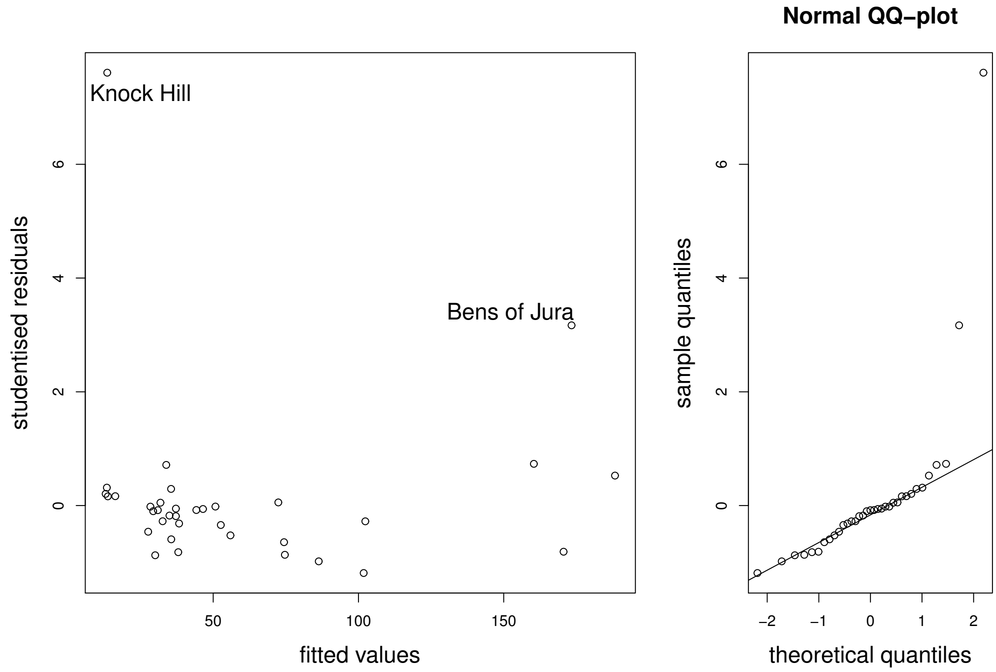

Figure 1: Hills data: diagnostic plots

Finally, we perform model diagnostics on the full model, by looking at studentised residuals versus fitted values, and the normal QQ-plot. The plots are displayed in Figure 1. Comment on possible sources of model misspecification. Is it possible that the problem lies with the data? If so, what do you suggest?

Paper 4, Section II

13J Statistical Modelling

Consider the general linear model $Y = X\beta + \epsilon$, where the $n \times p$ matrix X has full rank $p \leqslant n$, and where ϵ has a multivariate normal distribution with mean zero and covariance matrix $\sigma^2 I_n$. Write down the likelihood function for β, σ^2 and derive the maximum likelihood estimators $\hat{\beta}, \hat{\sigma}^2$ of β, σ^2. Find the distribution of $\hat{\beta}$. Show further that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

Paper 1, Section I

5J Statistical Modelling

Consider a binomial generalised linear model for data $y_1, ..., y_n$ modelled as realisations of independent $Y_i \sim \text{Bin}(1, \mu_i)$ and logit link $\mu_i = e^{\beta x_i}/(1 + e^{\beta x_i})$ for some known constants x_i, $i = 1, \ldots, n$, and unknown scalar parameter β. Find the log-likelihood for β, and the likelihood equation that must be solved to find the maximum likelihood estimator $\hat{\beta}$ of β. Compute the second derivative of the log-likelihood for β, and explain the algorithm you would use to find $\hat{\beta}$.

Paper 2, Section I

5J Statistical Modelling

Suppose you have a parametric model consisting of probability mass functions $f(y;\theta)$, $\theta \in \Theta \subset \mathbb{R}$. Given a sample $Y_1, ..., Y_n$ from $f(y;\theta)$, define the maximum likelihood estimator $\hat{\theta}_n$ for θ and, assuming standard regularity conditions hold, state the asymptotic distribution of $\sqrt{n}\,(\hat{\theta}_n - \theta)$.

Compute the Fisher information of a single observation in the case where $f(y;\theta)$ is the probability mass function of a Poisson random variable with parameter θ. If $Y_1, ..., Y_n$ are independent and identically distributed random variables having a Poisson distribution with parameter θ, show that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $S = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ are unbiased estimators for θ. Without calculating the variance of S, show that there is no reason to prefer S over $\overline{Y}$.

[You may use the fact that the asymptotic variance of $\sqrt{n}\,(\hat{\theta}_n - \theta)$ is a lower bound for the variance of any unbiased estimator.]

Paper 3, Section I

5J Statistical Modelling

Consider the linear model $Y = X\beta + \varepsilon$, where Y is a $n \times 1$ random vector, $\varepsilon \sim N_n(0, \sigma^2 I)$, and where the $n \times p$ nonrandom matrix X is known and has full column rank p. Derive the maximum likelihood estimator $\hat{\sigma}^2$ of σ^2. Without using Cochran's theorem, show carefully that $\hat{\sigma}^2$ is biased. Suggest another estimator $\tilde{\sigma}^2$ for σ^2 that is unbiased.

Paper 4, Section I

5J Statistical Modelling

Below is a simplified 1993 dataset of US cars. The columns list index, make, model, price (in \$1000), miles per gallon, number of passengers, length and width in inches, and weight (in pounds). The data are displayed in R as follows (abbreviated):

```
> cars
      make    model price mpg psngr length width weight
1    Acura  Integra  15.9  31     5    177    68   2705
2    Acura   Legend  33.9  25     5    195    71   3560
3     Audi       90  29.1  26     5    180    67   3375
4     Audi      100  37.7  26     6    193    70   3405
5      BMW     535i  30.0  30     4    186    69   3640
       ...                ...                       ...
92   Volvo      240  22.7  28     5    190    67   2985
93   Volvo      850  26.7  28     5    184    69   3245
```

It is reasonable to assume that prices for different makes of car are independent. We model the logarithm of the price as a linear combination of the other quantitative properties of the cars and an error term. Write down this model mathematically. How would you instruct R to fit this model and assign it to a variable "fit"?

R provides the following (slightly abbreviated) summary:

```
> summary(fit)

[...]

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.8751080  0.7687276   5.041 2.50e-06 ***
mpg         -0.0109953  0.0085475  -1.286 0.201724
psngr       -0.1782818  0.0290618  -6.135 2.45e-08 ***
length       0.0067382  0.0032890   2.049 0.043502 *
width       -0.0517544  0.0151009  -3.427 0.000933 ***
weight       0.0008373  0.0001302   6.431 6.60e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[...]
```

Briefly explain the information that is being provided in each column of the table. What are your conclusions and how would you try to improve the model?

Paper 1, Section II

13J Statistical Modelling

Consider a generalised linear model with parameter $\beta^{\top}$ partitioned as $(\beta_0^{\top}, \beta_1^{\top})$, where β_0 has p_0 components and β_1 has $p - p_0$ components, and consider testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Define carefully the deviance, and use it to construct a test for H_0.

[You may use Wilks' theorem to justify this test, and you may also assume that the dispersion parameter is known.]

Now consider the generalised linear model with Poisson responses and the canonical link function with linear predictor $\eta = (\eta_1, ..., \eta_n)^T$ given by $\eta_i = x_i^{\top} \beta$, $i = 1, ..., n$, where $x_{i1} = 1$ for every i. Derive the deviance for this model, and argue that it may be approximated by Pearson's χ^2 statistic.

Paper 4, Section II

13J Statistical Modelling

Every day, Barney the darts player comes to our laboratory. We record his facial expression, which can be either "mad", "weird" or "relaxed", as well as how many units of beer he has drunk that day. Each day he tries a hundred times to hit the bull's-eye, and we write down how often he succeeds. The data look like this:

```
>
Day Beer Expression  BullsEye
  1    3        Mad        30
  2    3        Mad        32
  ⋮    ⋮          ⋮         ⋮
 60    2        Mad        37
 61    4      Weird        30
  ⋮    ⋮          ⋮         ⋮
110    4      Weird        28
111    2    Relaxed        35
  ⋮    ⋮          ⋮         ⋮
150    3    Relaxed        31
```

Write down a reasonable model for $Y_1, \ldots, Y_n$, where $n = 150$ and where Y_i is the number of times Barney has hit bull's-eye on the ith day. Explain briefly why we may wish initially to include interactions between the variables. Write the R code to fit your model.

The scientist of the above story fitted her own generalized linear model, and subsequently obtained the following summary (abbreviated):

```
> summary(barney)
[...]

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.37258    0.05388  -6.916 4.66e-12 ***
Beer                   -0.09055    0.01595  -5.676 1.38e-08 ***
ExpressionWeird        -0.10005    0.08044  -1.244 0.213570
ExpressionRelaxed       0.29881    0.08268   3.614 0.000301 ***
Beer:ExpressionWeird    0.03666    0.02364   1.551 0.120933
Beer:ExpressionRelaxed -0.07697    0.02845  -2.705 0.006825 **

[...]
```

Why are `ExpressionMad` and `Beer:ExpressionMad` not listed? Suppose on a particular day, Barney's facial expression is weird, and he drank three units of beer. Give the linear predictor in the scientist's model for this day.

Based on the summary, how could you improve your model? How could one fit this new model in R (without modifying the data file)?

Paper 1, Section I

5I Statistical Modelling

Consider a binomial generalised linear model for data $y_1, \ldots, y_n$, modelled as realisations of independent $Y_i \sim \text{Bin}(1, \mu_i)$ and logit link, i.e. $\log \frac{\mu_i}{1-\mu_i} = \beta x_i$, for some known constants $x_1, \ldots, x_n$, and an unknown parameter β. Find the log-likelihood for β, and the likelihood equations that must be solved to find the maximum likelihood estimator $\hat{\beta}$ of β.

Compute the first and second derivatives of the log-likelihood for β, and explain the algorithm you would use to find $\hat{\beta}$.

Paper 2, Section I

5I Statistical Modelling

What is meant by an *exponential dispersion family*? Show that the family of Poisson distributions with parameter λ is an exponential dispersion family by explicitly identifying the terms in the definition.

Find the corresponding variance function and deduce directly from your calculations expressions for $\mathbb{E}(Y)$ and $\text{Var}(Y)$ when $Y \sim \text{Pois}(\lambda)$.

What is the canonical link function in this case?

Paper 3, Section I

5I Statistical Modelling

Consider the linear model $Y = X\beta + \varepsilon$, where $\varepsilon \sim N_n(0, \sigma^2 I)$ and X is an $n \times p$ matrix of full rank $p < n$. Suppose that the parameter β is partitioned into k sets as follows: $\beta^\top = (\beta_1^\top \; \cdots \; \beta_k^\top)$. What does it mean for a pair of sets β_i, β_j, $i \neq j$, to be *orthogonal*? What does it mean for all k sets to be *mutually orthogonal*?

In the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed, find necessary and sufficient conditions on $x_{11}, \ldots, x_{n1}, x_{12}, \ldots, x_{n2}$ for β_0, β_1 and β_2 to be mutually orthogonal.

If β_0, β_1 and β_2 are mutually orthogonal, what consequence does this have for the joint distribution of the corresponding maximum likelihood estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$?

Paper 4, Section I

5I Statistical Modelling

Sulphur dioxide is one of the major air pollutants. A dataset by Sokal and Rohlf (1981) was collected on 41 US cities/regions in 1969–1971. The annual measurements obtained for each region include (average) sulphur dioxide content, temperature, number of manufacturing enterprises employing more than 20 workers, population size in thousands, wind speed, precipitation, and the number of days with precipitation. The data are displayed in R as follows (abbreviated):

```
> usair
          region so2 temp manuf  pop wind precip days
1        Phoenix  10 70.3   213  582  6.0   7.05   36
2    Little Rock  13 61.0    91  132  8.2  48.52  100
          ...              ...                 ...
41     Milwaukee  16 45.7   569  717 11.8  29.07  123
```

Describe the model being fitted by the following R commands.

```
> fit <- lm(log(so2) ~ temp + manuf + pop + wind + precip + days)
```

Explain the (slightly abbreviated) output below, describing in particular how the hypothesis tests are performed and your conclusions based on their results:

```
> summary(fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2532456  1.4483686   5.008 1.68e-05 ***
temp        -0.0599017  0.0190138  -3.150  0.00339 **
manuf        0.0012639  0.0004820   2.622  0.01298 *
pop         -0.0007077  0.0004632  -1.528  0.13580
wind        -0.1697171  0.0555563  -3.055  0.00436 **
precip       0.0173723  0.0111036   1.565  0.12695
days         0.0004347  0.0049591   0.088  0.93066
Residual standard error: 0.448 on 34 degrees of freedom
```

Based on the summary above, suggest an alternative model.

Finally, what is the value obtained by the following command?

```
> sqrt(sum(resid(fit)^2)/fit$df)
```

Paper 1, Section II

13I Statistical Modelling

A three-year study was conducted on the survival status of patients suffering from cancer. The age of the patients at the start of the study was recorded, as well as whether or not the initial tumour was malignant. The data are tabulated in R as follows:

```
> cancer
    age malignant survive die
1   <50        no      77  10
2   <50       yes      51  13
3 50-69        no      51  11
4 50-69       yes      38  20
5   70+        no       7   3
6   70+       yes       6   3
```

Describe the model that is being fitted by the following R commands:

```
> total <- survive + die
> fit1 <- glm(survive/total ~ age + malignant, family = binomial,
+     weights = total)
```

Explain the (slightly abbreviated) output from the code below, describing how the hypothesis tests are performed and your conclusions based on their results.

```
> summary(fit1)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.0730     0.2812   7.372 1.68e-13 ***
age50-69      -0.6318     0.3112  -2.030   0.0424 *
age70+        -0.9282     0.5504  -1.686   0.0917 .
malignantyes  -0.7328     0.2985  -2.455   0.0141 *
----
    Null deviance: 12.65585  on 5  degrees of freedom
Residual deviance:  0.49409  on 2  degrees of freedom
AIC: 30.433
```

Based on the summary above, motivate and describe the following alternative model:

```
> age2 <- as.factor(c("<50", "<50", "50+", "50+", "50+", "50+"))
> fit2 <- glm(survive/total ~ age2 + malignant, family = binomial,
+     weights = total)
```

This question continues on the next page

Based on the output of the code that follows, which of the two models do you prefer? Why?

```
> summary(fit2)

Coefficients:

             Estimate Std. Error z value Pr(>|z|)

(Intercept)    2.0721     0.2811   7.372 1.68e-13 ***

age250+       -0.6744     0.3000  -2.248   0.0246 *

malignantyes  -0.7310     0.2983  -2.451   0.0143 *

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 12.656  on 5  degrees of freedom

Residual deviance:  0.784  on 3  degrees of freedom

AIC: 28.723
```

What is the final value obtained by the following commands?

```
> mu.hat <- inv.logit(predict(fit2))

> -2 * (sum(dbinom(survive, total, mu.hat, log = TRUE)

+     - sum(dbinom(survive, total, survive/total, log = TRUE)))
```

Paper 4, Section II

13I Statistical Modelling

Consider the linear model $Y = X\beta + \varepsilon$, where $\varepsilon \sim N_n(0, \sigma^2 I)$ and X is an $n \times p$ matrix of full rank $p < n$. Find the form of the maximum likelihood estimator $\hat{\beta}$ of β, and derive its distribution assuming that σ^2 is known.

Assuming the prior $\pi(\beta, \sigma^2) \propto \sigma^{-2}$ find the joint posterior of (β, σ^2) up to a normalising constant. Derive the posterior conditional distribution $\pi(\beta|\sigma^2, X, Y)$.

Comment on the distribution of $\hat{\beta}$ found above and the posterior conditional $\pi(\beta|\sigma^2, X, Y)$. Comment further on the predictive distribution of y^* at input x^* under both the maximum likelihood and Bayesian approaches.

1/I/5J **Statistical Modelling**

Consider the following Binomial generalized linear model for data $y_1, \dots, y_n$, with logit link function. The data $y_1, \dots, y_n$ are regarded as observed values of independent random variables $Y_1, \dots, Y_n$, where

$$Y_i \sim \text{Bin}(1, \mu_i), \quad \log \frac{\mu_i}{1 - \mu_i} = \beta^\top x_i, \quad i = 1, \dots, n,$$

where β is an unknown p-dimensional parameter, and where $x_1, \dots, x_n$ are known p-dimensional explanatory variables. Write down the likelihood function for $y = (y_1, \dots, y_n)$ under this model.

Show that the maximum likelihood estimate $\hat{\beta}$ satisfies an equation of the form $X^\top y = X^\top \hat{\mu}$, where X is the $p \times n$ matrix with rows $x_1^\top, \dots, x_n^\top$, and where $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_n)$, with $\hat{\mu}_i$ a function of x_i and $\hat{\beta}$, which you should specify.

Define the deviance $D(y; \hat{\mu})$ and find an explicit expression for $D(y; \hat{\mu})$ in terms of y and $\hat{\mu}$ in the case of the model above.

1/II/13J **Statistical Modelling**

Consider performing a two-way analysis of variance (ANOVA) on the following data:

```
> Y[,,1]              Y[,,2]                 Y[,,3]
      [,1] [,2]           [,1]    [,2]             [,1]   [,2]
 [1,] 2.72 6.66    [1,] -5.780  1.7200    [1,] -2.2900 0.158
 [2,] 4.88 5.98    [2,] -4.600  1.9800    [2,] -3.1000 1.190
 [3,] 3.49 8.81    [3,] -1.460  2.1500    [3,] -2.6300 1.190
 [4,] 2.03 6.26    [4,] -1.780  0.7090    [4,] -0.2400 1.470
 [5,] 2.39 8.50    [5,] -2.610 -0.5120    [5,]  0.0637 2.110
   .   .    .         .      .       .         .     .       .
   .   .    .         .      .       .         .     .       .
   .   .    .         .      .       .         .     .       .
```

Explain and interpret the R commands and (slightly abbreviated) output below. In particular, you should describe the model being fitted, and comment on the hypothesis tests which are performed under the `summary` and `anova` commands.

```
> K <- dim(Y)[1]
> I <- dim(Y)[2]
> J <- dim(Y)[3]
> c(I,J,K)
[1]  2  3 10
> y <- as.vector(Y)
> a <- gl(I, K, length(y))
> b <- gl(J, K * I, length(y))
> fit1 <- lm(y ~ a + b)
> summary(fit1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7673     0.3032   12.43  < 2e-16 ***
a2            3.4542     0.3032   11.39 3.27e-16 ***
b2           -6.3215     0.3713  -17.03  < 2e-16 ***
b3           -5.8268     0.3713  -15.69  < 2e-16 ***
> anova(fit1)
```

```
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
a          1 178.98  178.98  129.83 3.272e-16 ***
b          2 494.39  247.19  179.31 < 2.2e-16 ***
Residuals 56  77.20    1.38
```

The following `R` code fits a similar model. Briefly explain the difference between this model and the one above. Based on the output of the `anova` call below, say whether you prefer this model over the one above, and explain your preference.

```
> fit2 <- lm(y ~ a * b)
> anova(fit2)
Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
a          1 178.98  178.98 125.6367 1.033e-15 ***
b          2 494.39  247.19 173.5241 < 2.2e-16 ***
a:b        2   0.27    0.14   0.0963    0.9084
Residuals 54  76.93    1.42
```

Finally, explain what is being calculated in the code below and give the value that would be obtained by the final line of code.

```
> n <- I * J * K
> p <- length(coef(fit2))
> p0 <- length(coef(fit1))
> PY <- fitted(fit2)
> P0Y <- fitted(fit1)
> ((n - p)/(p - p0)) * sum((PY - P0Y)^2)/sum((y - PY)^2)
```

2/I/5J Statistical Modelling

Suppose that we want to estimate the angles α, β and γ (in radians, say) of the triangle ABC, based on a single independent measurement of the angle at each corner. Suppose that the error in measuring each angle is normally distributed with mean zero and variance σ^2. Thus, we model our measurements y_A, y_B, y_C as the observed values of random variables

$$Y_A = \alpha + \varepsilon_A, \quad Y_B = \beta + \varepsilon_B, \quad Y_C = \gamma + \varepsilon_C,$$

where $\varepsilon_A, \varepsilon_B, \varepsilon_C$ are independent, each with distribution $N(0, \sigma^2)$. Find the maximum likelihood estimate of α based on these measurements.

Can the assumption that $\varepsilon_A, \varepsilon_B, \varepsilon_C \sim N(0, \sigma^2)$ be criticized? Why or why not?

3/I/5J Statistical Modelling

Consider the linear model $Y = X\beta + \varepsilon$. Here, Y is an n-dimensional vector of observations, X is a known $n \times p$ matrix, β is an unknown p-dimensional parameter, and $\varepsilon \sim N_n(0, \sigma^2 I)$, with σ^2 unknown. Assume that X has full rank and that $p \ll n$. Suppose that we are interested in checking the assumption $\varepsilon \sim N_n(0, \sigma^2 I)$. Let $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimate of β. Write in terms of X an expression for the projection matrix $P = (p_{ij} : 1 \leqslant i, j \leqslant n)$ which appears in the maximum likelihood equation $\hat{Y} = X\hat{\beta} = PY$.

Find the distribution of $\hat{\varepsilon} = Y - \hat{Y}$, and show that, in general, the components of $\hat{\varepsilon}$ are not independent.

A standard procedure used to check our assumption on ε is to check whether the studentized fitted residuals

$$\hat{\eta}_i = \frac{\hat{\varepsilon}_i}{\tilde{\sigma}\sqrt{1 - p_{ii}}}, \quad i = 1, \ldots, n,$$

look like a random sample from an $N(0, 1)$ distribution. Here,

$$\tilde{\sigma}^2 = \frac{1}{n-p} ||Y - X\hat{\beta}||^2.$$

Say, briefly, how you might do this in `R`.

This procedure appears to ignore the dependence between the components of $\hat{\varepsilon}$ noted above. What feature of the given set-up makes this reasonable?

4/I/5J **Statistical Modelling**

A long-term agricultural experiment had $n = 90$ grassland plots, each 25m × 25m, differing in biomass, soil pH, and species richness (the count of species in the whole plot). While it was well-known that species richness declines with increasing biomass, it was not known how this relationship depends on soil pH. In the experiment, there were 30 plots of "low pH", 30 of "medium pH" and 30 of "high pH". Three lines of the data are reproduced here as an aid.

```
> grass[c(1,31, 61), ]
     pH   Biomass Species
1  high 0.4692972      30
31  mid 0.1757627      29
61  low 0.1008479      18
```

Briefly explain the commands below. That is, explain the models being fitted.

```
> fit1 <- glm(Species ~ Biomass, family = poisson)
> fit2 <- glm(Species ~ pH + Biomass, family = poisson)
> fit3 <- glm(Species ~ pH * Biomass, family = poisson)
```

Let H_1, H_2 and H_3 denote the hypotheses represented by the three models and fits. Based on the output of the code below, what hypotheses are being tested, and which of the models seems to give the best fit to the data? Why?

```
> anova(fit1, fit2, fit3, test = "Chisq")
Analysis of Deviance Table
Model 1: Species ~ Biomass
Model 2: Species ~ pH + Biomass
Model 3: Species ~ pH * Biomass
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        88     407.67
2        86      99.24  2   308.43 1.059e-67
3        84      83.20  2    16.04 3.288e-04
```

Finally, what is the value obtained by the following command?

```
> mu.hat <- exp(predict(fit2))
> -2 * (sum(dpois(Species, mu.hat, log = TRUE)) - sum(dpois(Species,
+     Species, log = TRUE)))
```

4/II/13J **Statistical Modelling**

Consider the following generalized linear model for responses $y_1, \ldots, y_n$ as a function of explanatory variables $x_1, \ldots, x_n$, where $x_i = (x_{i1}, \ldots, x_{ip})^\top$ for $i = 1, \ldots, n$. The responses are modelled as observed values of independent random variables $Y_1, \ldots, Y_n$, with

$$Y_i \sim \mathrm{ED}(\mu_i, \sigma_i^2), \quad g(\mu_i) = x_i^\top \beta, \quad \sigma_i^2 = \sigma^2 a_i,$$

Here, g is a given link function, β and σ^2 are unknown parameters, and the a_i are treated as known.

[*Hint: recall that we write* $Y \sim ED(\mu, \sigma^2)$ *to mean that* Y *has density function of the form*

$$f(y; \mu, \sigma^2) = a(\sigma^2, y) \exp\left\{ \frac{1}{\sigma^2} [\theta(\mu) y - K(\theta(\mu))] \right\}$$

for given functions a *and* θ.]

[You may use without proof the facts that, for such a random variable Y,

$$E(Y) = K'(\theta(\mu)), \quad \mathrm{var}(Y) = \sigma^2 K''(\theta(\mu)) \equiv \sigma^2 V(\mu).]$$

Show that the score vector and Fisher information matrix have entries:

$$U_j(\beta) = \sum_{i=1}^n \frac{(y_i - \mu_i) x_{ij}}{\sigma_i^2 V(\mu_i) g'(\mu_i)}, \qquad j = 1, \ldots, p,$$

and

$$i_{jk}(\beta) = \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\sigma_i^2 V(\mu_i) (g'(\mu_i))^2}, \qquad j, k = 1, \ldots, p.$$

How do these expressions simplify when the canonical link is used?

Explain briefly how these two expressions can be used to obtain the maximum likelihood estimate $\hat{\beta}$ for β.

1/I/5I **Statistical Modelling**

According to the *Independent* newspaper (London, 8 March 1994) the Metropolitan Police in London reported 30475 people as missing in the year ending March 1993. For those aged 18 or less, 96 of 10527 missing males and 146 of 11363 missing females were still missing a year later. For those aged 19 and above, the values were 157 of 5065 males and 159 of 3520 females. This data is summarised in the table below.

```
    age gender still total
1   Kid      M    96 10527
2   Kid      F   146 11363
3 Adult      M   157  5065
4 Adult      F   159  3520
```

Explain and interpret the `R` commands and (slightly abbreviated) output below. You should describe the model being fitted, explain how the standard errors are calculated, and comment on the hypothesis tests being described in the summary. In particular, what is the worst of the four categories for the probability of remaining missing a year later?

```
> fit <- glm(still/total ~ age + gender, family = binomial,
+            weights = total)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.06073    0.07216 -42.417  < 2e-16 ***
ageKid      -1.27079    0.08698 -14.610  < 2e-16 ***
genderM     -0.37211    0.08671  -4.291 1.78e-05 ***

Residual deviance:   0.06514  on 1  degrees of freedom
```

For a person who was missing in the year ending in March 1993, find a formula, as a function of age and gender, for the estimated expected probability that they are still missing a year later.

1/II/13I Statistical Modelling

This problem deals with data collected as the number of each of two different strains of *Ceriodaphnia* organisms are counted in a controlled environment in which reproduction is occurring among the organisms. The experimenter places into the containers a varying concentration of a particular component of jet fuel that impairs reproduction. Hence it is anticipated that as the concentration of jet fuel grows, the mean number of organisms should decrease.

The table below gives a subset of the data. The full dataset has $n = 70$ rows. The first column provides the number of organisms, the second the concentration of jet fuel (in grams per litre) and the third specifies the strain of the organism.

number	fuel	strain
82	0	1
58	0	0
45	0.5	1
27	0.5	0
29	0.75	1
15	1.25	1
6	1.25	1
8	1.5	0
4	1.75	0
.	.	.
.	.	.

Explain and interpret the `R` commands and (slightly abbreviated) output below. In particular, you should describe the model being fitted, explain how the standard errors are calculated, and comment on the hypothesis tests being described in the summary.

```
> fit1 <- glm(number ~ fuel + strain + fuel:strain,family = poisson)

> summary(fit1)

Coefficients:

             Estimate Std. Error z value Pr(>|z|)

(Intercept)  4.14443    0.05101  81.252  < 2e-16 ***

fuel        -1.47253    0.07007 -21.015  < 2e-16 ***

strain       0.33667    0.06704   5.022 5.11e-07 ***

fuel:strain -0.12534    0.09385  -1.336    0.182
```

The following `R` code fits two very similar models. Briefly explain the difference between these models and the one above. Motivate the fitting of these models in light of the summary from the fit of the one above.

```
> fit2 <- glm(number ~ fuel + strain, family = poisson)
> fit3 <- glm(number ~ fuel, family = poisson)
```

Denote by H_1, H_2, H_3 the three hypotheses being fitted in sequence above.

Explain the hypothesis tests, including an approximate test of the fit of H_1, that can be performed using the output from the following R code. Use these numbers to comment on the most appropriate model for the data.

```
> c(fit1$dev, fit2$dev, fit3$dev)
[1]  84.59557  86.37646 118.99503
> qchisq(0.95, df = 1)
[1] 3.841459
```

2/I/5I **Statistical Modelling**

Consider the linear regression setting where the responses Y_i, $i = 1, \ldots, n$ are assumed independent with means $\mu_i = x_i^{\mathrm{T}}\beta$. Here x_i is a vector of known explanatory variables and β is a vector of unknown regression coefficients.

Show that if the response distribution is Laplace, i.e.,

$$Y_i \sim f(y_i; \mu_i, \sigma) = (2\sigma)^{-1} \exp\left\{-\frac{|y_i - \mu_i|}{\sigma}\right\}, \quad i = 1, \ldots, n; \quad y_i, \mu_i \in \mathbb{R};\ \sigma \in (0, \infty);$$

then the maximum likelihood estimate $\hat{\beta}$ of β is obtained by minimising

$$S_1(\beta) = \sum_{i=1}^{n} |Y_i - x_i^{\mathrm{T}}\beta|.$$

Obtain the maximum likelihood estimate for σ in terms of $S_1(\hat{\beta})$.

Briefly comment on why the Laplace distribution cannot be written in exponential dispersion family form.

3/I/5I **Statistical Modelling**

Consider two possible experiments giving rise to observed data y_{ij} where $i = 1, \ldots, I,\ j = 1, \ldots, J$.

1. The data are realizations of independent Poisson random variables, i.e.,

$$Y_{ij} \sim f_1(y_{ij}; \mu_{ij}) = \frac{\mu_{ij}^{y_{ij}}}{y_{ij}!} \exp\{-\mu_{ij}\}$$

where $\mu_{ij} = \mu_{ij}(\beta)$, with β an unknown (possibly vector) parameter. Write $\hat{\beta}$ for the maximum likelihood estimator (m.l.e.) of β and $\hat{y}_{ij} = \mu_{ij}(\hat{\beta})$ for the (i,j)th fitted value under this model.

2. The data are components of a realization of a multinomial random 'vector'

$$Y \sim f_2((y_{ij}); n, (p_{ij})) = n! \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{p_{ij}^{y_{ij}}}{y_{ij}!}$$

where the y_{ij} are non-negative integers with

$$\sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} = n \quad \text{and} \quad p_{ij}(\beta) = \frac{\mu_{ij}(\beta)}{n} .$$

Write β^* for the m.l.e. of β and $y^*_{ij} = np_{ij}(\beta^*)$ for the (i,j)th fitted value under this model.

Show that, if

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \hat{y}_{ij} = n ,$$

then $\hat{\beta} = \beta^*$ and $\hat{y}_{ij} = y^*_{ij}$ for all i, j. Explain the relevance of this result in the context of fitting multinomial models within a generalized linear model framework.

4/I/5I Statistical Modelling

Consider the normal linear model $Y = X\beta + \varepsilon$ in vector notation, where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \varepsilon_i \sim \text{i.i.d. } N(0,\sigma^2),$$

where $x_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ is known and X is of full rank ($p < n$). Give expressions for maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$ of β and σ^2 respectively, and state their joint distribution.

Suppose that there is a new pair (x^*, y^*), independent of $(x_1, y_1), \ldots, (x_n, y_n)$, satisfying the relationship

$$y^* = x^{*\mathrm{T}}\beta + \varepsilon^*, \quad \text{where} \quad \varepsilon^* \sim N(0, \sigma^2).$$

We suppose that x^* is known, and estimate y^* by $\tilde{y} = x^{*\mathrm{T}}\hat{\beta}$. State the distribution of

$$\frac{\tilde{y} - y^*}{\tilde{\sigma}\tau}, \quad \text{where} \quad \tilde{\sigma}^2 = \frac{n}{n-p}\hat{\sigma}^2 \quad \text{and} \quad \tau^2 = x^{*\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x^* + 1.$$

Find the form of a $(1-\alpha)$–level prediction interval for y^*.

4/II/13I Statistical Modelling

Let Y have a Gamma distribution with density

$$f(y; \alpha, \lambda) = \frac{\lambda^\alpha y^{\alpha-1}}{\Gamma(\alpha)} \, e^{-\lambda y}.$$

Show that the Gamma distribution is of exponential dispersion family form. Deduce directly the corresponding expressions for $\mathbb{E}[Y]$ and $\mathrm{Var}[Y]$ in terms of α and λ. What is the canonical link function?

Let $p < n$. Consider a generalised linear model (g.l.m.) for responses y_i, $i = 1, \ldots, n$ with random component defined by the Gamma distribution with canonical link $g(\mu)$, so that $g(\mu_i) = \eta_i = x_i^{\mathrm{T}}\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of unknown regression coefficients and $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is the vector of known values of the explanatory variables for the ith observation, $i = 1, \ldots, n$.

Obtain expressions for the score function and Fisher information matrix and explain how these can be used in order to approximate $\hat{\beta}$, the maximum likelihood estimator (m.l.e.) of β.

[*Use the canonical link function and assume that the dispersion parameter is known.*]

Finally, obtain an expression for the deviance for a comparison of the full (saturated) model to the g.l.m. with canonical link using the m.l.e. $\hat{\beta}$ (or estimated mean $\hat{\mu} = X\hat{\beta}$).

1/I/5I **Statistical Modelling**

Assume that observations $Y = (Y_1, \ldots, Y_n)^T$ satisfy the linear model

$$Y = X\beta + \epsilon,$$

where X is an $n \times p$ matrix of known constants of full rank $p < n$, where $\beta = (\beta_1, \ldots, \beta_p)^T$ is unknown and $\epsilon \sim N_n(0, \sigma^2 I)$. Write down a $(1-\alpha)$-level confidence set for β.

Define Cook's distance for the observation (x_i, Y_i), where x_i^T is the ith row of X. Give its interpretation in terms of confidence sets for β.

In the above model with $n = 50$ and $p = 2$, you observe that one observation has Cook's distance 1.3. Would you be concerned about the influence of this observation?

[*You may find some of the following facts useful:*
(i) *If* $Z \sim \chi^2_2$, *then* $\mathbb{P}(Z \leqslant 0.21) = 0.1$, $\mathbb{P}(Z \leqslant 1.39) = 0.5$ *and* $\mathbb{P}(Z \leqslant 4.61) = 0.9$.
(ii) *If* $Z \sim F_{2,48}$, *then* $\mathbb{P}(Z \leqslant 0.11) = 0.1$, $\mathbb{P}(Z \leqslant 0.70) = 0.5$ *and* $\mathbb{P}(Z \leqslant 2.42) = 0.9$.
(iii) *If* $Z \sim F_{48,2}$, *then* $\mathbb{P}(Z \leqslant 0.41) = 0.1$, $\mathbb{P}(Z \leqslant 1.42) = 0.5$ *and* $\mathbb{P}(Z \leqslant 9.47) = 0.9$.]

1/II/13I **Statistical Modelling**

The table below gives a year-by-year summary of the career batting record of the baseball player Babe Ruth. The first column gives his age at the start of each season and the second gives the number of 'At Bats' (AB) he had during the season. For each At Bat, it is recorded whether or not he scored a 'Hit'. The third column gives the total number of Hits he scored in the season, and the final column gives his 'Average' for the season, defined as the number of Hits divided by the number of At Bats.

Age	AB	Hits	Average
19	10	2	0.200
20	92	29	0.315
21	136	37	0.272
22	123	40	0.325
23	317	95	0.300
24	432	139	0.322
25	457	172	0.376
26	540	204	0.378
27	406	128	0.315
28	522	205	0.393
29	529	200	0.378
30	359	134	0.373
31	495	184	0.372
32	540	192	0.356
33	536	173	0.323
34	499	172	0.345
35	518	186	0.359
36	534	199	0.373
37	457	156	0.341
38	459	138	0.301
39	365	105	0.288
40	72	13	0.181

Explain and interpret the R commands below. In particular, you should explain the model that is being fitted, the approximation leading to the given standard errors and the test that is being performed in the last line of output.

```
> Mod <- glm(Hits/AB~Age+I(Age^2),family=binomial,weights=AB)
> summary(Mod)


Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.5406713  0.8487687  -5.350 8.81e-08 ***
Age          0.2684739  0.0565992   4.743 2.10e-06 ***
I(Age^2)    -0.0044827  0.0009253  -4.845 1.27e-06 ***


Residual deviance: 23.345  on 19  degrees of freedom
```

Assuming that any required packages are loaded, draw a careful sketch of the graph that you would expect to see on entering the following lines of code:

```
> Coef <- coef(Mod)
> Fitted <- inv.logit(Coef[[1]]+Coef[[2]]*Age+Coef[[3]]*Age^2)
> plot(Age,Average)
> lines(Age,Fitted)
```

2/I/5I **Statistical Modelling**

Let $Y_1, \ldots, Y_n$ be independent Poisson random variables with means $\mu_1, \ldots, \mu_n$, for $i = 1, \ldots, n$, where $\log(\mu_i) = \beta x_i$, for some known constants x_i and an unknown parameter β. Find the log-likelihood for β.

By first computing the first and second derivatives of the log-likelihood for β, explain the algorithm you would use to find the maximum likelihood estimator, $\hat{\beta}$.

3/I/5I **Statistical Modelling**

Consider a generalized linear model for independent observations $Y_1, \ldots, Y_n$, with $\mathbb{E}(Y_i) = \mu_i$ for $i = 1, \ldots, n$. What is a *linear predictor*? What is meant by the *link function*? If Y_i has model function (or density) of the form

$$f(y_i; \mu_i, \sigma^2) = \exp\left[\frac{1}{\sigma^2}\{\theta(\mu_i)y_i - K(\theta(\mu_i))\}\right] a(\sigma^2, y_i),$$

for $y_i \in \mathcal{Y} \subseteq \mathbb{R}$, $\mu_i \in \mathcal{M} \subseteq \mathbb{R}$, $\sigma^2 \in \Phi \subseteq (0, \infty)$, where $a(\sigma^2, y_i)$ is a known positive function, define the *canonical link function*.

Now suppose that $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \mathrm{Bin}(1, \mu_i)$ for $i = 1, \ldots, n$. Derive the canonical link function.

4/I/5I **Statistical Modelling**

The table below summarises the yearly numbers of named storms in the Atlantic basin over the period 1944–2004, and also gives an index of average July ocean temperature in the northern hemisphere over the same period. To save space, only the data for the first four and last four years are shown.

Year	Storms	Temp
1944	11	0.165
1945	11	0.080
1946	6	0.000
1947	9	-0.024
⋮	⋮	⋮
2001	15	0.592
2002	12	0.627
2003	16	0.608
2004	15	0.546

Explain and interpret the `R` commands and (slightly abbreviated) output below.

```
> Mod <- glm(Storms~Temp,family=poisson)

> summary(Mod)

Coefficients:

             Estimate Std. Error z value Pr(>|z|)

(Intercept)  2.26061    0.04841  46.697  < 2e-16 ***

Temp         0.48870    0.16973   2.879  0.00399 **


Residual deviance: 51.499  on 59  degrees of freedom
```

In 2005, the ocean temperature index was 0.743. Explain how you would predict the number of named storms for that year.

4/II/13I **Statistical Modelling**

Consider a linear model for $Y = (Y_1, \ldots, Y_n)^T$ given by

$$Y = X\beta + \epsilon,$$

where X is a known $n \times p$ matrix of full rank $p < n$, where β is an unknown vector and $\epsilon \sim N_n(0, \sigma^2 I)$. Derive an expression for the maximum likelihood estimator $\hat{\beta}$ of β, and write down its distribution.

Find also the maximum likelihood estimator $\hat{\sigma}^2$ of σ^2, and derive its distribution.

[*You may use Cochran's theorem, provided that it is stated carefully. You may also assume that the matrix* $P = X(X^TX)^{-1}X^T$ *has rank* p, *and that* $I - P$ *has rank* $n - p$.]

1/I/5I Statistical Modelling

Suppose that $Y_1, \ldots, Y_n$ are independent random variables, and that Y_i has probability density function

$$f(y_i|\theta_i, \phi) = \exp\left[\frac{(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi)\right] .$$

Assume that $\mathbb{E}(Y_i) = \mu_i$ and that there is a known link function $g(.)$ such that

$$g(\mu_i) = \beta^T x_i ,$$

where $x_1, \ldots, x_n$ are known p-dimensional vectors and β is an unknown p-dimensional parameter. Show that $\mathbb{E}(Y_i) = b'(\theta_i)$ and that, if $\ell(\beta, \phi)$ is the log-likelihood function from the observations $(y_1, \ldots, y_n)$, then

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta} = \sum_1^n \frac{(y_i - \mu_i)x_i}{g'(\mu_i)V_i} ,$$

where V_i is to be defined.

1/II/13I Statistical Modelling

The Independent, June 1999, under the headline 'Tourists get hidden costs warnings' gave the following table of prices in pounds, called 'How the resorts compared'.

Algarve	8.00	0.50	3.50	3.00	4.00	100.00
CostaDelSol	6.95	1.30	4.10	12.30	4.10	130.85
Majorca	10.25	1.45	5.35	6.15	3.30	122.20
Tenerife	12.30	1.25	4.90	3.70	2.90	130.85
Florida	15.60	1.90	5.05	5.00	2.50	114.00
Tunisia	10.90	1.40	5.45	1.90	2.75	218.10
Cyprus	11.60	1.20	5.95	3.00	3.60	149.45
Turkey	6.50	1.05	6.50	4.90	2.85	263.00
Corfu	5.20	1.05	3.75	4.20	2.50	137.60
Sorrento	7.70	1.40	6.30	8.75	4.75	215.40
Malta	11.20	0.70	4.55	8.00	4.80	87.85
Rhodes	6.30	1.05	5.20	3.15	2.70	261.30
Sicily	13.25	1.75	4.20	7.00	3.85	174.40
Madeira	10.25	0.70	5.10	6.85	6.85	153.70

Here the column headings are, respectively: Three-course meal, Bottle of Beer, Suntan Lotion, Taxi (5km), Film (24 exp), Car Hire (per week). Interpret the R commands, and explain how to interpret the corresponding (slightly abbreviated) R output given below. Your solution should include a careful statement of the underlying statistical model, but you may quote without proof any distributional results required.

```
> price = scan("dresorts") ; price

> Goods = gl(6,1,length=84); Resort=gl(14,6,length=84)

> first.lm = lm(log(price) ~ Goods + Resort)

> summary(first.lm)

 Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)   1.8778     0.1629  11.527  < 2e-16

Goods2       -2.1084     0.1295 -16.286  < 2e-16

Goods3       -0.6343     0.1295  -4.900 6.69e-06

Goods4       -0.6284     0.1295  -4.854 7.92e-06

Goods5       -0.9679     0.1295  -7.476 2.49e-10

Goods6        2.8016     0.1295  21.640  < 2e-16

Resort2       0.4463     0.1978   2.257  0.02740

Resort3       0.4105     0.1978   2.076  0.04189

Resort4       0.3067     0.1978   1.551  0.12584

Resort5       0.4235     0.1978   2.142  0.03597

Resort6       0.2883     0.1978   1.458  0.14963

Resort7       0.3457     0.1978   1.748  0.08519

Resort8       0.3787     0.1978   1.915  0.05993

Resort9       0.0943     0.1978   0.477  0.63508

Resort10      0.5981     0.1978   3.025  0.00356

Resort11      0.3281     0.1978   1.659  0.10187

Resort12      0.2525     0.1978   1.277  0.20616

Resort13      0.5508     0.1978   2.785  0.00700

Resort14      0.4590     0.1978   2.321  0.02343


 Residual standard error: 0.3425 on 65 degrees of freedom

 Multiple R-Squared: 0.962
```

2/I/5I Statistical Modelling

You see below three *R* commands, and the corresponding output (which is slightly abbreviated). Explain the effects of the commands. How is the deviance defined, and why do we have d.f.=7 in this case? Interpret the numerical values found in the output.

```
> n = scan()
  3 5 16 12 11 34 37 51 56

> i = scan ()
  1 2 3 4 5 6 7 8 9

> summary(glm(n~i,poisson))
  deviance = 13.218
      d.f. = 7
  Coefficients:
                 Value     Std.Error
  (intercept)    1.363     0.2210
  i              0.3106    0.0382
```

3/I/5I Statistical Modelling

Consider the model $Y = X\beta + \epsilon$, where Y is an n-dimensional observation vector, X is an $n \times p$ matrix of rank p, ϵ is an n-dimensional vector with components $\epsilon_1, \ldots, \epsilon_n$, and $\epsilon_1, \ldots, \epsilon_n$ are independently and normally distributed, each with mean 0 and variance σ^2.

(a) Let $\hat{\beta}$ be the least-squares estimator of β. Show that

$$(X^T X)\hat{\beta} = X^T Y$$

and find the distribution of $\hat{\beta}$.

(b) Define $\hat{Y} = X\hat{\beta}$. Show that $\hat{Y}$ has distribution $N(X\beta, \sigma^2 H)$, where H is a matrix that you should define.

[*You may quote without proof any results you require about the multivariate normal distribution.*]

4/I/5I **Statistical Modelling**

You see below five R commands, and the corresponding output (which is slightly abbreviated). Without giving any mathematical proofs, explain the purpose of these commands, and interpret the output.

```
> Yes = c(12, 27,11,24)

> Total = c(117,170,52,118)

> Sclass = c("a","a","b","b")

> Sclass = factor(Sclass)

> summary(glm(Yes/Total~ Sclass, binomial, weights=Total))

Coefficients:

            Estimate Std. Error z value

(Intercept)  -1.8499     0.1723 -10.739

Sclassb       0.4999     0.2562   1.951


Residual deviance: 1.9369  on 2  degrees of freedom


Number of Fisher Scoring iterations: 4
```

4/II/13I **Statistical Modelling**

(i) Suppose that $Y_1, \ldots, Y_n$ are independent random variables, and that Y_i has probability density function

$$f(y_i|\beta, \nu) = \left(\frac{\nu y_i}{\mu_i}\right)^\nu \; e^{-y_i \nu/\mu_i} \; \frac{1}{\Gamma(\nu)} \, \frac{1}{y_i} \quad \text{for } y_i > 0$$

where

$$1/\mu_i = \beta^T x_i \; , \text{ for } \quad 1 \leqslant i \leqslant n,$$

and $x_1, \ldots, x_n$ are given p-dimensional vectors, and ν is known.

Show that $\mathbb{E}(Y_i) = \mu_i$ and that $\operatorname{var}(Y_i) = \mu_i^2/\nu$.

(ii) Find the equation for $\hat{\beta}$, the maximum likelihood estimator of β, and suggest an iterative scheme for its solution.

(iii) If $p = 2$, and $x_i = \begin{pmatrix} 1 \\ z_i \end{pmatrix}$, find the large-sample distribution of $\hat{\beta}_2$. Write your answer in terms of a, b, c and ν, where a, b, c are defined by

$$a = \sum \mu_i^2, \quad b = \sum z_i \mu_i^2, \quad c = \sum z_i^2 \mu_i^2 .$$

A1/13 **Computational Statistics and Statistical Modelling**

(i) Assume that the n-dimensional vector Y may be written as $Y = X\beta + \epsilon$, where X is a given $n \times p$ matrix of rank p, β is an unknown vector, and

$$\epsilon \sim N_n(0, \sigma^2 I).$$

Let $Q(\beta) = (Y - X\beta)^T(Y - X\beta)$. Find $\hat{\beta}$, the least-squares estimator of β, and state without proof the joint distribution of $\hat{\beta}$ and $Q(\hat{\beta})$.

(ii) Now suppose that we have observations $(Y_{ij}, 1 \leqslant i \leqslant I,\ 1 \leqslant j \leqslant J)$ and consider the model

$$\Omega : Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}\ ,$$

where $(\alpha_i), (\beta_j)$ are fixed parameters with $\Sigma\alpha_i = 0$, $\Sigma\beta_j = 0$, and (ϵ_{ij}) may be assumed independent normal variables, with $\epsilon_{ij} \sim N(0, \sigma^2)$, where σ^2 is unknown.

(a) Find $(\hat{\alpha}_i)$, $(\hat{\beta}_j)$, the least-squares estimators of (α_i), (β_j).

(b) Find the least-squares estimators of (α_i) under the hypothesis $H_0 : \beta_j = 0$ for all j.

(c) Quoting any general theorems required, explain carefully how to test H_0, assuming Ω is true.

(d) What would be the effect of fitting the model $\Omega_1 : Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$, where now (α_i), (β_j), (γ_{ij}) are all fixed unknown parameters, and (ϵ_{ij}) has the distribution given above?

A2/12 **Computational Statistics and Statistical Modelling**

(i) Suppose we have independent observations $Y_1, \ldots, Y_n$, and we assume that for $i = 1, \ldots, n$, Y_i is Poisson with mean μ_i, and $\log(\mu_i) = \beta^T x_i$, where $x_1, \ldots, x_n$ are given covariate vectors each of dimension p, where β is an unknown vector of dimension p, and $p < n$. Assuming that $\{x_1, \ldots, x_n\}$ span $\mathbb{R}^p$, find the equation for $\hat{\beta}$, the maximum likelihood estimator of β, and write down the large-sample distribution of $\hat{\beta}$.

(ii) A long-term agricultural experiment had 90 grassland plots, each 25m $\times$ 25m, differing in biomass, soil pH, and species richness (the count of species in the whole plot). While it was well-known that species richness declines with increasing biomass, it was not known how this relationship depends on soil pH, which for the given study has possible values "low", "medium" or "high", each taken 30 times. Explain the commands input, and interpret the resulting output in the (slightly edited) *R* output below, in which "species" represents the species count.

(The first and last 2 lines of the data are reproduced here as an aid. You may assume that the factor pH has been correctly set up.)

```
> species
     pH    Biomass Species
1  high 0.46929722      30
2  high 1.73087043      39
  ......................
  ......................
89  low 4.36454121       7
90  low 4.87050789       3

>  summary(glm(Species ~Biomass, family = poisson))
Call:
glm(formula = Species ~ Biomass, family = poisson)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.184094   0.039159   81.31  < 2e-16
Biomass     -0.064441   0.009838   -6.55 5.74e-11

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 452.35  on 89  degrees of freedom
Residual deviance: 407.67  on 88  degrees of freedom

Number of Fisher Scoring iterations: 4

> summary(glm(Species ~pH*Biomass, family = poisson))
Call:
glm(formula = Species ~ pH * Biomass, family = poisson)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76812    0.06153  61.240  < 2e-16
pHlow         -0.81557    0.10284  -7.931 2.18e-15
```

Question continues on next page.

```
pHmid         -0.33146    0.09217  -3.596 0.000323
Biomass       -0.10713    0.01249  -8.577  < 2e-16
pHlow:Biomass -0.15503    0.04003  -3.873 0.000108
pHmid:Biomass -0.03189    0.02308  -1.382 0.166954

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom

Number of Fisher Scoring iterations: 4
```

A4/14 **Computational Statistics and Statistical Modelling**

Suppose that $Y_1, \ldots, Y_n$ are independent observations, with Y_i having probability density function of the following form

$$f(y_i|\theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right]$$

where $\mathbb{E}(Y_i) = \mu_i$ and $g(\mu_i) = \beta^T x_i$. You should assume that $g(\ )$ is a known function, and β, ϕ are unknown parameters, with $\phi > 0$, and also $x_1, \ldots, x_n$ are given linearly independent covariate vectors. Show that

$$\frac{\partial \ell}{\partial \beta} = \sum \frac{(y_i - \beta_i)}{g'(\mu_i)V_i} x_i,$$

where ℓ is the log-likelihood and $V_i = \mathrm{var}\,(Y_i) = \phi b''(\theta_i)$.

Discuss carefully the (slightly edited) R output given below, and briefly suggest another possible method of analysis using the function glm ( ).

```
> s <- scan()
1: 33 63 157 38 108 159
7:
Read 6 items
> r <- scan()
1: 3271 7256 5065 2486 8877 3520
7:
Read 6 items
> gender <- scan(,"")
1: b b b g g g
7:
Read 6 items
> age <- scan(,"")
1: 13&under 14-18 19&over
4: 13&under 14-18 19&over
7:
Read 6 items
> gender <- factor(gender) ; age <- factor(age)
>  summary(glm(s/r ~ gender + age,binomial, weights=r))


       Coefficients:
```

Question continues on next page.

```
            Estimate  Std.Error  z-value Pr(>|z|)
(Intercept) -4.56479    0.12783 -35.710  < 2e-16
genderg      0.38028    0.08689   4.377 1.21e-05
age14-18    -0.19797    0.14241  -1.390    0.164
age19&over   1.12790    0.13252   8.511  < 2e-16

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 221.797542  on 5  degrees of freedom
Residual deviance:   0.098749  on 2  degrees of freedom

Number of Fisher Scoring iterations: 3
```

A1/13 **Computational Statistics and Statistical Modelling**

(i) Suppose Y_i, $1 \leqslant i \leqslant n$, are independent binomial observations, with $Y_i \sim Bi(t_i, \pi_i)$, $1 \leqslant i \leqslant n$, where $t_1, \ldots, t_n$ are known, and we wish to fit the model

$$\omega : \log \frac{\pi_i}{1-\pi_i} = \mu + \beta^T x_i \quad \text{for each } i,$$

where $x_1, \ldots, x_n$ are given covariates, each of dimension p. Let $\hat{\mu}$, $\hat{\beta}$ be the maximum likelihood estimators of μ, β. Derive equations for $\hat{\mu}$, $\hat{\beta}$ and state without proof the form of the approximate distribution of $\hat{\beta}$.

(ii) In 1975, data were collected on the 3-year survival status of patients suffering from a type of cancer, yielding the following table

		survive?	
age in years	malignant	yes	no
under 50	no	77	10
under 50	yes	51	13
50-69	no	51	11
50-69	yes	38	20
70+	no	7	3
70+	yes	6	3

Here the second column represents whether the initial tumour was not malignant or was malignant.

Let Y_{ij} be the number surviving, for age group i and malignancy status j, for $i = 1, 2, 3$ and $j = 1, 2$, and let t_{ij} be the corresponding total number. Thus $Y_{11} = 77$, $t_{11} = 87$. Assume $Y_{ij} \sim Bi(t_{ij}, \pi_{ij})$, $1 \leqslant i \leqslant 3$, $1 \leqslant j \leqslant 2$. The results from fitting the model

$$\log(\pi_{ij}/(1-\pi_{ij})) = \mu + \alpha_i + \beta_j$$

with $\alpha_1 = 0$, $\beta_1 = 0$ give $\hat{\beta}_2 = -0.7328$ (se $= 0.2985$), and deviance $= 0.4941$. What do you conclude?

Why do we take $\alpha_1 = 0$, $\beta_1 = 0$ in the model?

What "residuals" should you compute, and to which distribution would you refer them?

A2/12 **Computational Statistics and Statistical Modelling**

(i) Suppose $Y_1, \ldots, Y_n$ are independent Poisson variables, and

$$\mathbb{E}(Y_i) = \mu_i\,, \qquad \log \mu_i = \alpha + \beta t_i\,, \qquad \text{for} \quad i = 1, \ldots, n\,,$$

where α, β are two unknown parameters, and $t_1, \ldots, t_n$ are given covariates, each of dimension 1. Find equations for $\hat{\alpha}, \hat{\beta}$, the maximum likelihood estimators of α, β, and show how an estimate of $\mathrm{var}(\hat{\beta})$ may be derived, quoting any standard theorems you may need.

(ii) By 31 December 2001, the number of new vCJD patients, classified by reported calendar year of onset, were

$$8, 10, 11, 14, 17, 29, 23$$

for the years

$$1994, \ldots, 2000 \quad \text{respectively.}$$

Discuss carefully the (slightly edited) *R* output for these data given below, quoting any standard theorems you may need.

```
> year
year
[1] 1994 1995 1996 1997 1998 1999 2000
> tot
[1]  8 10 11 14 17 29 23
>first.glm _ glm(tot ~ year, family = poisson)
> summary(first.glm)
Call:
glm(formula = tot ~ year, family = poisson)
Coefficients
                    Estimate Std. Error z value Pr(>|z|)
      (Intercept) -407.81285   99.35366  -4.105 4.05e-05
      year           0.20556    0.04973   4.133 3.57e-05


      (Dispersion parameter for poisson family taken to be 1)


      Null deviance: 20.7753  on 6  degrees of freedom
  Residual deviance:  2.7931  on 5  degrees of freedom


  Number of Fisher Scoring iterations: 3
```

A4/14 **Computational Statistics and Statistical Modelling**

The nave height x, and the nave length y for 16 Gothic-style cathedrals and 9 Romanesque-style cathedrals, all in England, have been recorded, and the corresponding R output (slightly edited) is given below.

```
> first.lm _ lm(y ~ x + Style); summary(first.lm)
Call:
lm(formula = y ~ x + Style)


Residuals:
    Min      1Q  Median      3Q     Max
-172.67  -30.44   20.38   55.02   96.50


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.298     81.648   0.543   0.5929
x              4.712      1.058   4.452   0.0002
Style2        80.393     32.306   2.488   0.0209


Residual standard error: 77.53 on 22 degrees of freedom
Multiple R-Squared: 0.5384
```

You may assume that x, y are in suitable units, and that "style" has been set up as a factor with levels 1,2 corresponding to Gothic, Romanesque respectively.

(a) Explain carefully, with suitable graph(s) if necessary, the results of this analysis.

(b) Using the general model $Y = X\beta + \epsilon$ (in the conventional notation) explain carefully the theory needed for (a).

[*Standard theorems need not be proved.*]

A1/13 **Computational Statistics and Statistical Modelling**

(i) Suppose $Y_1, \ldots, Y_n$ are independent Poisson variables, and

$$\mathbb{E}(Y_i) = \mu_i \ , \ \log \mu_i = \alpha + \beta^T x_i \ , \ 1 \leqslant i \leqslant n$$

where α, β are unknown parameters, and $x_1, \ldots, x_n$ are given covariates, each of dimension p. Obtain the maximum-likelihood equations for α, β, and explain briefly how you would check the validity of this model.

(ii) The data below show $y_1, \ldots, y_{33}$, which are the monthly accident counts on a major US highway for each of the 12 months of 1970, then for each of the 12 months of 1971, and finally for the first 9 months of 1972. The data-set is followed by the (slightly edited) *R* output. You may assume that the factors 'Year' and 'month' have been set up in the appropriate fashion. Give a careful interpretation of this *R* output, and explain (a) how you would derive the corresponding standardised residuals, and (b) how you would predict the number of accidents in October 1972.

```
52 37  49  29  31  32  28  34  32  39  50  63
35 22  27  27  34  23  42  30  36  56  48  40
33 26  31  25  23  20  25  20  36
```

> first.glm _ glm(y~ Year + month, poisson) ; summary(first.glm)

Call:

glm(formula = y ~ Year + month, family = poisson)

Coefficients:

	Estimate	Std. Error	z value	Pr(> |z|)	
(Intercept)	3.81969	0.09896	38.600	< 2e − 16	***
Year1971	-0.12516	0.06694	-1.870	0.061521	.
Year1972	-0.28794	0.08267	-3.483	0.000496	***
month2	-0.34484	0.14176	-2.433	0.014994	*
month3	-0.11466	0.13296	-0.862	0.388459	
month4	-0.39304	0.14380	-2.733	0.006271	**
month5	-0.31015	0.14034	-2.210	0.027108	*
month6	-0.47000	0.14719	-3.193	0.001408	**
month7	-0.23361	0.13732	-1.701	0.088889	.
month8	-0.35667	0.14226	-2.507	0.012168	*
month9	-0.14310	0.13397	-1.068	0.285444	
month10	0.10167	0.13903	0.731	0.464628	
month11	0.13276	0.13788	0.963	0.335639	
month12	0.18252	0.13607	1.341	0.179812	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

(Dispersion parameter for poisson family taken to be 1)

```
    Null deviance:     101.143   on 32 degrees of freedom
Residual deviance:     27.273    on 19 degrees of freedom
```

Number of Fisher Scoring iterations: 3

A2/12 **Computational Statistics and Statistical Modelling**

(i) Suppose that the random variable Y has density function of the form

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right]$$

where $\phi > 0$. Show that Y has expectation $b'(\theta)$ and variance $\phi b''(\theta)$.

(ii) Suppose now that $Y_1, \ldots, Y_n$ are independent negative exponential variables, with Y_i having density function $f(y_i|\mu_i) = \frac{1}{\mu_i} e^{-y_i/\mu_i}$ for $y_i > 0$. Suppose further that $g(\mu_i) = \beta^T x_i$ for $1 \leqslant i \leqslant n$, where $g(\cdot)$ is a known 'link' function, and $x_1, \ldots, x_n$ are given covariate vectors, each of dimension p. Discuss carefully the problem of finding $\hat{\beta}$, the maximum-likelihood estimator of β, firstly for the case $g(\mu_i) = 1/\mu_i$, and secondly for the case $g(\mu) = \log \mu_i$; in both cases you should state the large-sample distribution of $\hat{\beta}$.

[*Any standard theorems used need not be proved.*]

A4/14 **Computational Statistics and Statistical Modelling**

Assume that the n-dimensional observation vector Y may be written as $Y = X\beta + \epsilon$, where X is a given $n \times p$ matrix of rank p, β is an unknown vector, with $\beta^T = (\beta_1, \ldots, \beta_p)$, and

$$\epsilon \sim N_n(0, \sigma^2 I) \qquad (*)$$

where σ^2 is unknown. Find $\hat{\beta}$, the least-squares estimator of β, and describe (without proof) how you would test

$$H_0 : \beta_\nu = 0$$

for a given ν.

Indicate briefly two plots that you could use as a check of the assumption $(*)$.

Continued opposite

Sulphur dioxide is one of the major air pollutants. A data-set presented by Sokal and Rohlf (1981) was collected on 41 US cities in 1969-71, corresponding to the following variables:

Y = sulphur dioxide content of air in micrograms per cubic metre

$X1$ = average annual temperature in degrees Fahrenheit

$X2$ = number of manufacturing enterprises employing 20 or more workers

$X3$ = population size (1970 census) in thousands

$X4$ = average annual wind speed in miles per hour

$X5$ = average annual precipitation in inches

$X6$ = average annual of days with precipitation per year.

Interpret the R output that follows below, quoting any standard theorems that you need to use.

> next.lm _ lm(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6)

> summary(next.lm)

Call: lm(formula = log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6)

```
Residuals:
       Min        1Q    Median        3Q       Max
  -0.79548  -0.25538  -0.01968  0.28328  0.98029
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	7.2532456	1.4483686	5.008	1.68e-05	***
X1	-0.0599017	0.0190138	-3.150	0.00339	**
X2	0.0012639	0.0004820	2.622	0.01298	*
X3	-0.0007077	0.0004632	-1.528	0.13580	
X4	-0.1697171	0.0555563	-3.055	0.00436	**
X5	0.0173723	0.0111036	1.565	0.12695	
X6	0.0004347	0.0049591	0.088	0.93066	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Residual standard error: 0.448 on 34 degrees of freedom

Multiple R-Squared: 0.6541

F-statistic: 10.72 on 6 and 34 degrees of freedom, p-value: 1.126e-06

A1/13 **Computational Statistics and Statistical Modelling**

(i) Assume that the n-dimensional observation vector Y may be written as

$$Y = X\beta + \epsilon \quad ,$$

where X is a given $n \times p$ matrix of rank p, β is an unknown vector, and

$$\epsilon \sim N_n(0, \sigma^2 I).$$

Let $Q(\beta) = (Y - X\beta)^T(Y - X\beta)$. Find $\widehat{\beta}$, the least-squares estimator of β, and show that

$$Q(\widehat{\beta}) = Y^T(I - H)Y \quad ,$$

where H is a matrix that you should define.

(ii) Show that $\sum_i H_{ii} = p$. Show further for the special case of

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \epsilon_i, \quad 1 \leqslant i \leqslant n,$$

where $\Sigma x_i = 0, \quad \Sigma z_i = 0$, that

$$H \quad = \quad \frac{1}{n}\mathbf{1}\mathbf{1}^T + axx^T + b(xz^T + zx^T) + czz^T \quad ;$$

here, $\mathbf{1}$ is a vector of which every element is one, and a, b, c, are constants that you should derive.

Hence show that, if $\widehat{Y} = X\widehat{\beta}$ is the vector of fitted values, then

$$\frac{1}{\sigma^2}\mathrm{var}(\widehat{Y}_i) = \frac{1}{n} + ax_i^2 + 2bx_iz_i + cz_i^2, \quad 1 \leqslant i \leqslant n.$$

A2/12 **Computational Statistics and Statistical Modelling**

(i) Suppose that $Y_1, \ldots, Y_n$ are independent random variables, and that Y_i has probability density function

$$f(y_i|\theta_i, \phi) = \exp[(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)].$$

Assume that $E(Y_i) = \mu_i$, and that $g(\mu_i) = \beta^T x_i$, where $g(\cdot)$ is a known 'link' function, $x_1, \ldots, x_n$ are known covariates, and β is an unknown vector. Show that

$$\mathbb{E}(Y_i) = b'(\theta_i),\ \operatorname{var}(Y_i) = \phi b''(\theta_i) = V_i,\ \text{say},$$

and hence

$$\frac{\partial l}{\partial \beta} \quad = \quad \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{g'(\mu_i)V_i},\ \text{where } l = l(\beta, \phi) \text{ is the log-likelihood.}$$

(ii) The table below shows the number of train miles (in millions) and the number of collisions involving British Rail passenger trains between 1970 and 1984. Give a detailed interpretation of the R output that is shown under this table:

	year	collisions	miles
1	1970	3	281
2	1971	6	276
3	1972	4	268
4	1973	7	269
5	1974	6	281
6	1975	2	271
7	1976	2	265
8	1977	4	264
9	1978	1	267
10	1979	7	265
11	1980	3	267
12	1981	5	260
13	1982	6	231
14	1983	1	249

```
Call:

glm(formula = collisions ~ year + log(miles), family = poisson)

Coefficients:
                    Estimate  Std. Error  z value  Pr(>|z|)
      (Intercept)  127.14453   121.37796    1.048     0.295
      year          -0.05398     0.05175   -1.043     0.297
      log(miles)    -3.41654     4.18616   -0.816     0.414

(Dispersion parameter for poisson family taken to be 1)

      Null deviance:  15.937 on 13 degrees of freedom

Residual deviance:  14.843 on 11 degrees of freedom

Number of Fisher Scoring iterations:  4
```

A4/14 **Computational Statistics and Statistical Modelling**

(i) Assume that independent observations $Y_1, \ldots, Y_n$ are such that

$$Y_i \sim \text{Binomial}(t_i, \pi_i), \ \log \frac{\pi_i}{1-\pi_i} = \beta^T x_i \quad \text{for } 1 \leqslant i \leqslant n \ ,$$

where $x_1, \ldots, x_n$ are given covariates. Discuss carefully how to estimate β, and how to test that the model fits.

(ii) Carmichael *et al.* (1989) collected data on the numbers of 5-year old children with "dmft", i.e. with 5 or more decayed, missing or filled teeth, classified by social class, and by whether or not their tap water was fluoridated or non-fluoridated. The numbers of such children with dmft, and the total numbers, are given in the table below:

dmft

Social Class	Fluoridated	Non-fluoridated
I	12/117	12/56
II	26/170	48/146
III	11/52	29/64
Unclassified	24/118	49/104

A (slightly edited) version of the *R* output is given below. Explain carefully what model is being fitted, whether it does actually fit, and what the parameter estimates and Std. Errors are telling you. (You may assume that the factors SClass (social class) and Fl (with/without) have been correctly set up.)

```
Call:

    glm(formula = Yes/Total ~ SClass + Fl, family = binomial,
          weights = Total)

Coefficients:

                Estimate  Std.  Error  z value
    (Intercept) -2.2716       0.2396   -9.480
    SClassII     0.5099       0.2628    1.940
    SClassIII    0.9857       0.3021    3.262
    SClassUnc    1.0020       0.2684    3.734
    Flwithout    1.0813       0.1694    6.383

(Dispersion parameter for binomial family taken to be 1)

     Null deviance:  68.53785 on 7 degrees of freedom

Residual deviance:  0.64225 on 3 degrees of freedom

Number of Fisher Scoring iterations:  3
```

Here 'Yes' is the vector of numbers with dmft, taking values $12, 12, \ldots, 24, 49$, 'Total' is the vector of Total in each category, taking values $117, 56, \ldots, 118, 104$, and SClass, Fl are the factors corresponding to Social class and Fluoride status, defined in the obvious way.