UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Tuesday 6 June 2006 1.30 to 4.30

PAPER 41

APPLIED STATISTICS

*Attempt **FOUR** questions.*

*There are **FIVE** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Suppose $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$, $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$, X is known $n \times p$ matrix with rank $p\,(< n)$, and $\boldsymbol{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_n)$, where $\epsilon_1, \ldots, \epsilon_n$ are independent normal random variables with mean 0 and variance σ^2. Obtain the least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, and derive the distribution of $\hat{\boldsymbol{\beta}}$.

Let $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$ be the vector of residuals, where the vector of fitted values is $\hat{\mathbf{Y}} = H\mathbf{Y}$ for a matrix H that you should identify. Show that the residuals are uncorrelated with $\hat{\mathbf{Y}}$, and uncorrelated with $\hat{\boldsymbol{\beta}}$.

[You may assume that for a $n \times 1$ random vector $\mathbf{Y}$ and $n \times n$ matrices A and B, $\operatorname{cov}(A\mathbf{Y}, B\mathbf{Y}) = A \operatorname{cov}(\mathbf{Y})B^T$].

For $i = 1, \ldots, n$, let Y_i be the yield in litres of a chemical experiment and suppose

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i \ ,$$

where $x_i = (t_i - 100)/10$ and t_i is the temperature in degrees celsius used in the i^{th} experiment. Suppose $n = 30$ and that $x_1 = \ldots = x_{10} = -1$, $x_{11} = \ldots = x_{20} = 0$, $x_{21} = \ldots = x_{30} = 1$. Obtain the covariance matrix of the least squares estimator $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})^T$, and find an expression for the variance of $\hat{\alpha} + \hat{\beta}x + \hat{\gamma}x^2$. Find the values of x in $[-1, 1]$ for which this variance is (i) maximised and (ii) minimised.

Given that the residual sum of squares is 2.43, derive a 95% confidence interval for the expected yield when the temperature is 100^{O}C.

[You may assume that for $b \neq 0$, $a \neq b$

$$\begin{bmatrix} a & 0 & b \\ 0 & b & 0 \\ b & 0 & b \end{bmatrix}^{-1} = \frac{1}{b(a-b)} \begin{bmatrix} b & 0 & -b \\ 0 & a-b & 0 \\ -b & 0 & a \end{bmatrix} . \quad]$$

2 Explain what is meant by a *factor* in analysis of variance. Suppose a response variable depends on two factors A and B. Define the interaction between A and B.

The edited S-Plus output below shows part of an analysis of the results of an experiment investigating how three factors affect the lustre value of plastic film. The three factors are the thickness of the film (with levels 0 and 1 corresponding to thin and thick), the temperature (0 and 1 corresponding to low and high) of the wash used in preparation of the film, and the length of the wash ($1, 2, 3, 4$ corresponding to $20, 30, 40, 60$ minutes). Assume that the S-Plus objects `Thickness`, `Temperature` and `Wash` have been correctly set up as these factors.

Give the algebraic form and assumptions for the model `film1.lm`. State the missing degrees of freedom in the output to the directive `anova(film1.lm)`, and explain carefully why the model `film2.lm` is fitted next.

Find the fitted values under this model and explain how to find the standard errors of these fitted values. Summarise briefly, with appropriate sketch graph(s) as required, how lustre depends on the three factors.

```
> lustre
 [1]  3.3  4.1  4.9  5.0  3.4  4.0  4.2  4.9 19.6 17.5 17.6 20.9 14.5 17.0 15.2
[16] 17.1  5.5  5.7  5.6  7.2  3.7  6.1  5.7  6.0 26.6 31.6 30.5 31.4 29.5 30.2
[31] 30.2 29.6
> Thickness
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
> Temperature
 [1] 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1
> Wash
 [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
> options(contrasts=c("contr.treatment","contr.poly"))
> film1.lm <- lm(lustre~Thickness*Temperature*Wash)
> anova(film1.lm)
Analysis of Variance Table

Response: lustre

Terms added sequentially (first to last)
                           Df Sum of Sq  Mean Sq  F Value     Pr(F)
                 Thickness  *   391.300  391.300  191.082 0.0000000
               Temperature  *  2806.878 2806.878 1370.671 0.0000000
                      Wash  *    16.443    5.481    2.677 0.0821499
     Thickness:Temperature  *   244.758  244.758  119.522 0.0000000
            Thickness:Wash  *     3.568    1.189    0.581 0.6360491
          Temperature:Wash  *     0.856    0.285    0.139 0.9349992
Thickness:Temperature:Wash  *     3.466    1.155    0.564 0.6464399
                 Residuals  *    32.765    2.048
> film2.lm <- lm(lustre~Thickness*Temperature)
> summary(film2.lm,cor=F)

Call: lm(formula = lustre ~ Thickness * Temperature)
Residuals:
   Min      1Q Median     3Q   Max
 -3.35 -0.3687 0.0125 0.5813 3.475

Coefficients:
                        Value Std. Error t value Pr(>|t|)
          (Intercept)  4.2250  0.5049     8.3683  0.0000
```

```
           Thickness  1.4625  0.7140     2.0483  0.0500
         Temperature 13.2000  0.7140    18.4871  0.0000
Thickness:Temperature 11.0625  1.0098    10.9555  0.0000

Residual standard error: 1.428 on 28 degrees of freedom
Multiple R-Squared: 0.9837
F-statistic: 562.8 on 3 and 28 degrees of freedom, the p-value is 0
```

3 The random variable Y has a Poisson distribution with mean μ. Show that $\mathbb{P}(Y = y)$ can be written in the form

$$\exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right\},$$

and identify $\theta, b(\theta)$ and ϕ. Verify that $b'(\theta) = \mathbb{E}(Y)$ and $\phi b''(\theta) = \operatorname{var}(Y)$. Explain what is meant by a generalised linear model for the distribution of Y. What is the canonical link function for the Poisson distribution?

Car insurance claims in a particular year are classified according to the class and merit rating of the policyholder. The four merit ratings are

3 no claims for 3 or more years
2 no claims for 2 years
1 no claims for 1 year
0 other.

There are five classes

1 non-business, no male driver under 25
2 non-business, secondary (but not principal) driver is male, under 25
3 business
4 non-business, principal driver is male, under 25 and unmarried
5 non-business, principal driver is male, under 25 and married

Let Y_{ij} and n_{ij} be the number of claims and the number of car years insured in class i with merit rating j, for $i = 1, \ldots, 5$, $j = 0, 1, 2, 3$. Consider the *rate* λ_{ij} of claims per car year insured in class i with merit rating j.

Write down the algebraic form, together with the assumptions, of the model `carins.glm` in the (slightly edited) S-Plus output below where `claims` and `insured` are the number of claims and the claim years insured respectively. Derive equations satisfied by the maximum likelihood estimates of the parameters in the model. Interpret carefully the rest of the S-Plus output.

```
> car <- read.table("carinsurance",header=T)
> car
   merit class insured claims
 1     3     1 2757520 217151
 2     3     2  130535  14506
 3     3     3  247424  31964
 4     3     4  156871  22884
 5     3     5   64130   6560
 6     2     1  130706  13792
 7     2     2    7233   1001
 8     2     3   15868   2695
 9     2     4   17707   3054
10     2     5    4039    487
11     1     1  163544  19346
12     1     2    9726   1430
13     1     3   20369   3546
```

```
14     1     4   21089   3618
15     1     5    4869    613
16     0     1  273944  37730
17     0     2   21504   3421
18     0     3   37666   7565
19     0     4   56730  11345
20     0     5    8601   1291
> attach(car)
> merit <- factor(merit)
> class <- factor(class)
> options(contrasts=c("contr.treatment","contr.poly"))
> carins.glm <- glm(claims~offset(log(insured))+merit+class,poisson)
> summary(carins.glm,cor=F)
Deviance Residuals:
       Min        1Q    Median       3Q      Max
 -10.79274 -3.007873 -1.575749 2.426679 11.62523

Coefficients:
                 Value  Std. Error    t value
(Intercept) -2.0357359 0.004311305 -472.18556
     merit1 -0.1377590 0.007172219  -19.20730
     merit2 -0.2206796 0.007997189  -27.59465
     merit3 -0.4929506 0.004502371 -109.48689
     class2  0.2998302 0.007258049   41.31003
     class3  0.4690550 0.005039141   93.08233
     class4  0.5258551 0.005364533   98.02439
     class5  0.2155504 0.010734511   20.08013

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: 33854.16 on 19 degrees of freedom

Residual Deviance: 579.5163 on 12 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

4 Let $Y_1, \ldots, Y_m$ be independent random variables with $Y_i \sim \mathrm{Bin}(n_i, p_i)$, $i = 1, \ldots, m$, and consider the model $\omega_1 : \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta}$, $i = 1, \ldots, m$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\mathbf{x}_i$ is a vector of covariate values for Y_i. Derive an expression for the deviance D_1 for this model. Find an expression for the deviance D_0 for the model $\omega_0 : \log\left(\frac{p_i}{1-p_i}\right) = \widetilde{\mathbf{x}}_i^T \widetilde{\boldsymbol{\beta}}$, $i = 1, \ldots, m$, where $\widetilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_k)^T$ for some fixed k, $1 \leqslant k < p$, and where $\tilde{\mathbf{x}}_i$ is the vector of corresponding covariate values. Explain how D_0 and D_1 may be used to test $H_0 : \beta_{k+1} = \ldots = \beta_p = 0$.

Anthers of a particular plant species are prepared under two different storage conditions ($1, 2$ are control, treatment respectively) and with different centrifuging forces ($40g, 150g, 350g$), and the number of embryogenic anthers is observed. Suppose n_{ij} anthers are prepared with storage condition i and centrifuging force x_j, $i = 1, 2$, $x_1 = 40$, $x_2 = 150$, $x_3 = 350$, where $n_{ij} \geqslant 70$ for all i, j. Let y_{ij} and p_{ij} be the number of embryogenic anthers and probability of the anthers being embryogenic with storage condition i in centrifuging force x_j, for $i = 1, 2$, $j = 1, 2, 3$.

The following models are fitted, resulting in the residual deviances as shown below. In each case $i = 1, 2$ and $j = 1, 2, 3$.

$$\omega_0 : \quad \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha, \qquad \text{residual deviance } D_0 = 10.452$$

$$\omega_1 : \quad \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i, \qquad \text{residual deviance } D_1 = 5.173$$

$$\omega_2 : \quad \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + \beta \log x_j, \qquad \text{residual deviance } D_2 = 8.092$$

$$\omega_3 : \quad \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + \beta \log x_j, \qquad \text{residual deviance } D_3 = 2.619$$

Interpret these four models. Carry out appropriate statistical tests to determine which is your preferred model.

[If $\chi^2_\nu(\alpha)$ is defined by $\mathbb{P}(Y > \chi^2_\nu(\alpha)) = \alpha$ where $Y \sim \chi^2_\nu$, then $\chi^2_1(0.05) = 3.84$, $\chi^2_2(0.05) = 5.99$, $\chi^2_3(0.05) = 7.82$, $\chi^2_4(0.05) = 9.49$, $\chi^2_5(0.05) = 11.07$]

5 Describe briefly the basic concept behind the E-M algorithm, explaining what the E and M steps correspond to in the algorithm and appropriately defining all notations used.

As part of a training session, a swimming instructor matches into 8 pairs a strong swimmer and a not so strong swimmer. The instructor asks the 8 swimming pairs to each line up behind one of the eight start platforms of a twenty five-metre pool. The instructor then tells the swimmers that there will be a fifty-metre relay race between the pairs. The order that the members of the pairs will race in is left up to each pair to decide. The instructor selects eight further members of his swimming class to record the swimming times for each member of each pair. The race is then started. Unfortunately, five of the eight members asked to record the swimming times of the swimmers only record the total times for their five designated pairs.

The swimming instructor comes to you with the data in the following form

$$X = (S_{11}, S_{12}, S_{21}, S_{22}, S_{31}, S_{32}, T_4, T_5, T_6, T_7, T_8)$$

where S_{i1}, S_{i2} correspond respectively to the swimming times of the strong swimmer and the not so strong swimmer in the ith pair, for $1 \leqslant i \leqslant 3$, and T_j corresponds to the total swim time for the jth pair, for $4 \leqslant j \leqslant 8$.

The instructor believes that all swimming times are independent, that those for the strong swimmers follow an exponential distribution with mean $1/\lambda_1$, and that those for the weaker swimmers follow an exponential distribution with mean $1/\lambda_2$, where $\lambda_1 > \lambda_2$. The instructor asks you to use X to estimate (λ_1, λ_2).

(a) Show that for $j = 4, \ldots 8$, T_j has probability density function

$$g(t) = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} \left(e^{-\lambda_2 t} - e^{-\lambda_1 t} \right) \quad \text{for} \quad t > 0.$$

(b) Write down $\log L(\lambda_1, \lambda_2 | X)$, the log-likelihood corresponding to the observed data X.

(c) Apply the E-M algorithm to this estimation problem, providing explicit expressions for the parameter updates obtained in the M-step of the algorithm to compute the maximum likelihood estimates of λ_1, λ_2.

[Note that

$$\int_0^t s\, e^{-\lambda s}\, ds = \frac{1}{\lambda^2} - \left(\frac{t}{\lambda} + \frac{1}{\lambda^2} \right) e^{-\lambda t}$$

for any $t > 0$ and any λ.]

END OF PAPER