UNIVERSITY OF CAMBRIDGE

M. PHIL. IN STATISTICAL SCIENCE

Thursday 30 May 2002 9 to 12

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) Define Ω as the linear model

$$\Omega : Y = \mu 1 + X\beta + \epsilon$$

where Y is an n-dimensional observation vector, 1 is the n-dimensional unit vector, μ and β are unknown parameters, X is a given $n \times p$ matrix of rank p, with $X^T 1 = \mathbf{0}$, and the components of ϵ are $\epsilon_1 \ldots, \epsilon_n$, distributed as $NID(0, \sigma^2)$, with σ^2 unknown. Define further

$$X\beta = X_1\beta_1 + X_2\beta_2,$$

where X is partitioned as $(X_1 : X_2)$, and β is similarly partitioned as $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$.

How would you test the hypothesis $\omega : \beta = 0$ against Ω? How would you test the hypothesis $\omega_1 : \beta_1 = 0$ against Ω? What does it mean to say that β_1, β_2 are orthogonal? (Standard theorems need not be proved but should be carefully quoted.)

(ii) Discuss carefully the S-Plus5 output for the data given below. How might you extend the analysis given?

```
From The Independent,
November 21, 2001, with the headline
'Supermarkets to defy bar on cheap designer goods'.

How prices compare: prices given in UK pounds.
```

Item	UK	Sweden	France	Germany	US
Levi 501 jeans	46.16	47.63	42.11	46.06	27.10
Dockers K1 khakis	58.00	54.08	47.22	46.20	32.22
Timberland women's boots	111.00	104.12	89.43	93.36	75.42
DieselKultar men's jeans	60.00	43.35	43.50	44.48	NA
Timberland cargo pants	53.33	48.58	43.54	58.66	31.70
Gap men's sweater	34.50	NA	26.93	27.26	28.76
Ralph Lauren polo shirt	49.99	42.04	36.41	40.26	32.48
H&M cardigan	19.99	17.31	18.17	15.28	NA

```
> p _ scan("pdata"); it _ 1:8; cou _ scan(,"")
UK Swe Fra Germ US

>x _ expand.grid(cou,it) ; country _ x[,1] ; item _ x[,2]
>item _ factor(item)
> first.lm _ lm(p~ country + item,na.action=na.omit)
> anova(first.lm)

Analysis of Variance Table

Response: p

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
  country  4   1115.56  278.890 10.57291 3.732294e-05
     item  7  16910.20 2415.743 91.58259 0.000000e+00
Residuals 25    659.44   26.378

> next.lm _ lm(p~ item + country, na.action=na.omit)
> anova(next.lm)

Analysis of Variance Table

Response: p

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
     item  7  16409.02 2344.146 88.86829 0.000000e+00
  country  4   1616.74  404.184 15.32293 1.859221e-06
Residuals 25    659.44   26.378
```

2 **(i)** Let $Y_1, \ldots, Y_n$ be independent binary random variables with

$$P(Y_i = 1) = p_i = 1 - P(Y_i = 0), \quad 1 \leqslant i \leqslant n,$$

where $p_1, \ldots, p_n$ are unknown probabilities. Describe briefly how to fit the model

$$\omega : \log \frac{p_i}{1 - p_i} = \beta^T x_i \quad , \quad 1 \leqslant i \leqslant n,$$

where $x_1, \ldots, x_n$ are given vectors, each of dimension p, and β is an unknown vector.

What is the maximised log-likelihood under the hypothesis $\Omega : 0 \leqslant p_i \leqslant 1$, $1 \leqslant i \leqslant n$? Why is the usual *deviance* not appropriate as a measure of the fit of ω?

(ii) Rousseauw *et al*, 1983, collected data on males in a heart-disease high-risk region of the Western Cape, South Africa. Our object is to predict chd = 1 or 0, i.e., coronary heart disease present or absent, from a set of covariates listed below

```
sbp          systolic blood pressure
tobacco      cumulative tobacco (kg)
ldl          low density lipoprotein cholesterol
adiposity
famhist      family history of heart disease (Present, Absent)
typea        type-A behaviour
obesity
alcohol      current alcohol consumption
age          age at onset
```

Interpret the corresponding S-Plus5 output, which makes use of the function

stepAIC

from library (MASS).

```
> SAheart.data[1:3,]
  sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
  1 160 12.00 5.73     23.11 Present    49   25.30   97.20  52   1
  2 144  0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
  3 118  0.08 3.48     32.28 Present    52   29.14    3.81  46   0
>table(famhist,chd)
         0  1
 Absent 206 64
Present  96 96

> first.glm _ glm(chd ~ sbp+tobacco+ldl+adiposity+famhist+typea+obesity+
+ alcohol + age, family = binomial)
> summary(first.glm,cor=F)

Coefficients:
                    Value  Std. Error     t value
(Intercept) -6.1506610935 1.306629106 -4.70727390
        sbp  0.0065040116 0.005727607  1.13555485
    tobacco  0.0793762052 0.026590779  2.98510268
        ldl  0.1739231824 0.059627387  2.91683387
  adiposity  0.0185864751 0.029270110  0.63499847
    famhist  0.9253661529 0.227736242  4.06332406
      typea  0.0395947051 0.012308368  3.21689313
    obesity -0.0629099612 0.044222058 -1.42259236
    alcohol  0.0001216154 0.004481130  0.02713944
        age  0.0452248070 0.012115699  3.73274426

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 596.1084 on 461 degrees of freedom

Residual Deviance: 472.14 on 452 degrees of freedom

Number of Fisher Scoring Iterations: 4
```

```
> stepAIC(first.glm)
Start:  AIC= 492.14
chd ~ sbp +tobacco +ldl +adiposity +famhist +typea +obesity +alcohol+
age

            Df Deviance      AIC
  - alcohol  1 472.1408 490.1408
- adiposity  1 472.5450 490.5450
      - sbp  1 473.4371 491.4371
     <none> NA 472.1400 492.1400
  - obesity  1 474.2332 492.2332
      - ldl  1 481.0701 499.0701
  - tobacco  1 481.6744 499.6744
    - typea  1 483.0466 501.0466
      - age  1 486.5284 504.5284
  - famhist  1 488.8851 506.8851

Step:  AIC= 490.14
chd ~ sbp +tobacco +ldl +adiposity +famhist +typea +obesity +age

            Df Deviance      AIC
- adiposity  1 472.5490 488.5490
      - sbp  1 473.4651 489.4651
     <none> NA 472.1408 490.1408
  - obesity  1 474.2404 490.2404
      - ldl  1 481.1541 497.1541
  - tobacco  1 482.0563 498.0563
    - typea  1 483.0604 499.0604
      - age  1 486.6412 502.6412
  - famhist  1 488.9925 504.9925

Step:  AIC= 488.55
 chd ~ sbp + tobacco + ldl + famhist + typea + obesity + age
```

```
          Df Deviance      AIC
    - sbp  1 473.9799 487.9799
   <none> NA 472.5490 488.5490
- obesity  1 474.6548 488.6548
- tobacco  1 482.5353 496.5353
    - ldl  1 482.9470 496.9470
  - typea  1 483.1925 497.1925
- famhist  1 489.3779 503.3779
    - age  1 495.4754 509.4754

Step:  AIC= 487.98
 chd ~ tobacco + ldl + famhist + typea + obesity + age

          Df Deviance      AIC
- obesity  1 475.6856 487.6856
   <none> NA 473.9799 487.9799
- tobacco  1 484.1760 496.1760
  - typea  1 484.2967 496.2967
    - ldl  1 484.5327 496.5327
- famhist  1 490.5818 502.5818
    - age  1 502.1120 514.1120

Step:  AIC= 487.69
 chd ~ tobacco + ldl + famhist + typea + age

          Df Deviance      AIC
   <none> NA 475.6856 487.6856
    - ldl  1 484.7143 494.7143
  - typea  1 485.4439 495.4439
- tobacco  1 486.0322 496.0322
- famhist  1 492.0948 502.0948
    - age  1 502.3788 512.3788
```

```
Call:
glm(formula = chd ~tobacco +ldl +famhist +typea +age,binomial)

Coefficients:
 (Intercept)    tobacco       ldl   famhist     typea        age
   -6.446392 0.08037506 0.1619908 0.9081708 0.0371149 0.05045984

Degrees of Freedom: 462 Total; 456 Residual
Residual Deviance: 475.6856

>summary(glm(chd ~tobacco+ldl+famhist+typea+age,binomial),cor=F)
Coefficients:
                  Value Std. Error   t value
(Intercept) -6.44639157 0.91929370 -7.012331
    tobacco  0.08037506 0.02586750  3.107183
        ldl  0.16199083 0.05493652  2.948691
    famhist  0.90817082 0.22560312  4.025524
      typea  0.03711490 0.01215529  3.053395
        age  0.05045984 0.01019143  4.951201

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 596.1084 on 461 degrees of freedom

Residual Deviance: 475.6856 on 456 degrees of freedom

Number of Fisher Scoring Iterations: 4
```

3 The table below shows the number of road accidents at eight different locations, over a number of years, before and after installation of some traffic control measures. The question of interest is whether there has been a significant change in the rate of accidents. Let

y_{ij} = number of accidents in location i under 'treatment' j

with $j = 1$ corresponding to 'before', and $j = 2$ to 'after'

installation of traffic control.

Let p_{ij} be the corresponding period of observation, so that for example $p_{11} = 9$ years, during which a total of $y_{11} = 13$ accidents were observed. (The total of 'Before' accidents was 114 over 68 years (rate 1.676/year), and the total of 'after' accidents was 15 over 18 years (rate 0.833/year).)

(i) Write down the equations to find the maximum likelihood estimates of the unknown parameters in the model in which y_{ij} are assumed independent Poisson variables with

$$\mathbb{E}(y_{ij}) = p_{ij}\mu_{ij}, \text{ and}$$
$$\log \mu_{ij} = \mu + \alpha_i + \beta_j, \qquad 1 \leqslant i \leqslant 8,\ 1 \leqslant j \leqslant 2,$$

and $\alpha_1 = \beta_1 = 0$.

Indicate briefly how glm( ) solves the corresponding equations, and interpret the attached S-Plus output.

(ii) Let e_{ij} be the corresponding 'fitted values' in this model. Show that

$$\sum_j e_{ij} = \sum_j y_{ij} \text{ for each } i, \text{ and}$$
$$\sum_i e_{ij} = \sum_i y_{ij} \text{ for each } j.$$

	Before		After	
Location	Years	Accidents	Years	Accidents
1	9	13	2	0
2	9	6	2	2
3	8	30	3	4
4	8	20	2	0
5	9	10	2	0
6	8	15	2	6
7	9	7	2	1
8	8	13	3	2

```
>summary(glm(acc ~ treat + site,poisson,offset=log(year)),cor=F)


Call:glm(formula =acc~treat+site,family=poisson,offset=log(year))
Deviance Residuals:
       Min        1Q      Median        3Q      Max
 -2.027386 -0.591431 -0.02094977 0.3122669 2.141791


Coefficients:
                 Value Std. Error    t value
(Intercept)  0.2707792  0.2784869  0.9723229
      treat -0.7806616  0.2751810 -2.8369024
      site2 -0.4855078  0.4493122 -1.0805578
      site3  1.0176088  0.3263931  3.1177397
      site4  0.5370828  0.3562308  1.5076822
      site5 -0.2623643  0.4205764 -0.6238207
      site6  0.5858730  0.3528776  1.6602725
      site7 -0.4855078  0.4493133 -1.0805552
      site8  0.1992985  0.3791789  0.5256054

(Dispersion Parameter for Poisson family taken to be 1 )


    Null Deviance: 132.9485 on 15 degrees of freedom


Residual Deviance: 16.27524 on 7 degrees of freedom


Number of Fisher Scoring Iterations: 4
```

4 A client has come to two statisticians (Dr. Mean and Dr. Variance) with data collected from a one-academic year randomised-controlled study on m students, known for their tendency to get into fights in school. The study randomised students to receive, at the beginning of the academic year, either the new Counselling and Managing Behaviour (CAMB) therapy treatment or the standard Warning treatment (which is administered at the time of a fight) in order to determine whether the new treatment procedure was effective in reducing the number of fight episodes seen during the academic year.

The client has brought the fight-episode data in the form of counts $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})$, $1 \leqslant i \leqslant m$, recorded for each term in the academic year. Additional information on a student is recorded in covariate vectors $\mathbf{x}_{ij}$, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant 3$, which includes information on what treatment was received.

Both Drs. Mean and Variance realise that there will be a correlation between the components of $\mathbf{Y}_i$. Dr. Mean decides to model the data as follows. He assumes that

$$\begin{aligned} \log E(Y_{ij} \,|\, \mathbf{x}_{ij}) &= \beta_0 + \beta^T \mathbf{x}_{ij} = \log \mu_{ij} \\ \mathrm{Var}\,(Y_{ij} \,|\, \mathbf{x}_{ij}) &= \mu_{ij} \\ \mathrm{Corr}\,(Y_{ij}, Y_{ik} \,|\, \mathbf{x}_{ij}, \mathbf{x}_{ik}) &= \rho\,(j \neq k). \end{aligned}$$

However, Dr. Variance decides to adopt the following alternative approach. She assumes that conditional on b_i, the responses Y_{ij}'s on the ith student are independent Poisson random variables with

$$\begin{aligned} E(Y_{ij} \,|\, \mathbf{x_{ij}}; b_i) &= \eta_{ij} \\ \mathrm{Var}\,(Y_{ij} \,|\, \mathbf{x}_{ij}; b_i) &= \eta_{ij} \\ \mathrm{Cov}\,(Y_{ij}, Y_{ik} \,|\, \mathbf{x_{ij}}, \mathbf{x_{ik}}; b_i) &= 0, (j \neq k) \\ \log \eta_{ij} &= b_i + \beta_0 + \beta^T \mathbf{x_{ij}} \end{aligned}$$

She also assumes that the $\exp(b_i)$'s are independent and identically distributed Gamma(τ^2/θ, τ/θ) (i.e. with mean τ and variance θ).

(i) What are the differences between the two approaches?

(ii) How would you interpret, for the client, the intercept parameter, β_0, and the treatment parameters, say β_1, from the two models? How would you interpret the parameter θ?

(iii) Find $\log \mathbb{E}(Y_{ij} \,|\, x_{ij})$ for Dr. Variance's model and compare it with the expression given in Dr. Mean's model. If Dr. Variance's model was correct in this situation, would Dr. Mean be *consistently* estimating what *he thinks* he is estimating? Explain your answer.

(iv) If the variance and correlation structures in Dr. Mean's model were incorrectly specified, but the mean structure was correctly specified, how would Dr. Mean be able to make valid inferences about the parameters of interest?

5 **(i)** Suppose that $y_1, \ldots, y_n$ are independent Poisson random variables, and $\mathbb{E}(y_i) = \mu_i$, $1 \leqslant i \leqslant n$. We wish to fit the model ω, defined as

$$\omega : \log \mu_i = \mu + \beta^T x_i, \quad 1 \leqslant i \leqslant n,$$

where μ, β are unknown parameters and $x_1, \ldots, x_n$ are given covariates. Show that the deviance D for testing the fit of ω may be written as

$$D = 2 \sum y_i \log(y_i / e_i)$$

where (e_i) are the "expected values" under ω, and show that $D \simeq \sum (y_i - e_i)^2 / e_i$.

(ii) Now suppose that $y_1, \ldots y_n$ are independent negative binomial variables, and that y_i has frequency function

$$f(y_i \,|\, \theta, \mu_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(\theta) y_i!} \quad \frac{\mu_i^{y_i} \theta^\theta}{(\mu_i + \theta)^{\theta + y_i}}$$

for $y_i = 0, 1, 2, \ldots$, thus $\mathbb{E}(y_i) = \mu_i$, $\mathrm{var}(y_i) = \mu_i + \mu_i^2 / \theta$.

Assume that θ is known. Show that the deviance for testing

$$\omega_n : \log \mu_i = \beta^T x_i \quad , \quad 1 \leqslant i \leqslant n$$

is say D_n, where

$$D_n = 2 \sum y_i \log \frac{y_i}{e_i} - 2 \sum (y_i + \theta) \log \frac{(y_i + \theta)}{(e_i + \theta)}$$

where (e_i) are the "expected values" under ω_n.