UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Tuesday 3 June 2003 9 to 12

PAPER 38

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 The Table below shows you the percentage of people with "excessive" alcohol consumption, classified by sex, age and year. Thus, for example, in 1996, 7% of women aged 65 and over had excessive alcohol consumption, that is, they consumed more than 14 units per week.

Health related behaviour: prevalence of alcohol consumption above 21/14 units a week for men/women ages 18 and over, in England,

	1986	1990	1992	1994	1996
men (above 21 units)					
18-24	39	37	38	36	42
25-44	22	33	30	30	31
45-64	24	26	24	27	27
65+	13	14	15	17	18
women (above 14 units)					
18-24	19	18	19	20	22
25-44	13	13	14	16	16
45-64	8	10	12	13	14
65+	4	5	5	8	7

Explain carefully (quoting any standard theorems necessary) the S-Plus analysis that follows below. What do you expect would be the result of the final S-Plus command?

```
>p
[1] 39 37 38 36 42 33 33 30 30 31 24 26 24 27 27 13 14 15 17 18 19 18
 19 20 22
[26] 13 13 14 16 16  8 10 12 13 14  4  5  5  8  7

> Sex _ scan(," ")
1: men women
3:
> Year _ scan(,"")
1: 1986 1990 1992 1994 1996
6:
> Age _ scan(,"")
1: 18-24 25-44 45-64 65+
5:
> x _ expand.grid(Year,Age,Sex)
> YEAR _ x[,1] ; AGE_ x[,2] ; SEX _ x[,3]
> is.factor(YEAR)
[1] T
> first.lm _ lm(p~ YEAR + SEX*AGE) ; summary(first.lm,cor=F)

Call: lm(formula = p ~ YEAR + SEX * AGE)
Residuals:
    Min      1Q  Median    3Q   Max
 -3.025 -0.6563 -0.1125 0.825 2.725
```

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  37.2750   0.8031    46.4128   0.0000
   YEAR1990   0.3750   0.7331     0.5115   0.6130
   YEAR1992   0.5000   0.7331     0.6820   0.5008
   YEAR1994   1.7500   0.7331     2.3870   0.0240
   YEAR1996   3.0000   0.7331     4.0920   0.0003
        SEX -18.8000   0.9274   -20.2726   0.0000
   AGE25-44  -7.0000   0.9274    -7.5483   0.0000
   AGE45-64 -12.8000   0.9274   -13.8026   0.0000
     AGE65+ -23.0000   0.9274   -24.8015   0.0000
SEXAGE25-44   1.8000   1.3115     1.3725   0.1808
SEXAGE45-64   4.6000   1.3115     3.5075   0.0015
  SEXAGE65+   9.2000   1.3115     7.0149   0.0000

Residual standard error: 1.466 on 28 degrees of freedom
Multiple R-Squared: 0.9858
F-statistic: 177.1 on 11 and 28 degrees of freedom, the p-value is 0
>interaction.plot(AGE,SEX,p)
```

2 The numbers of UK new vCJD patients classified by calendar year of onset, for the years 1999 and 2002, are given in the following 2×2 table

	males	females
1999	20	9
2000	12	11

Discuss carefully the (slightly edited) S-Plus output that follows below. Any general theorems needed may be used without proof.

How would you interpret the above table to a non-statistician?

```
> a _ c(20,9)
> b _ c(12, 11)
> r _ c(a,b)
> Row _ c(1,1,2,2) ; Col _ c(1,2,1,2)
> Row _ factor(Row); Col _ factor(Col)
> first.glm _ glm(r~ Row*Col,poisson)
>summary(first.glm,cor=F)
Call: glm(formula = r ~ Row * Col, family = poisson)

Coefficients:
                Value Std. Error    t value
(Intercept)  2.9957323  0.2236068  13.397322
        Row -0.5108256  0.3651484  -1.398954
        Col -0.7985077  0.4013865  -1.989374
    Row:Col  0.7114963  0.5790972   1.228630

    Null Deviance: 5.016056 on 3 degrees of freedom

Residual Deviance: 0 on 0 degrees of freedom

Number of Fisher Scoring Iterations: 1

>next.glm _ glm(r~ Row+Col, poisson)
>summary(next.glm,cor=F)

Call: glm(formula = r ~ Row + Col, family = poisson)

Coefficients:
                Value Std. Error    t value
(Intercept)  2.8817880  0.2156372  13.364058
        Row -0.2318016  0.2791960  -0.830247
        Col -0.4700036  0.2850183  -1.649030

    Null Deviance: 5.016056 on 3 degrees of freedom

Residual Deviance: 1.527855 on 1 degrees of freedom
```

```
Number of Fisher Scoring Iterations: 3

>fisher.test(rbind(a,b))

Fisher's exact test

data:  rbind(a, b)
p-value = 0.2597
alternative hypothesis: two.sided
```

3 (a) Suppose $y_1, \ldots, y_n$ are independent binary observations, with

$$\pi_i = P(Y_i = 1) = 1 - P(Y_i = 0)$$

and we wish to fit the model $H_0 : \text{logit}\, \pi_i = \beta^T x_i$, $1 \leqslant i \leqslant n$, where $x_1, \ldots, x_n$ are given covariate values, each of dimension p. Take H_1 as the "saturated" model $0 \leqslant \pi_i \leqslant 1$, $1 \leqslant i \leqslant n$. Show that the maximised loglikelihood, under H_1, is always 0, regardless of the values of $y_1, \ldots, y_n$.

(b) Comment on the S-Plus output for the data-set described below. You should describe the models being fitted, and interpret the corresponding terms in the output. (You may assume that the logistic model is taken with $\pi_i = P(Y_i = 1) = P$ (response = "yes").)

The data set

J.W. Smith et al (1988), "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", published a data-set relating to a population of women who were at least 21 years old, of Pima Indian heritage, and living near Phoenix, Arizona. Each woman was tested for diabetes according to World Health Organization criteria. The first few lines of the data are given in the Table below. The reported variables are

npreg = number of pregnancies,
glu = plasma glucose concentration in an oral glucose tolerance test,
bp = diastolic blood pressure (mm Hg)
skin = triceps skinfold thickness (mm)
bmi = body mass index (weight in kg/(height in m)^2),
ped = diabetes "pedigree" function
age = age in years
type = Yes (ie diabetic) or No (ie not diabetic)

npreg	glu	bp	skin	bmi	ped	age	type
5	86	68	28	30.2	0.364	24	No
7	195	70	33	25.1	0.163	55	Yes
5	77	82	41	35.8	0.156	35	No
0	165	76	43	47.9	0.259	26	No
0	107	60	25	26.4	0.133	23	No
5	97	76	27	35.6	0.378	52	Yes
3	83	58	31	34.3	0.336	25	No

```
Call: glm(formula = type ~ npreg + glu + bp + skin + bmi + ped + age,
family =
binomial)
Deviance Residuals:
       Min         1Q     Median         3Q      Max
 -1.982974 -0.6772605 -0.3680958 0.6439307 2.315364

Coefficients:
                   Value  Std. Error     t value
(Intercept) -9.772793573 1.764308691 -5.53916308
      npreg  0.103180903 0.064586211  1.59756860
```

```
         glu  0.032115958 0.006768506  4.74491126
          bp -0.004766793 0.018502716 -0.25762664
        skin -0.001916782 0.022450798 -0.08537703
         bmi  0.083620686 0.042733255  1.95680592
         ped  1.820337113 0.663665204  2.74285453
         age  0.041182353 0.022051102  1.86758707

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 256.4142 on 199 degrees of freedom

Residual Deviance: 178.3907 on 192 degrees of freedom

Number of Fisher Scoring Iterations: 4

Call: glm(formula = type ~ glu, family = binomial)
Deviance Residuals:
       Min        1Q     Median       3Q     Max
 -1.971406 -0.779478 -0.5291695 0.849138 2.26331

Coefficients:
                  Value  Std. Error   t value
(Intercept) -5.50363485 0.835824892 -6.584675
        glu  0.03778371 0.006275751  6.020588

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 256.4142 on 199 degrees of freedom

Residual Deviance: 207.3727 on 198 degrees of freedom

Number of Fisher Scoring Iterations: 4
```

4 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

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

where μ and 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

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

where (e_i) are the "expected values" under ω, and show that $\Sigma e_i = \Sigma y_i$. How is D used to check ω?

(ii) Suppose $y_1, \ldots, y_n$ is a random sample from the frequency function

$$f(y|\mu, \theta) = \frac{\Gamma(\theta + y)}{\Gamma(\theta) y!} \quad \frac{\mu^y \theta^\theta}{(\mu + \theta)^{\theta + y}} \quad \text{for } y = 0, 1, \ldots \ .$$

Show that $\mathbb{E}(Y) = \mu, \text{var}(Y) = \mu + \frac{\mu^2}{\theta}$, and that if $(\hat{\mu}, \hat{\theta})$ is the maximum likelihood estimator of (μ, θ) obtained from $(y_1, \ldots, y_n)$, then the asymptotic correlation of $\hat{\mu}, \hat{\theta}$ is zero.

5 Depression is a serious mental disorder that ranks as one of the leading causes of disability in developed countries.

A psychiatrist has collected data from a randomised-controlled trial on m subjects in the community who suffer from clinical depression. The study was designed to assess the effectiveness of a new anti-depression drug in reducing the recurrence of clinical depression, as compared to the standard prescribed drug treatment. The trial was conducted over a six-month period. At two-month intervals, a validated depression questionnaire, SAD (Schedule for the Assessment of Depression), was administered, which recorded information on depression tendencies over the prior two-month period. The information from the questionnaire was summarised into a binary outcome indicating whether or not the patient was depressed during the previous two months. The outcome data for the ith subject was recorded as a vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})$ taken over the three time intervals. Baseline information on each patient, i, was recorded in a covariate vector $\mathbf{x}_i$. The treatment variable is denoted by the binary variable z_i, and its parameter is denoted by ϕ. The variable $t_j (j = 1, \ldots, 3)$ records the time interval under observation and takes the values 2, 4 or 6 months. Unfortunately, as with many other psychiatric studies, patients dropped-out during the six-month period and consequently there were missing outcome data after dropout.

The psychiatrist has attempted to analyse the data by assuming that Y_{ij}'s are independent Bernoulli random variables with means modelled as

$$\log \frac{E(Y_{ij}|z_i; \mathbf{x}_i; t_j)}{1 - E(Y_{ij}|z_i; \mathbf{x}_i; t_j)} = \alpha + \phi Z_i + \beta^T \mathbf{x}_i + \delta t_j, \quad (i = 1, \ldots, m; j = 1, \ldots, 3).$$

However the psychiatrist being hesitant of publishing incorrectly analysed data approaches you with the data set, and with the results obtained from fitting the model above.

(i) Will the results obtained from the psychiatrist's analysis be correct? Explain your answer.

(ii) How would you "correctly" model the data in each of the following 2 cases?

(a) The psychiatrist is interested in making public health recommendations for the treatment of clinical depression in the community.

(b) The psychiatrist is interested in determining the potential individual-specific effect of the new anti-depression drug on individual patient's response profile.

You need to write out in full the models you suggest, defining all new notations used and stating all assumptions made.

(iii) What are the differences (if any) between your models, in terms of interpretation of parameters (e.g. the intercept, treatment parameter and the time slope parameter), and validity under different missing data mechanisms?

(iv) What would you do if the missing data mechanism was thought to be informative?