

ESTIMATION OF PROSTATE CANCER MORTALITY RATES IN SMALL AREAS USING A NONLINEAR MODEL

N. Elkum and D. Murdoch¹

ABSTRACT

Data on the prostate cancer mortality rates in all the Canadian provinces has been collected since at least 1950. In small provinces, the few cases lead to unreliable estimates of the mortality rate. In this paper we implement and evaluate small area estimation of a nonlinear regression model, using a hierarchical Bayes approach. The Gibbs sampler is used to obtain the posterior parameter densities.

KEY WORDS: Nonlinear models; Small area estimation; Hierarchical Bayes; Metropolis-within-Gibbs algorithm.

RÉSUMÉ

Les données sur les taux de mortalité du cancer de la prostate ont été ramassées depuis les années 1950 dans toutes les provinces canadiennes. Dans les petites provinces, le nombre peu élevé de cas mène à des estimations peu fiables du taux de mortalité. Dans cet article, nous mettons en action et évaluons l'estimation d'un modèle de régression non linéaire dans les petites régions, en utilisant une approche hiérarchique bayésienne. L'échantillonneur de Gibbs est utilisé pour obtenir les paramètres de la densité a posteriori.

MOTS CLÉS: Modèles non-linéaires; estimation d'un modèle; hiérarchique bayésienne; l'algorithme Metropolis-within-Gibbs.

1. INTRODUCTION

Disease mortality data are commonly available as summary counts or rates for formally defined regions such as counties, districts, or census tracts. Often regions are small and the information available about them is not enough to provide accurate estimates of local mortality rates.

Mortality due to prostate cancer is one such example where the mortality in a small or average-sized area may be low and information from a single area is limited. Because mortality records include all prostate cancer deaths, the sample size cannot be increased. We need to use small area estimation (SAE) techniques to borrow strength from similar areas when we want to make accurate estimates and predictions of local mortality rates.

Much previous work has been done in small area estimation of mortality rates (Tsutakawa 1985 and 1988; and Desouza 1992). However, the use of nonlinear models in developing reliable small area statistics has received relatively little attention. Our

goal is to implement and evaluate a Hierarchical Bayes (HB) model-based methodology for estimation of small area nonlinear models and to obtain an estimate of prostate cancer mortality rates by province, year and age group.

A simple way to estimate prostate cancer mortality rates is to use nonlinear least squares, but this approach ignores the common area characteristics. In this paper, we will propose and apply a HB nonlinear model, for estimating and predicting prostate cancer mortality rates. In section 2, we will present a discussion of the data used. The general structure of the model will be presented in section 3. Parameter estimates will be discussed in section 4. Section 5 will apply this approach to the prostate cancer data and section 6 will contain our conclusions.

2. THE DATA

Prostate cancer mortality data has been collected for all Canadian provinces (National Cancer Institute of

¹ Naser Elkum and Duncan Murdoch, Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada, K7L 3N6.

Canada, 1994). Within province mortality counts were observed for subgroups of populations defined by five year age group from 40-49 to 85+. For each age group within each province, counts were collected annually from 1950 to 1990.

The raw mortality rates are quite variable, particularly among the small provinces. Thus direct estimates of individual small province characteristics are likely to be extremely unreliable.

3. THE MODEL

Typically rates of rare diseases are modeled with the Poisson distribution. However, the prostate cancer data appear to exhibit extra Poisson variation (*i.e.*, the variation appears larger than one would expect from a Poisson distribution), and we chose instead to use a power transformation of the raw data which made it appear to be roughly normally distributed. Specifically, our first stage model is the nonlinear regression model

$$y_{ijk} = \eta_{ijk}(\theta_i) + \varepsilon_{ijk}.$$

Here the transformed observation is $y_{ijk} = n_{ijk}^{\nu} d_{ijk}^{\lambda}$, where n_{ijk} is the size of the population for the i -th province ($i = 1, \dots, I$), the j -th age group ($j=1, \dots, J$) and the k -th year ($k = 1, \dots, K$), d_{ijk} is the observed death rate in this cell, and ν and λ are chosen to achieve roughly constant variance and normality of y_{ijk} . The transformed mean is

$$\eta_{ijk}(\theta_i) = n_{ijk}^{\nu} \{(\alpha_i + \beta_i \text{year}_k) \exp(\gamma_i \text{age}_j + \delta_i \text{age}_j^2)\}^{\lambda}.$$

The linear-exponential factor involving province-specific parameters α_i , β_i , γ_i , and δ_i will accommodate both the expected nonlinear relationship with age and either a decline or increase in time. Finally, we assume $\varepsilon_{ijk} \sim N(0, \tau^{-1})$, where τ is the measurement precision.

The first stage model above describes the observations, given full knowledge of the parameters for that particular cell. In the second stage, we model the between area variability of the parameters of the first stage. This is where the borrowing of strength occurs.

We assume that the vector of area parameters $\theta_i = (\alpha_i, \beta_i, \gamma_i, \delta_i)$ comes from a multivariate normal distribution with mean μ and variance-covariance matrix Σ , *i.e.*, $\theta_i \sim MVN(\mu, \Sigma)$. The parameter μ represents what our estimate of a province's parameters would be if we had no data from that province. The

parameter Σ describes the amount of variation in the θ_i s between the provinces. The measurement precision τ is assumed to come from a diffuse $\text{Gamma}(a/2, b/2)$ distribution.

In our third stage, we specify a prior distribution on μ and Σ . This should mirror prior belief concerning the type of relationship of the θ_i s. In a case like ours, we have very little prior information, and we use vague priors to reflect our uncertainty about μ and Σ . We assume $\mu \sim MVN(\xi, \Gamma)$. To make a diffuse prior for μ , we may assign large values to Γ or formally set $\Gamma^{-1} = 0$. With this choice, the value used for ξ has little effect on the outcome. Finally, we assume Σ comes from a Wishart distribution with degrees of freedom ρ and scale matrix R , *i.e.*, $E(\Sigma^{-1}) = \rho R^{-1}$.

To summarize, the joint density may be written as

$$[Y, \theta, \tau, \Sigma, \mu] = [Y|\theta, \tau] [\theta|\mu, \Sigma] [\tau|\mu] [\Sigma].$$

Here Y is the vector of all IK observations, θ is the vector of $4I$ province-specific parameters, and square brackets denote the joint and conditional probability densities.

4. PARAMETER ESTIMATION

In the proposed hierarchical Bayes approach, inferences are based on the posterior distribution; in particular, a parameter of interest is estimated by its posterior mean and its precision is measured by its posterior standard deviation. We are particularly interested in the posterior distribution $[\theta|Y]$ of province-specific parameter estimates and the posterior distribution $[Y^*|Y]$ of a vector of predictions Y^* under future conditions. We hope to strengthen the knowledge of a given area by borrowing strength from other areas. Because of the complex nature of our model, the posterior distributions can not be expressed in closed form, and Gibbs sampling (Gelfand and Smith, 1990) is necessary.

The idea of Gibbs sampling is to replace sampling from the full posterior $[\theta, \tau, \Sigma, \mu|Y]$ with sequential sampling from the conditional distributions of each parameter, *i.e.*, first from $[\theta|\tau, \Sigma, \mu, Y]$, then from $[\tau|\theta, \Sigma, \mu, Y]$, and so on in a cyclic fashion through all the parameters. The resulting Markov chain of parameter values has the joint distribution as its steady state, and inferences can be made by taking sample properties of the observed chain. For our purposes, sample means and standard deviations of θ were calculated, and individual values were used to simulate Y^*

for future predictions.

The implementation of the Gibbs sampler required sampling from several conditional distributions. The density $[\theta | \tau, \Sigma, \mu, Y]$ is proportional to

$$\exp\{-(\tau/2)\sum_{jk}[y_{ijk} - \eta_{ijk}(\theta_i)]^2 - (1/2)(\theta_i - \mu)^T \Sigma^{-1}(\theta_i - \mu)\}.$$

This is not a standard distribution, and sampling required the use of the Metropolis-within-Gibbs algorithm (Gilks *et al.*, 1995). The density $[\tau | \theta, \Sigma, \mu, Y]$ is Gamma with parameters $(a+n)/2$ and $b + \sum_{ijk}[y_{ijk} - \eta_{ijk}(\theta_i)]^2$. The density $[\Sigma | \theta, \tau, \mu, Y]$ is Wishart with degrees of freedom and scale matrix $\sum_i(\theta_i - \mu)(\theta_i - \mu)^T + R$. Finally, the density $[\mu | \theta, \tau, \Sigma, Y]$ is multivariate normal with mean $(I\Sigma^{-1} + \Gamma^{-1})^{-1}(I\Sigma^{-1}\theta + \Gamma^{-1}\xi)$ where θ is the mean of the θ_i values, and variance-covariance matrix $(I\Sigma^{-1} + \Gamma^{-1})^{-1}$.

5. APPLICATION

We fitted the model described above to the prostate cancer mortality data. Different values for λ and ν were tried; residual plots indicated a choice of $\lambda = 0.5$ and $\nu = 0.75$. The usual square root transformation of Poisson data would correspond to $\lambda = \nu = 0.5$.

To represent prior knowledge regarding the parameters τ, μ and R , we selected a subset of the data from 1950 to 1959, and used nonlinear least squares to estimate the θ_i s. We used the mean of these values as ξ and the inverse of the observed covariance matrix as ρR^{-1} . The parameter ρ was taken to be 4, for the vaguest possible Wishart prior. The parameters a and b for the prior for τ were both taken to be 0.

The Gibbs sampler algorithm was run for 4000 iterations, with starting values chosen at the mean of the prior distribution. To assess the performance of the Gibbs sampler, we used the method of Raftery and Lewis (1992). We used SPlus code for implementing the Raftery and Lewis convergence diagnostic based on the *gibbsit* function contributed to the Statlib archive by Steven Lewis. Applying the Raftery and Lewis measure to each parameter separately indicated that the Gibbs sampler was operating well with respect to all the key parameters. Small burn-in values for all parameters suggest that the chain converged very quickly. The whole process took a total of 12 hours on a Sun SPARC-10 workstation.

To illustrate the results, we consider the largest and smallest provinces (Ontario and P.E.I.) in detail. The parameter estimates and estimated errors are given in

Tables 1 and 2. It appears that the hierarchical model has an advantage over the least squares method in that the posterior standard deviations are often substantially less than the least squares standard errors. (This behaviour was not universal; standard errors for some Newfoundland parameters increased.)

Table 1: Comparison between least squares standard errors and the posterior standard deviations of parameter estimates for the Ontario data.

	Least Square Estm. (s. e.)	Posterior Mean (s. d.)	$\frac{s.d}{s.e.}$
α_i	98.8 (1.6)	98.2 (0.8)	0.5
β_i	0.82 (0.12)	0.81 (0.06)	0.5
γ_i	0.15 (0.001)	0.14 (0.0007)	0.7
δ_i	-0.003 (0.0001)	-0.002 (0.0005)	0.5

Table 2: Comparison between least squares standard errors and the posterior standard deviations of parameter estimates for the P.E.I. data.

	Least Square Estm. (s. e.)	Posterior Mean (s. d.)	$\frac{s.d}{s.e.}$
α_i	92.3 (8.3)	97.0 (8.0)	0.96
β_i	1.4 (0.7)	1.3 (0.4)	0.57
γ_i	0.18 (0.01)	0.14 (0.002)	0.2
δ_i	-0.005 0.0009	-0.003 0.0002	0.2

The results are similar for predictions. HB standard deviations of predicted values are generally smaller than least squares standard errors (Figure 1), though there are exceptions.

and assume a certain matrix G exists such that $Gy_t = GZ_t\alpha_t$, i.e., $GZ_t\alpha_t$ is observed without error. Then, the estimated state $\hat{\alpha}_{t|t}$ satisfies the constraint: $GZ_t\hat{\alpha}_{t|t} = GZ_t\alpha_t$.

To prove the assertion, notice that $GZ_t\hat{\alpha}_{t|t} = GZ_t\hat{\alpha}_{t|t-1} - GZ_t\text{Cov}(\alpha_t - \hat{\alpha}_{t|t-1}, v_t)[\text{Var}(v_t)]^{-1}v_t = Gy_t = GZ_t\alpha_t$, where $v_t = y_t - Z_t\hat{\alpha}_{t|t-1}$ (note that $Z_t\text{Cov}(\alpha_t - \hat{\alpha}_{t|t-1}, v_t) = \text{Var}(v_t)$). By taking $G = \begin{bmatrix} \mathbf{0} & I \end{bmatrix}$ in the above, it follows that the predictions (3.5) satisfy the benchmark constraints (3.2).

3.3 Estimation of Hyperparameters

We propose estimating the hyperparameters (the elements of Γ_ϵ , Γ_η and Γ_ζ) by minimizing the SSPE penalty function

$$L(\Gamma_\epsilon, \Gamma_\eta, \Gamma_\zeta; y_1, \dots, y_T; x_1, \dots, x_{T+1}) = \sum_{t=d+1}^T \|\hat{y}_{t|t-1}^* - y_t\|^2,$$

where $\|\cdot\|$ is the Euclidean norm and where d is the smallest integer such that $d \geq \dim(\alpha_t)/\dim(y_t^*)$.

4. APPLICATION TO FORECASTING SUB-PROVINCIAL POPULATIONS

4.1 Generalized Harvey's Holt-Winters Method

Let the estimate of the j th sub-provincial area at time t be denoted by $y_t^{(j)}$, let $y_t = (y_t^{(1)}, \dots, y_t^{(j)})'$ (where J is the number of sub-provincial regions) and let $x_t = \sum_{j=1}^J y_t^{(j)}$ be the provincial total.

We assume the model given by (3.1). Further, we assume the provincial total $x_t = \sum_{j=1}^J y_t^{(j)}$ is observed without error, i.e., $x_t = \sum_{j=1}^J y_t^{(j)} = \sum_{j=1}^J L_t^{(j)}$. There is no seasonality in our model, since the series is annual.

Note that the variance-covariance matrix Γ_ϵ is singular since $\sum_j y_t^{(j)}$ is observed without error yielding $\sum_j \epsilon_t^{(j)} = 0$. The data that become available at year t consist of the sub-provincial estimates for year $t-2$ and the provincial estimated total x_t , and may be written as a vector $y_t^* = (y_{t-2}^*, x_t)'$. If the state at time t is defined as $\alpha_t = (L_{t-2}', L_{t-1}', L_t', R_t)'$ then we have the state space model (3.3a-b) where the observation and transformation matrices Z_t and T_t are given by

$$Z_t = \begin{pmatrix} I & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}' & \mathbf{0} \end{pmatrix} \text{ and } T_t = \begin{pmatrix} \mathbf{0} & I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I & I \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I \end{pmatrix} \quad (4.1)$$

where I is the $J \times J$ identity matrix, and $\mathbf{1}$ is a vector of 1's (of appropriate dimensions), and where $\mathbf{0}$ in T_t and in the first row of Z_t is a zero $J \times J$ matrix and zero $1 \times J$ row vector in the second row of Z_t . The disturbances are given by $\epsilon_t^* = (\epsilon_t', 0)'$ and $\eta_t^* = (\mathbf{0}, \mathbf{0}, \eta_t', \zeta_t)'$.

A small number of hyper-parameters was used because of the short length of the series so that the estimates of Γ_ϵ , Γ_η and Γ_ζ may be stable. We made the assumption that all three matrices have the form $\sigma^2 [I + \rho(\mathbf{1}\mathbf{1}' - I)]$ for some constants $\sigma^2 > 0$, $-1 < \rho < 1$ (i.e., equal variances on its diagonal and equal correlations between the components). Since Γ_ϵ is singular, it must have the form

$$\Gamma_\epsilon = \sigma_\epsilon^2 \left[I - \frac{1}{J} \mathbf{1}\mathbf{1}' \right]. \quad (4.2)$$

We also have,

$$\Gamma_\eta = \sigma_\eta^2 [I + \rho_\eta(\mathbf{1}\mathbf{1}' - I)], \quad \Gamma_\zeta = \sigma_\zeta^2 [I + \rho_\zeta(\mathbf{1}\mathbf{1}' - I)]. \quad (4.3)$$

4.2 The Ruler Method

The method currently used by Statistics Canada is the ruler method. Its predictions are given by

$$\hat{y}_{T+k|T}^{(j)} = x_{T+k} \left(\sum_i \tilde{y}_{T+k|T}^{(i)} \right)^{-1} \tilde{y}_{T+k|T}^{(j)}$$

where

$$\tilde{y}_{T+k|T}^{(j)} = (k+1)y_T^{(j)} - k y_{T-1}^{(j)}, \quad k=1, 2.$$

These predictions tend to be quite unstable as they solely depend on the two last-known time points. The method does, however, perform well when the population growth has been approximately linear in the most recent 4 years. Thus, the proposed generalized HHW method may be viewed as being more stable than the ruler method to departures from linearity because it takes advantage of information from the past. However it may be non-robust (see Introduction) if the underlying model does not hold. In the following subsection, we therefore, propose a composite GHHW method to address this problem.

4.3 The Composite GHHW Method

The ruler method is not a time series method. As mentioned before, while the ruler method is generally unstable, it does perform well when the recent growth has been linear. On the other hand, time series

methods, such as the generalized HHW which was introduced above, would generally perform better when the model (3.3) adequately describes the series. Thus, a combination of these two methods would be robust when abrupt changes in the series occur. The predictions in the proposed composite GHHW are convex combinations of the generalized HHW predictions and the ruler method's predictions. The coefficient of the convex combination is also determined by minimizing the SSPE penalty function. Empirical results comparing predictions from the ruler method and the two methods proposed in this paper are given in the next section.

5. NUMERICAL RESULTS

5.1 Description of the Data and the Evaluation Measure

The data consist of estimates of economic regions (ERs) for the provinces of Nova Scotia and Manitoba for the years 1986-94 and provincial totals for 1986-96. The ERs are non-overlapping and their union covers the whole province. (As well, the 1986-94 ER estimates add up to the provincial totals.) We have applied the GHHW and the composite GHHW (c-GHHW in the sequel) methods and the ruler method to the 1984-91 part of the ER series, together with the 1992-3 provincial totals, and predicted the (known) 1992 and 1993 ER estimates (*i.e.*, $T=7$, see Tables 1a and 1b). We have also applied these methods to the 1984-92 part of the ER series, together with the 1993-4 provincial totals, and predicted the (known) 1993 and 1994 ER estimates ($T=8$, Tables 2a and 2b). The results were evaluated by comparing the predictions by each method to the known values. We have calculated for each province the mean absolute relative one and two year ahead prediction errors of ER populations, as well as the maximum absolute relative prediction error.

Table 1a.

Mean and Maximum Absolute Relative Error (%) of the Ruler, GHHW and the Composite GHHW Methods, $T=7$, for the Province of Nova Scotia

	One Step		Two Step	
	Mean	Max.	Mean	Max.
Ruler	0.18	0.37	0.37	0.81
GHHW	0.20	0.53	0.38	1.00
c-GHHW	0.13	0.24	0.25	0.49

Table 1b.

Mean and Maximum Absolute Relative Error (%) of the Ruler, GHHW and the Composite GHHW Methods, $T=7$ for the Province of Manitoba.

	One Step		Two Step	
	Mean	Max.	Mean	Max.
Ruler	0.33	0.96	0.82	1.64
GHHW	0.48	0.96	0.86	1.64
c-GHHW	0.40	1.04	0.83	1.49

Table 2a.

Mean and Maximum Absolute Relative Error (%) of the Ruler, GHHW and the Composite GHHW Methods, $T=8$ for the Province of Nova Scotia

	One Step		Two Step	
	Mean	Max.	Mean	Max.
Ruler	0.05	0.10	0.47	0.74
GHHW	0.23	0.50	0.33	0.63
c-GHHW	0.15	0.32	0.32	0.53

Table 2b.

Mean and Maximum Absolute Relative Error (%) of the Ruler, GHHW and the Composite GHHW Methods, $T=8$ for the Province of Manitoba

	One Step		Two Step	
	Mean	Max.	Mean	Max.
Ruler	0.50	1.47	0.86	2.12
GHHW	0.57	1.51	0.82	2.10
c-GHHW	0.44	0.94	0.51	1.12

5.2 Summary of the Results

The c-GHHW method performed best in both provinces in two-step ahead predictions both in predicting for 1993, from the 1986-91 part of the series, and for 1994 using 1986-92 part. However, in the case of one step ahead predictions, the c-GHHW method showed mixed performance. Note that in the cases where the c-GHHW was not best, the relative error was still small, showing the robustness of the method.

6. REMARKS

A method of prediction of multivariate series under linear constraints, motivated by a generalization of Harvey's Holt-Winters method is presented. This method requires only semi-parametric assumptions and is as easily applicable as the traditional Holt-Winters method. The currently used ruler method is simple but unstable. A composite method, combining the generalized HW and the ruler method predictions, is also introduced. The two new methods are applied to the problem of prediction of sub-provincial population counts for the CLFS. Empirical results look promising. Future work will include testing the proposed method with simulated data and with real data when the 1996 data become available.

ACKNOWLEDGEMENT

The authors wish to thank R. Bender from the Demography Division, and A. Delisle, Labour and Household Surveys Analysis Division, both at Statistics Canada, for useful discussions and provision of the data. The second author's research was supported in part by a grant from the National Science and Engineering Research Council of Canada held at Carleton University, Ottawa under an adjunct research professorship.

REFERENCES

- Delisle, A. (1993). "Subprovincial sampled population estimation programme: Methodology, limitations, present predicament and possible solutions," Statistics Canada, Ottawa, Canada.
- Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press.
- Holt, C.C. (1957). "Forecasting seasonals and trends by exponentially weighted moving averages," ONR Research Memorandum 52, Carnegie Institute of Technology, Pittsburgh, Pennsylvania.
- Pfeffermann, P. and Allon, J. (1989). "Multivariate exponential smoothing: Method and practice," *International Journal of Forecasting* 5, 83-98.
- Winters, P.R. (1960). "Forecasting sales by exponentially weighted moving averages," *Management Sciences* 6, 324-342.