ESTIMATION OF STANDARD ERRORS IN NON-LINEAR REGRESSION MODELS: SPATIAL VARIATION IN RISK AROUND PUTATIVE SOURCES.

Ramis, Rebeca[1,2,3]; Diggle, Peter[1]; López-Abente, Gonzalo[2,3]

[1] *Department of Medicine, Lancaster University, UK.*
[2] *Cancer and Environmental Epidemiology Area, National Centre for Epidemiology. Carlos III Institute of Health, Madrid, Spain.*
[3] *CIBERESP*

Background

We consider the problem of investigating spatial variation in the risk of non-infectious diseases in populations exposed to pollution from one or more point sources.

The data most commonly available to study this question include case-counts (O_i) in each of a set of areas that partition the geographical region of interest, suitable denominators, Ei, proportional to the expected number of cases in each area, and the locations of the relevant point sources, from which we can compute distances d_{ij} between the *j*th focus and a reference location, typically the centroid, within the *I*th area. Also available in most applications are covariates relating to socio-economic status or other risk-factors associated with each area, which we denote by Z_k.

The standard approach to the analysis of data of this kind is a log-linear regression of the case-counts on the covariates, with log-transformed denominators as an offset variable. To model distance-related point source effects, a log-linear formulation is unrealistic because of the need to combine an elevated risk close to the source with a neutral long-distance effect. We therefore extend the model by including a non-linear distance function, $f(d_{ij})$, hence [1,2]:

$$O_i \sim Po(E_i \mu_i)$$

$$\mu_i = \rho \sum_k (\vartheta_k Z_{ik}) \prod_j f(d_{ij}); \qquad f(d_{ij}) = 1 + \alpha_j \exp\left[-(d_{ij} / \beta_j)^2\right]$$

Parameter estimation and standard error calculations

Generic functions available in R to fit non-lineal regression models include the *“gnlm”* library by J. K. Lindsey [3], which in turn uses the *nlm* function of Bates and Pinheiro to estimate the parameters. These functions use a numerical estimate of the Hessian matrix evaluated at the parameter estimate to calculate standard errors.

We have found that, for point source models like the one described above, even when numerically accurate values are returned for the maximum likelihood parameter estimates, the associated standard errors derived by inverting the estimated Hessian can be unreliable. As an alternative strategy, we obtain standard errors by combining an R function for direct maximisation of the likelihood with replicated Monte Carlo simulations of the fitted model.

Results

We have carried out a simulation study to compare the estimators yielded by the two methodologies and to asses the performance of the Hessian and Monte Carlo methods for calculating approximate standard errors. As expected, parameter estimates obtained from the two methods are almost identical. However, standard errors for the non-linear parameters (α_j, β_j) are estimated more reliably by Monte Carlo than by inversion of the estimated Hessian.

REFERECES:

1. Diggle, P., Elliott, P., Morris, S., Shaddick, G., 1997. Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society, Series A* 160, 491:505.
2. Diggle, P., Rowlingson, B., 1994. A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A* 157 (3), 433:440.
3. Lindsey, J. K. *Nonlinear Models in Medical Statistics.* Oxford University Press (2001)