P.M.E.Altham **Lent 2005**

STATISTICAL MODELLING

Example Sheet 2

1. If Y_i are independent Poisson, means $\exp \beta^T x_i$, $1 \leq i \leq n$, how would you evaluate $\hat{\beta}$ and its asymptotic covariance matrix? What does 'scale parameter taken as 1.000' mean in the corresponding glm output?

2 * If the loglikelihood can be written

$$\ell(\beta) = (\beta^T t - \psi(\beta))/\phi \qquad \text{where } \phi > 0,$$

and $t = t(\mathbf{y})$ is a p-dimensional vector, show that the covariance matrix of $t(\mathbf{y})$ is

$$\phi \left(\frac{+\partial^2 \psi}{\partial \beta \partial \beta^T} \right)$$

and hence that $\ell(\beta)$ is a strictly concave function of β. What is the practical application of this result in estimation of β? Illustrate your answer for either the binomial or the Poisson distribution.

3. We can say that an asymptotic 90% confidence region for $\hat{\beta}$ is derived from

$$(\hat{\beta} - \beta)^T (V(\hat{\beta}))^{-1} (\hat{\beta} - \beta) \sim \chi^2_p \quad \text{(approx)}.$$

Show that if $p = 2$, the resulting region is an ellipse, and finds its equation in the case where

$$\frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_2} \equiv 0\,.$$

Why is the remark $\int_c^\infty e^{-x} dx = e^{-c}$ relevant in this context?

4. In the least-squares fit of the model

$$y_{ij} = \mu + \theta_i + \varepsilon_{ij}, \qquad 1 \leq j \leq n_i, 1 \leq i \leq t$$

with $\theta_1 = 0$ (i.e. the glm constraint), and the usual assumption that ε_{ij} are $NID(0, \sigma^2)$, show that

$$\hat{\mu} = y_{1+}/n_1$$
$$\hat{\theta}_i = -y_{1+}/n_1 + y_{i+}/n_i \qquad (i \neq 1)$$

and

$$\text{var}\,(\hat{\theta}_i) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_i} \right) \qquad (i \neq 1)$$

Find $\text{cov}\,(\hat{\theta}_i, \hat{\theta}_\ell)$ for $i \neq \ell$, $(i, \ell \neq 1)$. (Here y_{i+} is defined as $\sum_{j=1}^{n_i} y_{ij}$). Show that the 'fitted value of y_{ij}' is y_{i+}/n_i, for $i = 1, \ldots, t$.
Hint: it's easier to do the algebra with the 'sum to zero' constraint, and then transform back to the 'corner-point' constraint'.

5. In the least-squares fit of the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},\ 1 \leq k \leq r, 1 \leq i \leq I, 1 \leq j \leq J$$

with $\alpha_1 = 0$, $\beta_1 = 0$, and the usual assumption that ε_{ijk} are $NID(0, \sigma^2)$, show that, for $i \neq 1$,

$$\hat{\alpha}_i = \frac{1}{Jr} (\sum_{j,k} y_{ijk} - \sum_{j,k} y_{1jk}),$$

with the corresponding expression for $\hat{\beta}_j$. How is your answer affected if the condition $1 \le k \le r$ is replaced by the condition $1 \le k \le r_{ij}$?

6. Why would you expect, in an experiment with IJ observations, that the model

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij},$$

where

$$\varepsilon_{ij} \sim NID(0, \sigma^2), \qquad 1 \le i \le I, 1 \le j \le J$$

gives a perfect fit to the data (i.e. zero deviance)?

7. In the model

$$\begin{array}{ccccccc} & Y & = & X & \beta & + & \varepsilon \\ & \swarrow & & \downarrow & \searrow & & \searrow \\ n \times 1 & & & & p \times 1 & & n \times 1 \\ & & & n \times p & & & \end{array} \qquad \varepsilon_i \sim NID(0, \sigma^2)$$

explain why displaying the estimated covariance matrix of $\hat{\beta}$ is a method of finding out about $(X^T X)^{-1}$.

8. The data-set below is taken from the Minitab student Handbook(1976) by Ryan,T., Joiner, B. and Ryan, B. and is also discussed in the book by Aitkin et al. (See Chapter 3 of the book.) For the 31 cherry trees, the table below shows d, h, v. These are defined by d is the diameter(in inches), at a height of 4.5 feet from the ground, h is the height(in feet) of the trees, v is the volume of useable wood, in cubic feet.
Reminder: 1 foot= 12 inches.
This is one of the datasets already in R: try
data(trees); attach(trees); trees[1,] (but you need to take 'Girth' as d, which is confusing, I know).
The order is (d, h, v).
8.3 70 10.3, 8.6 65 10.3, 8.8 63 10.2, 10.5 72 16.4,
10.7 81 18.8, 10.8 83 19.7, 11.0 66 15.6, 11.0 75 18.2,
11.1 80 22.6, 11.2 75 19.9, 11.3 79 24.2, 11.4 76 21.0,
11.4 76 21.4, 11.7 69 21.3, 12.0 75 19.1, 12.9 74 22.2,
12.9 85 33.8, 13.3 86 27.4, 13.7 71 25.7, 13.8 64 24.9,
14.0 78 34.5, 14.2 80 31.7, 14.5 74 36.3, 16.0 72 38.3,
16.3 77 42.6, 17.3 81 55.4, 17.5 82 55.7, 17.9 80 58.3,
18.0 80 51.5, 18.0 80 51.0, 20.6 87 77.0.
Take $lv = log(v)$, with ld, lh defined similarly.
Verify that the following models give the estimates(with se's) and deviances below, and discuss the fit of these models.

$$M_0 : E(lv) = \mu$$

for which

$$\hat{\mu} = 3.273(0.0945)$$

and deviance = 8.3087(df=30);

$$M_1 : E(lv) = \mu + \beta\ ld$$

for which

$$\hat{\mu} = -2.353(0.2307), \hat{\beta} = 2.200(0.0898)$$

and deviance= 0.38324(df=29);

$$M_2 : E(lv) = \mu + \beta\ ld + \gamma\ lh$$

for which

$$\hat{\mu} = -6.632(0.7998), \hat{\beta} = 1.983(0.0750), \hat{\gamma} = 1.117(0.2044)$$

and deviance= 0.18546(df=28);

$$M_3 : E(lv) = \mu + \gamma\, lh$$

for which

$$\hat{\mu} = -13.96(3.755), \hat{\gamma} = 3.982(0.8677)$$

and deviance=4.8130(df=29).

What are the numerical consequences of the non-orthogonality of the parameters β, γ?

The volume of a cylinder of length ℓ, diameter d, is $(\pi d^2 \ell)/4$, and the volume of a cone of height ℓ and base diameter d is $(\pi d^2 \ell)/12$. Are these cherry trees more like cylinders than cones?

9. ' The Independent' (21/12/88) gave the 'League Table of football-related arrests', printed in the table below. This list details a total of 6147 football-related arrests in the 1987-8 season, and is compiled by the Association of Chief Constables. It does not differentiate between Home-fans and Away-fans. There are 4 Divisions (with Division 1 containing the best clubs) and these have 21, 23, 24 and 24 clubs respectively, each Division corresponding to a pair of columns in the Table below.
The columns below give (a,n) where a= attendance, in thousands, and n= number of arrests, for each of the 4 soccer divisions, with the order (reading across the rows) being (a,n) for Division 1, (a,n) for Division 2, (a,n) for Division 3, (a,n) for Division 4.
YOU ARE NOT INTENDED TO TYPE IN THIS DATA: ASK ME TO EMAIL THE SET TO YOU.

325	282	404	308	116	99	71	145
409	271	286	197	401	80	227	132
291	208	443	184	105	72	145	90
350	194	169	149	77	66	56	83
598	153	222	132	63	62	77	53
420	149	150	126	145	50	74	46
396	149	321	110	84	47	102	43
385	130	189	101	128	47	39	38
219	105	258	99	71	39	40	35
266	91	223	81	97	36	45	32
396	90	211	79	205	34	53	29
343	86	215	78	106	32	51	28
518	74	108	68	43	28	51	27
160	49	210	67	59	22	115	21
291	43	224	60	88	22	52	21
783	38	211	57	226	21	67	21
792	33	168	55	61	21	52	17
314	32	185	44	91	21	52	17
556	24	158	38	140	20	72	15
174	14	429	35	85	18	49	12
162	1	226	29	127	11	101	10
NA	NA	150	20	59	5	90	8
NA	NA	148	19	87	4	50	5
NA	NA	NA	NA	79	3	41	0

Let (n_{ij}, a_{ij}) be the observations for the ith club of the jth Division, for $j = 1, ..., 4$. Making the standard assumption that the errors are $NID(0, \sigma^2)$ consider how to fit the following models in Splus, and sketch graphs to show what these models represent:
a) $E(n_{ij}) = \mu$, b) $E(n_{ij}) = \mu + \alpha\, a_{ij}$, c) $E(n_{ij}) = \mu + \beta_j + \alpha\, a_{ij}$, d) $E(n_{ij}) = \mu + \beta_j + \alpha_j\, a_{ij}$.
You will find that the deviances for these 4 models are, respectively 371056 (91 df), 296672 (90 df), 272973 (87 df), 241225 (84 df).
Now plot n against a. What do you conclude about your assumption of constant error variances?
Now repeat the model-fitting exercise with n, a replaced by $log(n), log(a)$. Can you now think of a way of identifying certain clubs as 'rogues', or indeed as 'saints' within their Division?

10. The data-set below which is also discussed by Agresti (1995, p101) is based on a study of British doctors by R.Doll and A.B.Hill(1966) and gives the number of coronary deaths for smokers and non-smokers, for each of 5 different age-groups, with the corresponding 'person-years', ie the total time at risk. Thus for example, in the youngest age-group, in the non-smoking category, the total time at risk was 18793 years, and during this time there were 2 coronary deaths in this particular class. The age-groups are 35-44, 45-54, 55-64, 65-74, 75-84 years.
Define d_{ij} as the number of deaths in age group i, smoking group j, where $i = 1, ..., 5, j = 1, 2$ with $j = 1$ for non-smokers, $j = 2$ for smokers. Assume that (d_{ij}) are distributed as independent Poisson variables, with $E(d_{ij}) = \mu_{ij}$ and $\mu_{ij} = \theta_{ij} p_{ij}$ where p_{ij} = total person-years at risk, for age i, smoking group j. Hence $log(\mu_{ij}) = log(\theta_{ij}) + log(p_{ij})$ for all i, j. This is why we take $log(p_{ij})$ as the 'offset' in the glm analysis below. We model $log(\theta_{ij})$, the parameter of interest.
Verify and interpret the results from the models fitted below.

```
pyears < − scan()
18793 52407
10673 43248
5710 28612
2585 12663
1462 5317
BLANK LINE
deaths < − scan()
2 32
12 104
28 206
28 186
31 102
BLANK LINE
sm < −rep(c(1,2),times=5)
age < − c(1,1,2,2,3,3,4,4,5,5)
(these are crude but easy-to-understand ways of setting up the factor levels)
sm < −factor(sm) ; age < − factor(age)
prop< −deaths/pyears ; tapply(prop,list(age,sm),mean)
l< −log(pyears)
summary(glm(deaths ~ age + offset(l),poisson),cor=F)
summary(glm(deaths ~ age + sm + offset(l),poisson),cor=F)
```

It should now be obvious to you that (i) smoking is bad for you and (ii) so is getting old. But you will also see that this final model does not fit well (its deviance is 12.134, with 4 df). Can you suggest a way of improving the fit?

11. With $y_1, \ldots, y_n$ independent observations, and y_i having pdf from the usual glm family, with $g(\mu_i) = \beta^T x_i$ as usual, find the expectation of the second derivative of the log-likelihod function with respect to β, ϕ, and hence show that $\hat{\beta}, \hat{\phi}$ are asymptotically independent.

12. New for 2005: introduction to the inverse Gaussian distribution.
Consider the density

$$f(y|\theta, \phi) = exp[(y\theta - b(\theta))/\phi + c(y, \phi)], \quad y > 0,$$

where

$$\phi = \sigma^2, b(\theta) = -(-2\theta)^{1/2} \quad ,$$

and

$$-2c(y, \phi) = log(2\pi\phi y^3) + (1/\phi y).$$

Show that $E(Y) = (-2\theta)^{-1/2} = \mu$, say. Check that $1/\mu^2$ is the canonical link function for this glm, and that $var(Y) = \mu^3\sigma^2$. **nb, no fancy integration required, at all.**