IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Prediction of age-specific cancer mortality by using multinomial time series model

Seongyong Kim[1], Saebom Jeon[2]

Assistant Professor, Division of Global Management Engineering, Hoseo University, Asan, Korea [1]

Assistant Professor, Department of Marketing Information Consulting, Mokwon University, Daejeon, Korea[2]

Abstract: Cancer is the largest cause of death in Korea, and its proportion is increasing. Meanwhile, the cancer mortality rates vary over time as well as age. With the increased life expectancy in Korea, the proportion of the elderly age among cancer deaths has increased over time, while that of the young age has decreased. To reflect the proportions of the categories with such dynamic structures of age and time, a multinomial time series model can be used as a prediction model. However, there is a difficulty in estimating the parameters through the Markov Chain Monte Carlo (MCMC) method when some cell counts are very small relative to others, such as the number of deaths from cancer of young age group. In order to predict the age-specific cancer mortality by reflecting its dynamic structure and by overcoming estimation problems in MCMC, a power transformation is adopted as a link function of multinomial time series model instead of a logit link function, and forecasts the age-specific cancer mortality of male in Korea by 2040 using the proposed method.

Keywords: multinomial time series model, MCMC, link function, power transformation.

I. INTRODUCTION

Cancer is the largest cause of death in Korea. In 2005, the percentage of male deaths from cancer is 30.63%, and 21.78% for women. As shown in Table 1, the cancer mortality rates vary over time as well as age groups. The age-specific cancer mortality under 40 years has been decreased, while that over 40 years has been increased. It implies that, in order to establish future medical policy or welfare policy, it is necessary to investigate the change of the number of deaths over time not only by disease but also by age.

TABLE 1 AGE-SPECIFIC CANCER MORTALITY OF MALE

Year	0-19	20-39	40-59	60-79	80+
1997	1.3%	4.5%	32.7%	56.1%	5.5%
1998	1.2%	4.4%	32.2%	55.8%	6.6%
1999	1.1%	3.7%	31.5%	57.5%	6.2%
2000	1.0%	3.8%	30.3%	56.6%	8.3%
2001	0.9%	3.4%	27.1%	59.6%	9.2%
2002	0.7%	3.7%	27.5%	58.1%	10.0%
2003	0.7%	2.7%	26.1%	59.7%	10.8%
2004	0.5%	2.8%	27.3%	58.4%	11.0%
2005	0.6%	2.8%	24.9%	59.6%	12.1%
2006	0.5%	2.3%	24.9%	59.2%	13.1%
2007	0.7%	1.8%	23.7%	60.7%	13.2%

The most simple and easy way to predict the age-specific deaths from cancer is univariate time series model for each age group. However, it leads to an internal discrepancy problem that the summation of estimated number of deaths from the each univariate time series model for each age group does not match to the total number of deaths [1][2]. Moreover, for the categories with small cell counts, the univariate model is well-known to be inappropriate because of the normality assumption [3][4].

To overcome these problems, [5], [4] and [6] proposed a multinomial time series model. They assume that cell counts in categories follow multinomial distribution with a fixed total sum for each time, and the logit link functions transformed from the cell probabilities have dynamic structures. [4] Proposed a multinomial time series model in which a vector of logit link functions follows a linear model and the coefficient vector of this model has a first order

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

autocorrelation. [6] suggested the unit root model for the mean of the vectors of logit link functions and used the MCMC method to estimate the parameters of each model.
However, for the categories with a small cell counts such as a small number of cancer deaths for young age groups, the asymptotic variance of the logit link function is too large which makes impossible to calculate the posterior distribution in MCMC [7][8]. In this respect, we first aim to forecast the age-specific cancer mortality among males in Korea using the multinomial time series model to reflect the dynamic structure of cancer mortality over time and age groups, and second aim to use a new link function to overcome the estimation problem using the power transformation [9]. Finally, we forecast the 10 year age-specific cancer mortality by 2040 in Korea. In Section 2, we introduce a multinomial time series model using a power transformation as a link function. In Section 3, the age-specific cancer mortality in Korea are forecasted. Section 4 includes conclusion remarks.

II. MULTINOMIAL TIME SERIES MODEL

Let the total number of age group be I, and total period of observation be T. We also let the number of deaths belonging to ith age group be y_{ti} and total number of deaths be N_t at time t. Then, we assume that y_{ti} follow multinomial distribution at t given N_t. That is, for $t = 1,2,\cdots,T$,

$$\boldsymbol{y_t} \sim Multinomial(N_t, \boldsymbol{\pi_t})$$

where $\boldsymbol{y_t} = (y_{t1},\cdots,y_{tI})'$ and $\boldsymbol{\pi_t} = (\pi_{t1},\cdots,\pi_{tI})'$. To incorporate covariates with π_{ti} as a linear model, a power link [10] function given by

$$\eta_{ti} = \frac{(\frac{\pi_{ti}}{\pi_{tI}})^{\alpha_i} - 1}{\alpha_i}$$

is considered as a link function where α_i is a power. Note that [9] suggested to set the category with the largest number as the reference category in order to reduce the asymptotic variance of a link function. They showed that this make the implementation of MCMC possible by simulation studies. For a power transformation link function, a linear model is assumed as following.

$$\begin{aligned}\eta_{ti} &= \boldsymbol{x_{ti}'}\boldsymbol{\beta_i} + \epsilon_{ti} \\ &= \beta_{i,0} + \beta_{i,1}time + \sum_{j=1}^{q}\beta_{i,j+1}\eta_{t-j,i} + \epsilon_{ti}, \\ \epsilon_{ti} &\sim N(0,\psi_i)\end{aligned}$$

where time is time effect which have same values with indicator t, $\boldsymbol{x_{ti}}$ is a $p(=q+2)$ dimensional covariate vector, and $\boldsymbol{\beta_i}$ is a p dimensional coefficient vector. We assume that ψ_i is a known scalar, and diffuse prior for $\boldsymbol{\beta_i}$ as a hyper prior distribution. Then, we have a following full posterior distribution,

$$L_{pos} \propto \prod_{t=1}^{T}\prod_{i=1}^{I}\pi_{ti}^{y_{ti}}\prod_{t=1}^{T}|\boldsymbol{\Psi}|^{-0.5}exp[-0.5(\boldsymbol{\eta_t} - \boldsymbol{X_t\beta})'\boldsymbol{\Psi}^{-1}(\boldsymbol{\eta_t} - \boldsymbol{X_t\beta})]$$

where $\boldsymbol{X_t} = diag(\boldsymbol{x_{t1}'},\cdots,\boldsymbol{x_{t,I-1}'})$, $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_{I-1})$, and $\boldsymbol{\Psi}$ is a diagonal matrix whose ith element is ψ_i. For inference of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}_t$ from the above full posterior distribution, Gibbs sampling is implemented. Each step for Gibbs sampling is provided as follows:

Step 1: Sample $\boldsymbol{\eta}_1,\cdots,\boldsymbol{\eta}_t$ successively from each conditional posterior distribution of $\boldsymbol{\eta}_t$ given $\boldsymbol{\eta}_{t-1},\cdots,\boldsymbol{\eta}_{t-q},\boldsymbol{\beta},\boldsymbol{y_t}$.
Step 2: Sample $\boldsymbol{\beta}$ from the conditional posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\eta}_1,\cdots,\boldsymbol{\eta}_t$.
Step 3: Repeat Step 1 and Step 2 until samples are converged.

The conditional posterior distribution of $\boldsymbol{\eta}_t$ is

$$\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1},\cdots,\boldsymbol{\eta}_{t-q},\boldsymbol{\beta},\boldsymbol{y_t} \propto \prod_{i=1}^{I}\pi_{ti}^{y_{ti}}\prod_{t=1}^{T}|\boldsymbol{\Psi}|^{-0.5}exp[-0.5(\boldsymbol{\eta_t} - \boldsymbol{X_t\beta})'\boldsymbol{\Psi}^{-1}(\boldsymbol{\eta_t} - \boldsymbol{X_t\beta})].$$

Since the above distribution is not a standard form of distributions, Metropolis-Hastings algorithm is implemented [11]. For a proposal distribution of Metropolis-Hastings algorithm, normal approximated multinomial likelihood as described in [4]. Note that the sampled η_{ti} from Metropolis-Hastings algorithm can be transformed to cell probability as following.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

$$\pi_{ti} = \begin{cases} \dfrac{(\alpha_i \eta_{ti} + 1)^{1/\alpha_i}}{1 + \sum_{i=1}^{I-1} (\alpha_i \eta_{ti} + 1)^{1/\alpha_i}} & i \neq I \\ \dfrac{1}{1 + \sum_{i=1}^{I-1} (\alpha_i \eta_{ti} + 1)^{1/\alpha_i}} & i = I \end{cases}$$

By multiplying sampled π_{ti} s by a given N_t, the estimates of y_{ti} s can be calculated. The conditional posterior distribution of $\boldsymbol{\beta}$ is multivariate normal distribution as following.

$$\boldsymbol{\beta} | \boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_t \sim N(\mu^*, \Sigma^*)$$

where

$$\Sigma^* = (\sum_t X_t' \Psi^{-1} X_t)^{-1}, \mu^* = \Sigma^* \sum_t X_t' \Psi^{-1} \eta_t.$$

By using the above conditional posterior distributions and Gibbs sampling, the age-specific cancer mortality rates are forecasted in the following section.

III.FORECASTING OF AGE-SPECIFIC DEATHS FROM CANCER IN KOREA

To forecast the age-specific cancer mortality, we use monthly data from January 1997 to December 2007, by 10-year age groups. Linear models to incorporate covariates to power link functions are given by

$$\eta_{ti} = \beta_{i,0} + \beta_{i,1} \text{time} + \beta_{i,2} \eta_{t-1,i} + \beta_{i,2} \eta_{t-2,i} + \epsilon_{ti}.$$

Note that the reference age group is 60-69 years whose mortality rate is the highest. The power α_i's are set to be 0.5 for age groups below 40 years and above 90 years, but those for other age groups are set to be near 0 which is equivalent to a logit link function.

In this paper, 10,000 iterations of Gibbs samplings are implemented, and the convergence is evaluated by autocorrelations of Gibbs samples, Gelman-Rubin's PSRF (potential scale reduction factor), and acceptance rates of samples from a proposal distribution in Metropolis-Hastings algorithm [12]. After performing Gibbs sampling, autocorrelations of β_{ij}'s were between 0.2 and 0.4 for all i and j, and the acceptance rates of η_{ti}'s were also appropriate because all rates were between 25% and 50% for all t and i [8]. The PSRF's for all β_{ij}'s also showed values less than 1.2, implying Gibbs samplings were well performed.

Table 2 presents the 95% confidence interval and median of the values obtained from the MCMC sample. Note that the numbers in parenthesis are a lower limit and upper limit. Here, the medians in all categories under 65 are negative for the time effect, meaning that the cancer mortality rates under 60 have been decreased over time relative to the baseline age 60-69 years. On the other hand, the medians in all categories over 70 are positive, meaning that the cancer mortality rate over 70 have been increased relative to the baseline age 60-69 years. In particular, the 80s are expected to have a much higher mortality rate than other age groups.

TABLE 2 MEDIAN AND 95% CONFIDENCE INTERVAL OF B_{I0} AND B_{I1}

Age	β_{i0}			β_{i1}		
	Median	**Lower limit**	**Upper limit**	**Median**	**Lower limit**	**Upper limit**
0-9	-1.85415	-2.5337	-1.1578	-0.00055	-0.0011	-0.00015
10-19	-1.6559	-2.3503	-0.9865	-0.00065	-0.00115	-0.00015
20-29	-1.68195	-2.35255	-1.02525	-0.0011	-0.00175	-0.00055
30-39	-1.3223	-1.83965	-0.7933	-0.0021	-0.00305	-0.00115
40-49	-1.0941	-1.5534	-0.64505	-0.0032	-0.0049	-0.0016
50-59	-0.33715	-0.4866	-0.18945	-0.0019	-0.00325	-0.0005
70-79	-0.27315	-0.4067	-0.1453	0.0011	0.0004	0.0019
80-89	-1.7936	-2.55785	-1.03535	0.0056	0.00305	0.00825
90-99	-1.6243	-2.37645	-0.88335	0.00085	0.0003	0.0014
100+	-1.9913	-3.0647	-0.8934	0	-0.0005	0.0005

Table 3 presents the predictive distribution of the 10-year age-specific cancer mortality rate of male in Korea by2040. As shown in Table 3, the age-specific cancer mortality rates under 70 years are decreasing over time, while those over 70 year are increasing over time. In particular, the mortality rate in 2030 for 80-89 years is expected to be about four times of that in 2005. While the mortality rate for 80-89 is expected to increase explosively, that over 90 is not expected

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

to increase significantly. This differs from the expectation that the mortality rate over 90 years will also increase due to the advancement of medical technology and the increase in life expectancy. Such results seem that the dynamic structures are not reflected over the age of 90 years, due to its small number of cancer deaths.

TABLE 3 AGE-SPECIFIC CANCER MORTALITY OF MALE

Year	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-100	100+
2000	0.3%	0.7%	0.9%	3.0%	8.9%	21.2%	31.5%	24.9%	7.8%	0.4%	0.0%
2005	0.3%	0.4%	0.6%	2.1%	8.2%	16.9%	30.8%	28.7%	11.0%	1.0%	0.0%
2010	0.2%	0.2%	0.4%	1.1%	7.3%	14.2%	27.9%	31.8%	15.1%	1.0%	0.0%
2020	0.1%	0.1%	0.0%	0.1%	5.0%	8.8%	21.8%	34.4%	26.7%	1.7%	0.0%
2030	0.0%	0.0%	0.0%	0.0%	3.1%	4.8%	15.1%	31.6%	42.0%	2.0%	0.0%
2040	0.0%	0.0%	0.0%	0.0%	1.7%	2.4%	10.2%	30.7%	52.0%	3.0%	0.0%

IV. CONCLUSION

In this paper, we applied a multinomial time series model using power transforms as a link function to the age-specific cancer mortality, in order to not only ensure the internal consistency but also overcome the difficulty of estimation in MCMC. The predicted cancer mortality rates for old age groups are expected to be increasing, while those for younger ages are decreasing. However, the proposed multinomial time series model can be used to under the given total number at each time. Therefore, if the total number in future - such as the total number of cancer deaths in the future - is unknown, this model cannot be used to forecast the future number of cancer deaths for each age group. In order to solve this problem, it is necessary to incorporate the model for predicting total cancer deaths in the future into the multinomial time series model, as suggested by [1][2]. Furthermore, as indicated by the predicted results of the age-specific mortality rate in Section 3, the underestimated prediction results over 90 years old, should be considered to improve in the future work.

REFERENCES

[1] Y. Park, J. W. Choi, and H. Y. Kim, Forecasting Cause-age Specific Mortality using Two Random Processes. Journal of the American Statistical Association, vol. 101, pp. 472-483, 2006a.
[2] Y. Park, J. W. Choi, and D. H. Lee, A Parametric Approach for Measuring the Effect of the 10th Revision of the International Classification of Diseases. Journal of the Royal statistical Society, Series C, vol. 55, pp. 677-697, 2006b.
[3] A. Agresti. Categorical Data Analysis. 2nd ed., New York: John Wiley & Sons, Inc, 2002.
[4] C. Cargoni, R. Muller, and M. West, Bayesian Forecasting of Multinomial Time Series through Condtional Gaussian Dynamic Models. Journal of the American Statistical Association, vol. 92, pp. 640-647, 1997.
[5] M. West, and P. J. Harrison, Bayesian Forecasting and Dynamic Models. New York: Springer-Verlag, 1989.
[6] C. Reilly, A. Gelman, and J. Katz, Poststratification without Population Level Information on the Poststratifying Variable, with Application to Political Polling. Journal of the American Statistical Association, vol. 96, pp. 1-11, 2001.
[7] C. J. Geyer, Practical Markov Chain Monte Carlo (with discussion). Statistical Science, vol. 7, pp. 473-511, 1992.
[8] S. M. Lynch, Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. New York: Springer Science+Business Media, LLC, 2007.
[9] S. Kim, A Power Transformation Method for Categorical Data. Korea University, 2011.
[10] G. E. P. Box, and D. R. Cox, An analysis of transformations. Journal of the Royal Statistical Society, Series B, vol. 26, pp. 211–252, 1964.
[11] L. Tierney, Markov Chains for Exploring Posterior Distributions. The Annals of Statistics, vol. 22, pp. 1701-1728, 1994.
[12] A. Gelman, J. P. Carlin, H. S. Stern, and D. B. Rubin, Bayesian Data Analysis. 2nd Edition. New York: Chapman and Hall/CRC, 2004.