Statistical Modelling, Lent 2005, Solutions to Ex 2

P.M.E.Altham, Statistical Laboratory, University of Cambridge.

February 20, 2007

1. Here Y_i are distributed as independent $Po(\mu_i)$, with $\log(\mu_i) = \beta^T x_i$. Thus

$$f(y_i|\beta) \propto (\exp -\mu_i)\ (\mu_i)^{y_i}$$

so that

$$\log f(y|\beta) = \sum_1^n [-\exp(\beta^T x_i) + y_i \beta^T x_i] + constant$$

ie $L_n = -\sum_1^n [\exp(\beta^T x_i)] + \beta^T \sum_1^n y_i x_i + constant.$
Thus

$$\frac{\partial}{\partial \beta} L_n = \sum_i^n [-x_i \exp(\beta^T x_i) + y_i x_i]$$

and minus the matrix of second derivatives of L_n is say J, where

$$J = +\sum_1^n x_i x_i^T \exp(\beta^T x_i).$$

This is $+\sum_1^n x_i x_i^T \mu_i$, where $\mu_i > 0$.
Hence J is a positive definite matrix, and so $\hat{\beta}$ (the mle) is the solution of

$$\frac{\partial}{\partial \beta} L_n = 0,$$

which we may write as

$$\sum x_i \mu_i = \sum x_i y_i$$

ie the observed and the expected values of the sufficient statistics $\sum x_i y_i$ agree exactly at $\beta = \hat{\beta}$.
The equation $\frac{\partial}{\partial \beta} L_n = 0$ has to be solved iteratively. The asymptotic covariance matrix of $\hat{\beta}$ is the inverse of $E(J)$, which here is

$$\sum_1^n \mu_i x_i x_i^T.$$

(There is no simple formula for the inverse.)
"Scale parameter taken as 1.000 " means that in writing the pdf in glim formulation, ie as

$$f(y_i|\theta_i, \phi) = \exp\left[(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)\right]$$

then we take ϕ as 1.

2. The loglikelihood is

$$l(\beta) = (\beta^T t - \psi(\beta))/\phi \text{ where } \phi > 0.$$

Thus

$$\frac{\partial}{\partial \beta} l(\beta) = [t - \frac{\partial}{\partial \beta}\psi(\beta)]/\phi$$

and$E[\frac{\partial}{\partial\beta}l] = 0$ (as always) so that $E(t) = \frac{\partial}{\partial\beta}\psi(\beta)$. Further the matrix of 2nd derivatives of ℓ is - ϕ^{-1} times the matrix of 2nd derivatives of ψ and, using

$$\int f(y|\beta)dy = 1 \; for \; all \; \beta$$

again, we have

$$E(-matrix \; of \; 2nd \; derivatives \; of \; l) = E(UU^T), \; where \; U = \frac{\partial}{\partial\beta}l$$

Thus covariance matrix of t(y) is

$$\phi \times \; matrix \; of \; 2nd \; derivatives \; of \; \psi$$

which must therefore be a positive definite matrix, so that $l(\beta)$ is a STRICTLY CONCAVE function. Thus any solution $\hat{\beta}$ say, of$\frac{\partial}{\partial\beta}l = 0$ must be THE MAXIMUM of $l(\beta)$. Thus, eg if we solve $U = 0$ by the Newton-Raphson algorithm, we will unerringly home in on the right solution, & not something nasty & irrelevant like a local minimum.

3. We know (using $\sim$ to mean "approx'ly dist'd as")

$$\hat{\beta} \sim N(\beta, V(\beta))$$

and so

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

and hence

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

Hence from χ^2 tables we can choose c, given α, such that

$$Pr[(\hat{\beta} - \beta)^T (V(\hat{\beta}))^{-1} (\hat{\beta} - \beta) \leq c] \simeq 1 - \alpha.$$

Write $V = V(\hat{\beta}))$ for short; our $(1 - \alpha)$ confidence region is

$$(\hat{\beta} - \beta) V^{-1} (\hat{\beta} - \beta) \leq c$$

which (because V^{-1} is pos def) is an ELLIPSE, centred on $\hat{\beta}$. For

$$\frac{\partial^2}{\partial\beta_1 \partial\beta_2}\ell = 0$$

we have V a diagonal matrix, with diagonal entries v_1, v_2 say and the ellipse is

$$(\beta_1 - \hat{\beta}_1)^2/v_1 + (\beta_2 - \hat{\beta}_2)^2/v_2 = c.$$

The relevance of the final remark is that χ^2_p has pdf $(1/2)exp(-x/2)$ for $p = 2$.

4. Given

$$y_{ij} = \mu + \theta_i + \epsilon_{ij} \; ,$$

for $i = 1 \ldots t, \; j = 1 \ldots n_i$ with $\theta_1 = 0$
equivalently

$$y_{ij} = \mu_i + \epsilon_{ij}$$

with $\mu_1 = \mu$, and $\mu_i = \mu + \theta_i$ for $i > 1$. We see that

$$\Sigma\Sigma(y_{ij} - \mu - \theta_i)^2$$

is minimised, equivalently $\Sigma\Sigma(y_{ij}-\mu_i)^2$ is minimised w.r.t. (μ_i) by

$$\hat{\mu}_i = \frac{y_{i+}}{n_i}.$$

giving

$$\hat{\mu} = \frac{y_{1+}}{n_1} \; and$$

$$\hat{\theta}_i = -\frac{y_{1+}}{n_1} + \frac{y_{i+}}{n_i} \; for \; i > 1,$$

as req'd. Clearly, since ϵ_{ij} are NID(0,σ^2),

$$var(\frac{y_{i+}}{n_i}) = \frac{\sigma^2}{n_i}$$

and

$$cov(y_{1+}, y_{i+}) = 0 \; for \; i > 1$$

hence $var(\hat{\theta}_i)$ is as given, and

$$cov(\hat{\theta}_i, \hat{\theta}_l) = var(\frac{y_{1+}}{n_1}) \; for \; i \neq l.$$

The fitted value of y_{ij} is $\hat{\mu}_i$, ie $\frac{y_{i+}}{n_i}$

5.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, k = 1\ldots r, \; i = 1\ldots I, j = 1, \ldots J.$$

We can reparametrise this as

$$y_{ijk} = m + a_i + b_j + \epsilon_{ijk} \; ,$$

with $\Sigma a_i = 0, \Sigma b_j = 0$. Then $\mu = m + a_1 + b_1$, $\mu + \alpha_i = m + a_i + b_1$, $\mu + \beta_j = m + a_1 + b_j$, so that $\alpha_i = a_i - a_1$ *, and* $\beta_j = b_j - b_1$.
Straightforward minimisation of

$$\Sigma\Sigma\Sigma(y_{ijk} - m - a_i - b_j)^2$$

subject to the constraints $\Sigma a_i = 0$, $\Sigma b_j = 0$ gives

$$\hat{m} = \bar{y}, \hat{a}_i = \frac{y_{i++}}{n_i} - \bar{y} \; , etc.$$

Hence, returning to the original model, we see that

$$\hat{\alpha}_i = \frac{y_{i++}}{n_i} - \frac{y_{1++}}{n_1} \; .$$

If now $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$, with

$$k = 1\ldots r_{ij}, \; i = 1, \ldots, I, \; j = 1, \ldots, J$$

then we find that $\hat{\alpha}_i$ is no longer given by the above simple formula .
This is best seen by the following simple example. Suppose the observations (y_{ijk}) are
23.9 , 7.2 for $(i, j) = (1, 1)$,
3.6 for $(i, j) = (1, 2)$,
10.4 for $(i, j) = (2, 1)$,
29.7 for $(i, j) = (2, 2)$.
Note that you will different estimates for, say α_i, depending on whether or not β_j is in the model, because the design is unbalanced.
6. Here the model is

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$$

for $i = 1\ldots I, j = 1\ldots J$, which is equivalent to

$$y_{ij} = \mu_{ij} + \epsilon_{ij},$$

and the lse(=mle) of (μ_{ij}) are obtained by minimising

$$\Sigma\Sigma(y_{ij} - \mu_{ij})^2 \text{ wrt } \mu_{ij}.$$

This gives $\mu_{ij} = y_{ij}$ so that the minimised residual sum of squares is 0, trivially.
Further, we have no degrees of freedom left to estimate σ^2. This is called a "saturated" model: it is saturated with parameters.
7.

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

with ϵ distributed as N(0,$\sigma^2 I$), gives

$$(Y - X\beta)^T(Y - X\beta)$$

is min'd by $\hat{\beta}$ such that $X^T X\hat{\beta} = X^T Y$

$$ie\ \hat{\beta} = (X^T X)^{-1}X^T Y.$$

Hence $\hat{\beta}$ is distributed as $N(0, \sigma^2(X^T X)^{-1})$.
If we display the the estimated covariance matrix of $\hat{\beta}$ it is of course

$$s^2(X^T X)^{-1}$$

where $s^2 = deviance/df$. Since s^2 is therefore known, we can evaluate $(X^T X)^{-1}$.
8. Using an obvious notation, you need to try

```
lm(lv ~ 1) ; lm(lv ~ lh) ; lm(lv ~ ld) ; lm(lv~lh + ld)
```

Observe that we never expected M_0 to fit: we always start by fitting it as our "baseline". Observe also that M_1 fits much better than M_0, as we would expect, and that $\hat{\beta}$ is clearly significant. (The ratio 2.2/.0898 is much greater than 2).
Observe that the estimate of β changes between M_1 and M_2 because β and γ are not mutually orthogonal. [There is a non-zero correlation between the vectors (ld),(lh)].
The coefficients $\hat{\beta}$ and$\hat{\gamma}$ in M_2 are clearly both significant.
Observe: the reduction in deviance in moving from M_0 to M_1 the "ss due to γ"
is different from the reduction in deviance in moving from M_1 to M_2 , the "ss due to γ, allowing for β ".
This is another consequence of the non-orthogonality of these 2 parameters. Our final model is thus

$$lv_i = -6.632(.7998) + 1.117(.2044)\ lh_i +\ 1.983(.07501)\ ld_i$$

Observe that both a cylinder & a cone would give

$$lv_i = C + 1\ log(h_i) + 2\ log(d_i/12)$$

where $C = log(\pi/4)$ for a cylinder, $log(\pi/12)$ for a cone. [warning : d_i is actually measured in inches. Do today's metric students know about feet and inches?]
The rest of the solution is up to you. Not surprisingly, trees do turn out to be more like cones than cylinders.
9. Note that the linear regression of n, the number of arrests on a, the attendance, although having a significant positive slope (as we would expect) is really rather a poor fit: it has R^2, the proportion of the total deviance "explained by" the regression as only 0.2005. (By definition, R^2 lies between 0 and 1, with R^2=1 for a perfect fit.)
If we plot the $residuals = (observed - fitted\ values)$ against the corresponding $fitted\ values$, we get a very pronounced 'fanning-out' effect, suggesting that $var(n_i)$ increases as $E(n_i)$ increases, which is what we would expect, since n_i might well be Poisson-like in its behaviour. Thus the analysis presented, which fails to START by doing some simple plots, is not very sensible. Its big drawback is that it relies on the assumption that $var(n_i)$ is constant over i.
This is what the analysis does. First fit

$$n_{ij} = \mu + \epsilon_{ij},\ for\ j = 1\ldots 4, i = 1\ldots n_j\ .$$

with the usual assumption that ϵ_{ij} is $NID(0, \sigma^2)$.
Then fit

$$n_{ij} = \mu + \alpha\ a_{ij} + \epsilon_{ij}$$

ie the same line for all 4 divisions. Then fit

$$n_{ij} = \mu + \beta_j + \alpha\ a_{ij} + \epsilon_{ij}$$

ie parallel lines for all 4 divisions. Lastly try

$$n_{ij} = \mu + \beta_j + \alpha_j\ a_{ij} + \epsilon_{ij}.$$

There isn't any appreciable improvement in fit after the second model.
And, as we have already said, the assumption of homogeneity of error variance looks very implausible.
However, if we try

$$log(n_{ij}) = \mu + \alpha\ log(a_{ij}) + \epsilon_{ij}$$

it does now seem reasonable to assume $var(\epsilon_{ij})$ is constant. It seems that the same line will fit all 4 divisions, namely

$$log(n_{ij}) = -0.01365(.6475) +\ 0.7506(.1285) log(a_{ij})$$

which has $R^2 = .2748$ (just a little better than before). Indeed the constant can be dropped, to show that a straight line through the origin fits quite well.
The question about 'rogues and saints' clearly relates to picking out those clubs with high (negative or positive) standardised residuals.
10. Let p_{ij} ='pyears' = person-years at risk, and let d_{ij}= the number of deaths,
for i=1, …, 5, (i corresponds to age), $j = 1$(for non-smoker), $j = 2$ (for smoker).
We assume d_{ij} is distributed as indep. Po(μ_{ij}), with the default link for μ, which is log( ). Further, we take $\mu_{ij} = \theta_{ij}\ p_{ij}$, so that

$$log(\mu_{ij}) = log(\theta_{ij}) + log(p_{ij}).$$

This is the interpretation of $log(p_{ij})$ as the *offset*. We try various models for $log(\theta_{ij})$.
The table suggests that θ_{ij} increases with i, for each fixed j.
The first model fitted is in effect

$$log(\theta_{ij}) = \mu +\ \alpha_i, for\ i = 1 \ldots 5, j = 1, 2$$

with α_1 =0. This model includes an age effect but not a smoking effect. This model has deviance= 23.99($df = 5$), but the fit is not good enough.
But clearly this model shows that $\hat{\alpha}_i$ increases with i: this makes sense.
Next we try the model

$$log(\theta_{ij}) = \mu + \alpha_i + \beta_j.$$

In this model, both smoking and age effects are present, but they operate additively (thus the difference between smokers and non-smokers is constant over all 5 ages). The fit is now much improved, and shows a clear effect of smoking($\hat{\beta}_2/(its\ se) > 0$), but the fit is still not acceptable (refer 12.13 to ${\chi^2}_4$).
The model corresponding to

```
glm(deaths ~ age:sm + offset(l), poisson)
```

would give a perfect fit (since it is the saturated model) so therefore to improve the fit, we include an interaction term in a 'weaker' way as our final model. We have
$log(\theta_{ij}) = \mu + \alpha_i$ for sm =1, ie non-smoker,
$log(\theta_{ij}) = \mu + \alpha_i + \beta_j + \gamma\ i$ for sm=2, ie smoker.
This model fits very well (compare the deviance of 1.54 with ${\chi^2}_3$). It shows that although smoking is bad for you (compute 1.445/.3729), there is an interaction between smoking and age: the discrepancy between the mortality rates for smokers and non-smokers DECLINES with age.
11. Write down $\ell(\beta, \phi)$, and find

$$\frac{\partial^2 \ell}{\partial\beta\partial\phi}.$$

It is easy to show that this has expectation zero.
12. Use the glm facts that $E(Y) = b'(\theta) = \mu$ and $var(Y) = \phi b''(\theta)$.