UNIVERSITY OF
CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Tuesday 5 June 2007 1.30 to 4.30

PAPER 43

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 The data in the table below refer to an investigation into the effectiveness of two new teaching methods (B and C) compared to the existing teaching method (A). Twenty-one children are given an initial aptitude test and then each child is taught using one of the three teaching methods. Their scores in an achievement test at the end of the investigation are also recorded.

Teaching Method

A		B		C	
final achievement score	initial aptitude score	final achievement score	initial aptitude score	final achievement score	initial aptitude score
6	3	8	4	6	3
4	1	9	5	7	2
5	3	7	5	7	2
3	1	9	4	7	3
4	2	8	3	8	4
3	1	5	1	5	1
6	4	7	2	7	4

The aim of the analysis in the S-Plus output below is to investigate any differences in effectiveness of the teaching methods, taking into account differences in initial aptitudes between the three groups of children. Interpret the commands and output in detail. Which is your preferred model, and what do you conclude about the different teaching methods? Illustrate with suitable sketch graph(s).

```
> teachingdata <- read.table("teachingdata", header=T)
> attach(teachingdata)
> score

  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
  6  4  5  3  4  3  6  8  9  7  9  8  5  7  6  7  7  7  8  5  7

> aptitude

  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
  3  1  3  1  2  1  4  4  5  5  4  3  1  2  3  2  2  3  4  1  4

> group <- rep(1:3, each=7)
> Group <- factor(group)
> options(contrasts=c("contr.treatment","contr.poly"))
> teach1.lm <-
  lm (score {\catcode‘\˜ = 11 ˜} aptitude * Group)
> anova(teach1.lm)

Analysis of Variance Table

Response: score
Terms added sequentially (first to last)

                Df  Sum of Sq   Mean Sq   F Value      Pr(F)
      aptitude   1   36.57548  36.57548  56.94218  0.0000018
         Group   2   16.93200   8.46600  13.18021  0.0004969
aptitude:Group   2    0.66714   0.33357   0.51932  0.6052426
     Residuals  15    9.63490   0.64233

> teach2.lm <-
  lm (score {\catcode‘\˜ = 11 ˜} aptitude + Group)
> anova(teach2.lm)

Analysis of Variance Table

Response: score
Terms added sequentially (first to last)

           Df  Sum of Sq   Mean Sq  F Value         Pr(F)
 aptitude   1   36.57548  36.57548  60.35534  0.0000005428
    Group   2   16.93200   8.46600  13.97024  0.0002578664
Residuals  17   10.30204   0.60600
```

```
> summary(teach2.lm, cor=F)

Call:
lm (formula = score {\catcode‘\˜ = 11 ˜}
                    aptitude  +  Group)

Residuals:
    Min       1Q   Median      3Q    Max
 -1.739  -0.5796  0.07347  0.4898  1.004

Coefficients:

             Value  Std.Error  t value  Pr(>|t|)
(Intercept)  2.8367  0.4235     6.6988   0.0000
   aptitude  0.7429  0.1421     5.2267   0.0001
     Group2  2.1878  0.4545     4.8139   0.0002
     Group3  1.8612  0.4240     4.3901   0.0004

Residual standard error: 0.7785 on 17 degrees of freedom
Multiple R-Squared: 0.8386
F-statistic:  29.43 on 3 and 17 degrees of freedom,
$\phantom {-} \qquad \qquad \qquad$
              the p-value is 5.889e-07
```

2 Let $Y_1, \ldots, Y_n$ be independent random variables with the density of Y_i given by

$$f_{Y_i}(y_i) \;=\; \exp\left\{\frac{y_i\theta_i - b\,(\theta_i)}{\phi} \;+\; c\,(y_i, \phi)\right\}, \qquad (*)$$

and suppose that $g(\mu_i) \;=\; \boldsymbol{\beta}^T\boldsymbol{x}_i$ where $\mu_i = \mathbb{E}(Y_i)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{x}_i$ is a known $p \times 1$ vector of covariates, $i = 1, \ldots, n$. Write down expressions for $\mathbb{E}(Y_i)$ and $\mathrm{var}(Y_i)$ in terms of derivatives of $b(\theta_i)$ and ϕ.

Explain the terms *canonical link function*, *variance function*, *scaled deviance* and *deviance.*

Now suppose that Y_i has a normal distribution with mean μ_i and variance σ^2, $i = 1, \ldots, n$. Show that the density of Y_i can be written as in $(*)$, and identify θ_i, $b(\theta_i)$ and ϕ. Suppose $g(\mu_i) \;=\; \boldsymbol{\beta}^T\boldsymbol{x}_i$ where $g(\cdot)$ is the canonical link function. Derive the relationship between the deviance and the residual sum of squares.

The S-Plus commands below refer to serum cholesterol levels (millimoles per litre) in `chol`, age (years) in `age`, and body mass index (weight (kg) divided by the square of height (m)) in `bmi` for thirty women. Four models are fitted to the data as shown. Models 1, 2, 3 and 4 have deviances 27.10, 31.64, 35.82 and 49.70 respectively. Explain carefully how to test whether serum cholesterol levels are associated with body mass index, taking account of the possible effect of age, and give your conclusions.

```
>  chol1.lm <- lm (chol {\catcode‘\~= 11 ~} age + bmi) # Model 1
>  chol2.lm <- lm (chol {\catcode‘\~= 11 ~} age)       # Model 2
>  chol3.lm <- lm (chol {\catcode‘\~= 11 ~} bmi)       # Model 3
>  chol4.lm <- lm (chol {\catcode‘\~= 11 ~} 1)         # Model 4
```

[*Hint: If* $F_{n_1,n_2}(0.05)$ *and* $\chi^2_\nu(0.05)$ *denote the upper five percentage points of an* F_{n_1,n_2} *and* χ^2_ν *distribution respectively then*

$$F_{1,28}(0.05) \;=\; 4.196\;, \qquad F_{1,27}(0.05) \;=\; 4.210\;, \qquad \chi^2_1(0.05) \;=\; 3.841\;]$$

3 Let Y_{ij} , $i = 1, \ldots, n$ and $j = 1, \ldots, k$, be independent random variables where Y_{ij} has a $\text{Bin}(m, p_{ij})$ distribution.

Suppose that

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \mu + \alpha_i \qquad i = 1, \ldots, n, \quad j = 1, \ldots, k,$$

where $\alpha_1 = 0$. Derive equations satisfied by the maximum likelihood estimators $\hat{\mu}, \hat{\alpha}_2, \ldots, \hat{\alpha}_n$. When $n = 2$, find the asymptotic distribution of $\hat{\mu}$.

As part of an investigation into the effect of radiation on cancer cells, 27 dishes, each containing 400 cells, were irradiated. This experiment was spread over nine days, with three dishes being irradiated each day. In the (edited) S-Plus output below, the number of cells surviving in each dish is in `survived`, and `day` contains the day on which the corresponding dish was irradiated.

(i) Explain the S-Plus commands and output carefully. Write down in detail the models fitted in `model1.glm` and `model2.glm`, defining any notation you use.

(ii) Determine the values of the four numbers that have been replaced by asterisks in the output. Investigate whether there is a day effect, and comment on the fit of your preferred model.

```
> day

  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
  1  1  1  2  2  2  3  3  3  4  4  4  5  5  5

 16 17 18 19 20 21 22 23 24 25 26 27
  6  6  6  7  7  7  8  8  8  9  9  9

> survived

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
178 193 217 109 112 115  66  75  80 118 125 137 123 146 170

 16  17  18  19  20  21  22  23  24  25  26  27
115 130 133 200 189 173  88  76  90 121 124 136

> Day <- factor (day)
> ncells <- rep (400,27)
> prop <- survived / ncells
> options(contrasts = c ("contr.treatment",
                          "contr.poly"))

> model1 <- glm (prop ~ 1, family = binomial,  weights = ncells)
> model2 <- glm (prop ~ Day, family = binomial,  weights = ncells)
> anova (model2)

Analysis of Deviance Table

Binomial model
Response: prop
Terms added sequentially (first to last)

      Df  Deviance   Resid.  Df  Resid. Dev
NULL                          *    495.6308
 Day   *     *                *     32.7945
```

4 The S-Plus output below refers to an experiment to investigate whether cross-fertilised plants for a particular species of corn plant tend to be taller than self-fertilised plants. In the experiment, there were fifteen pots, each containing one cross-fertilised plant and one self-fertilised plant. In the output, the vectors `cross` and `self` contain the heights of the fifteen cross-fertilised and self-fertilised plants in the order of the pots.

A statistician has used three different directives, each giving a different p-value, as shown in the output, and has asked you for advice about how to proceed. In each of the three cases, explain which test has been carried out, stating the null and alternative hypotheses carefully. Without doing any calculations, explain how the test statistics V and Z have been obtained in tests 1 and 3 respectively, and indicate briefly how the p-values have been obtained in all three tests. Write a paragraph for the statistician explaining which test(s) is (are) most relevant for these data, and giving your conclusions about heights of cross-fertilised and self-fertilised plants.

```
# Test 1

> wilcox.test (cross,self, paired = T, alternative = "greater")

      Exact Wilcoxon signed-rank test

data: cross and self
signed-rank statistic V = 96, n = 15, p-value = 0.0206
alternative hypothesis: mu is greater than 0

# Test 2

> wilcox.test (cross, self, paired = T, alt = "greater", exact = F)

      Wilcoxon signed-rank test

data: cross and self
signed-rank normal statistic with correction Z = 2.0163,
                                       p-value = 0.0219
alternative hypothesis: mu is greater than 0

# Test 3

> wilcox.test (cross, self, alt = "greater", exact = F)

      Wilcoxon rank-sum test

data: cross and self
rank-sum normal statistic with correction Z = 3.0105,
                                    p-value = 0.0013
alternative hypothesis: mu is greater than 0
```

5 (a) Explain what is meant by a generalised additive model for independent normally distributed random variables.

Suppose that $n \geqslant 3$ and consider a set of points $\{(x_i, y_i) : i = 1, \ldots, n\}$ where $x_1 < \ldots < x_n$.

(b) What is meant by a natural cubic spline interpolating these points?

(c) Now consider

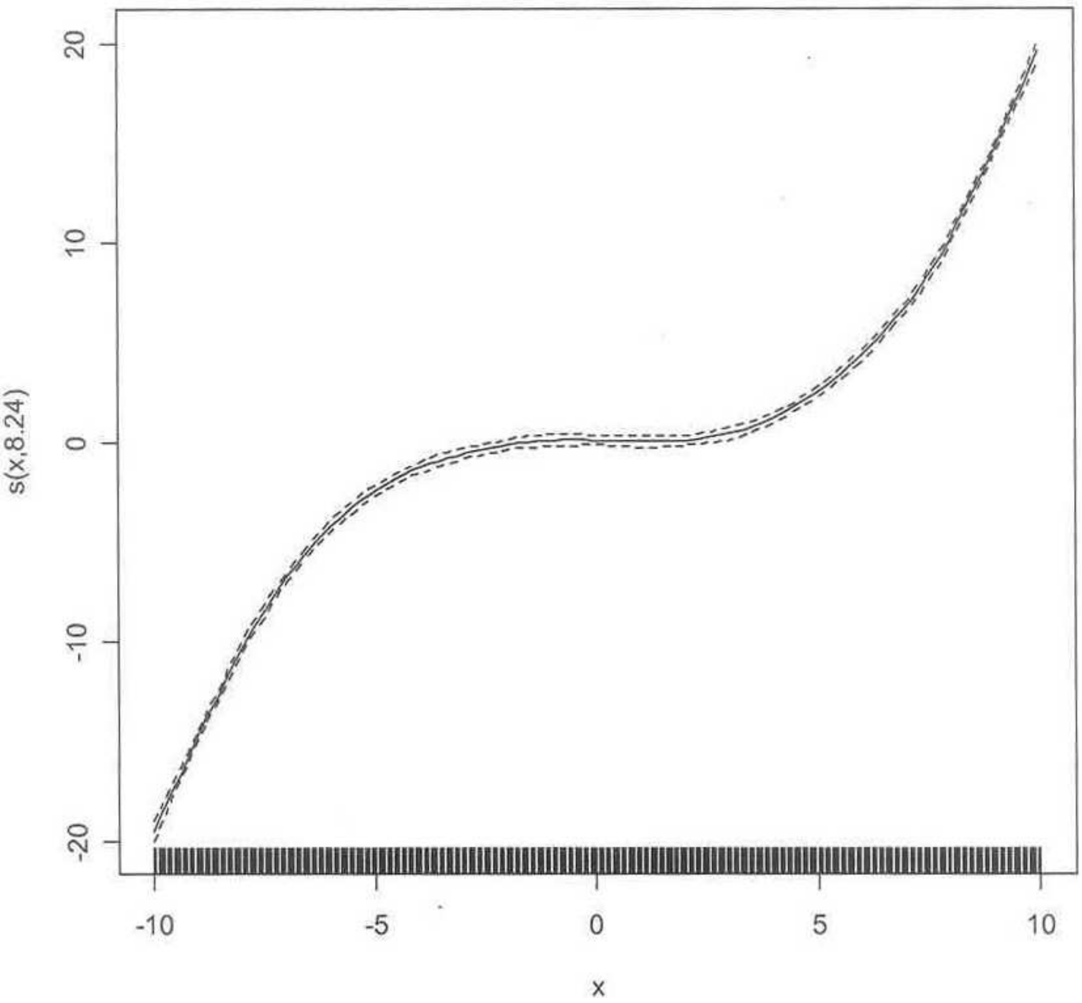
$$y_i = g(x_i) + \epsilon_i \,, \qquad i = 1, \ldots, n \,,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables. Define the penalised sum of squares with smoothing parameter $\lambda > 0$, and the cubic smoothing spline $\hat{g}_\lambda(\cdot)$.

(d) Define the generalised cross-validation score in terms of the influence matrix A , and explain how it can be used to choose λ .

(e) Below is the R output (with figure) from an analysis of simulated data $\{(x_i, y_i) : i = 1, \ldots, 401\}$. Describe carefully the output produced. You are required to write out the form of the model fitted. What is the range of x? From the figure, suggest an alternative simpler model to fit.

```
> library (mgcv)
> model.gam <- gam (y ~ s (x, bs = "cr"))

> summary (model.gam) # Slightly abbreviated output

Family: gaussian
Link function: identity

Formula: y ~ s (x, bs = "cr")

Parametric coefficients:

             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -0.13241     0.05426    -2.44    0.0151  *

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

Approximate significance of smooth terms:

        edf  Est.rank  p-value
s(x)  8.237     9.000   <2e-16  ***

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

R-sq.(adj) = 0.983 Deviance explained = 98.4 %
GCV score  = 0.99018 Scale est. = 0.96737 n = 401
```



END OF PAPER