The Korean Communications
in Statistics Vol. 4, No. 1, 1997
pp. 33-39

Mixed Model with Time Effect for Analyzing Geographic Variability in Mortality Rates

Yong-Chul Kim[1]

Abstract

Tsutakawa(1988) proposed a mixed model for using empirical Bayes method to study the geographic variability in mortality rates of a disease. In particular cases of the analysis in mortality rate, we need to consider the effect of time. If observed data are collected annually for the time period, then time effect will be emphasized. Here, an extended model for estimating the geographic effect and the mortality rates of the disease with time effect is proposed.

1. Introduction

One goal of building a model is a good representation of observed data which have already been collected. In particular cases of the analysis in mortality rate, a great deal of work has been done with homogeneous Poisson model. Geographic variability in mortality rates has been studied by Breslow and Day(1975) who used a multiplicative Poisson model to estimate mortality rates for different populations. They assume that the populations are sufficiently large and events are rare, that the data are well represented by the Poisson model. They give a particularly simple mathematical model for the rate structure which is given as a product of rate factor and region factor. The parameters are estimated by the method of maximum likelihood. Osborne(1975) and Gail(1978) have also used the fixed effect multiplicative models to study the cancer mortality. Manton, Woodbury and Stallard(1981) proposed a mixed categorical-continuous variable model for the analysis of mortality rates. This model assumes that the number of cancer deaths for a group of individuals at the same risk level is a Poisson variable and individual risks are assumed to be gamma distributed. Then the distribution of the number of cancer deaths in a given cell is negative binomial. They used the scale and shape parameters of the gamma distribution to explain the demographic and geographic variability, respectively, as fixed effects. For example, period cancer mortality data for a given population could be modeled with a gamma shape parameter which is constant

[1] Full-time Lecturer, Department of Computer Science and Statistics, Yongin University, Yongin, Kyunggido, 449-714, Korea

over age and a scale parameter which is changed over age. Tsutakawa(1985) used a full Bayesian approach to estimate cancer mortality rate of disease with its logit. Tsutakawa(1988) used a gamma prior for the mortality rate, in contrast to a normal prior for the logit of the mortality rate previously used. The gamma distribution is parameterized by fixed demographic-effects parameters and a random geographic-effects parameter having an inverse gamma prior with an unknown hyperparameter. The hyperparameter and demographic parameters are estimated by maximum likelihood. Relative risks and mortality rates are then. estimated by their posterior means conditionally on fixed parameters estimated by maximum likelihood.

Brillinger(1986) suggested that the time effect can be used in extending the Poisson model with a point process theory. Here, a nonhomogeneous Poisson gamma model based on the nonhomogeneous Poisson process theory is derived. For the given time interval I_k, the proposed model will be a good representation of observed data for expressing the variability of the geographic and demographic effects, in addition to the time effect.

2. Model and Derivation of Estimation

This article is motivated by an earlier study of the mixed model for analyzing geographic variability in mortality rates in Missouri by Tsutakawa(1988).

2.1 Interpretation of parameters and model

Consider I geographic regions and J demographic groups for K time intervals. Let n_{ijk} be the size of the population and y_{ijk} be the frequency of death within the j-th demographic group in the i-th geographic region for given k-th time interval. z_i is the random geographic parameter for the i-th county and p_{ijk} is the mortality rate per individual within the j-th demographic group in the i-th geographic region for given k-th time interval.

Assume we now have demographic parameter $(\theta_{1j}, \theta_{2j}, \sigma_j^2)$ with $\theta_{1j} > 0$, $\theta_{2j} > 0$, $\sigma_j^2 > 0$;j=1,2,...,J, and random geographic parameter z_i with $z_i > 0$; i=1,2,...,I, such that conditionally on these parameters p_{ijk} is independent gamma random variable with mean and variance that may be expressed as

$$E(p_{ijk} | \theta_{1j}, \theta_{2j}, \sigma_j^2, z_i) = z_i \theta_{1j} e^{\theta_{2j} t_k} \tag{2.1}$$

and

$$Var(p_{ijk} | \theta_{1j}, \theta_{2j}, \sigma_j^2, z_i) = \sigma_j^2 (z_i \theta_{1j} e^{\theta_{2j} t_k})^2, \tag{2.2}$$

where t_k is the mid point of a fixed k-th time period.

Thus p_{ijk} has the gamma distribution with shape parameter σ_j^{-2} and scale parameter $\sigma_j^2 z_i \theta_{1j} e^{\theta_{2j} t_k}$. Assume that z_i is independent and identically distributed over counties with an inverse gamma density with unknown parameter $\gamma > 0$ and the mean of z_i equals 1. We can obtain the nonhomogeneous Poisson gamma model which is equal to Poisson gamma model with $\theta_{2j} = 0$; j=1,2,...,J. Also the logarithm of equation (2.1) has form of the classical mixed linear model, since

$$\log p_{ijk} = \log \theta_{1j} + \log z_i + \theta_{2j} t_k + \varepsilon_{ijk.}$$

where $\log z_i$'s are random and ε_{ijk}'s are the error terms with some means and variances structure(Tsutakawa(1988)).

2.2 Estimation

An empirical Bayes approach with the random parameter p_{ijk} and z_i having a distribution that depends on the unknown hyperparameters θ_{1j}, θ_{2j}, σ_j^2, and γ will be considered. After hyperparameter is estimated by maximum likelihood estimator, random effect parameters p_{ijk} and z_i are estimated by the posterior mean as follows

$$E(p_{ijk} | y_{ijk}, \theta_{1j}, \theta_{2j}, \sigma_j^2, z_i) = \frac{y_{ijk} \overline{p_{ijk}} + \overline{p_{ijk}} / \sigma_j^2}{n_{ijk} \overline{p_{ijk}} + \sigma_j^{-2}},$$

where $\overline{p_{ijk}} = z_i \theta_{1j} e^{\theta_{2j} t_k}$.

Under the assumed model, the likelihood function of hyperparameters ϕ=(θ_{1j}, θ_{2j}, σ_j^2, γ) is given by

$$q(\phi | y) = \prod_{i=1}^{I} \int \prod_{j=1}^{J} \prod_{k=1}^{K} f(y_{ijk} | \sigma_j^2, n_{ijk} \sigma_j^{-2} \theta_{1j} e^{\theta_{2j} t_k}) h(z_i | \gamma) dz_i ,$$

where f is a negative binomial and h is an inverse gamma with $E(z_i) = 1$.

2.3 Computation for estimation of hyperparameters

The log likelihood function of ϕ is given by

$$\log q(\phi | y) = \sum_{i=1}^{I} \log \int \prod_{j=1}^{J} \prod_{k=1}^{K} f(y_{ijk} | \sigma_j^2, n_{ijk} \sigma_j^{-2} \theta_{1j} e^{\theta_{2j} t_k}) h(z_i | \gamma) dz_i. \qquad (2.3)$$

The maximum likelihood estimate of ϕ will be computed by Marquardt(1963) method. The estimation of the hyperparameter needs the calculation of multiple integrals for the derivatives of equation (2.3) with respective to γ. We use the profile likelihood function of γ (Richards(1961)) : We maximize the equation (2.3) with respect to θ_{1j}, θ_{2j}, and σ_j^2 at fixed value γ. Then we can select the value γ that maximizes the profile likelihood function. The evaluation of equation (2.3) and its derivatives require numerical integration of z. After changing the variable of integration from z to log z, the Gauss-Hermite quadrature method is used.

3. Numerical Example and Conclusion

The numerical example here uses female lung cancer deaths for age groups, 45-64, 65-74, 75 over, for years 1973-1984, in all counties of Missouri. The population measurements for the 115 counties are based on census reports and are those from the Missouri Department of Health. In this model, it is difficult to get the estimation of hyperparameter based on yearly data because of the small frequency of deaths in the yearly data. So the data are combined in terms of three year periods.

Figure 1 shows a scatter plot of the raw annual rates of age groups per million for female lung cancer using three periods(1973-1975, 1976-1978, 1979-1981 and 1982-1984). In Figure 1, the observed rate shows an increasing trend with respect to age.

The maximum likelihood estimate $\hat{\phi} = (\hat{\sigma}^2, \hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{13}, \hat{\theta}_{21}, \hat{\theta}_{22}, \hat{\theta}_{23}, \hat{\gamma})$ for mixed effect model with temporal effect is obtained as follows ;

$$\hat{\sigma}^2 = 0.0011921 \qquad \hat{\theta}_{11} = 0.000758$$

$$\hat{\theta}_{12} = 0.001128 \qquad \hat{\theta}_{13} = 0.001168$$

$$\hat{\theta}_{21} = 0.183109 \qquad \hat{\theta}_{22} = 0.238777$$

$$\hat{\theta}_{23} = 0.215682 \qquad \hat{\gamma} = 22.0.$$

In the case of the mixed effect model without time effect, (2.1) and (2.2) are

$$E(p_{ijk} | \theta^*_{jk}, \sigma^2, z_i) = z_i \theta^*_{jk}$$

and

$$Var(p_{ijk} | \theta^*_{jk}, \sigma^2, z_i) = \sigma^2 (z_i \theta^*_{jk})^2,$$

where k denotes the k-th time period, respectively.

Figure 2 shows the comparison of the estimates of annual rates of age groups for mixed model without time, mixed model with time and the observed rates. The estimates of annual ra es of age groups with each model are smaller than the observed rates. The reason for this is that the rate values of counties with large populations dominate the rate values of the counties with small population. This is due to extra-Poisson variation of mixed model with time effect, which cannot show up under the multiplicative Poisson model but is explained by the randomness of p_{ijk} and z_t. Note a slight decline in the highest age group may be due to cohort effects of other diseases even though rates generally increase over age. So this model should be better with respect to estimation on the annual mortality rates.

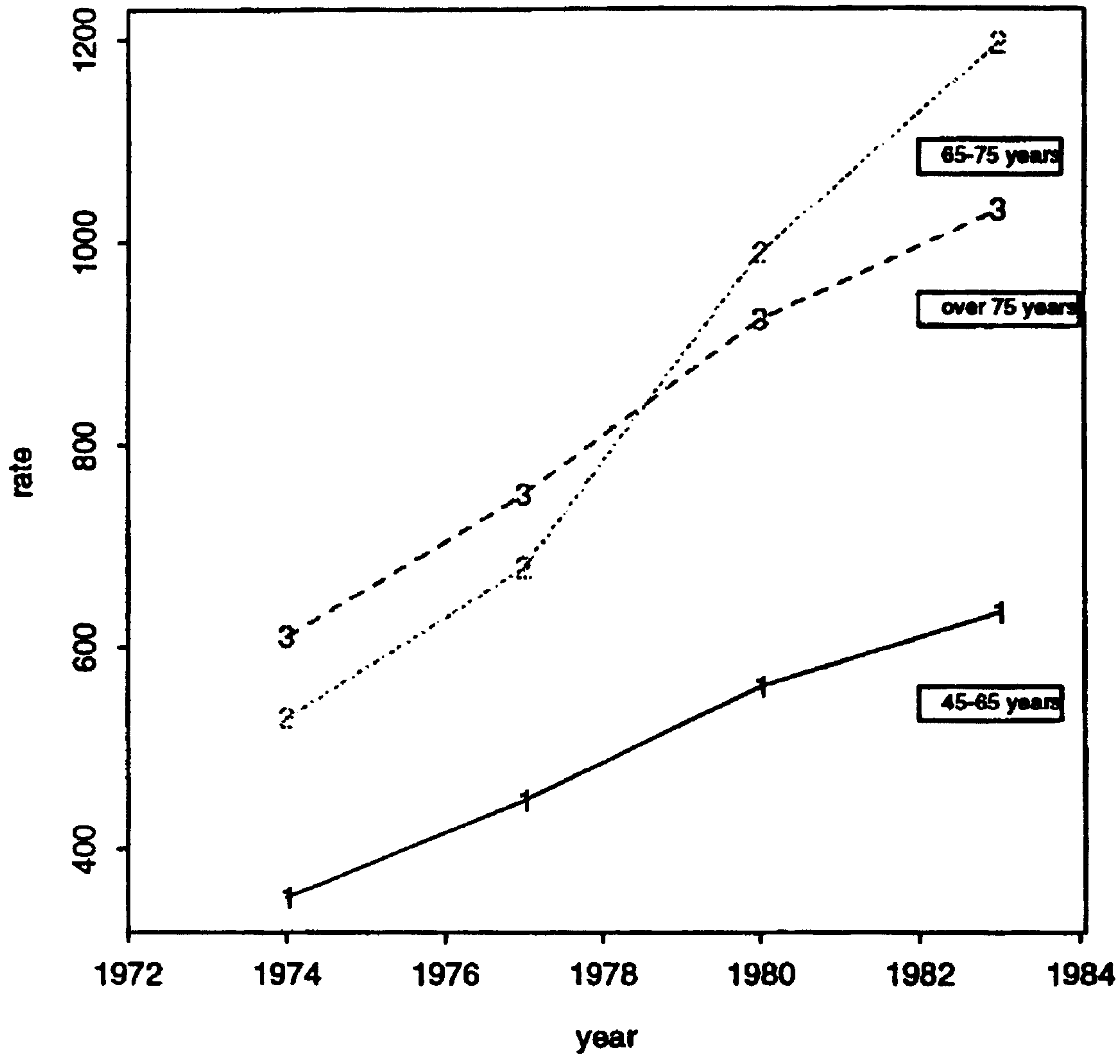
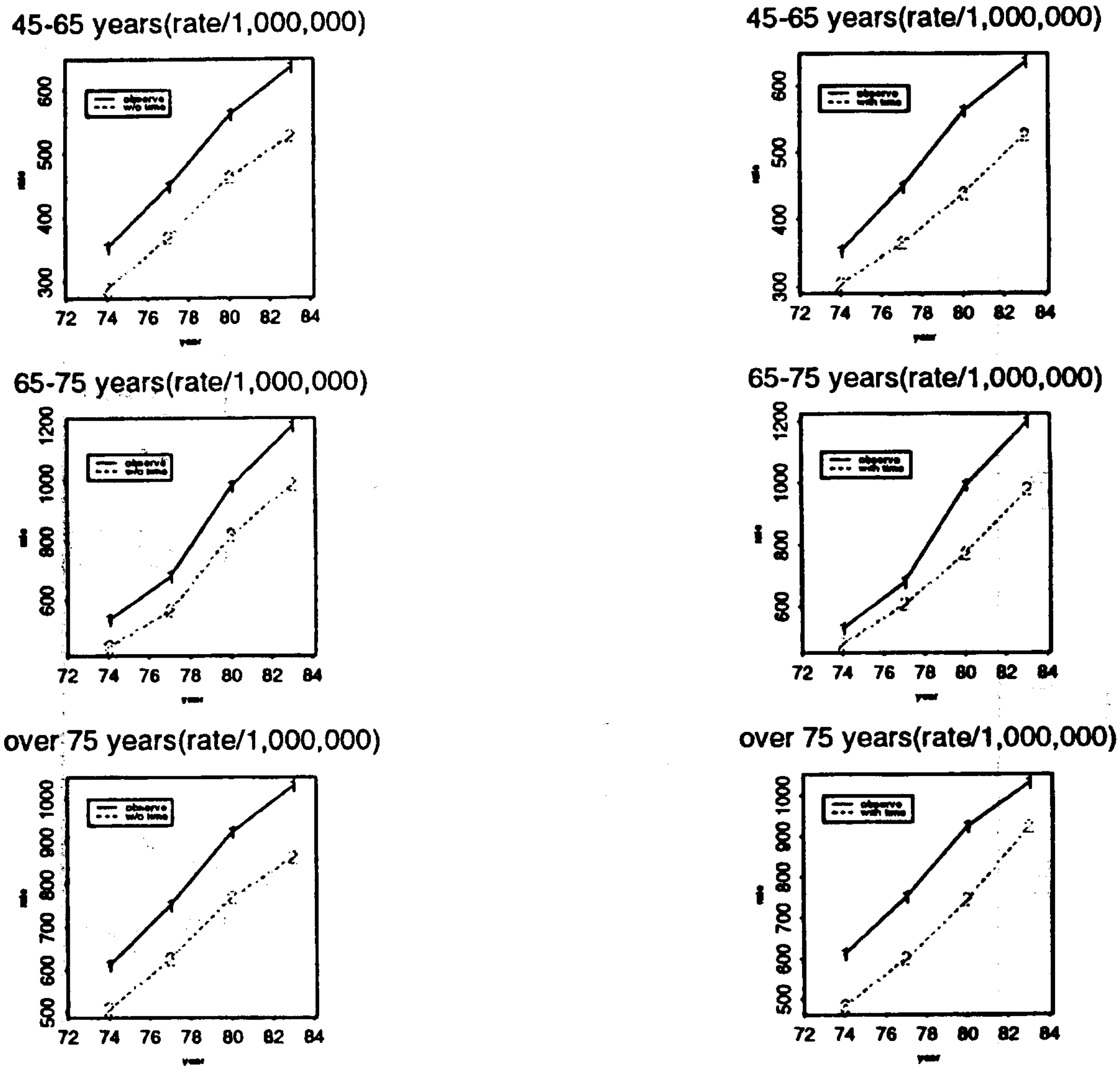


Figure 1 : Raw rates(rate/1,000,000)


45-65 years(rate/1,000,000)

45-65 years(rate/1,000,000)

65-75 years(rate/1,000,000)

65-75 years(rate/1,000,000)

over 75 years(rate/1,000,000)

over 75 years(rate/1,000,000)

Figure 2 : Observed rates vs Estimated rates(rate/1,000,000)

References

[1] Bresolw, N.E. and Day, N.E. (1975). Indirect Standardization and Multiplicative Models for Rates, With Reference to the Age Adjustment of Cancer Incidence and Relative Frequency Data, *Journal of Chronic Diseases*, 28, 289-303.

[2] Brillinger, D.R. (1986). The Natural Variability of Vital Rates and Associated Statistics, *Biometrics*, 42, 693-734.

[3] Gail, M. (1978). The Analysis of Heterogeneity for Indirect Standardized Mortality Rates, *Journal of the Royal Statistical Society*, Ser.A, 141, 224-234.

[4] Laird, N. (1978). Empirical Bayes Methods for Two-Way Contingency Tables, *Biometrika*, 65, 581-590.

[5] Leonard, T. (1975). Bayesian Estimation Methods for Two-Way Contingency Tables, *Journal of the Royal Statistical Society*, Ser.B, 37, 23-37.

[6] Manton, K.G., Woodbury, M.A. and Stallard, E. (1981). A Variance Components Approach to Categorical Data Models With Heterogeneous Mortality Rates in North Carolina Counties, *Biometrics*, 37, 259-269.

[7] Marquardt, D.W. (1963). An Algorithm for Least-Squares Estimation of Non-Linear Parameters, *Journal of the Society of Applied Mathematics*, 11, 431-441.

[8] Osborne, J. (1975). A Multiplicative Model for the Analysis of Vital Statistics Rates, *Applied Statistics*, 24, 75-84.

[9] Richards, F.S.G. (1961). A Method of Maximum-likelihood Estimation, *Journal of the Royal Statistical Society*, Ser. B, 23, 469-475.

[10] Tsutakawa, R.K. (1985). Estimation of Cancer Mortality Rates : A Bayesian Analysis of Small Frequencies, *Biometrics*, 41, 69-79.

[11] Tsutakawa, R.K. (1988). Mixed Model for Analyzing Geographic Variability in Mortality Rates, *Journal of American Statistical Association*, 83, 37-42.