UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Wednesday, 3 June, 2009 9:00 am to 12:00 pm

PAPER 38

APPLIED STATISTICS

*Attempt no more than **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 (edited) R output below shows part of an analysis of an investigation into the distances (in mm) travelled by paper aeroplanes made using a high performance design and using a simple design, coded 1 and 2 respectively in `plane`. The aeroplanes were made using two different weights of paper, 80 g per square m and 50 g per square m, coded 1 and 2 respectively in `paper`, and two angles of launch, horizontal and 45 degrees upwards, coded 1 and 2 respectively in `angle`. In the R output, `paper`, `angle` and `plane` are factors, and corner-point constraints are used.

Write down the algebraic form of the model fitted in `model1.lm`, defining your notation carefully and writing down the constraints explicitly. Test whether the three-factor interaction is needed. Explain what the `stepAIC` directive does. Write down the algebraic form of the model fitted in `model2.lm`, giving the parameter estimates and standard errors. Explain the final line of the output to the directive `summary(model2.lm)`.

Which combination of design, weight of paper and angle of launch would you choose in order to maximise the distance travelled? Briefly summarise the results of the analysis in words. What plots would you have carried out before fitting any models to these data, and what further plots would you have made after the analysis shown in the output?

```
> paperplanes
   Distance paper angle plane
1      2160     1     1     1
2      1511     1     1     1
3      4596     1     1     2
4      3706     1     1     2
5      3854     1     2     1
6      1690     1     2     1
7      5088     1     2     2
8      4255     1     2     2
9      6520     2     1     1
10     4091     2     1     1
11     2130     2     1     2
12     3150     2     1     2
13     6348     2     2     1
14     4550     2     2     1
15     2730     2     2     2
16     2585     2     2     2
> attach(paperplanes)
> model1.lm <- lm(Distance~paper*angle*plane)
> anova(model1.lm)
Analysis of Variance Table
Response: Distance
                  Df   Sum Sq  Mean Sq F value   Pr(>F)
paper              1  1718721  1718721  1.6384 0.236412
angle              1   654481   654481  0.6239 0.452377
plane              1   385641   385641  0.3676 0.561111
paper:angle        1   419904   419904  0.4003 0.544599
paper:plane        1 23386896 23386896 22.2940 0.001499
angle:plane        1    73441    73441  0.0700 0.798013
```

```
paper:angle:plane  1    21025    21025  0.0200 0.890919
Residuals          8  8392178  1049022
> library(MASS)
> stepAIC(model1.lm,list(upper=~paper*angle*plane,lower=~1))
Start:  AIC= 226.72
 Distance ~ paper * angle * plane
                    Df Sum of Sq     RSS     AIC
- paper:angle:plane  1     21025 8413203     225
<none>                           8392178     227
Step:  AIC= 224.76
 Distance ~ paper + angle + plane + paper:angle + paper:plane +
    angle:plane
# output omitted
Call:
lm(formula = Distance ~ paper + plane + paper:plane)
Coefficients:
  (Intercept)         paper2         plane2  paper2:plane2
         2304           3074           2108          -4836

> model2.lm <- lm(Distance~paper*plane)
> summary(model2.lm)
Call:
lm(formula = Distance ~ paper * plane)
Residuals:
    Min      1Q  Median      3Q     Max
-1286.3  -636.6  -103.8   545.1  1550.3
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2303.8      446.3   5.162 0.000236
paper2          3073.5      631.2   4.870 0.000385
plane2          2107.5      631.2   3.339 0.005899
paper2:plane2  -4836.0      892.6  -5.418 0.000156
Residual standard error: 892.6 on 12 degrees of freedom
Multiple R-Squared: 0.7272
F-statistic: 10.66 on 3 and 12 DF,  p-value: 0.001057
```

2 Suppose that Y has density

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

where θ and ϕ are real parameters. Show that $\mathbb{E}(Y) = b'(\theta)$ and $\mathrm{var}(Y) = \phi b''(\theta)$.

[*You may assume that* $\mathbb{E}(\partial l_1/\partial\theta) = 0$ *and* $\mathbb{E}(-\partial^2 l_1/\partial\theta^2) = \mathbb{E}\big((\partial l_1/\partial\theta)^2\big)$, *where* l_1 *is the loglikelihood for the single observation* Y.]

Suppose that observations $Y_1,\ldots,Y_n$ are independent and that Y_i has density $f(\cdot;\theta_i,\phi)$. Suppose further that $\theta_i = x_i^T\beta$, where β is a p-dimensional vector of unknown parameters and x_i is a p-dimensional known vector. Let $\mathbb{E}(Y_i) = \mu_i(= \mu_i(\beta))$, with maximum likelihood estimate $\hat{\mu}_i = \mu_i(\hat{\beta})$ where $\hat{\beta}$ is the maximum likelihood estimator of β. Let l_n be the loglikelihood for all n observations. Using $\frac{\partial l_n}{\partial\beta} = \sum_{i=1}^n \frac{\partial l_n}{\partial\theta_i}\frac{\partial\theta_i}{\partial\beta}$, show that

$$X^T(Y - \hat{\mu}) = 0,$$

where $X = (x_1,\ldots,x_n)^T$, $Y = (Y_1,\ldots,Y_n)^T$, and $\hat{\mu} = (\hat{\mu}_1,\ldots,\hat{\mu}_n)^T$. Show further that

$$\mathbb{E}\Big(-\frac{\partial^2 l_n}{\partial\beta\partial\beta^T}\Big) = X^T V X,$$

where V is a diagonal matrix which you should specify. [*You may assume that* $\mathbb{E}\Big(-\frac{\partial^2 l_n}{\partial\beta\partial\beta^T}\Big) = \mathbb{E}\Big(\big(\frac{\partial l_n}{\partial\beta}\big)\big(\frac{\partial l_n}{\partial\beta}\big)^T\Big)$.]

Suppose that Y_i has a Poisson distribution with mean μ_i where

$$\log(\mu_i) = \begin{cases} \alpha & \text{for } i = 1,\ldots,n \\ \alpha + \nu & \text{for } i = n+1,\ldots,2n, \end{cases}$$

and let $\beta = (\alpha,\nu)^T$ where α and ν are unknown. Find X and V. Hence, quoting a general theorem for the asymptotic distribution of $\hat{\beta}$, find the approximate distribution of the maximum likelihood estimator $\hat{\nu}$ of ν for large n.

3 A total of 371 coal miners were examined for severe lung disease, and the number of years of exposure to coal dust was recorded for each miner. The grouped data are given in the table below. For example, the final group consists of 11 miners with an average of 51.5 years of exposure per miner.

Years of exposure	Number of miners	Number with disease
5.8	98	0
15.0	54	1
21.5	43	3
27.5	48	8
33.5	51	9
39.5	38	8
46.0	28	10
51.5	11	5

The R output below contains an analysis of these data. Write down the model fitted in `miners1.glm` in algebraic form, including any assumptions, and defining your notation. Using deviances, test whether the probability of severe disease depends on exposure times. Comment on the fit of the model `miners1.glm`, and briefly explain both the input and the output to `predict`.

Show that increasing the exposure by one year affects the odds of disease by multiplication by a fixed amount. Explain how to estimate this amount and how to find a 95% confidence interval for this amount. Write down the corresponding multiplier and confidence interval when the exposure is increased by 10 years.

What model is fitted in `miners2.glm`? In the analysis of deviance table, nine entries have been replaced by asterisks. Find these values. Briefly compare the two models, saying which you think is better and why.

```
> years
[1]  5.8 15.0 21.5 27.5 33.5 39.5 46.0 51.5
> cases
[1]  0  1  3  8  9  8 10  5
> miners
[1] 98 54 43 48 51 38 28 11
> pcases <- cases/miners
> miners1.glm <- glm(pcases~years,binomial,weights=miners)
> summary(miners1.glm)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6625  -0.5746  -0.2802   0.3237   1.4852
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.79648    0.56859  -8.436  < 2e-16
years        0.09346    0.01543   6.059 1.37e-09
---
(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 56.9028  on 7  degrees of freedom
Residual deviance:  6.0508  on 6  degrees of freedom
Number of Fisher Scoring iterations: 4
> predict(miners1.glm, data.frame(years=40),type="response",se.fit=T)
$fit
[1] 0.2576988
$se.fit
[1] 0.03637976

> years2 <- years*years
> miners2.glm <- glm(pcases~years+years2,binomial,weights=miners)
> anova(miners2.glm, test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: pcases
Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       *        *
years   *    *             *        *    9.96e-13
years2  *    *             *      3.282     0.096
```

4 Describe how to carry out a one-sided Wilcoxon rank sum test for two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, and a one-sided Wilcoxon signed rank test for independent pairs $(U_1, V_1), \ldots, (U_n, V_n)$. In each case, state the assumptions, give the null and alternative hypotheses and the test statistic. For the rank sum test, describe how the null distribution can be calculated.

The times (in minutes) taken to run a 2.5km distance by 26 men who have been participating in a particular fitness programme for more than a year are analysed in the (edited) S-Plus output below. Thirteen of the men have been classified as $RHR1$ and thirteen have been classified as $RHR2$, according to their recovery heart rates after a specified step excercise in the gym. Recovery heart rates are a measure of fitness, with $RHR1$ corresponding to a higher level of fitness than $RHR2$. In the output below, `rhr1` and `rhr2` contain the running times for the $RHR1$ and $RHR2$ participants, respectively. What are the hypotheses, and what is the value of the test statistic? What is the result of the hypothesis test?

```
> median(rhr1)
[1] 11.52
> median(rhr2)
[1] 12.78
> wilcox.test(rhr1,rhr2,alt="less")
data:  rhr1 and rhr2
W = 135, n=13, m=13, p-value = 0.0193
alternative hypothesis: mu is less than 0
```

There are 17 additional men who are new to the fitness programme this year, with 8 of them being classified $RHR1$ and the rest are $RHR2$. In the output below, `newrhr1` and `newrhr2` contain the running times for the new $RHR1$ and $RHR2$ participants respectively. Describe in detail the two tests that have been carried out in the S-Plus output below. Write a paragraph summarising your conclusions, based on all three tests, about the running times for all the men.

```
> wilcox.test(newrhr1,newrhr2,alt="less")
data:  newrhr1 and newrhr2
W = 58, n=8, m=9, p-value = 0.0998
alternative hypothesis: mu is less than 0

> wilcox.test(c(rhr1,newrhr1),c(rhr2,newrhr2),alt="less",exact=F)
data:  c(rhr1, newrhr1) and c(rhr2, newrhr2)
normal statistic with correction Z=-2.2233, p-value = 0.0131
alternative hypothesis: mu is less than 0
```

5 Fatigue is an important symptom in many chronic diseases. It is defined as an overwhelming, sustained sense of exhaustion and decreased capacity for physical and mental work. A study of Psoriatic Arthritis patients followed up over a number of clinic visits focused on the longitudinal pattern of fatigue and the predictive relationship between various demographic, clinical and laboratory variables and subsequent change in fatigue level.

The level of fatigue was classified into three possible states: mild (including having no fatigue), moderate or severe (coded 1, 2 or 3 respectively in `state`). Some of the explanatory variables were gender (SEX = 0 corresponds to female; SEX = 1 corresponds to male), duration of arthritis (ARTHDUR; in years), physical functioning (measured using the Health Assessment Questionnaire (HAQ) on a scale from 0 to 3, with higher scores indicating higher level of physical disability) and Haemoglobin (HGB). The patient identity number is in `id`.

Various multi-state models, with time from entry into the study (in years, in `time`) as the time scale, were fitted to the data using the msm package in R and the following (edited) R output obtained.

```
> fatigue.msm # multi-state model with no covariates fitted

Call:
msm(formula = state ~ time, subject = id, data = fatigue.dat, qmatrix = Qmat)

Maximum likelihood estimates:
Transition intensity matrix

        State 1                 State 2                State 3
State 1 -0.173 (-0.2145,-0.1396)  0.173 (0.1396,0.2145)   0
State 2  0.4684 (0.3804,0.5768)  -0.8519 (-1.005,-0.7219) 0.3835 (0.2972,0.495)
State 3  0                     0.3655 (0.2899,0.4609)   -0.3655 (-0.4609,-0.2899)

-2 * log-likelihood:  1787.295

> totlos.msm(fatigue.msm, start=1, fromt=0, tot=10)
  State 1  State 2  State 3
7.014800  1.724686 1.260514

> fatigue.msm1 # multi-state model with covariates included

Call:
msm(formula = state ~ time, subject = id, data = fatigue.dat, qmatrix = Qmat, covari-
                                                ates = ~ SEX + ARTHDUR + HAQ + HGB)

Maximum likelihood estimates:
Transition intensity matrix with covariates set to their means

        State 1                  State 2                 State 3
State 1  -0.2344 (-0.2986,-0.184)  0.2344 (0.184,0.2986)    0
State 2  0.5509 (0.4368,0.6946)    -0.9756 (-1.189,-0.8008) 0.4247 (0.3064,0.5887)
State 3  0                     0.5822 (0.4222,0.8028)     -0.5822 (-0.8028,-0.4222)
```

```
Log-linear effects of SEX

            State 1              State 2                  State 3
State 1  0                       -0.7695 (-1.328,-0.2108)   0
State 2  -0.3285 (-0.8292,0.1722)  0                        0.3811 (-0.2649,1.027)
State 3  0                        0.1712 (-0.4381,0.7805)   0

Log-linear effects of ARTHDUR

            State 1                 State 2                     State 3
State 1  0                           0.001331 (-0.02564,0.0283)  0
State 2  -0.01047 (-0.03466,0.01372)  0                          -0.0226 (-0.05003,0.004833)
State 3  0                           -0.0264 (-0.05229,-0.000508) 0

Log-linear effects of HAQ

        State 1                  State 2                 State 3
State 1  0                       0.9913 (0.5232,1.459)    0
State 2  -0.3006 (-0.7835,0.1823) 0                      -0.05399 (-0.5534,0.4454)
State 3  0                       -1.045 (-1.529,-0.5618)  0

Log-linear effects of HGB

        State 1                     State 2                     State 3
State 1  0                           0.01368 (-0.003527,0.03088)    0
State 2  -0.002136 (-0.01674,0.01247) 0                            -0.03375 (-0.05384,-0.01366)
State 3  0                          -0.02616 (-0.04506,-0.007265)   0

-2 * log-likelihood:  1622.442
```

(i) This part of the question applies to the model in `fatigue.msm`. Draw the transition diagram, including on it the estimated transition intensities corresponding to each type of transition. What is the estimated mean time spent (with 95% confidence interval) in the moderate fatigue state before making a transition out of it? Given that a transition out of the moderate fatigue state is made, what is the probability that the transition is to the mild fatigue state?

(ii) Explain the output from the R command `totlos.msm`.

(iii) Write out mathematically the multi-state model corresponding to `fatigue.msm1`, making sure to define all notation used and stating all assumptions being made.

(iv) Interpret the effects of the statistically significant covariates on the mild to moderate transition.

END OF PAPER