

**Fixed Effects models for Dilution Series data via EM**

Murray Jorgensen

Department of Statistics, University of Waikato, Hamilton, New Zealand

The statistical analysis of dilution series data is has quite a long history. In his clear introduction to the topic Cochran [1] gives McCrady [2] as his earliest reference. The microbiological method named by McCrady [2] as the method of “Most Probable Number” can be seen anachronistically as an application of Maximum Likelihood Estimation. In many data sets where microbiological variables have been measured using dilution series it is common to find that the original tube results are unavailable and only the calculated MPNs remain, or in cases where no MPN exists, an upper or lower bound on the microbial density (count/volume) may be stated. This is unfortunate because the precision of the MPNs varies with the tube results pattern. A rough analysis may be carried out by ignoring issues of weighting and by imputing fixed low or high values to the censored microbial log densities. When the tube results are available we can do better. We can treat the unknown true microbial log-densities for each observation as “missing data” for an application of the EM algorithm. A model can be built for the true log-densities and linked to the observed tube results using the complementary log-log GLM. We will illustrate this approach using a linear model for the unknown true log-densities.

Suppose that we have data for  $I$  sites, there being  $J_i$  observations on site  $i$ ,  $i = 1, 2, \dots, I$ . Denote by  $\xi_{ij}$  the unknown log-density on the  $j$ th visit to site  $i$ ,  $j = 1, 2, \dots, J_i$ . Suppose that  $X$  is a design matrix for covariate observations and factors.

Now consider the model

$$\xi_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij}$$

where  $\mathbf{x}_{ij}^T$  is the  $ij$  row of the  $n \times p$  design matrix  $X$  and  $(n = \sum J_i)$ .

We will show the details of the EM algorithm [3] for models of this kind and how the information matrix may be calculated using the method of Oakes [4].

Finally we apply the method to the analysis of a large survey of *Campylobacter* in New Zealand streams.

**References**

- [1] W. G. Cochran. Estimation of bacterial densities by means of the “Most Probable Number”. *Biometrics*, 6:105–106, 1950.
- [2] M. H. McCrady. The numerical interpretation of fermentation-tube results. *J. Infec. Dis.*, 17: 183–212, 1917.
- [3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B*, 39:1–38, 1977.
- [4] D. Oakes. Direct calculation of the information matrix via the EM algorithm. *J. Roy. Statist. Soc. B*, 61:479–482, 1999.