UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday 27 May, 2004 9:00 to 12:00

PAPER 37

Applied Statistics

Attempt **FOUR** *questions.*

There are **five** *questions in total.*

The questions carry equal weight.

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

1 (i) Let $Y_1, \ldots, Y_n$ be independent Poisson variables, with

$$\mathbb{E}(Y_i) = \mu_i, \text{ and } \log \mu_i = \beta^T x_i, \quad \text{for } 1 \leqslant i \leqslant n.$$

Discuss carefully the estimation of the unknown p-dimensional vector β. (You may assume that $x_1, \ldots, x_n$ are known covariate vectors of the same dimension as β.)

(ii) Suppose now that the observations $Y_1, \ldots, Y_n$ are independent, with $E(Y_i) = \mu_i$, and $\text{var}(Y_i) = \phi\mu_i$, and $\log(\mu_i) = \beta x_i$, for some unknown ϕ and unknown scalar parameter β. Let β_0 be the true value of this unknown parameter.

Our aim is to estimate β, but ϕ is an unknown 'dispersion' parameter. Clearly $\phi > 1$ will correspond to over-dispersion relative to the Poisson. In the absence of knowledge of ϕ, we choose our estimator $\hat{\beta}$ to maximise the function $l_p(\beta)$, where

$$l_p(\beta) = -\Sigma\mu_i + \beta\Sigma x_i y_i + \text{ constant } .$$

(Thus $l_p()$ is in general not the 'correct' loglikelihood function: we work out below whether this is a serious problem.)

By expanding

$$\frac{\partial l_p(\beta)}{\partial \beta}$$

evaluated at $\hat{\beta}$, about β_0, show that $(\hat{\beta} - \beta_0)$ is approximately equal to $(I(\beta_0))^{-1}U(\beta_0)$, where

$$U(\beta) = \frac{\partial l_p(\beta)}{\partial \beta}, \text{ and } I(\beta) = \Sigma x_i^2 \exp \beta x_i.$$

and hence show that, approximately,

$$E(\hat{\beta}) = \beta_0, \text{ and } \text{var}(\hat{\beta}) = \phi(I(\beta_0))^{-1}.$$

2 'Commissioned analysis of surgical performance by using routine data: lessons from the Bristol inquiry' is a paper published in J.R. Statistical Soc. A in 2002, by David J. Spiegelhalter and others. It includes the data given in the table below, which refers to mortality due to cardiac surgery at each of 12 UK centres, of which the Bristol Royal Infirmary is Centre 1. Thus, in 1984-87, there were 63 babies under 1 year old for example, who received cardiac surgery at Centre 1, and of these 63 babies, 16 tragically died as a result of surgery. Centre 1 is of special interest for this inquiry.

(i) If you restrict attention to the years 1984-7 only, how would you test whether the mortality rate is constant over the 12 Centres?

(ii) Now restrict attention to the second and third of the three time periods, and describe briefly how to fit the model

$$g(\pi_{ij}) = \mu + \alpha_i + \beta_j, \qquad \text{for } i = 1, \ldots, 12 \text{ and } j = 2, 3 \ ,$$

where $g(\cdot)$ is a suitable link function, π_{ij} is the probability of death for a baby at Centre i during the time period j, and $\alpha_1 = 0$, $\beta_2 = 0$.

Discuss carefully the results of the model-fitting, given in the S-Plus output at the end of the question. (You may assume that the factors Centre, Year have been set up correctly.)

Mortality due to cardiac surgery for babies under 1 year
Table

	1984-87		1988-90		1991-Mar95	
Centre	r1	t1	r2	t2	r3	t3
1	16	63	31	108	43	181
2	11	66	22	107	27	200
3	10	36	35	135	26	157
4	0	0	14	45	15	142
5	23	83	26	104	36	217
6	48	242	34	198	49	417
7	19	186	25	184	27	253
8	55	236	57	362	57	369
9	15	68	11	79	28	214
10	28	109	34	90	31	184
11	30	77	57	438	67	740
12	28	187	21	121	32	268

```
> summary(glm(r/tot ~ Centre + Year,binomial, weights = tot), cor=F)
Call: glm(formula = r/tot ~ Centre + Year, family = binomial,
weights = tot)
Deviance Residuals:
       Min        1Q      Median        3Q      Max
 -1.454029 -0.356005 0.002994555 0.3386306 1.773282

Coefficients:
                  Value Std. Error    t value
(Intercept) -0.80958468 0.14281655 -5.6687035
    Centre2 -0.58775796 0.20675867 -2.8427247
    Centre3 -0.30568326 0.19815078 -1.5426801
    Centre4 -0.57335447 0.24379181 -2.3518201
    Centre5 -0.34389804 0.19609795 -1.7537055
    Centre6 -0.77363166 0.17978907 -4.3029961
    Centre7 -0.96239712 0.20072879 -4.7945146
    Centre8 -0.67930844 0.17002204 -3.9954139
    Centre9 -0.76597500 0.21928367 -3.4930782
   Centre10 -0.08223982 0.19661703 -0.4182741
   Centre11 -1.07947416 0.16546053 -6.5240585
   Centre12 -0.75768827 0.20071357 -3.7749729
       Year -0.42675763 0.07880675 -5.4152418

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 119.0659 on 23 degrees of freedom

Residual Deviance: 15.73891 on 11 degrees of freedom

Number of Fisher Scoring Iterations: 5
```

3 The S-Plus output below gives 2 standard statistical tests for the small data sets x, y. Describe carefully how the statistics and the p-values are calculated, for the 2 tests, and compare their outcomes for the data given.

```
>x _ scan()
3.7 2.1 4.5 7.1

>y_scan()
 6.1 7.9 10.3 11.4 13.7

>summary(x)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
  2.1     3.3    4.1 4.35    5.15  7.1

>summary(y)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
  6.1     7.9   10.3 9.88    11.4 13.7
>t.test(x,y, alt="less")

 Standard Two-Sample t-Test

data:  x and y
t = -3.1364, df = 7, p-value = 0.0082
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        NA -2.189557
sample estimates:
 mean of x mean of y
      4.35      9.88

> rank(c(x,y))
[1] 2 1 3 5 4 6 7 8 9

>wilcox.test(x,y, alt="less")

 Exact Wilcoxon rank-sum test

data:  x and y
rank-sum statistic W = 11, n = 4, m = 5, p-value = 0.0159
alternative hypothesis: true mu is less than 0
```

4 On December 5, 2003, the Times Business News published the following table, under the headline "Serving up a Sterling Christmas". This shows the forecasts for Europe's online Christmas sales, for the countries UK, France, Germany and the rest of Europe, for each of 13 different types of goods, ie Software, Books , . . . , Leisure Travel. The sums given are in millions of Euros.

Discuss carefully the S-Plus analysis which follows, interpreting the commands and the (slightly edited) output.

How would the analysis have been affected if 2 entries of the table of data had been unavailable?

Table of data

	UK	France	Germany	RestofE
Software	88.5	13.0	73.5	76.2
Books	378.9	72.2	425.7	393.2
Music	300.2	59.2	169.4	178.5
Videos/DVDs	249.8	62.5	129.5	134.4
EventTickets	115.0	26.6	82.6	142.3
Clothing	394.2	118.7	456.7	197.4
Toys	66.7	17.1	96.0	42.2
VideoGames	93.9	13.6	56.3	56.0
SportsEquip	17.2	1.0	18.6	13.0
ElectronicGds	309.1	90.3	244.2	206.4
Groceries	599.7	83.3	308.0	107.0
Housewares	129.0	30.1	97.4	41.5
LeisureTravel	436.1	72.6	279.0	324.3

```
> spend <- scan("o.data")
> goods <- 1:13
> country <- c("UK","France","Germany","restofE")
> z <- expand.grid(country,goods) ; z[1:9,]
> Goods <- z[,2] ; Goods <- factor(Goods)
> Country <- z[,1]
> first.lm <- lm(spend ~ Goods + Country)
> summary(first.lm, cor=F)
Call: lm(formula = spend ~ Goods + Country)
Residuals:
    Min     1Q Median    3Q   Max
 -157.2 -37.82  1.202 37.67 238.2
```

```
Coefficients:
                   Value Std. Error   t value  Pr(>|t|)
   (Intercept)  149.8269   44.9205     3.3354    0.0020
        Goods2  254.7000   57.2626     4.4479    0.0001
        Goods3  114.0250   57.2626     1.9913    0.0541
        Goods4   81.2500   57.2626     1.4189    0.1645
        Goods5   28.8250   57.2626     0.5034    0.6178
        Goods6  228.9500   57.2626     3.9982    0.0003
        Goods7   -7.3000   57.2626    -0.1275    0.8993
        Goods8   -7.8500   57.2626    -0.1371    0.8917
        Goods9  -50.3500   57.2626    -0.8793    0.3851
       Goods10  149.7000   57.2626     2.6143    0.0130
       Goods11  211.7000   57.2626     3.6970    0.0007
       Goods12   11.7000   57.2626     0.2043    0.8393
       Goods13  215.2000   57.2626     3.7581    0.0006
 CountryFrance -193.7000   31.7636    -6.0982    0.0000
CountryGermany  -57.0308   31.7636    -1.7955    0.0810
CountryrestofE  -97.3769   31.7636    -3.0657    0.0041

Residual standard error: 80.98 on 36 degrees of freedom
Multiple R-Squared: 0.7743
F-statistic: 8.232 on 15 and 36 degrees of freedom,
the p-value is 1.254e-07

> next.lm<- lm(log(spend) ~ Goods + Country)
>summary(next.lm, cor=F)
Call: lm(formula = log(spend) ~ Goods + Country)
Residuals:
   Min      1Q  Median     3Q    Max
 -1.01 -0.1807 0.04115 0.1978 0.5048

Coefficients:
                  Value Std. Error  t value Pr(>|t|)
   (Intercept)   4.5267   0.1980    22.8625   0.0000
        Goods2   1.6415   0.2524     6.5039   0.0000
        Goods3   1.1059   0.2524     4.3817   0.0001
        Goods4   0.9354   0.2524     3.7062   0.0007
        Goods5   0.4298   0.2524     1.7029   0.0972
        Goods6   1.6210   0.2524     6.4226   0.0000
        Goods7  -0.0831   0.2524    -0.3294   0.7438
        Goods8  -0.1176   0.2524    -0.4658   0.6442
        Goods9  -1.8364   0.2524    -7.2759   0.0000
       Goods10   1.3465   0.2524     5.3349   0.0000
       Goods11   1.3858   0.2524     5.4906   0.0000
       Goods12   0.2226   0.2524     0.8818   0.3837
       Goods13   1.5243   0.2524     6.0393   0.0000
 CountryFrance  -1.6800   0.1400   -11.9998   0.0000
CountryGermany  -0.2447   0.1400    -1.7477   0.0890
CountryrestofE  -0.5033   0.1400    -3.5951   0.0010
```

```
Residual standard error: 0.3569 on 36 degrees of freedom
Multiple R-Squared: 0.9381
F-statistic: 36.34 on 15 and 36 degrees of freedom,
the p-value is 0
>anova(next.lm)
Analysis of Variance Table

Response: log(spend)

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
    Goods 12  47.85186 3.987655 31.29875 1.998400e-15
  Country  3  21.60350 7.201168 56.52133 1.071365e-13
Residuals 36   4.58662 0.127406
```

5 Data have been collected from a multi-centre randomised-controlled trial on 1000 early-stage breast cancer patients, who after having successful surgery to remove a lump were randomised to receive either tamoxifen (chemotherapy) alone (coded: trt = 0) or a combination of tamoxifen and radiotherapy (coded: trt = 1). The patients were followed up every two years for a ten-year period and the events recorded were local recurrence, distant metastasis and death. (Notification of the date of death for patients in the study, *recorded to within a day*, was obtained from an outside organisation.) A patient may be observed in any of the following states (or stages) during the follow-up period: free of cancer (state 1), local recurrence only (state 2), distant metastasis only (state 3), both local recurrence and distant metastasis together (state 4) and death (state 5).

(i) Below are two patients' follow-up data recorded in the form (time, in years, from surgery, state):

```
Patient
 1        (0,1)  (2,1)  (4,3)  (4.2,5)
 2        (0,1)  (2,1)  (4,1)  (6,1)   (8,2)  (10,2)
```

Construct the likelihood contributions of these two patients, defining any terms that you use.

(ii) Below is the edited *R* output from a multi-state model analysis of the data from the study. (The follow-up times are measured in years.)

```
>breastcancer.msm

Multi-state Markov models in continuous time

Maximum likelihood estimates:

* Matrix of transition intensities with covariates set
    to their means

          Stage 1  Stage 2  Stage 3  Stage 4  Stage 5
Stage 1   -.085     0.025    0.056     0       0.004
Stage 2     0      -0.181     0       0.164    0.017
Stage 3     0        0      -0.369    0.064    0.305
Stage 4     0        0        0      -0.513    0.513
Stage 5     0        0        0        0        0
```

```
corresponding standard errors


          Stage 1     Stage 2    Stage 3  Stage 4    Stage 5
Stage 1    .0036      0.0021      0.0034     0         0.0018
Stage 2     0         0.0211      0         0.0251    0.0174
Stage 3     0          0          0.0246    0.0131    0.0236
Stage 4     0          0          0         0.0678    0.0678
Stage 5     0          0          0          0        0

 *No covariates on transition intensities

-2* log-likelihood: 10845.84

> pmatrix.msm(breastcancer.msm, t=5)

          Stage 1  Stage 2  Stage 3  Stage 4  Stage 5
Stage 1    0.65     0.06      0.10    0.02     0.16
Stage 2     0       0.40       0      0.16     0.43
Stage 3     0        0        0.16    0.04     0.80
Stage 4     0        0        0       0.08     0.92
Stage 5     0        0        0        0        1
```

a) Draw the transition diagram for this model, including on it the transition intensity corresponding to each type of transition.

b) Show how to calculate the mean sojourn time (in years), and the corresponding 95% confidence intervals.

c) Interpret the first row of the 5-year transition probability matrix provided above. Why are the first 4 elements in the last row of this matrix all zeroes?