ESTIMATION OF DISEASE RATES IN SMALL AREAS: A NEW MIXED MODEL FOR SPATIAL DEPENDENCE*

BRIAN G. LEROUX †, XINGYE LEI† , AND NORMAN BRESLOW†

Abstract. In this paper, a new model is proposed for spatial dependence that includes separate parameters for overdispersion and the strength of spatial dependence. The new dependence structure is incorporated into a generalized linear mixed model useful for the estimation of disease incidence rates in small geographic regions. The mixed model allows for log-linear covariate adjustment and local smoothing of rates through estimation of the spatially correlated random effects. Computer simulation studies compare the new model with the following sub-models: intrinsic autoregression, an independence model, and a model with no random effects. The major finding was that regression coefficient estimates based on fitting intrinsic autoregression to independent data can have very low precision compared with estimates based on the full model. Additional simulation studies demonstrate that penalized quasi-likelihood (PQL) estimation generally performs very well although the estimates are slightly biased for very small counts.

Key words. Random effect, log-linear model, penalized quasi-likelihood, Gaussian intrinsic auto-regression, generalized linear mixed model, Monte Carlo simulation.

AMS(MOS) subject classifications. Primary 62M40, 62F11, 62J12, 62M30, 92D30.

1. Introduction.

Epidemiologists often use maps to illustrate geographic variation in disease incidence or mortality rates in small regions such as U.S. counties. Producing such maps is complicated by the fact that raw incidence rates are typically unstable because of small incidence counts, and also by the presence of spatial correlation in the rates. Note that spatial dependence may exist in rates of non-infectious diseases such as cancer, possibly because of the presence of environmental risk factors which are themselves spatially correlated.

Statistical models for use in producing stable estimates of rates must be able to accommodate all of the features of the data, including non-normal distributions of count data, the effects of explanatory variables, and spatial correlation. One approach involves the use of generalized linear mixed models (GLMMs) in which a generalized linear model (GLM) is augmented by unobserved normally distributed random effects that explain spatial correlation as well as overdispersion (Clayton and Kaldor, 1987). Conditional on a random effects vector b, the observed incidence counts y_i follow a log-linear GLM with conditional means μ_i given by

$$\log \mu_i = \log E_i + x_i'\alpha + b_i, \tag{1.1}$$

*This research was supported in part by United States Public Health Service Grants CA40644 and CA09168.

†Department of Biostatistics, University of Washington, Seattle WA 98195, USA, E-mail: leroux@biostat.washington.edu

M. E. Halloran et al. (eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*

where x_i is a vector of explanatory variables for the ith region, α is the vector of regression coefficients, and E_i is the expected count, which may be based on the age distribution in the region and a set of standard rates. Possible models for the random effects have been discussed by Besag, York, and Mollié (1991).

In this article, we propose a new spatial dependence model that includes separate parameters for overdispersion and the strength of the spatial dependence. We study the performance of this model as a basis for estimation of parameters and prediction of SMRs in individual regions through computer simulation and compare it to intrinsic autoregression, as well as to an independence model and a model with no random effects. An additional purpose of the article is to examine the performance of the penalized quasi-likelihood (PQL) method of parameter estimation (Breslow and Clayton, 1993).

2. A new model for spatial dependence. Under Gaussian intrinsic autoregression (Besag et al., 1991), b has a (singular) multivariate normal distribution with mean 0 and a covariance matrix D with Moore-Penrose generalized inverse $D^- = R/\sigma^2$, where R is determined by the neighbourhood structure of the regions. The typical element of R is

$$R_{ij} = \begin{cases} n_i, & i = j \\ -I\{i \sim j\}, & i \neq j \end{cases},$$

where n_i is the number of neighbours of region i, $i \sim j$ indicates that regions i and j are neighbours, and I is the indicator function. Typically, neighbours are those regions which share a border, although other ways of defining neighbours may be used instead. Under intrinsic autoregression, the conditional mean of the random effect for any region given all the other random effects is equal to the mean of the random effects for the neighbouring regions, and the conditional variance is inversely proportional to the number of such neighbours (Besag, 1974).

A limitation of intrinsic autoregression is that the parameter σ^2 serves both to represent overdispersion and spatial dependence. A few ways of separating overdispersion and spatial dependence have been suggested. In one approach, a proportionality constant is introduced into the conditional mean, but the form of the conditional variance is unchanged (Cressie, 1991). Besag et al. (1991) note that this model can have unappealing properties when the proportionality constant is close to 0; note that in this model the conditional variance is inversely proportional to the number of neighbours even in the independence case. In Clayton and Kaldor's (1987) model, this problem is avoided by using a constant conditional variance, but the conditional mean then becomes proportional to the sum (rather than the mean) of the neighbours. An alternative approach uses two additive random effect components, one an intrinsic autoregression and the other an independence (white noise) process (Besag et al., 1991).