MATH382301

This question paper consists of 6 printed pages, each of which is identified by the reference **MATH3823**.

New Cambridge Elementary Statistical Tables are provided. Only approved basic scientific calculators may be used.

Examination for the Module MATH3823
(May / June 2004)

GENERALIZED LINEAR MODELS

Time allowed: **2 hours**

Attempt not more than THREE questions.
All questions carry equal marks.

1. Consider a generalized linear model representing a response variable Y in terms of a set of explanatory variables $X_1, \ldots, X_p$.

(a) A generalized linear model contains three main components. Name these components and briefly describe them.

(b) Write down the three components from part (a) for the usual linear regression model with normal errors and constant variance assuming there are no interaction terms.

(c) Explain how each component can be extended beyond the linear regression model, giving examples where appropriate.

(d) The normal linear model in matrix form can be written $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^T + \boldsymbol{\varepsilon}$; here $\mathbf{X}$ is called the design matrix. Explain why representing a qualitative variable X taking k levels by k dummy variables leads to aliasing. How is this problem overcome?

(e) Let X_1 be a factor taking two levels, X_2 a factor taking four levels, and X_3 a quantitative variable. Suppose that in a sample of $n = 8$ individuals, these variables take the following values: $X_1 = (1,1,1,1,2,2,2,2)$, $X_2 = (1,2,3,4,1,2,3,4)$, and $X_3 = (17,14,12,19,23,12,14,16)$.

Show that the design matrix for the model $Y \sim X_1 + X_3$ may be represented in the form

$$\begin{bmatrix} 1 & 0 & 17 \\ 1 & 0 & 14 \\ 1 & 0 & 12 \\ 1 & 0 & 19 \\ 1 & 1 & 23 \\ 1 & 1 & 12 \\ 1 & 1 & 14 \\ 1 & 1 & 16 \end{bmatrix},$$

and find the design matrix for the model $Y \sim X_1 * X_3 + X_2$.

2. A study was carried out on drug-related complaints at a hospital accident and emergency unit. Each patient in a sample of $n = 2100$ was classified by

S: sex ($S = 1$: male; S = 2: female),

M: marital status ($M = 1$: divorced; $M = 2$: widowed; $M = 3$: married), and

C: drug-related complaint ($C = 1$: overdose; $C = 2$: suicide; $C = 3$: psychiatric; $C = 4$: addiction).

The table below gives the counts for this three-way classification.

Drug-related complaints by sex and marital status

	Marital Status					
	Divorced		Widowed		Married	
Complaint	Male	Female	Male	Female	Male	Female
Overdose	96	208	46	82	266	330
Suicide	14	100	18	96	72	156
Psychiatric	38	24	22	10	116	36
Addiction	52	68	14	12	146	78
Total	200	400	100	200	600	600

(a) Suppose the count Y_{ijk} in cell (i, j, k) is modelled by a Poisson distribution $Po(\lambda_{ijk})$. Describe how λ_{ijk} can be modelled to depend on the factors S, M, and C and their interactions through a log-linear model. Formulate this model as a generalized linear model.

(b) In the log-linear model, it is also possible to regard the values of S, M, and C as random for each patient through a suitable conditioning argument. Derive the joint distribution for (S, M, C) in terms of the λ_{ijk}.

For the log-linear model denoted by $Y \sim S + M * C$, derive the joint distribution of (S, M, C) in terms of the parameters representing S, M, C, and the parameters for the interactions between M and C. For this model, describe the independence properties between S, M, and C.

(c) For the data given above, the purpose of the study is to see how the complaint C depends on sex S and marital status M. The data, as given, cannot be used to draw conclusions about drug-related complaints for all people. What further data would be required before such conclusions could be made?

A variety of log-linear models have been fitted with the deviances and degrees of freedom given below. Give the missing values (i) to (iv).

Choose which model you would use to describe the data, giving reasons for your choice. What further analysis might you do to confirm the adequacy of this model?

Model		Deviance	d.f.
A.	$Y \sim S + M + C$	316.0	(i)
B.	$Y \sim S + M * C$	255.2	11
C.	$Y \sim M + S * C$	117.4	(ii)
D.	$Y \sim S * M + S * C$	58.5	(iii)
E.	$Y \sim S * C + M * C$	56.5	(iv)
F.	$Y \sim S * M + S * C + M * C$	5.5	6

(d) For the model you have chosen in part (c), why might it be dangerous to ignore the data on the sex of the patients and simply analyse the two-way table of data on the counts broken down by marital status and complaint?

Briefly explain what the problem of overdispersion is and why it might affect this data set. If you believed that overdispersion was a problem here, how might you allow for it in your analysis of the data?

3. Let Y be a random variable following an exponential family distribution with cannonical parameter θ, scale parameter ϕ and density function

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}. \tag{1}$$

(a) Show that $E(Y) = b'(\theta)$ and $\text{var}(Y) = b''(\theta)a(\phi)$.

(b) Given n independent observations $\mathbf{y} = (y_1, \ldots, y_n)$ from an exponential family distribution with density function (1), find the equation for the maximum likelihood estimator $\widehat{\theta}$ of θ.

Let θ_0 denote the true value of θ. From the Taylor expansion of $b'(\widehat{\theta})$ about θ_0, show that $\widehat{\theta}$ is approximately unbiased.

(c) Let Y follow a gamma distribution with density function

$$f(y; \lambda, \alpha) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha} y^{\alpha-1} e^{-\lambda y}, \qquad y \geqslant 0,$$

where λ and α are positive parameters.

Show that Y has an exponential family distribution with scale parameter $1/\alpha$. Find the canonical parameter θ and the functions $a(\phi), b(\theta)$, and $c(y, \phi)$. Use this representation of Y and the results in part (a) to find $E(Y)$ and $\text{var}(Y)$.

Now assume that $Y_1, \ldots, Y_n$ are independent $\text{Gamma}(\lambda, \alpha)$ random variables. Write down the maximum likelihood estimator $\widehat{\theta}$ in this case.

4. Consider a dose-response experiment where female mice were given food containing varying doses of a contraceptive drug, conestrathane. For each of n dosage levels x_i (in μg), m_i mice were tested, of which y_i did not conceive in the three-month period of the study.

(a) Describe how a generalized linear model with a logit link function can be used to model this data set, where the systematic part of the model is given by $\eta = \beta_0 + \beta_1 x$. In particular, express the probability, p, that a mouse does not conceive in terms of the dosage and the regression parameters β_0 and β_1.

(b) The following data were gathered from the study:

Mouse contraception data

Dosage, x_i	0	3	6	9	12	15	18	21	24	27	30
No. of mice on trial, m_i	113	123	121	125	122	127	118	115	116	124	125
No. of mice not conceiving, y_i	6	11	37	56	78	102	109	115	112	123	125

A logistic regression model was fitted to these data in R, with the following results:

```
> response <- cbind(y, m - y)
> glm1 <- glm(response ~ x, family = binomial)
> summary(glm1)

Call:
glm(formula = response ~ x, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50368  -0.56869  -0.03272   0.48659   2.86103

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.86737    0.18579  -15.43   <2e-16 ***
x            0.29453    0.01606   18.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 870.513  on 10  degrees of freedom
Residual deviance:  15.030  on  9  degrees of freedom
AIC: 56.374

Number of Fisher Scoring iterations: 4
```


QUESTION 4 CONTINUED...

Comment on the quality of the fit. What else might you do to check the adequacy of this model? (You do not need to do any calculations or supply any R commands, merely indicate what you would do.)

(c) Write down an expression for p in terms of the dosage x. It is desired to choose x so that 95% of female mice on this dosage do not conceive in a three month period. What value of x would you use?

(d) Assume that each female mouse has a fertility modelled by a random variable Z with cumulative distribution function $F_Z(z)$ so that the mouse does not conceive if and only if $Z < 0$. Also assume that this fertility is reduced by a dose of conestrathane to $W = Z - \beta_0 - \beta_1 x$.

Show that this represents a generalized linear model and express the link function $g(\mu)$ in terms of F_Z. Show that if Z has a logistic distribution with distribution function

$$F_Z(z) = \frac{e^z}{1 + e^z},$$

this formulation leads to the logit link function. What distributions for Z lead to the probit and complementary log-log link functions respectively?

END