P.M.E.ALTHAM, November 1998.

Here are some extra problems on generalized linear modelling. These problems are constructed from extracts from recent examination questions for Part IIA of the Cambridge University Mathematics Tripos, which is an examination taken by third-year mathematics undergraduates, and the Diploma in Mathematical Statistics, which was an examination taken by first year graduate students in statistics, now replaced by the M.Phil. in Statistical Science.

MATHEMATICAL TRIPOS

1994.A1.no11.
Suppose $Y_1, \ldots, Y_n$ are independent observations, with Y_i distributed as Poisson with mean μ_i, where

$$\log(\mu_i) = \beta^T x_i, \ i = 1, \ldots, n,$$

and where ${x_1}^T, \ldots, {x_n}^T$ are the rows of a known $n \times p$ matrix X of rank p. Write down the log-likelihood $\ell(\beta)$ and find $\frac{\partial \ell}{\partial \beta}$ and $\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}$.

Show that the matrix $\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}$ is negative-definite. How is this relevant to the problem of finding the maximum likelihood estimator $\hat{\beta}$ of β?

1994.A2.no10.
A.C. Atkinson (1986) analysed data on the record times in 1984 for 35 Scottish hill races. The three variables recorded were x, the distance on the map (in miles), z, the total height gained during the route (in feet), and y, the record time (in minutes). Consider the linear model

$$H : E(y_i) = \beta_1 + \beta_2 x_i + \beta_3 z_i, \quad i = 1, \ldots, 35$$

with the usual assumption of independent normal errors with common unknown variance σ^2. The following three models were fitted:

(a) $E(y_i) = \beta_1$, giving $s = 50.04, df = 34$,

(b) $E(y_i) = \beta_1 + \beta_2 x_i$, giving $\hat{\beta}_2 = 8.3305 (se = 0.6194), s = 19.96, df = 33$,

(c) $E(y_i) = \beta_1 + \beta_2 x_i + \beta_3 z_i$, giving $\hat{\beta}_2 = 6.2180 (se = 0.6011), \hat{\beta}_3 = 0.0110 (se = 0.0021), s = 14.68, df = 32$.
Here, in each of the 3 cases, s^2 is defined as (Residual sum of squares)/df.

On the basis of these three fits, which is your preferred model for the record time? Give reasons for your answer.

For these data $\Sigma_i (x_i - \bar{x})(z_i - \bar{z}) > 0$. In what respect would the above analysis have been simpler if in fact $\Sigma_i (x_i - \bar{x})(z_i - \bar{z}) = 0$?

1994.A4.no13.
In 1974 the University of Chicago National Research Center asked 1305 male respondents, of varying educational background, whether they agreed or disagreed with the following statement:

"*Women should take care of running their homes and leave running the country up to men.*"

For $i = 0, 1, \ldots, 20$, let us denote by t_i the number of respondents with i years of education, by a_i the number of these who agreed with the statement, and by d_i the number who disagreed. Thus $t_i = a_i + d_i$.

Assume that $a_i, i = 0, \ldots, 20$, are independent binomial variables with parameters (t_i, p_i). Write down the log-likelihood $\ell(p_0, \ldots, p_{20})$.

Describe an iterative method to fit the hypothesis

$$\log(p_i/(1-p_i)) = \alpha + \beta\ i,\ i = 0, \ldots, 20$$

and explain how to test the hypothesis $\beta = 0$. Give the corresponding S-Plus or GLIM commands.

In fact the result of fitting the model was as follows:
$\hat{\alpha} = 2.098, (se = 0.2355)$,
$\hat{\beta} = -0.234, (se = 0.02019)$,
deviance $= 18.95, df = 19$.

How do you interpret this?

Sketch the fitted values of $\hat{p}_i$ as a function of i.

1995.A1.no11.
The table below comes from a study of British doctors by R. Doll and A.B. Hill (1966), and gives the number of coronary deaths for smokers, for 5 different age-groups, with the corresponding 'person-years', ie total time at risk.

Age	35–44	45–54	55–64	65–74	75–84
person-years	52407	43248	28612	12663	5317
number of deaths	32	104	206	186	102

Let y_i be the number of deaths from age-group i, and let t_i be the corresponding number of person-years, for $1 \leq i \leq 5$. In fitting the model

$$y_i \sim Po(\mu_i t_i),$$

with

$$\log(\mu_i) = \alpha + \beta\ i + \gamma\ i^2,\ 1 \leq i \leq 5,$$

we obtain

deviance $= 0.297, df = 2$,
$\hat{\alpha} = -9.289 (se = .3025), \hat{\beta} = 2.026 (se = .1936), \hat{\gamma} = -0.1910 (se = .02925)$.

Interpret these results stating any general properties of generalized linear modelling to which you appeal.

Would you expect the model

$$\log(\mu_i) = \alpha + \beta i,\ 1 \leq i \leq 5$$

to be a good fit?

If you had the corresponding two rows of data for the non-smokers, what models would you consider for the full data set?

1995.A2.no10.
Norton and Dunn (1985) presented the following data, based on an epidemiological survey to investigate snoring as a possible risk factor for heart disease. Those surveyed were classified according to their spouses' report of how much they snored.

The individuals were classified by Snoring, as

Never ($x = 0$)
Occasional ($x = 2$)
Nearly every night ($x = 4$), or
Every night ($x = 5$).

They were also classified according to whether they had a heart attack, for which the corresponding observed frequencies were

24, 35, 21, 30, respectively,

or did *not* have a heart attack, for which the corresponding observed frequencies were

1355, 603, 192, 224 respectively.

Thus the observed proportion of those having a heart attack was

0.017,0.055, 0.099, 0.118 respectively.

Using the x-values given above, Agresti (1996) obtained, with the binomial 'error function', the regression equation

$$\log(p(x)/(1-p(x)) = -3.866(.166) + 0.397(.050)\ x$$

(standard errors in brackets) where $p(x) = P(\text{heart attack} \mid \text{snoring}= x)$. The corresponding deviance was $2.809, df = 2$.

Give a careful interpretation of these results. (A detailed mathematical exposition is *not* sought for this part of the question.)

1996.A1.no11.
(i) The linear model

$$y_i = \beta^T x_i + \epsilon_i,\ 1 \le i \le n,$$

with ϵ_i normally and independently distributed, mean 0, unknown variance σ^2, may be rewritten as

$$y = X\beta + \epsilon,$$

where X is a $n \times p$ matrix, which you may assume to be of rank p. Let

$$R(\beta) = (y - X\beta)^T(y - X\beta).$$

Derive an expression for $\hat{\beta}$, the maximum likelihood estimate of β, and state without proof the joint distribution of $(\hat{\beta}, R(\hat{\beta}))$.

(ii) Consider the following special case of the above model

$$y_i = \alpha + \beta(x_i - \bar{x}) + \gamma(z_i - \bar{z}) + \epsilon_i,\ 1 \leq i \leq n,$$

where now α, β, γ are unknown scalar parameters, where $\epsilon_i \sim NID(0, \sigma^2)$ with σ^2 known, and where

$$\bar{x} = n^{-1}\Sigma x_i, \bar{z} = n^{-1}\Sigma z_i.$$

Find $\hat{\beta}$ and var$(\hat{\beta})$.

Let β^* be the maximum likelihood estimate of β under the model

$$H_0 : y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i,\ 1 \leq i \leq n.$$

Find β^* and var(β^*), and show that

$$var(\beta^*) \leq var(\hat{\beta}).$$

When is this inequality an equality? How would you test H_0?

You may be interested to know that the above inequality is a (very) special case of the result stated by Altham (1994).

The message of this result can loosely be stated as this: the fewer parameters you fit, the more accurate these parameter estimates will be. The result provides one reason for fitting models which are *parsimonious* in parameters, if the data permit.

1996.A2.no10.

The data in the table below, slightly modified from Crawley (1993), is from a field study on insect parasitism. The number d_i of parasitized caterpillars in a total of n_i lepidopteran caterpillars was counted, in 6 independent random samples from each of 3 different habitats, labelled $h = 1, 2, 3$. At the time the insects were collected, an estimate x_i of the corresponding insect population density was recorded.

x	d	n	h
3	7	14	1
5	10	22	1
11	9	22	1
12	8	17	1
22	6	10	1
57	8	11	1
2	3	10	2
7	2	5	2
17	20	31	2
23	17	20	2
29	9	11	2
33	18	22	2
4	8	17	3
7	10	22	3
7	7	15	3
10	6	9	3
12	22	43	3
14	5	11	3

With the model d_i independent Binomial, parameters n_i, p_i, and

$$\log(p_i/(1-p_i)) = \mu + h_j + \beta\ x$$

where j is the level of the factor h, so that j has possible values $1, 2, 3$, standard glm software finds that the deviance is $8.434(df = 14)$, and the parameter estimates for μ, h_2, h_3, β are respectively

$-0.5255(0.2698)$, $0.5372(0.3207)$, $0.1512(0.2857)$, $0.03844(0.01297)$

(with the standard errors in brackets).

Give a careful interpretation of this output, with a suitable sketch-graph.

The next step in the analysis was to fit the model

$$\log(p_i/(1-p_i)) = \mu + \beta\ x.$$

This caused the deviance to increase by $3.261(df = 2)$, so that the resulting model had deviance $11.695(df = 16)$. What does this mean?

1996.A4.no13.
Suppose that $y_1, \ldots, y_k$ are independent Poisson variables and, for $1 \leq i \leq k$,

$$E(y_i) = \mu_i,$$
$$\log(\mu_i) = \mu' + \beta^T x_i,$$

where (x_i) are known and (μ', β) are unknown.

(a) Show that $(\Sigma y_i, \Sigma x_i y_i)$ is sufficient for (μ', β), and that the observed and expected value of this vector coincide at $(\hat{\mu}', \hat{\beta})$, the maximum likelihood estimate of (μ', β).

(b) Show that the asymptotic covariance matrix of $\hat{\beta}$ is the inverse of the matrix

$$\Sigma \mu_i x_i {x_i}^T \ - \ (\Sigma \mu_i)^{-1} (\Sigma \mu_i x_i)\ (\mu_i {x_i}^T),$$

where $\mu_i = exp(\mu' + \beta^T x_i),\ 1 \leq i \leq k.$

HINT: added May 1998

You will need to invert a partitioned matrix, which is most easily done by considering how to solve

$$a u_1 + b^T u_2 = v_1$$
$$b u_1 + C u_2 = v_2$$

for u_1, u_2 as functions of v_1, v_2.

(Here a is a scalar, b is a vector, and C is a square matrix.)

DIPLOMA IN MATHEMATICAL STATISTICS

Diploma Paper A.1994.no10.
We plan to carry out a medical study on a large number of patients, to investigate the

possible association between a disease D and patients' covariate values x(e.g. age, sex, smoking status etc.). Let

$D_i = 1$ if the i^{th} patient has the disease, $D_i = 0$ otherwise,
$x_i =$ the vector of covariate values for the i^{th} patient (fixed and known),
$S_i = 1$ if the i^{th} patient is *selected* for the study, $S_i = 0$ otherwise.

Assume that for $i = 1, \ldots, n$,

$$\log(P(D_i = 1|x_i)/P(D_i = 0|x_i)) = \alpha + \beta^T x_i,$$

and

$$P(S_i = 1|D_i = 1) = \rho_1,\ P(S_i = 1|D_i = 0) = \rho_0,$$

where ρ_1, ρ_0 are known. Show that

$$\log(P(D_i = 1|x_i, S_i = 1)/(P(D_i = 0|x_i, S_i = 1)) = \alpha^* + \beta^T x_i$$

where α^* is to be defined.

Hence write down the loglikelihood $\ell(\alpha^*, \beta)$ for those patients for whom $S_i = 1$, and discuss briefly the estimation of β.

Diploma Paper A.1995.no10.

(a) In an experiment to compare 3 brands of instant coffee, A_1, A_2 and A_3, a number of student volunteers are available, each able to compare exactly 2 brands, and to say which brand he or she prefers. The data are obtained as follows: n_{ij} students are given A_i and A_j, and of these r_{ij} students prefer A_i to A_j, for $1 \le i < j \le 3$. Assuming that the trials are conducted so that

$$r_{ij} \sim \text{independent } Bi(n_{ij}, p_{ij}),\ 1 \le i < j \le 3,$$

discuss carefully how to fit the model

$$H_0 : \log(p_{ij}/(1 - p_{ij})) = \alpha_i - \alpha_j,\ 1 \le i < j \le 3.$$

Why do we need to impose a constraint on $\alpha_1, \alpha_2, \alpha_3$?

(b) With data $r_{12} = 7$, $n_{12} = 10$, $r_{23} = 6, n_{23} = 11$, $r_{13} = 9, n_{13} = 12$, we find:
deviance $= .004, df = 1$,
$\hat{\alpha}_1 = 1.075(se = .5380)$
$\hat{\alpha}_2 = 0.2020(se = .5124)$
with the constraint $\alpha_3 = 0$.

If we now fit the model H_0 as above, with the restrictions $\alpha_2 = \alpha_3 = 0$, we find:
deviance $= .159, df = 2$,
$\hat{\alpha}_1 = 0.9808(se = .4787)$.

What is your conclusion about the students' preferences?

Diploma Paper A.1996.no9.

(a) Suppose data (y_{ij}) is such that $y_{ij} \sim$ independent Poisson, mean μ_i, for $1 \le j \le n_i,\ 1 \le i \le k$.

Derive an expression for the deviance used in testing the fit of the hypothesis

$$H_0 : \log(\mu_i) = \beta^T x_i, 1 \leq i \leq k$$

(where $x_1, \ldots, x_k$ are given covariates) against the alternative

$$H : \mu_1, \ldots, \mu_k \text{ any positive numbers.}$$

(b) If the full dataset (y_{ij}) is replaced by $(\Sigma_j y_{ij}, 1 \leq i \leq k)$, how does this affect

(i) the estimation of β, and the corresponding standard errors?
(ii) the deviance for testing H_0 against H_1?

How should you ask your glm software to make use of the information (n_i)?

Diploma 1996 Paper A.no10.
Suppose $y_1, \ldots, y_n$ are independent, with

$$f(y_i|\mu_i) = \frac{1}{\mu_i} \exp -(y_i/\mu_i),\ y_i > 0.$$

(a) Show that if $\log(\mu_i) = \beta^T x_i$ for $1 \leq i \leq n$ where $x_1, \ldots, x_n$ are given covariates, then the asymptotic covariance matrix of $\hat{\beta}$, the mle of β, is free of β.
(b) Discuss the estimation of β if we assume, instead of (a), that $1/\mu_i = \beta^T x_i$ for $1 \leq i \leq n$.

References

Agresti, A. (1996).*An Introduction to Categorical Data Analysis.* New York: Wiley.
Altham, P.M.E. (1994). Improving the precision of estimation by fitting a generalized linear model, and Quasi-likelihood. *Glim Newsletter* **23**,43-45.
Atkinson, A.C. (1986). Comment: Aspects of diagnostic regression analysis. *Statistical Science* **1**, 397-402.
Collett, D. (1991). *Modelling Binary Data.* London:Chapman and Hall. (for the 1994.A4.no13 data)
Crawley, M.J. (1993). *GLIM for Ecologists.* Oxford: Blackwell Scientific Publications.
Doll, R. and Hill, A.B. (1996)*Natl. Cancer Inst. Monogr.* **19**, 205-268.
Norton, P.G. and Dunn, E.V. (1985), *Brit. Med. J.* **291**, 630-632.

Statistics : Analysis of Drink-Drive data

For this project, it will be helpful to have attended the IIA course, Computational Statistics and Statistical Modelling, but other candidates who are familiar with GLIM (or similar suitable statistical software) may also attempt the project.

The Independent on Sunday (2 January 1994) printed the table below under the headline "Sharp increase in road accidents over Christmas. Police condemn 'hard core' of risk takers". The data are available in the file DRINKDAT. Analyse these data in any way you think might be appropriate, bearing in mind the likely points of interest for:

(i) the typical UK driver;

(ii) the Minister of Transport;

(iii) the Association of Chief Police Officers.

What (if any) additional information might have been useful in analysing this data?

[Suggestions for this essay:

(a) good graphic displays of data;

(b) are any particular *regions* obvious 'outliers'?

(c) changes from 1992 to 1993;

(d) are any regressions helpful?]

Your answer, which may be handwritten, must not exceed 10 pages in length, including any relevant tables, print-outs and graphs. It should be clear from your answer precisely which tests you have used, but you do not need to describe the theory underlying these tests.

Do not attempt to write a polished report; you should think of your answer as providing an organised collection of statistical analyses, graphs, tables, and comments that would be useful to someone else who did wish to write a full report on the data.

Drink–Drive Figures for England and Wales during Christmas Campaign

Key to Column Headings

tst: number of breath tests
+ve: number of positive tests
acc: number of accidents involving injury

	tst 93	tst 92	+ve 93	+ve 92	acc 93	acc 92	Region
1.	317	514	49	85	102	150	Avon and Somerset
2.	965	702	54	59	57	85	Bedfordshire
3.	1700	1558	53	39	89	81	Cambridgeshire
4.	1124	826	154	78	96	69	Cheshire
5.	236	86	41	9	1	2	City of London
6.	474	621	48	52	49	33	Cleveland
7.	632	757	46	33	64	29	Cumbria
8.	2088	807	73	61	93	92	Derbyshire
9.	1172	1263	85	119	124	135	Devon and Cornwall
10.	541	626	41	62	86	41	Dorset
11.	742	1015	71	67	41	35	Durham
12.	661	776	35	41	27	27	Dyfed–Powys
13.	2786	2754	119	105	144	171	Essex
14.	367	408	35	25	61	36	Gloucester
15.	7591	5126	350	297	324	299	Greater Manchester
16.	906	734	55	34	38	24	Gwent
17.	2314	1982	114	134	137	137	Hampshire
18.	646	428	66	49	73	68	Hertfordshire
19.	522	525	76	65	78	87	Humberside
20.	1609	2029	99	109	166	153	Kent
21.	1222	1423	104	127	141	153	Lancashire
22.	1356	1086	76	54	62	59	Leicestershire
23.	1034	941	55	49	58	51	Lincolnshire
24.	867	566	105	130	226	230	Merseyside
25.	8792	12379	461	804	729	789	Metropolitan
26.	643	917	27	35	69	66	Norfolk
27.	645	728	43	33	36	41	Northamptonshire
28.	494	383	108	118	138	153	Northumbria
29.	1599	1284	97	76	80	32	North Wales
30.	730	665	52	38	112	63	North Yorks
31.	448	342	67	80	131	109	Nottinghamshire
32.	916	1953	98	149	119	110	South Wales
33.	920	831	121	95	59	62	South Yorks
34.	560	505	76	69	135	127	Staffordshire
35.	751	638	35	68	70	59	Suffolk
36.	1666	1188	98	43	136	96	Surrey
37.	844	741	95	71	123	122	Sussex
38.	3798	3856	137	131	139	184	Thames Valley
39.	304	205	37	24	65	53	Warwickshire
40.	948	1184	65	98	133	103	West Mercia
41.	1500	1160	246	205	344	278	West Midlands
42.	1215	1638	180	188	190	223	West Yorks
43.	1436	919	78	40	113	41	Wiltshire