Analyzing Irish suicide rates with mixture models

Nick Sofroniou[1], Jochen Einbeck[2] and John Hinde[2]

[1] Educational Research Centre, St Patrick's College, Dublin, Ireland
[2] Department of Mathematics, National University of Ireland, Galway, Ireland

Abstract: In the analysis of morbidity and mortality data, variance component models are commonly used to provide an improvement in the estimation of rates for small regions which typically show large variability. This article investigates Irish suicide data using Poisson mixed models. The random effect distributions are estimated using Nonparametric Maximum likelihood which allows the calculation of shrinkage estimates from the posterior probability estimates of the EM algorithm, as well as the construction of 'league tables'. As these models are inefficient in the case of spatial dependency, we investigate the addition of spatial autocorrelation terms based on neighboring average crude rates and standardized mortality ratios, as well as gender-specific versions of these. We consider models for the average crude rate as well as for the relative risk. A close correspondence between fitted values from both types of models suggests that information concerning within-region variability, incorporated in the parameters of the average crude rate model, appears indirectly in the simpler relative risk model by means of the expected values used in the offset term of the latter.

Keywords: Generalized linear models; random effects; nonparametric maximum likelihood; spatial autocorrelation; suicide rates.

1 Introduction

The use of generalized linear models with random effects is already well established in the analysis of morbidity and mortality data. Administrative regions defined in geographical terms have reported counts of cause of death or illness and the aim is to model the variation in these. Calculating separate estimates of risk for each area may result in small regions having estimates with large variability, leading to small-area estimation problems (e.g., Longford, 2005). Variance component models enable the generation of empirical Bayes shrinkage estimates to improve the estimation of local risk (Aitkin, 1996b).

Assume we have a division of some region into m districts with population sizes $n_i, i = 1, \ldots, m$ and counts $Y_i, i = 1, \ldots, m$, and a further division into subpopulations $j = 1, \ldots, J$ (e.g., certain gender/age groups) with explanatory vectors x_{ij}, observed counts Y_{ij}, and sizes n_{ij}, such that

$\sum_j Y_{ij} = Y_i$ and $\sum_j n_{ij} = n_i$. The observed mortality/morbidity counts Y_{ij} are commonly assumed to follow a Poisson distribution with mean μ_{ij}, which can be either specified using the rate λ_{ij}, giving $\mu_{ij} = n_{ij}\lambda_{ij}$, or the relative risk θ_{ij}, implying that $\mu_{ij} = E_{ij}\theta_{ij}$, with E_{ij} being the expected number of cases obtained from some reference population (Ahlbom, 1993). Alternatively, the models can be based on a binomial distribution as in Aitkin (1996b). The Poisson distribution is the more natural choice if the occurrence of the death/disease is a rather rare event.
If the number m of districts is quite high (say, more than six or seven), modelling the regional heterogeneity as a fixed effect would require a quite large number $(m-1)$ of additional model parameters. This can be avoided by using random effects $z_i, i = 1, \ldots, m$. For the two cases mentioned above, we use a log-linear model for the parameter of interest,

$$\left\{ \begin{array}{c} \log(\lambda_{ij}) \\ \log(\theta_{ij}) \end{array} \right\} = \beta' x_{ij} + z_i, \tag{1}$$

yielding the generalized random effect models

$$\log(\mu_{ij}) = \text{offset} + \beta' x_{ij} + z_i, \tag{2}$$

with offsets $\log(n_{ij})$ or $\log(E_{ij})$, respectively. One observes from (2) that both families of models actually only differ by the offset, and hence can be represented within a larger family where the two offsets are present in the linear predictor, each multiplied by an indicator variable to select the relevant offset. Thus, although the autocorrelation terms and offsets being compared may differ, the likelihoods will still be on the same scale and they can be compared with each other using their disparities (i.e., $-2\log L$, with L being the likelihood).
In this article we analyze Irish suicide data using both models specified in (1). The rates λ_{ij} correspond in our setting to death rates due to "suicide or intentional self-harm". The tool used for analysis of the data is the Nonparametric Maximum Likelihood (Aitkin, 1996a). The paper can be seen on the one hand as an methodological addition to the findings in Aitkin(1996b), as we extend the idea of empirical Bayes shrinkage to situations where one or more covariates are present, and on the other hand as a complement to the models introduced in Biggeri et al. (2000), as we explore the ample ground between modelling a *random* spatial autocorrelation term and not modelling spatial autocorrelation at all.

2 Irish suicide data

The data considered here describe the mortality due to suicide and intentional self-harm in the Republic of Ireland from 1989–1998, obtained from the All Ireland Mortality Database (Institute of Public Health in Ireland,

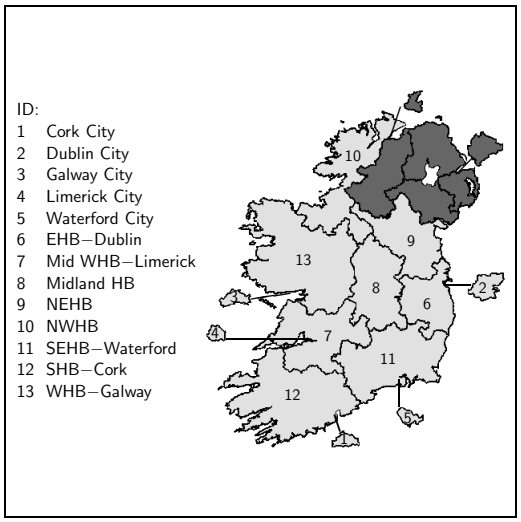

FIGURE 1. *Map of Health Boards and Cities for the Republic of Ireland. The excluded regions of Northern Ireland are shown in dark grey. The '−' sign indicates that a city is excluded from its health board.*

2005). This database divides the Republic of Ireland into 13 'health regions' (the 8 former health boards which existed during this period, and the cities Cork, Dublin, Galway, Limerick, and Waterford extracted from these health boards; see Fig. 1). The data are graphically displayed in Fig. 2 (left) and are part of the R package `npmlreg` (Einbeck et al., 2006). We will use the explanatory variables gender, age, a suitable measure of regional autocorrelation, and a cluster-level random effect to account for the regional heterogeneity (e.g., arising from regions with big/small populations, outliers etc.). This leads to a two-level model, also called a variance component model, where the clustering variable is the health region ID. The age variable is a factor with four categories from 0–29 (reference category), 30–39, 40–59, and 60+ years.

For each region $i = 1, \ldots, 13$ and each subpopulation $j = 1, \ldots, 8$ (defined by a certain gender/age combination), we have a total count of suicides Y_{ij} over the 10 years. Further, the subpopulation sizes n_{ij} are available, as well as the standardized mortality ratios (SMR), i.e., the ratio observed/expected number of deaths, from which the E_{ij} are immediately obtained.

3 Modelling suicide rates

We firstly focus on the model for the rate λ. The 'core' model

$$\log(\lambda_{ij}) = \alpha + \beta_1 \cdot \text{sex}_{ij} + \beta_2 \cdot \text{age}_{2,ij} + \beta_3 \cdot \text{age}_{3,ij} + \beta_4 \cdot \text{age}_{4,ij} \equiv \alpha + \eta_{ij} \quad (3)$$

gives a disparity of $-2 \log L = 793.8$, with all five estimated parameters being highly significant. Next, we replace the constant intercept α by a regional random effect z_i, assuming that all individuals living within one health region share a common intercept. The NPML approach approximates the unknown and unspecified distribution of the random effects by a discrete mixture, yielding mass points $z_1, \ldots, z_k$ and masses $\pi_1, \ldots, \pi_k$. Fitting a model with $k = 3$ mass points, the disparity already drops to 697.2, and does not fall significantly when increasing k further.
To improve this result, we construct average 'neighboring crude rates' $r_i = \sum_{\ell \in S_i} Y_\ell / \sum_{\ell \in S_i} n_\ell$, where S_i is the set of regions adjacent to the $i-$th region. Including this variable as a fixed effect, the model can be formulated as

$$\log(\lambda_{ij}) = \eta_{ij} + \gamma \cdot r_i + z_i,$$

and the disparity drops to 691.3. Using a random instead of a fixed coefficient for the autocorrelation term, one achieves a slight additional decrease in disparity down to 689.8. However, using gender-specific neighboring rates r_{ij} and a fixed parameter γ, nearly the same reduction can be achieved, yielding the disparity 690.1. Combining these ideas using a random gender-specific autocorrelation term one gets a further slight improvement towards 688.8 (all models using $k = 3$). The addition of interaction terms between age and sex as fixed effects yields a further improvement in the model, with the deviance dropping to 646.8 for fixed γ, and to 645.1 for a random coefficient γ. Summarizing, largely independent of the order of the inclusion of the terms, we get a disparity reduction of about 95 points for the regional random effect, of about 45 points for the interaction, and of about 10 points for the regional autocorrelation.
One nice feature of NPML estimation is that the posterior probability that unit i stems from cluster k corresponds to the weights in the final iteration of the EM algorithm. Firstly, this enables us to calculate empirical Bayes predictions from posterior estimates of the random effects combined with the fixed part of the linear predictor (Aitkin, 1996b, did this for the case of posterior means from a binomial model without covariates), shrinking the – highly variable – crude observations by 'borrowing' information from similar regions. The results are shown in Fig. 2 (right) for the fixed gender-specific autocorrelation model with age.gender interaction terms, which we consider as our favorite model from the average crude rate family of models, for this data. One observes that particularly the rates for the cities – based on small population sizes compared to the health boards – are considerably shrunk. Secondly, one can classify the regions into clusters according to the

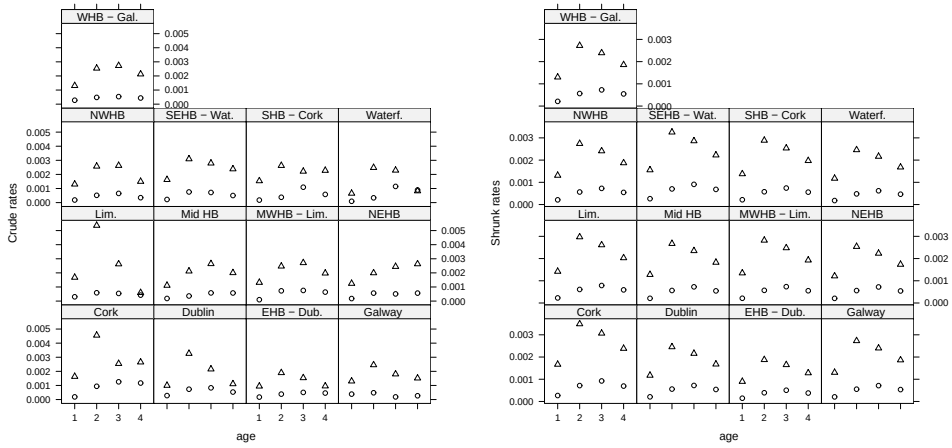

FIGURE 2. *Irish Suicide Rates, crude (left) and shrunk (right), for men (△) and women (○), and for four groups of increasing age as outlined in Section 2*

mass point with the highest posterior probability (for use in 'league table' type comparisons and performance monitoring). In the case of a random intercept model with covariates we use the term *posterior intercept* for the posterior estimate of the random effect term. We use a convention of assigning a cluster to a mass point component if $p \geq 0.90$ and confidently excluding it from a component if $p < 0.01$ (see Table 1). For a model which incorporates relevant explanatory variables, these classifications are informative about clusters with either excess or exemplary rates of mortality, conditional on the fitted model and the observed data. It turns out that only the health region 'EHB minus Dublin' is assigned to mass point 1 (as a region with very few suicides), the city Cork and the region 'SEHB minus Waterford' are classified to mass point 3 as regions with a large number of suicides, and all other regions are assigned to the intermediate mass point 2, except the cities of Waterford and Limerick of which we are limited to inferences regarding the mass points that they can be *excluded* from (Limerick is excluded from the low suicide rate and Waterford from the high suicide rate).

4 Modelling relative risks

A similar analysis as in Section 3 is conducted using the relative risk parameter θ as model parameter. We make use of the simpler 'core' model $\log(\theta_{ij}) = \alpha$, giving the disparity 754.4. One might wonder at this point why we get a *better* disparity compared to model (3), using a model *without* covariates. The reason for this is that the information about the explana-

TABLE 1. *Posterior probabilities for suicide data modelled with crude rate (left) and relative risk (right) as model parameters.*

	Average crude rate				Relative risk			
	Posterior intercept	Masspoints			Posterior intercept	Masspoints		
Intercept		−8.910	−8.533	−8.313		−0.576	−0.192	0.026
Proportion		0.118	0.703	0.179		0.121	0.702	0.178
SEHB − Waterford	−8.31			1.00	0.03			1.00
Cork City	−8.31			1.00	0.02		0.01	0.99
Limerick City	−8.46		0.68	0.32	−0.12		0.69	0.31
NEHB	−8.53		1.00		−0.19		1.00	
Dublin City	−8.53		1.00		−0.19		1.00	
SHB − Cork	−8.53		1.00		−0.19		1.00	
Mid WHB − Limerick	−8.53		1.00		−0.19		1.00	
Midland HB	−8.53		1.00		−0.19		1.00	
NWHB	−8.53		1.00		−0.19		1.00	
WHB − Galway	−8.53		1.00		−0.19		1.00	
Galway City	−8.56	0.07	0.92	0.01	−0.21	0.06	0.93	0.01
Waterford City	−8.71	0.47	0.53		−0.38	0.50	0.50	
EHB − Dublin	−8.91	1.00			−0.58	1.00		

Posterior probabilities: $p \geq 0.95$, $0.90 \leq p < 0.95$, $p < 0.90$.

tory variables – including the interaction – is essentially contained in the expected values E_{ij}, which goes into the model as an offset according to equation (2). Hence, the 40 points improvement compared to model (3) stems from the indirect inclusion of main effects and an interaction term. Carrying out an analysis along the same lines as in Section 3, our favorite model,

$$\log(\theta_{ij}) = \gamma \cdot r_{ij} + z_i,$$

again turns out to contain a random intercept z_i for regions and a fixed gender-specific autocorrelation term. Note that the r_{ij} are now computed as average neighboring SMRs. The disparity 647.5 of this model is only very slightly worse than in Section 3 (646.8), given that we save 7 degrees of freedom, just by employing another offset!

The fitted values are very similar to those of the final average crude rate model and show considerable shrinkage for the city regions. The strong agreement between the fitted values of the models (Fig. 3), despite the omission of the age and sex variables for the relative risk model, supports our statement above that information concerning variation in rates within a region is incorporated in the expected value offset used with the latter family of models. Classifying the regions into mass point components based on their posterior probabilities (shown in Table 1) again indicates that Cork City and 'SEHB minus Waterford' are assigned to the high suicide rate mass point 1, and 'EHB minus Dublin' is the only region assigned to the low suicide rate mass point 3. The other regions are assigned to the intermediate rate mass point 2, apart from Waterford City which is excluded from the high rate and Limerick City which is excluded from the low rate, but both of which have posterior probabilities spread across two mass points.

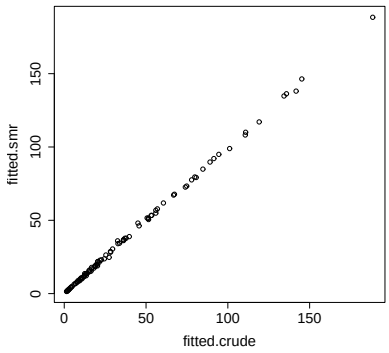

FIGURE 3. *Empirical Bayes Predictions for the crude rate model (x-axis) and relative risk model (y-axis).*

5 Conclusion

In the context of the Irish suicide mortality dataset, a spatial autocorrelation appears to be separately identifiable, in addition to the random-effect for regional heterogeneity. Summarizing the present findings, we conclude i) Modelling regional heterogeneity with spatial random effects improves the model fits greatly. ii) Further improvements can be gained including a spatial autocorrelation term. iii) Unlike Biggeri et al. (2000), we do not observe much gain in using a *random* coefficient for the autocorrelation term. iv) The relative-risk models incorporate information about variation within regions through the expected values, rather than through the additional covariate terms. The average crude rate modelling approach may be preferable when there is interest in evaluating the effects of explanatory variables, e.g., to inform our understanding of the data generating process. Further the average crude rate approach allows continuous covariates and finer groupings into factors, since the problem of counts with corresponding SMR values of zero does not arise in that case.

Due to differences in administration and health policy, the bordering regions of Northern Ireland were omitted in the calculation of the autocorrelation terms for the adjacent regions in the Republic of Ireland. Further development along the lines of the present analysis might incorporate the 6 regions of the North and examine whether autocorrelations which include cross-border effects improve the fit of the model. National differences in the rates across the two sets of counties could be allowed for by means of additional interaction terms, though the number of regions in the North is small.

We finish with a word of caution: though we did not observe computational problems in fitting these models neither using a GLIM nor using an R implementation of NPML, certain problems can arise in jointly modelling both

heterogeneity and spatial dependence between regions, as noted by Aitkin (1999), e.g., in some cases a joint distribution of spatial random effects for each region may be singular given a very high intra-area correlation.

Acknowledgments: This work was partly supported by Science Foundation Ireland Basic Research Grant 04/BR/ M0051.

References

Ahlbom, A. (1993). *Biostatistics for Epidemiologists*. Boca Raton: Lewis Publishers.

Aitkin, M. (1996a). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–262.

Aitkin, M. (1996b). Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In: *Proceedings of the 11th IWSM 1996*. 87–94, Orvieto, Italy.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–128.

Biggeri, A., Marchi, M., Lagazio, C., Martuzzi, M. and Böhning, D. (2000). Non-parametric maximum likelihood estimators for disease mapping. *Statistics in Medicine*, **19**, 2539–2554.

Einbeck, J., Darnell, R. and Hinde, J. (2006). R package `npmlreg`. Nonparametric maximum likelihood estimation for random effect models, `www.nuigalway.ie/maths/je/npml.html`.

Institute of Public Health in Ireland (2005). All Ireland Mortality Database. Retrieved August 8, 2005, from `http://mapserver1.cdc-ni.com/iph/index.htm`

Longford, N. T. (2005). *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician.* New York: Springer.