Journal of the Korean
Statistical Society
Vol. 25, No. 4, 1996

Multiprocess Dynamic Survival Models with Numbers of Deaths[†]

Joo Yong Shim, Joong Kweon Sohn and Sang Gil Kang [1]

Abstract

The multiprocess dynamic survival model is proposed for the application of the regression model on the analysis of survival data with time-varying effects of covariates, where the survival data consists of numbers of deaths at certain time-points. The algorithm for the recursive estimation of a time-varying parameter vector is suggested. Also the algorithm of forecasting of numbers of deaths of each group in the next time interval based on the information gathered until the end of current time interval is suggested.

Key Words: Multiprocess dynamic models; Survival data; Covariates; Recursive estimation

[†] This Study was supported by the Basic Science Research Institute Program, Ministry of Education, 1996, Project No. BSRI-96-1403.

[1] Department of Statistics, Kyungpook National University, Taegu, 702-701, Korea.

1. INTRODUCTION

Cox(1972) proposed the regression model for the survival analysis to separate and estimate effects of covariates which are assumed to be fixed in time. In certain cases (Stablein *et al.*,1981), it has been observed that effects of covariates vary in time. Gamerman(1991) developed the dynamic Bayesian model for the survival data consisted of observed survival times by incorporating the hazard rate of a time-varying parameter vector modeling effects of covariates into the dynamic generalized linear model(West *et al.*,1985).

Now, provided only numbers of deaths of different groups at certain time points are available, intuitively we consider binomial trials to analyze this data to estimate the failure proportion in each time interval. To introduce the regression model for the survival data into this case, we assume that the failure proportion is the product of a covariate vector and a time-varying parameter vector modeling effects of different covariates. We develop the multiprocess dynamic survival model by incorporating the failure proportion of a time-varying parameter vector into the multiprocess dynamic generalized linear model(Bolstad,1988) with a distributional assumption of numbers of deaths, which generalizes the application of the multiprocess dynamic approach(Harrison *et al.*,1976) to the regression model for the survival data.

The multiprocess dynamic survival model is described in Section 2. The algorithm for recursive estimations of a parameter vector under the multiprocess dynamic model is provided in Section 3. Also the algorithm for the forecasting of numbers of deaths in the next time interval is provided in Section 4. The performance of the estimation and forecasting under the proposed model is illustrated via the simulated data in Section 5.

2. MODEL DESCRIPTIONS

Here we assume that there is no censoring and that the number of deaths(failures) of individuals of the j-th covariate vector Z_j, $j = 1, \cdots, J$, follows a piecewise binomial distribution which has a constant failure proportion in each time interval as

$$\theta_j(t) = \theta_{i(j)} \text{ for } t \in I_i = (\tau_{i-1}, \tau_i], i = 1, \cdots, s,$$

where τ_0 is usually set to 0 and $I_s = (\tau_{s-1}, \infty)$. We denote it by

$$Y_{ij} \sim Binomial(n_{ij}, \theta_{i(j)}),$$

where Y_{ij} is the number of deaths of a covariate vector Z_j in the time interval I_i, n_{ij} is the number of the alive of a covariate vector Z_j at the beginning of the time interval I_i, and $\theta_{i(j)}$ is the failure proportion of individuals of a covariate vector Z_j in the time interval I_i. We denote the set of numbers of deaths in the time interval I_i by $Y_{i\cdot}$, and the corresponding set of observed numbers of deaths by $y_{i\cdot}$. Let D_i be a set of information of the time interval I_i which can be represented as the set of numbers of deaths in previous time intervals including I_i and let $D_{i-1(j)}$ be a set containing D_{i-1} and $y_{i1},\cdots, y_{ij}$, where y_{ij} is the observed number of deaths of individuals of the covariate vector Z_j in the time interval I_i. Likelihood L_i of $(\theta_{i1},\cdots,\theta_{iJ})$ in time interval I_i is obtained as

$$L_i(\theta|y_{i\cdot}, D_{i-1}) = \prod_{j=1}^{J} \binom{n_{ij}}{y_{ij}} \theta_{i(j)}^{y_{ij}}(1-\theta_{i(j)})^{n_{ij}-y_{ij}}.$$

Likelihood L_i is used to update prior distributions of β_i which is linearly related to $\theta_{i(j)}$ by $\theta_{i(j)} = Z_j\beta_i$ in each time interval I_i. Here we define the perturbation of the time interval as a set of perturbations of each parameters of a parameter vector. We define the number of perturbations of a parameter vector as the possible number of combinations of parameters of a parameter vector. We assume that the model selection probability of the time interval I_i, $\pi_i^{(\cdot)}$, is fixed prior to obtaining any informations from individuals alive in the time interval I_i. Let α_i be the perturbation index variable of interval I_i then

$$\pi_i^{(k)} = P(\alpha_i = k|D_l) \text{ for } k = 1,\cdots,K, \quad l = 0,\cdots,i-1.$$

Then the multiprocess dynamic survival model is defined as follows.

i) Observation equation :

$$Y_{ij} \sim Binomial(n_{ij}, \theta_{i(j)}) \text{ for } i = 1,\cdots,s, \ j = 1,\cdots,J.$$

ii) Guide relationship:

$$\theta_{i(j)} = Z_j\beta_i \quad \text{for} \quad i = 1,\cdots,s, \ j = 1,\cdots,J.$$

iii) Evolution equation:

$$\beta_i = G_i\beta_{i-1} + w_i \quad \text{for} \quad i = 1,\cdots,s,$$

where G_i is a known transition matrix of the time interval I_i, and w_i is the evolution error vector whose distribution is specified by mean vector

0 and the variance-covariance matrix $W_i^{(k)}$ which depends on the value of the perturbation index variable of the time interval I_i, $\alpha_i = k$, for $k = 1, \cdots, K$, and is independent of the distribution of the parameter vector of the previous time interval.

3. RECURSIVE ESTIMATION OF A PARAMETER VECTOR

The process is started with the initial distribution of a parameter vector at time 0, β_0, specified by mean vector $\widehat{\beta}_0$ and variance-covariance matrix V_0, where $\widehat{\beta}_0$ and V_0 are given prior to time interval I_1, which do not affect distributional behaviors of the parameter vector in future time intervals for a certain extent of time elapsing.

At the beginning of each time interval I_i, each of K posterior distributions of β_{i-1} obtained in time interval I_{i-1} leads K prior distributions of β_i which is specified by mean vector $a_i^{(kl)}$ and variance-covariance matrix $R_i^{(kl)}$ as

$$(\beta_i | alpha_{i-1} = k, alpha_i = l, D_{i-1}) sim[a_i^{(kl)}, R_i^{(kl)}],$$

where

$$a_i^{(kl)} = G_i \widehat{\beta}_{i-1}^{(k)}, \;\; R_i^{(kl)} = G_i V_{i-1}^{(k)} G_i' + W_i^{(l)}.$$

With informations from first (j-1) observations, the joint prior distribution of β_i and $\theta_{i(j)}$ is obtained by the guide relationship,

$$\left(\begin{matrix} \beta_i \\ \theta_{i(j)} \end{matrix} \mid \alpha_{i-1} = k, alpha_i = l, D_{i-1(j-1)} \right) \sim \left[\begin{pmatrix} a_{ij}^{(kl)} \\ f_{ij}^{(kl)} \end{pmatrix}, \begin{pmatrix} R_{ij}^{(kl)} & S_{ij}^{(kl)} \\ S_{ij}^{(kl)\prime} & q_{ij}^{(kl)} \end{pmatrix} \right] \tag{3.1}$$

where

$$f_{ij}^{(kl)} = Z_{ij} a_{ij}^{(kl)}, \;\; S_{ij}^{(kl)} = Z_{ij} R_{ij}^{(kl)} \;\text{ and }\; q_{ij}^{(kl)} = S_{ij}^{(kl)} Z_{ij}',$$

with $a_{i1}^{(kl)} = a_i^{(kl)}$ and $R_{i1}^{(kl)} = R_i^{(kl)}$. Here the prior distribution of $\theta_{i(j)}$ is assumed to be a conjugate beta distribution $(b_{ij}^{(kl)}, r_{ij}^{(kl)})$, where $b_{ij}^{(kl)}$ and $r_{ij}^{(kl)}$ are estimated in terms of the mean and the variance of the distribution of $\theta_{i(j)}$, such as, respectively,

$$q_{ij}^{(kl)^{-1}} f_{ij}^{(kl)^2} (1 - f_{ij}^{(kl)}) - f_{ij}^{(kl)},$$
$$q_{ij}^{(kl)^{-1}} f_{ij}^{(kl)} (1 - f_{ij}^{(kl)})^2 - (1 - f_{ij}^{(kl)}).$$

With information from the *j-th* observation the posterior distribution of $\theta_{i(j)}$ is obtained as

$$(\theta_{i(j)}|\alpha_{i-1}=k,\alpha_i=l,D_{i-1(j)}) \ \sim \ Beta(b_{ij}^{(kl)}+y_{ij},r_{ij}^{(kl)}+n_{ij}-y_{ij}).(3.2)$$

Using (3.2) and applying the linear Bayes estimation on (3.1) the updated distribution of β_i given $D_{i-1(j)}$ is obtained as

$$(\beta_i|\alpha_{i-1}=k,\alpha_i=l,D_{i-1(j)}) \sim [\widehat{\beta}_{ij}^{(kl)},V_{ij}^{(kl)}],$$

where

$$\begin{aligned}
\widehat{\beta}_{ij}^{(kl)} &= a_{ij}^{(kl)} + S_{ij}^{(kl)} q_{ij}^{(kl)^{-1}} \left(\frac{b_{ij}^{(kl)}+y_{ij}}{b_{ij}^{(kl)}+r_{ij}^{(kl)}+n_{ij}} - f_{ij}^{(kl)} \right), \\
V_{ij}^{(kl)} &= R_i^{(kl)} - S_i^{(kl)} S_i^{(kl)'} q_{ij}^{(kl)^{-1}} \\
&\quad + S_i^{(kl)} S_i^{(kl)'} q_{ij}^{(kl)^{-2}} \left(\frac{(b_{ij}^{(kl)}+y_{ij})(r_{ij}^{(kl)}+n_{ij}-y_{ij})}{(b_{ij}^{(kl)}+r_{ij}^{(kl)}+n_{ij})^2(b_{ij}^{(kl)}+r_{ij}^{(kl)}+n_{ij}+1)} \right).
\end{aligned}$$

Since there is no parametric evolution in each time interval, the joint prior distribution of β_i and $\theta_{i(j+1)}$ is given as

$$\left(\begin{array}{c} \beta_i \\ \theta_{i(j+1)} \end{array} \mid \alpha_{i-1}=k,\alpha_i=l,D_{i-1(j)} \right) \ \sim \ \left[\left(\begin{array}{c} a_{i,j+1}^{(kl)} \\ f_{i,j+1}^{(kl)} \end{array} \right), \left(\begin{array}{cc} R_{i,j+1}^{(kl)} & S_{i,j+1}^{(kl)} \\ S_{i,j+1}^{(kl)'} & q_{i,j+1}^{(kl)} \end{array} \right) \right],$$

where

$$\begin{aligned}
a_{i,j+1}^{(kl)} &= \widehat{\beta}_{ij}^{(kl)}, f_{i,j+1}^{(kl)} = Z_{i,j+1} a_{i,j+1}^{(kl)} \\
R_{i,j+1}^{(kl)} &= V_{ij}^{(kl)}, S_{i,j+1}^{(kl)} = Z_{i,j+1} R_{i,j+1}^{(kl)} \\
q_{i,j+1}^{(kl)} &= S_{i,j+1}^{(kl)} Z'_{i,j+1}.
\end{aligned}$$

When all observations in time interval I_i are processed K^2 posterior distributions of β_i are obtained as

$$(\beta_i|\alpha_{i-1}=k,\alpha_i=l,D_i) \sim [\widehat{\beta}_i^{(kl)},V_i^{(kl)}],$$

where

$$D_i = D_{i-1(J)}, \quad \widehat{\beta}_i^{(kl)} = \widehat{\beta}_{iJ}^{(kl)} \quad \text{and} \quad V_i^{(kl)} = V_{iJ}^{(kl)}.$$

Here the posterior distribution of $(\beta_i|\alpha_i=l,D_i)$ is represented as the mixture of K posterior distributions of $(\beta_i|\alpha_{i-1}=k,\alpha_i=l,D_i)$ with the posterior index probability $p_i^{(kl)}$. Using that

$$\begin{aligned} & p(y_{i\cdot}|\alpha_{i-1}=k, \alpha_i = l, D_{i-1}) \\ = & \prod_{j=1}^{J} \binom{n_{ij}}{y_{ij}} \frac{\Gamma(b_{ij}^{(kl)}+y_{ij})\Gamma(r_{ij}^{(kl)}+n_{ij}-y_{ij})}{\Gamma(b_{ij}^{(kl)}+r_{ij}^{(kl)}+n_{ij})}. \end{aligned}$$

the posterior index probability is obtained as

$$\begin{aligned} p_i^{(kl)} &= P(\alpha_{i-1}=k, \alpha_i = l|D_i) \\ &\propto p(y_{i1}, \cdots, y_{iJ}|\alpha_{i-1}=k, \alpha_i=l, D_{i-1}) p_{i-1}^{(k)} \pi_i^{(l)}, \end{aligned}$$

where $p_{i-1}^{(k)} = P(\alpha_{i-1}=k|D_{i-1})$. Thus the posterior distribution of β_i given $\alpha_i = l$ and D_i is specified by mean vector $\widehat{\beta}_i^{(l)}$ and variance-covariance matrix $V_i^{(l)}$, where

$$\begin{aligned} \widehat{\beta}_i^{(l)} &= \sum_{k=1}^{K} \widehat{\beta}_i^{(kl)} p_i^{(kl)} / p_i^{(l)}, \\ V_i^{(l)} &= \sum_{k=1}^{K} [V_i^{(kl)} + (\widehat{\beta}_i^{(l)} - \widehat{\beta}_i^{(kl)})(\widehat{\beta}_i^{(l)} - \widehat{\beta}_i^{(kl)})'] p_i^{(kl)} / p_i^{(l)}. \end{aligned}$$

And the distribution of β_i given D_i is specified by mean vector $\widehat{\beta}_i$ and variance-covariance matrix V_i, where

$$\begin{aligned} \widehat{\beta}_i &= \sum_{l=1}^{K} \widehat{\beta}_i^{(l)} p_i^{(l)}, \\ V_i &= \sum_{l=1}^{K} [V_i^{(l)} + (\widehat{\beta}_i - \widehat{\beta}_i^{(l)})(\widehat{\beta}_i - \widehat{\beta}_i^{(l)})'] p_i^{(l)}. \end{aligned}$$

4. FORECASTING OF THE NUMBER OF DEATHS

In this section we obtain the forecasted value of $Y_{i+1,j}$ in terms of the value of $E[Y_{i+1,j}|D_i]$, $j = 1, \cdots, J$, based on informations gathered until the end of time interval I_i.

At the beginning of time interval I_{i+1}, we obtain K prior distributions of β_{i+1} from each of K posterior distributions of β_i through the evolution equation, which are

$$(\beta_{i+1}|\alpha_i = k, \alpha_{i+1} = l, D_i) \sim [a_{i+1}^{(kl)}, R_{i+1}^{(kl)}], \quad k, l = 1, \cdots, K.$$

By the evolution equation and the guide relationship, the joint prior distribution of β_{i+1} and $\theta_{i+1(j)}$, $j = 1, \cdots, J$, is obtained as

$$\left(\begin{array}{c} \beta_{i+1} \\ \theta_{i+1(j)} \end{array} \mid \alpha_i = k, \alpha_{i+1} = l, D_i \right) \sim \left[\left(\begin{array}{c} a_{i+1}^{(kl)} \\ f_{i+1,j}^{(kl)} \end{array} \right), \left(\begin{array}{cc} R_{i+1}^{(kl)} & S_{i+1,j}^{(kl)} \\ S_{i+1,j}^{(kl)'} & q_{i+1,j}^{(kl)} \end{array} \right) \right], \quad (4.1)$$

where

$$f_{i+1,j}^{(kl)} = Z_j a_{i+1}^{(kl)}, \quad S_{i+1,j}^{(kl)} = Z_j R_{i+1}^{(kl)}, \quad q_{i+1,j}^{(kl)} = S_{i+1,j}^{(kl)} Z_j'.$$

Here the prior distribution of $\theta_{i+1(j)}$ is assumed to be a conjugate beta distribution $(b_{i+1,j}^{(kl)}, r_{i+1,j}^{(kl)})$. Note that $b_{i+1,j}^{(kl)}$ and $r_{i+1,j}^{(kl)}$ are estimated to be expressed in the mean and the variance of the distribution of $\theta_{i+1(j)}$ in (4.1) such as, respectively,

$$\begin{array}{l} q_{i+1,j}^{(kl)^{-1}} f_{i+1,j}^{(kl)^2}(1 - f_{i+1,j}^{(kl)}) - f_{i+1,j}^{(kl)}, \\ q_{i+1,j}^{(kl)^{-1}} f_{i+1,j}^{(kl)}(1 - f_{i+1,j}^{(kl)})^2 - (1 - f_{i+1,j}^{(kl)}). \end{array}$$

Thus the distribution of $(Y_{i+1,j} | \alpha_i = k, \alpha_{i+1} = l, D_i)$ is specified by the mean $\mu_{i+1,j}^{(kl)}$ and the variance $V_{i+1,j}^{(kl)}$ as, respectively,

$$\begin{aligned} \mu_{i+1,j}^{(kl)} &= \frac{n_{i+1,j} b_{i+1,j}^{(kl)}}{b_{i+1,j}^{(kl)} + r_{i+1,j}^{(kl)}}, \\ V_{i+1,j}^{(kl)} &= \frac{n_{i+1,j} b_{i+1,j}^{(kl)} r_{i+1,j}^{(kl)}}{(b_{i+1,j}^{(kl)} + r_{i+1,j}^{(kl)})(b_{i+1,j}^{(kl)} + r_{i+1,j}^{(kl)} + 1)} \\ &+ \frac{n_{i+1,j}^2 b_{i+1,j}^{(kl)} r_{i+1,j}^{(kl)}}{(b_{i+1,j}^{(kl)} + r_{i+1,j}^{(kl)})^2 (b_{i+1,j}^{(kl)} + r_{i+1,j}^{(kl)} + 1)}. \end{aligned}$$

By collapsing we obtain the forecasted distribution of $Y_{i+1,j}$ given D_i,

$$(Y_{i+1,j} | D_i) \sim [\mu_{i+1,j}, V_{i+1,j}], \quad i = 1, \cdots, s-1, \quad (4.2)$$

where

$$\begin{aligned} \mu_{i+1,j} &= \sum_{k,l=1}^{K} \mu_{i+1,j}^{(kl)} p_i^{(k)} \pi_{i+1}^{(l)}, \\ V_{i+1,j} &= \sum_{k,l=1}^{K} [V_{i+1,j}^{(kl)} + (\mu_{i+1,j} - \mu_{i+1,j}^{(kl)})(\mu_{i+1,j} - \mu_{i+1,j}^{(kl)})'] p_i^{(k)} \pi_{i+1}^{(l)}. \end{aligned}$$

The mean of $Y_{i+1,j}$ given D_i, $\mu_{i+1,j}$, is used as the forecasted value of $Y_{i+1,j}$, $j = 1, \cdots, J$, in time interval I_{i+1}.

5. ILLUSTRATIONS

In this section we consider the performance of the estimation and the forecasting proposed in Section 3 and 4 through the Monte Carlo simulation studies. The data consists of numbers of deaths of individuals in 50 time intervals, where individuals are divided into two groups. The first part of the data which is assumed to be the set of numbers of deaths of the group 1 consists of 50 simulated random samples of the starting size 200, from the population of the failure proportion is 0.09 in each time interval. The other part of the data which is assumed to be the set of numbers of deaths of individuals of the group 2 consists of 50 simulated random samples of the starting size 200, from the population where the failure proportion is 0.05 for the first 30 samples, increases by 0.03 in each of next 5 samples, and remains at 0.2 in last 15 samples. Thus the sampling distribution of the number of deaths of individuals of the group j in the time interval I_i is a binomial distribution such as,

$$Y_{ij} \sim Binomial(n_{ij}, \theta_{i(j)}), \text{ for } i = 1, \cdots, 51, \ j = 1, 2,$$

where n_{ij} is the number of individuals of the group j alive at the beginning of the time interval I_i with n_{1j}=200. In the multiprocess dynamic survival model we proposed, the failure proportion of the individuals of the group j alive at the beginning of the time interval I_i assumed to be the linear function of the parameter vector β_i,

$$\theta_{i(j)} = Z_j \beta_i$$

with Z_j=$(1,z_j)$ and β_i=$(\beta_{0i},\beta_{1i})'$, where z_j=0 for individuals of the group 1 and z_j=1 for individuals of the group 2 alive at the beginning of the time interval I_i, β_{1i} is the parameter modeling the effect of the difference of the group 2 from the group 1 on the survival pattern. We assume that there are two perturbations for the parameter β_{1i} in each time interval, steady change and sudden slope change numbered by 1 and 2 respectively. The model selection probability and the transition matrix in each time interval I_i is assumed by

$$\pi_i^{(1)} = 0.9, \quad \pi_i^{(2)} = 0.1, \quad G_i = \mathbf{I}_2 \ \text{ for } \ i = 1, \cdots, 50.$$

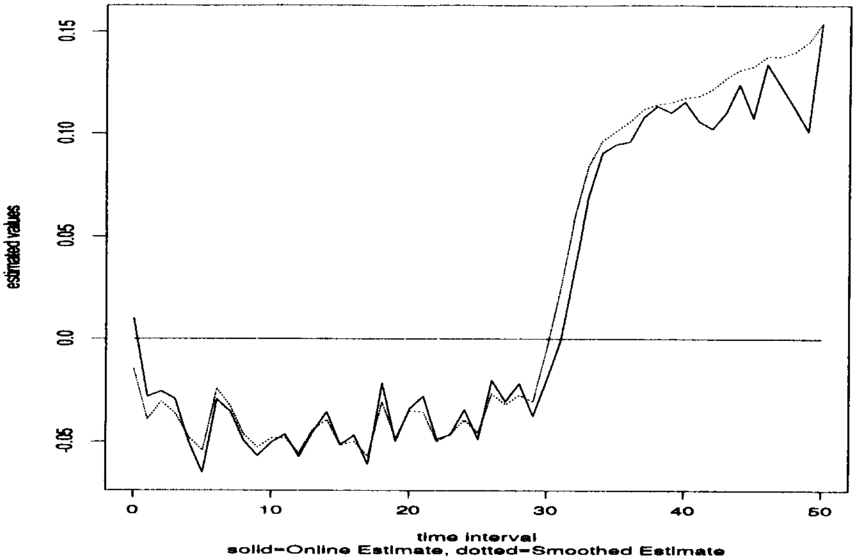

Figure 1. Estimated Mean of The Parameter β_{1i}

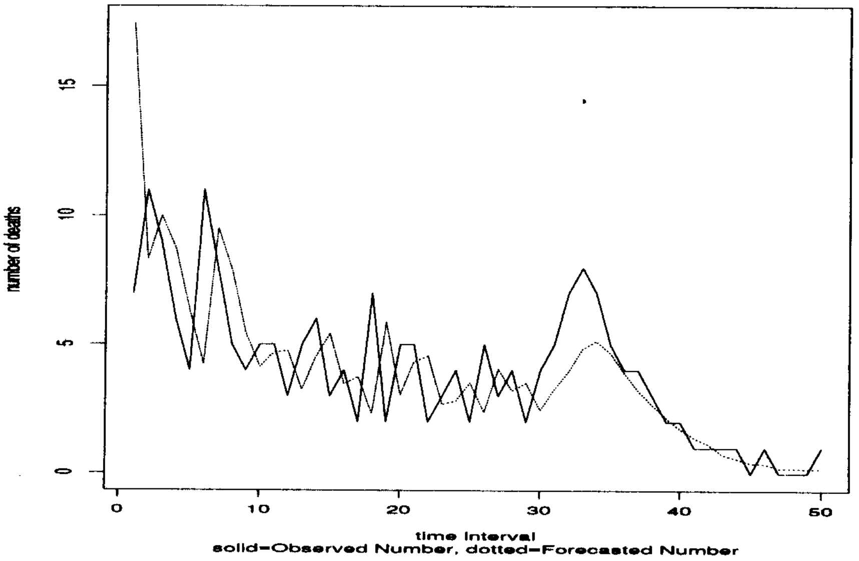

Figure 2. Forecasted Number of Deaths of Group 2

We start the analysis with the prior distribution,

$$(\beta_0|D_0) \sim [0.01\mathbf{1}, 0.01\mathbf{I}_2],$$

and the variance-covariance matrix of the evolution error vector by

$$W_i^{(1)} = \begin{pmatrix} 0 & 0 \\ 0 & 0.001 \end{pmatrix}, \quad W_i^{(2)} = \begin{pmatrix} 0 & 0 \\ 0 & 0.01 \end{pmatrix}.$$

Figure 1 shows the considerable time variation of the mean of the parameter β_{1i} for the effect of the difference of group 2 from the group 1. We notice that, the mean of the parameter retains steady values around -0.05 nearly until the time interval I_{30}, gets sudden increases to nearly 0.1 in next 5 time intervals and slow increases in last 15 time intervals. Figure 2 shows the forecasted number of deaths and the observed number of deaths of individuals of the group 2 in the next time interval. In figures one can see that the multiprocess dynamic survival model gives quick responses to real changes but it is not quite sensitive to abrupt changes.

REFERENCES

(1) Bolstad, W. M. (1988), Estimation in the Multiprocess Dynamic Generalized Linear Model, *Communication Statistics: Theory and Method*, **17**(12), 4179-4204.

(2) Cox, D. R. (1972), Regression Models and Life-Tables(with discussions), *Journal of the Royal Statistical Society* B, **34**, 187-220.

(3) Gamerman, D. (1991), Dynamic Bayesian Models for Survival Data, *Applied Statistics*, **40**, 63-79.

(4) Harrison, P. J. and Stevens, C. F. (1976), Bayesian Forecasting(with discussions), *Journal of the Royal Statistical Society* B, **38**, 205-247.

(5) Stablein, D. M., Carter, W. H. and Novak, J. W. (1981). Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials*, **2**, 149-159.

(6) West, M., Harrison, P. J. and Migon, H. S. (1985), Dynamic Generalized Linear Models and Bayesian Forecasting, *Journal of the American Statistical Association*, **80**, 73-97.