Solutions to "Statistical Models", May 30, 2012

Problem 1

(1a) All model forms correspond to generalized linear models, based on the Poisson, thus having a log link as canonical link. From the link function point of view the first three forms are preferable to the last one, which corresponds to a noncanonical link. The only little difference between the second and third forms is in the interpretation of γ, where this parameter in the third form corresponds to $\log\gamma$ in the second form. The first choice differs from the third only in the quantitative meaning of x_{ij}. Writing $x_{ij} = \exp(\log x_{ij})$, we see that $\log x_{ij}$ in the first form corresponds to x_{ij} in the second and third forms. So for both statistical and notational simplicity, we choose the third form. The resulting model has canonical link and linear predictor variables (canonical statistic) with the row sums and column sums $\{y_{i\cdot}\}$ and $\{y_{\cdot j}\}$ as components together with $\sum x_{ij}y_{ij}$.

(1b) By exponential family theory it is just to put the canonical statistic equal to its mean value:

$$\begin{aligned} y_{i\cdot} &= \sum_j \alpha_i\,\beta_j\,e^{\gamma\,x_{ij}} = \alpha_i \sum_j \beta_j\,e^{\gamma\,x_{ij}}\,, \\ y_{\cdot j} &= \sum_i \alpha_i\,\beta_j\,e^{\gamma\,x_{ij}} = \beta_j \sum_i \alpha_i\,e^{\gamma\,x_{ij}}\,,\,, \\ \sum_{ij} x_{ij}y_{ij} &= \sum_{ij} \alpha_i\,\beta_j\,e^{\gamma\,x_{ij}}\,, \end{aligned}$$

cf. the multiplicative Poisson example on page 34, of which the present model is an extension.

Since $\sum_i y_{i\cdot} = \sum_j y_{\cdot j}$, there is a linear dependence between the components of the canonical statistic that reduces the effective parameter dimension. A simple way to deal with this, and get a full family, is to specify the value of one of the α_is, naturally to give it the value 1. The corresponding equation is then deleted from the equation system, cf. Example 2.5 cont'd on page 34.

(1c) The test asked for is the same as the test of H_0: $\gamma = 0$. The only large-sample test statistic that does not involve the MLE of γ is the score test statistic. Since $\dim\gamma = 1$, this test statistic is approximately $\chi^2(1)$ distributed. What we need to be able to write down the statistic, see formula (5.7) page 59 or (A.37) page 213, is first v and its expected value, where $v = \sum_{ij} x_{ij}y_{ij}$. This expected value was given in the likelihood equations above, but now we need it only under H_0, so the last factor $e^{\gamma\,x_{ij}}$ is replaced by the factor $e^0 = 1$. We also need the covariance matrix of all components, to be first inverted, and next its v-part to be picked out. Except for covariances involving $v = \sum_{ij} x_{ij}y_{ij}$, the elements of the covariance matrix are given on

page 34, Additionally needed is:

$$Var(v) = \sum_{ij} Var(y_{ij})\, x_{ij}^2 \;\; = \sum_{ij} \alpha_i\, \beta_j\, x_{ij}^2\,,$$

and

$$Cov(y_{i.}, v) = \sum_{j} Cov(y_{ij}, x_{ij} y_{ij}) = \alpha_i \sum_{j} \beta_j x_{ij}\,,$$

and similarly for column sums instead of row sums. For α_i and β_j we insert the explicit MLEs given in formula (3.22).

Problem 2.

(2a, part 1) H_0 is a model reduction from $(y_1,\, y_2)$ to $y_1 + y_2$, and the latter is then binomial of length $n_1 + n_2$;. As nuisance parameter, that is a parameter under H_0, we may take the probability p_2 of the second binomial, or equivalently its logit θ_2, say (since $\theta_1 = \theta_2$ under H_0). The MLE of the former under H_0 becomes $p_2 = (y_1 + y_2)/(n_1 + n_2)$.

(2a, part 2) The exact test statistic is for example $v = y_1$, and its conditional distribution under H_0 given $u = y_1 + y_2$, is hypergeometric. This is seen by writing down the conditional probability, $f(y_1)f(y_2)/f(y_1 + y_2)$, or by referring to the test being a standard test for homogeneity of two binomials, where data may be set up as a 2×2 table, see Lecture Notes (2012) Ex. 5.3, page 57 (r_0 and r_1 corresponding to n_1 and n_2)

(2b, part 1) Here $H_0 : \theta_1 = \sqrt{2}\,\theta_2$ is a model reduction from $(y_1,\, y_2)$ to $u = \sqrt{2}\, y_1 + y_2$, as seen by writing the exponent as

$$\theta_1\, y_1 + \theta_2\, y_2 = (\theta_1 - \sqrt{2}\,\theta_2)\, y_1 + \theta_2\, (\sqrt{2}\, y_1 + y_2)$$

As nuisance parameter, that is a parameter under H_0, we may still take the probability p_2 of the second binomial, or its logit θ_2. The likelihood equation under H_0 is obtained by putting $(\sqrt{2}\, y_1 + y_2)$ equal to its expected value, which is now $\sqrt{2}\, n_1 p_1 + n_2 p_2$ where p_1 can be expressed in terms of p_2 by

$$p_1 = \frac{e^{\theta_1}}{1 + e^{\theta_1}} = \frac{e^{\sqrt{2}\theta_2}}{1 + e^{\sqrt{2}\theta_2}} = \frac{p_2^{\sqrt{2}}}{(1 - p_2)^{\sqrt{2}} + p_2^{\sqrt{2}}}$$

.

(2b, part 2) The problem is that for given $u = \sqrt{2}\, y_1 + y_2$, because $\sqrt{2}$ is an irrational number, there cannot be more than (at most) one pair $(y_1,\, y_2)$ that has a specified value of u, and that must be the observed $(y_1,\, y_2)$. Thus, there is no other $v = y_1$ than the observed v that can have the observed u-value, so the conditional distribution of v is degenerate, concentrated in one single point. In other words, the exact test is meaningless in this situation, because it demands a too severe conditioning.