Journal of the Korean Data & Information Science Society
2009, **20**(2), 411–423

한국데이터정보과학회지

삼각분할표 자료에서 베이지안 모형을 이용한 예측

이주미[1] · 임요한[2] · 한규섭[3] · 이경은[4]

[1]경북대학교 병원 임상시험센터 · [2]서울대학교 통계학과
[3]연세대학교 언더우드 국제학부 · [4]경북대학교 통계학과

접수 2009년 1월 22일, 수정 2009년 3월 19일, 게재확정 2009년 3월 23일

요 약

본 논문은 삼각 분할표 자료의 예측문제에 있어 Verrall (1990)의 발생연도효과와 경과년도효과만 있는 베이지안 선형모형을 절대연도효과가 있는 모형으로 확장한 모형을 제시하고 이에 대한 추정 방법으로 마르코프 연쇄 몬테칼로 방법을 제안한다. 제안된 모형과 추정 방법은 세 가지 실제 예를 통하여 기존의 방법들에 비해서 일반적으로 작은 상대 예측오차를 제공함을 보였다.

주요용어: 마르코프 연쇄 몬테칼로 방법, 베이지안 선형모형, 삼각분할표 자료.

1. 머리말

우리가 접하는 시계열 또는 생존시간 자료는 많은 경우 절대시간 (또는 절대년도)와 경과시간 (또는 경과년도)의 두 가지 시간 축을 지니고 있다. 여기서 절대시간은 물리학적인 현재의 시간을 의미하고 경과시간이란 어떤 사건이 발생한 발생시간 (또는 발생연도)이후 현재까지 경과한 시간을 의미한다. 본 논문에서 다루게 될 삼각 분할표는 이러한 두 개의 시간축을 가진 자료를 가장 효율적으로 표현하는 자료형태로 가로축에는 경과시간을 세로축에는 발생시간을 표시하게 되고 각 칸에는 해당 발생시간과 경과시간에 해당되는 관측값을 표시하게 된다.

삼각분할표에 대한 보다 정확한 이해를 위해 본 논문에서 사용하게 될 두 종류의 예제를 살펴보자. 먼저 보험자료에서 지급준비금 (loss-reserve)의 예를 살펴보고자 한다. 이 자료에서는 보험 증권(policy)이 작성된 해가 발생연도 (accidental year)가 되고 발생년도 이후 보험회사가 보험 증권에 따라 보험클레임을 받을 수 있게 된다. 여기서 사건은 보험 클레임의 청구이고 발생연도로부터 보험클레임이 들어온 시간까지가 경과년도 (developmental year)가 된다. 따라서 발생년도 i, 경과년도 j와 관련된 보험 클레임 총액을 Z_{ij}라 하면, 자료 $\{Z_{ij}; i = 1, 2, \cdots, m, j = 1, 2, \cdots, m - i + 1\}$의 형태는 다음 표 1.1과 같은 삼각형 분할표 형태를 가지게 된다. 다른 예로는 이혼율 자료를 생각할 수 있다. 이혼율 자료에서는 어느 쌍이 결혼에 도달한 연도를 발생연도로 생각할 수 있고 결혼 상태를 유지하고 있는 기간을 경과 년도로 생각할 수 있다. 이 때 각 발생년도 i와 경과년도 j에 해당하는 이혼비율 Y_{ij}를 관측했다고 하면 이혼율 자료 또한 위의 삼각분할표 형태로 표현 될 수 있다. 아래의 표 1.2는 실제 본 논문에서 분석하게 될 자료를 보여준다.

[1] (700-721) 대구광역시 중구 동덕로 200, 경북대학교 병원 임상시험센터, 연구원.
[2] (151-747) 서울특별시 관악구 관악로 599, 서울대학교 통계학과, 부교수.
[3] (120-712) 서울특별시 서대문구 성산로 262, 연세대학교 언더우드 국제학부, 조교수.
[4] 교신저자: (702-701) 대구광역시 북구 산격동 1370 경북대학교 통계학과, 조교수.
E-mail: artlee@knu.ac.kr

두 개의 시간을 가진 자료의 분석을 위한 대다수의 기존 방법들은 분석의 용이성을 위하여 절대시간의 효과를 무시할 수 있다는 가정을 하게 된다. 이 가정 하에서는 경과시간을 반응변수로 생각하고 잘 알려진 다양한 통계 모형을 사용할 수 있다는 장점이 있다. 하지만, 많은 경우에 있어 절대시간의 효과가 무시할 수 없음을 쉽게 인지 할 수 있다. 한 예로 표 1.2의 이혼율 자료를 보면 97년에서 99년에 있었던 대한민국의 경제위기로 인하여 모든 경과년도에 해당하는 이혼율의 증가현상을 읽을 수 있고 97년-99년의 절대 시간효과를 무시 할 수 없음을 알 수 있다. 따라서 삼각분할표 자료의 분석에 있어서 주된 관심인 경과시간에 관련된 모수를 추정함에 있어 절대시간의 효과를 보정하는 방법을 필요로 하게 된다.

표 1.1 삼각 분할표의 표준 형태

	경과년도						
발생년도	1	2	$\cdots$	$m-i+1$	$\cdots$	$m-1$	m
1	Z_{11}	Z_{12}	$\cdots$	$\cdots$	$\cdots$	$\cdots$	Z_{1m}
2	Z_{21}	Z_{22}	$\cdots$	$\cdots$	$\cdots$	$Z_{2,m-1}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
i	Z_{i1}	Z_{i2}	$\cdots$	$Z_{i,m-i+1}$			
$\vdots$	$\vdots$	$\vdots$	$\ddots$				
$\vdots$	$\vdots$	$Z_{m-1,2}$					
m	Z_{m1}						

표 1.2 대한민국 이혼률 자료

		경과년도									
발생년도	혼인건수	1	2	3	4	5	6	7	8	9	10
1990	399312	0.91	1.05	1.06	1.03	1.01	0.97	1.06	1.11	1.35	1.31
1991	416872	0.85	1.01	1.06	1.04	0.96	1.05	1.10	1.29	1.26	1.20
1992	419774	0.88	1.06	1.13	1.05	1.14	1.15	1.32	1.27	1.22	1.40
1993	402593	0.95	1.19	1.18	1.26	1.24	1.43	1.34	1.30	1.48	1.60
1994	393121	1.03	1.24	1.30	1.29	1.48	1.44	1.37	1.50	1.61	1.83
1995	398484	1.06	1.42	1.47	1.66	1.54	1.50	1.60	1.65	1.84	
1996	434911	1.17	1.59	1.73	1.57	1.48	1.60	1.61	1.81		
1997	388591	1.39	1.98	1.90	1.78	1.86	1.91	2.06			
1998	375616	1.77	2.05	1.96	2.90	2.05	2.19				
1999	362673	1.77	2.11	2.18	2.17	2.28					
2000	334030	2.04	2.44	2.48	2.56						
2001	320063	2.17	2.44	2.68							
2002	306573	2.42	2.75								

본 논문은 삼각분할표 자료에 관련된 여러 가지의 목적 중 특별히 예측문제에 대하여 이야기 하고자 한다. 예측문제는 많은 삼각분할표 자료의 중요한 목적이다. 보험자료 의 예를 살펴보면, 지급 준비금은 다음 회계 연도에 지급해야할 보험 클레임의 총액으로 보험회사의 부채 중 가장 큰 부분을 차지한다. 이 지급 준비금은 항상 현금화 가능 상태로 유지해야 하는 부분으로, 이에 대한 정확한 예측은 보험회사의 재무 설계에 중요한 과제가 된다. 다른 예인 이혼율 자료를 살펴보면, 이혼율은 사회의 변화에 대한 중요한 척도로 여겨져 매년 통계청이 계산하여 보고하는 자료 중 하나 (이상복, 2007)이다. 이혼율에 대

한 다양한 정의가 존재하는데 최근에 발표된 황형태 등 (2005) 에서와 같이 올해의 이혼율을 올해 결혼한 부부 중에서 언젠가는 이혼에 이르게 되는 비율이라 정의하면, 현재 이혼율의 계산은 위의 삼각분할표에서 다음 절대시간에 해당되는 칸의 값들을 예측하는 문제로 표현이 가능하게 되고 이혼율의 계산 문제는 삼각분할표에서의 예측 문제로 나타내지게 된다.

삼각분할표의 예측문제는 보험에서의 지급준비금예측과 관련하여 많은 연구가 진행 되어 왔다. 기존의 많은 연구들이 절대연도의 효과를 고려하지 않은 발생연도와 경과년도의 효과만을 포함하는 모형을 이용한 예측을 제안하였다 (Brosisu, 1992; Mack, 1993, 1994; Murphy, 1994; Zehnwirth, 1994; England와 Verrall, 2002). 이들과는 다르게 De Vylder와 Goovaerts (1979)는 절대연도의 효과까지를 포함한 일반적인 모형의 사용을 생각하였고, 특히 Barnett과 Zehnwirth (2000)는 아래와 같은 확률경향모형 (probabilistic trend family, PTF)을 제안하기도 했다.

$$\sum_{k=1}^{j} Y_{i,k} = \alpha_i + \sum_{k=1}^{j-1} \beta_k + \sum_{t=1}^{i+j-2} \gamma_t + \epsilon_{i,j}$$

여기서, α는 발생년도 효과, β는 경과년도 효과, γ는 통화 팽창과 같은 극적인 경향의 변화를 나타내는 절대연도의 효과를 나타낸다.

절대시간 (또는 절대연도)의 효과까지 포함한 모형을 통한 예측의 주된 어려움은 자료의 수보다 추정해야 할 모수의 수가 많아진다는 데에 있고 이러한 이유로 절대연도 효과에 강한 가정을 하게 된다. 한 예로 Barnett과 Zehnwirth (2000)는 절대연도의 효과가 선형함수 또는 구간별 상수함수임을 가정하게 된다. 이러한 가정들은 연구자의 주관에 의해 정해지고 모형을 통해 얻어진 예측 값이 이러한 가정에 상당한 의존성을 보인다는 문제가 있다.

위에서 언급된 절대시간 (또는 절대연도)의 어려움을 해결하기 위하여 본 논문에서는 Verrall (1990)의 발생시간과 경과시간만을 고려한 베이지안 선형 모형을 확률적으로 변하는 물리적 시간요인을 포함한 모형으로 확장하고 확장된 모형을 추정하기 위한 마코브 연쇄 몬테칼로 (Markov Chain Monte Carlo)방법을 소개한다. 추가로 삼각분할표에서의 예측은 주어진 모형의 추정절차로부터 자연스럽게 얻어지게 된다.

본 논문의 구성은 다음과 같다. 2절에서는 Verrall (1990)의 모형을 확장한 모형을 제시하고 마코브 연쇄 몬테칼로 방법을 이용한 추정 방법을 설명한다. 3절에서는 2절에서 제안된 방법을 3개의 실제 자료에 적용하여 본다. 3절에서 다룰 3개의 예는 1) Verrall (1990) 논문에서 소개된 일반 보험에 관한 자료, 2) 국내 A 생명보험 회사의 입원환자 지급준비금 관련자료, 그리고 3) 본 서론에서 소개된 대한민국의 이혼율 자료이다. 마지막으로 4절에서는 제안된 방법의 추가 확장에 관한 짤막한 논의와 함께 본 논문을 마무리 하고자 한다.

2. 모형

이 절에서는 Verrall (1990)의 확장된 모형을 정의하고 이를 추정하기 위한 마코브 연쇄 몬테칼로 방법을 소개 한다. 이후의 설명에서는 설명의 편이를 위하여 보험예를 통하여 모형을 설명한다.

먼저 아래 표 2.1과 같은 삼각분할표 자료가 관측되었다고 하자. 여기서 변수 Z_{ij}는 발생년도 i별 ($i = 1, \cdots, m$) j경과년도 ($j = 1, \cdots, m$)에 따른 지급보험금을 나타내고 낸다. 이 지급액들이 Verrall (1990)에서와 같이 log-normal 분포를 따른다고 가정한다. 따라서 $Y_{ij} = \log Z_{ij}$라 두면 Y_{ij}는 정규분포를 따르게 되고 본 논문에서는 다음과 같은 절대연도 효과를 포함한 회귀 모형을 제안한다.

표 2.1 시간요인을 추가한 베이지안 선형 모형 구조

	경과년도(j)						
발생년도(i)	1	2	$\cdots$	$m-i+1$	$\cdots$	$\cdots$	m
1	Z_{11}	Z_{12}	$\cdots$	$\cdots$	$\cdots$	$\cdots$	Z_{1m}
2	Z_{21}	Z_{22}	$\cdots$	$\cdots$	$\cdots$	$Z_{2,m-1}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
i	Z_{i1}	Z_{i2}	$\cdots$	$Z_{i,m-i+1}$			
$\vdots$	$\vdots$	$\vdots$	$\ddots$				
$\vdots$	$\vdots$	$Z_{m-1,2}$					
m	Z_{m1}						

* 위의 색칠된() 부분은 발생년도 i를 기준으로 경과년도 $m-i+1$시점 이후로 지급되지 않은 보험금을 나타냄.
* 같은 색깔의 cell들은 같은 해에 일어났다는 것을 의미 함.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ij}, \tag{2.1}$$

여기서, $i = 1, \ldots, m, j = 1, \ldots, m, k = i + j - 1 \leq m - 1$이고, 오차들은 평균이 0이고 분산이 σ^2인 서로 독립이고 동일한 정규분포를 따른다고 가정한다. 모수에 관한 일반적인 제약조건으로 $\alpha_1 = \beta_1 = \gamma_1 = 0$으로 두고 특별히 추정 가능한 (estimable) 발생연도 효과 모수를 위하여 제약조건 $\gamma_2 = \gamma_3 = 0$더 추가하였다.

모수들을 다음과 같이 $\boldsymbol{\beta} = (\mu, \alpha_2, \ldots, \alpha_m, \beta_2, \ldots, \beta_m)'$과 $\boldsymbol{\gamma} = (\gamma_4, \gamma_5, \ldots, \gamma_m)'$의 벡터형태로 표기하자. 전체 표본의 수는 $m(m+1)/2$이 되고 이를 편이성을 위하여 n으로 표기한다.

$$E(\mathbf{Y}) = \begin{pmatrix} \mu + \alpha_1 + \beta_1 + \gamma_1 \\ \mu + \alpha_2 + \beta_1 + \gamma_2 \\ \vdots \\ \mu + \alpha_m + \beta_1 + \gamma_m \\ \mu + \alpha_1 + \beta_2 + \gamma_2 \\ \mu + \alpha_2 + \beta_2 + \gamma_3 \\ \vdots \\ \mu + \alpha_{m-1} + \beta_2 + \gamma_m \\ \vdots \\ \mu + \alpha_1 + \beta_m + \gamma_m \end{pmatrix} = \begin{pmatrix} 1\,0\cdots0\,0\cdots0\,0\cdots0 \\ 1\,1\cdots0\,0\cdots0\,0\cdots0 \\ \vdots\;\vdots\;\ddots\;\vdots\;\vdots\;\ddots\;\vdots\;\vdots\;\ddots\;\vdots \\ 1\,0\cdots1\,0\cdots0\,0\cdots1 \\ 1\,0\cdots0\,1\cdots0\,0\cdots0 \\ 1\,0\cdots0\,1\cdots0\,0\cdots0 \\ \vdots\;\vdots\;\ddots\;\vdots\;\vdots\;\ddots\;\vdots\;\vdots\;\ddots\;\vdots \\ 1\,0\cdots1\,1\cdots0\,0\cdots1 \\ \vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots \\ 1\,0\cdots0\,0\cdots1\,0\cdots1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_m \\ \beta_2 \\ \vdots \\ \beta_m \\ \gamma_4 \\ \vdots \\ \gamma_m \end{pmatrix} + \begin{pmatrix} 0\,0\cdots0 \\ 0\,0\cdots0 \\ \vdots\;\vdots\;\vdots \\ 0\,0\cdots1 \\ 0\,0\cdots0 \\ \vdots\;\vdots\;\vdots \\ 0\,0\cdots1 \\ \vdots\;\vdots\;\vdots \\ 0\,0\cdots1 \end{pmatrix} \begin{pmatrix} \gamma_4 \\ \gamma_5 \\ \vdots \\ \gamma_m \end{pmatrix}$$

$$= \mathbf{X}_1\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\gamma}$$

여기서,

$$\mathbf{Y}=\begin{pmatrix} Y_{11}\\ Y_{21}\\ \vdots\\ Y_{m1}\\ Y_{12}\\ Y_{22}\\ \vdots\\ Y_{m-1,2}\\ \vdots\\ Y_{1m}\end{pmatrix},\mathbf{X_1}=\begin{pmatrix} 1&0&\cdots&0&0&\cdots&0&0&\cdots&0\\ 1&1&\cdots&0&0&\cdots&0&0&\cdots&0\\ \vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots\\ 1&0&\cdots&1&0&\cdots&0&0&\cdots&1\\ 1&0&\cdots&0&1&\cdots&0&0&\cdots&0\\ 1&0&\cdots&0&1&\cdots&0&0&\cdots&0\\ \vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots\\ 1&0&\cdots&1&1&\cdots&0&0&\cdots&1\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\ 1&0&\cdots&0&0&\cdots&1&0&\cdots&1\end{pmatrix},\mathbf{X_2}=\begin{pmatrix} 0&0&\cdots&0\\ 0&0&\cdots&0\\ \vdots&\vdots&\vdots&\vdots\\ 0&0&\cdots&1\\ 0&0&\cdots&0\\ \vdots&\vdots&\vdots&\vdots\\ 0&0&\cdots&1\\ \vdots&\vdots&\vdots&\vdots\\ 0&0&\cdots&1\end{pmatrix}$$

이고 각각의 크기는 $n\times 1, n\times(2m-1)$ 이고 $n\times(m-3)$이다. 모형 (2.1) 을 다음과 같이 나타낼 수 있다.

$$\mathbf{Y}=\mathbf{X}_1\boldsymbol{\beta}+\mathbf{X}_2\gamma+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim N_n(\mathbf{0},\sigma^2\mathbf{I})$$

Verrall (1990)은 절대 연도의 효과를 표현하는 γ항이 존재하지 않는 발생연도 효과와 경과연도 효과만을 지닌 선형모형에서 다음과 같은 계층적 사전분포들을 가정하고 있다.

$$\boldsymbol{\beta}|\boldsymbol{\theta}_1\sim N_{2m-1}(\mathbf{A}_1\boldsymbol{\theta}_1,\mathbf{C}_1)$$
$$\boldsymbol{\theta}_1\propto\mathbf{1}$$

여기서,

$$\mathbf{C_1}=\begin{pmatrix}\sigma_\mu^2&&&&&&&\\ &\sigma_\alpha^2&&&&&&\\ &&\ddots&&&&&\\ &&&\sigma_\alpha^2&&&&\\ &&&&\sigma_\beta^2&&&\\ &&&&&\ddots&&\\ &&&&&&\sigma_\beta^2\end{pmatrix},\boldsymbol{\theta}_1=\begin{pmatrix}w\\ \theta_\alpha\\ \xi_2\\ \vdots\\ \xi_m\end{pmatrix},\mathbf{A_1}=\begin{pmatrix}1&0&\cdots&&0\\ 0&1&0&&\vdots\\ \vdots&1&\vdots&&\\ \vdots&\vdots&\vdots&&\\ 0&1&0&&\\ \vdots&0&1&&\vdots\\ \vdots&\vdots&&\ddots&0\\ 0&\cdots&\cdots&0&1\end{pmatrix}$$

이며 각각의 크기는 $(2m-1)\times(2m-1), (m+1)\times 1$이고 $(2m-1)\times(m-1)$이다.

또한 $\sigma_\mu^{-2}\to 0, \sigma_\beta^{-2}\to 0$이고 $v\lambda/\sigma^2$와 $v_\alpha\lambda_\alpha/\sigma_\alpha^2$가 각각 $\chi^2(v)$와 $\chi^2(v_\alpha)$를 따른다고 가정하면 w와 $\xi_2,\ldots,\xi_m$의 3단계 사전분포에 대한 가정은 필요 없게 된다.

제안된 모형 (2.1)의 절대연도 효과를 설명하는 모수 γ들에 대하여 다음의 1차 자기회귀모형관계를 가정한다:

$$\gamma_k|\gamma_{k-1}\sim N(\rho\gamma_{k-1},\sigma_\gamma^2),\quad |\rho|<1 \tag{2.2}$$

이 관계는 벡터 표현을 사용하면 공분산 행렬

$$\Sigma_\gamma = \frac{\sigma_\gamma^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{m-2} \\ \rho & 1 & \rho & \cdots & \rho^{m-3} \\ \rho^2 & \rho & 1 & \cdots & \rho^{m-4} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{m-2} & \rho^{m-3} & \rho^{m-4} & \cdots & 1 \end{pmatrix}$$

에 대하여 $\boldsymbol{\gamma} \sim N_{m-3}(\mathbf{0}, \Sigma_\gamma)$로 표현 할 수 있다.

마지막으로 초모수 $v_\gamma\lambda_\gamma/\sigma_\gamma^2$와 ρ는 $\chi^2(v_\gamma)$와 $U(0,1)$를 각각 따른다고 가정함으로써 제안된 베이지안 계층적 모형의 정의를 마치게 된다.

모수들에 대한 결합 사후 분포가 명확한 형태를 가지지 않기 때문에 깁스 샘플링과 메트로 해스팅스 알고리즘을 이용하려고 한다. 모수들 $\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \sigma_\alpha^2, \sigma_\gamma^2, \rho, \theta_\alpha$의 각각의 완전 사후 확률(full conditional distribution)을 계산하면 다음과 같다. 자세한 내용은 부록을 참조하기 바란다.

- $\boldsymbol{\beta}$에 대한 사후확률분포는 다음과 같이 정규분포를 따른다:

$$\boldsymbol{\beta}|\cdot \sim N_{2m-1}\left(\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{\sigma^2}+\mathbf{C}_1^{-1}\right)^{-1}\left[\frac{1}{\sigma^2}\left(\mathbf{X}_1'(\mathbf{y}-\mathbf{X}_2\boldsymbol{\gamma})+\mathbf{C}_1^{-1}\mathbf{A}_1\boldsymbol{\theta}_1\right)\right], \left(\frac{\mathbf{X}_1'\mathbf{X}_1}{\sigma^2}+\mathbf{C}_1^{-1}\right)^{-1}\right)$$

- 절대연도 효과를 나타내는 모수 $\boldsymbol{\gamma}$에 대한 사후 확률 분포는 다음과 같이 정규분포를 따른다:

$$\boldsymbol{\gamma}|\cdot \sim N_{m-3}\left(\left(\frac{\mathbf{X}_2'\mathbf{X}_2}{\sigma^2}+\Sigma_\gamma^{-1}\right)^{-1}\mathbf{X}_2'(\mathbf{y}-\mathbf{X}_1\beta), \left(\frac{\mathbf{X}_2'\mathbf{X}_2}{\sigma^2}+\Sigma_\gamma^{-1}\right)^{-1}\right)$$

- 분산 σ^2에 대한 사후확률분포는 다음과 같이 역감마분포를 따른다:

$$\sigma^2|\cdot \sim IG\left(\frac{n+v}{2}, \frac{1}{2}\left[v\lambda+(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})'(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})\right]\right)$$

- α의 분산 σ_α^2에 대한 사후확률분포는 다음과 같이 역감마분포를 따른다:

$$\sigma_\alpha^2|\cdot \sim IG\left(\frac{(m-1)+v_\alpha}{2}, \frac{\sum_{i=2}^m(\alpha_i-\theta_\alpha)^2+v_\alpha\lambda_\alpha}{2}\right)$$

- γ의 분산 σ_γ^2에 대한 사후확률분포는 다음과 같이 역감마분포를 따른다:

$$\sigma_\gamma^2|\cdot \sim IG\left(\frac{(m-3)+v_\lambda}{2}, \frac{v_\lambda\lambda_\gamma+\gamma'\Sigma_\gamma^{-1}\gamma}{2}\right)$$

- α의 평균 θ_α에 대한 사후확률분포는 다음과 같이 정규분포를 따른다:

$$\theta_\alpha|\cdot \sim N\left(\bar{\alpha}, \frac{\sigma_\alpha^2}{(m-2)}\right)$$

- 절대연도 효과를 나타내는 모수 γ들의 상관계수 ρ에 대한 사후 확률 분포는 다음과 같은 비례관계를 따른다:

$$p(\rho|\cdot) \propto \Sigma_\gamma^{-1} \exp\left\{-\frac{1}{2}\gamma'\Sigma_\gamma^{-1}\gamma\right\} \cdot I(0,1)$$

마지막으로 본 논문의 주된 주제인 다음 절대연도의 지급준비금을 예측하는 절차를 소개 하고자 한다. 다음 절대연도 $k = m+1$일 때 지급준비금 $\mathbf{Y}_{new} = (Y_{m,2}, \ldots, Y_{2,m})'$은 다음과 같다.

$$\mathbf{Y}_{new} = \begin{pmatrix} Y_{m,2} \\ Y_{m-1,3} \\ \vdots \\ Y_{2,m} \end{pmatrix} = \begin{pmatrix} \mu+\alpha_m+\beta_2+\gamma_{m+1} \\ \mu+\alpha_{m-1}+\beta_3+\gamma_{m+1} \\ \vdots \\ \mu+\alpha_2+\beta_m+\gamma_{m+1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{m,2} \\ \varepsilon_{m-1,3} \\ \vdots \\ \varepsilon_{2,m} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 1 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_m \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \gamma_{m+1} \\ \gamma_{m+1} \\ \vdots \\ \gamma_{m+1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{m,2} \\ \varepsilon_{m-1,3} \\ \vdots \\ \varepsilon_{2,m} \end{pmatrix}$$

$$\therefore \mathbf{Y}_{new} = \mathbf{X}_{new}\boldsymbol{\beta} + \mathbf{1}\cdot\gamma_m + \mathbf{1} + \boldsymbol{\varepsilon}$$

여기서

$$\mathrm{X}_{new} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 1 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

이다. 따라서 $\mathbf{Y}_{new}$에 대한 자연스러운 예측 값은 $\mathbf{Y}_{new}$의 사후 확률 분포의 평균인 $E(\mathbf{Y}_{new}|\underline{\mathbf{y}})$ (김영화, 2007) 이고 이는 β의 사후 확률 분포의 평균인 $\hat{\boldsymbol{\beta}}$과 γ_m+1의 사후확률분포의 평균인 $\hat{\rho}\hat{\gamma}_m$을 이용하여 계산될 수 있다.

$$E(\mathbf{Y}_{new}|\mathbf{y}) = \mathbf{X}_{new} \times E(\boldsymbol{\beta}|\mathbf{y}) + E(\gamma_{m+1}|\mathbf{y})\mathbf{1} = \mathbf{X}_{new} \times \hat{\boldsymbol{\beta}} + \hat{\rho}\hat{\gamma}_m\mathbf{1}$$

와 같이 계산될 수 있다.

3. 실제 자료 분석 및 비교

본 절에서는 2절에서 소개된 절대연도 효과를 고려한 베이지안 선형모형(BMT)을 세 개의 실제 예에 적용하여 본다. 본 절에서 사용할 세 가지 자료는 Verrall (1990)에서 소개된 일반 보험에 관한 자료, 국내 A 생명보험회사의 입원환자 지급준비금 관련자료, 그리고 통계청에서 제공한 대한민국 이혼율 자

료이다. 제안된 모형의 적합성을 검증하기 위하여 BMT의 결과를 Kremer (1982)의 선형모형 (LM)과 Verrall (1990)의 베이지안 선형모형 (BLM)와 비교하였고 모형 적합성의 척도로는 아래에서 정의될 평균상대예측제곱오차 (Average Relative Mean Squared Prediction Error: ARMSPE)를 사용하였다.

본 절의 전체적인 자료 분석 과정은 다음과 같다.

- (1 단계) 삼각분할표 자료에서 왼쪽 윗부분의 절대 시간 $k+1$까지에 해당되는 자료 $\{Z_{ij} : i = 1, \ldots, k, j = 1, \ldots, k-i+1\}$를 이용하여 모형을 적합 시키고 다음 절대시간 $k+2$에 해당하는 $\mathbf{Z}_k = (Z_{k,2}, Z_{k-1,3}, \ldots, Z_{2,k})'$를 예측하고 이를 실제 값과 비교한다. 예측값을 $\hat{\mathbf{Z}}_\mathbf{k} = \left(\hat{Z}_{k,2}, \hat{Z}_{k-1,3}, \ldots, \hat{Z}_{2,k}\right)'$라 하면 상대예측제곱오차는 다음과 같이 정의 된다.

$$\mathrm{RMSPE}(k) = \frac{1}{k-1} \sum_{i+j=k+2,\ i,j>1} \left| \frac{\hat{Z}_{ij} - Z_{ij}}{Z_{ij}} \right|^2$$

- (2 단계) $k = m_0, m_0+1, \ldots, m-1$에 관하여 (1 단계)를 반복한다.
- (3 단계) (1 단계)와 (2 단계)의 결과를 이용하여 평균상대예측제곱오차

$$\mathrm{ARMSPE} = \frac{1}{m-m_0} \sum_{k=m_0}^{m-1} \mathrm{RMSPE}(k)$$

를 계산한다.

3.1. 일반보험자료

첫 번째 자료는 Verrall (1990)의 논문에서 소개된 자료로써 일반 보험 자료이고 자료(표 3.1)는 아래와 같다. 자료에 대한 자세한 배경 설명은 Verrall (1990)을 참조하기 바란다.

표 3.1 일반보험자료 (Verrall, 1990)

	경과년도									
발생년도	1	2	3	4	5	6	7	8	9	10
1	357848	766940	610542	482940	527326	574398	146342	139950	227229	67948
2	352118	884021	933894	1183289	445745	320996	527804	266172	425046	
3	290507	1001799	926219	1016654	750816	146923	495992	280405		
4	310608	1108250	776189	1562400	272482	352053	206286			
5	443160	693190	991983	769488	504851	470639				
6	396132	937085	947498	805037	705960					
7	440832	847631	1131398	1063269						
8	359480	1061648	1443370							
9	376686	986608								
10	344014									

표 3.2는 절대시간 효과를 가진 베이지안 선형모형(BMT)은 선형모형(LM)보다는 좋은 결과를 보이고 있으나 베이지안 선형모형(BLM)보다 약간 높은 평균상대제곱오차를 가짐을 보여주고 있다. 본 자료의 경우는 절대시간 효과가 존재하지 않음을 추측 할 수 있고 복잡한 모형인 BTM가 단순한 모형인 BLM이 약간 높은 제곱오차를 보여주나 각 k값 별 상대제곱오차를 살펴보면 두 방법사이에 유의한 차이가 있지는 않다.

표 3.2 일반보험자료의 상대예측제곱오차

k	LM	BLM	BMT
5	0.306	0.202	**0.175**
6	0.406	0.393	**0.378**
7	1.544	**1.344**	1.458
8	0.067	**0.059**	0.089
9	0.115	0.128	**0.116**
ARMSPE	0.4876	**0.4252**	0.4432

3.2. 국내 A 생명보험회사 자료

다음은 국내 A 생명보험회사의 9개월간의 건강 상품 입원비 담보자료이다.

표 3.3 건강상품 입원비 담보자료

	경과개월								
발생개월	1	2	3	4	5	6	7	8	9
1	10961200	45662084	36424620	15785546	13398744	3908128	8804891	530017	246170
2	24469507	64707845	52472347	17882315	13361215	5645996	7377687	2414278	
3	19110976	83260504	65105234	29474550	13299793	15608077	33151024		
4	27249876	78397484	55772701	29711262	12569220	24985675			
5	20118065	106379365	61877823	49733574	17138262				
6	33492112	96008180	85061882	**58552375**					
7	25900369	155821060	**99751192**						
8	49011549	**165671393**							
9	33481135								

본 자료에 있어서는 아래 표 3.4에서 볼 수 있듯이 BMT가 가장 작은 평균상대제곱오차를 가지고 있음을 볼 수 있고 약간의 절대시간의 효과의 변동이 존재함을 추측하여 볼 수 있다. 추가로 각 k값 별 상대예측제곱오차를 살펴보아도 전반적으로 BMT가 BLM보다 작은 오차를 보인다.

표 3.4 건강상품 입원비 담보자료의 상대예측제곱오차

k	LM	BLM	BMT
5	0.124	0.121	**0.120**
6	0.068	**0.038**	0.064
7	0.222	0.255	**0.220**
8	0.208	0.252	**0.198**
ARMSPE	0.155	0.167	**0.151**

3.3. 이혼율 자료

통계청에서 제공한 이혼율자료(표1.2)는 1990년부터 2002년까지 혼인한 부부가 2003년까지 이혼에 이르게 된 비율을 나타낸 표이다.아래에서 변수 Y_{ij}는 i년에 결혼한 부부가 경과년도 j년에 이혼하게 되는 비율을 나타낸다. 본 자료에 로그 변환을 취하지 않았고 각 발생연도를 기준으로 이혼건수의 상대적 비율이므로 발생연도 효과를 고려할 수 없게 되므로 다음의 모형을 고려하였다.

$$Y_{ij} = \mu + \beta_j + \gamma_k + \varepsilon_{ij}$$

제안된 모형에서 모수와 오차에 대한 가정은 앞 절에서 소개한 가정들과 동일하다.

본 이혼율 자료는 97년에서 99년까지 대한민국의 경제위기와 함께 이혼율의 증가라는 절대연도의 효과가 존재하는 자료라 사료된다. 이 자료의 경우 위의 표 3.5에서 볼 수 있듯이 절대시간효과을 추가한 베이지안 선형모형(BMT)이 다른 두 모형, 베이지안 선형모형(BLM)과 선형모형(LM)보다는 월등히 좋은 결과를 보이고 있음을 알 수 있다.

표 3.5 이혼율자료의 상대예측제곱오차

k	LM	BLM	BMT
7	0.031	0.031	**0.019**
8	0.076	0.076	**0.059**
9	0.051	0.051	**0.046**
10	0.043	0.043	**0.024**
11	0.074	0.074	**0.023**
12	0.086	0.086	**0.030**
13	0.121	0.121	**0.052**
ARMSPE	0.069	0.0691	**0.036**

4. 결 론

삼각 분할표 형태의 자료(run-off triangle data)는 많은 응용분야에서 발생할 수 있으며 여러 사회 현상으로 인한 절대연도 효과를 고려하는 것이 바람직하다. 본 논문에서는 Verrall (1990)의 발생시간과 경과시간의 효과만을 고려한 베이지안 선형모형을 절대연도 효과를 포함한 모형으로 확장하였으며 본 논문에서 제안한 모형이 전반적으로 다른 두 모형보다 좋은 결과를 보이고 있다. 추가적으로, 절대연도 효과가 매 시점마다 바뀐다는 가정을 좀 더 완화한 다중 변화점 모형으로 확장할 수 있을 것으로 기대한다.

부록

- $\boldsymbol{\beta}$에 대한 사후확률분포는 다음과 같이 정규분포를 따른다:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\gamma}) &\propto p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma})p(\boldsymbol{\beta}|\boldsymbol{\theta}_1) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})'(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1-\mathbf{X}_2\boldsymbol{\gamma})\right\} \\ &\quad\times\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{A}_1\boldsymbol{\theta}_1)'\mathbf{C}_1^{-1}(\boldsymbol{\beta}-\mathbf{A}_1\boldsymbol{\theta}_1)\right\} \end{aligned}$$

$$\therefore \boldsymbol{\beta}_1|\cdot \sim N\left(\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{\sigma^2}+\mathbf{C}_1^{-1}\right)^{-1}\left[\frac{1}{\sigma^2}\left(\mathbf{X}_1'(\mathbf{y}-\mathbf{X}_2\boldsymbol{\gamma})+\mathbf{C}_1^{-1}\mathbf{A}_1\boldsymbol{\theta}_1\right)\right],\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{\sigma^2}+\mathbf{C}_1^{-1}\right)^{-1}\right)$$

- 절대연도 효과를 나타내는 모수 $\boldsymbol{\gamma}$에 대한 사후확률분포는 다음과 같이 정규분포를 따른다:

$$\begin{aligned} p(\boldsymbol{\gamma}|\cdot) &\propto p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma}) \cdot p(\boldsymbol{\gamma}) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})'(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})\right\} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}'\Sigma_\gamma^{-1}\boldsymbol{\gamma}\right\} \\ &\therefore \boldsymbol{\gamma}|\cdot \sim N\left(\left(\frac{\mathbf{X}_2'\mathbf{X}_2}{\sigma^2}+\Sigma_\gamma^{-1}\right)^{-1}\mathbf{X}_2'\left(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1\right), \left(\frac{\mathbf{X}_2'\mathbf{X}_2}{\sigma^2}+\Sigma_\gamma^{-1}\right)^{-1}\right) \end{aligned}$$

- 분산 σ^2에 대한 사후확률분포는 다음과 같이 역감마분포를 따른다:

$$\begin{aligned} p(\sigma^2|\cdot) &\propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})'(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})\right\} \\ &\quad \times (\sigma^2)^{-\frac{v}{2}-1} \exp\left\{-\frac{v\lambda}{2\sigma^2}\right\} \\ &\propto (\sigma^2)^{-\frac{n+v}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left(v\lambda+(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})'(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})\right)\right\} \\ &\therefore \sigma^2|\cdot \sim IG\left(\frac{n+v}{2}, \frac{1}{2}\left[v\lambda+(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})'(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma})\right]\right) \end{aligned}$$

- α의 분산 σ_α^2에 대한 사후확률분포는 다음과 같이 역감마분포를 따른다:

$$\begin{aligned} p(\sigma_\alpha^2|\cdot) &\propto \left(\sigma_\alpha^2\right)^{-\frac{(m-1)}{2}} \exp\left\{-\frac{1}{2\sigma_\alpha^2}\sum_{i=2}^{m}(\alpha_i-\theta_\alpha)^2\right\} \cdot \left(\sigma_\alpha^2\right)^{-\frac{v_\alpha}{2}-1} \exp\left\{-\frac{v_\alpha\lambda_\alpha}{2\sigma_\alpha^2}\right\} \\ &\propto \left(\sigma_\alpha^2\right)^{-\frac{(m-1)+v_\alpha}{2}-1} \exp\left\{-\frac{1}{2\sigma_\alpha^2}\left[\sum_{i=2}^{m}(\alpha_i-\theta_\alpha)^2+v_\alpha\lambda_\alpha\right]\right\} \\ &\therefore \sigma_\alpha^2|\cdot \sim IG\left(\frac{(m-1)+v_\alpha}{2}, \frac{\sum_{i=2}^{m}(\alpha_i-\theta_\alpha)^2+v_\alpha\lambda_\alpha}{2}\right) \end{aligned}$$

- γ의 분산 σ_γ^2에 대한 사후확률분포는 다음과 같이 역감마분포를 따른다:

$$\begin{aligned} p(\sigma_\gamma^2|\cdot) &\propto |\Sigma_\gamma^{-1}|^{1/2} \exp\left\{-\frac{1}{2}\gamma'\Sigma_\gamma^{-1}\gamma\right\} \cdot \left(\sigma_\gamma^2\right)^{-\frac{v_\gamma}{2}-1} \exp\left\{-\frac{v_\gamma\lambda_\gamma}{2\sigma_\gamma^2}\right\} \\ &\therefore \sigma_\gamma^2|\cdot \sim IG\left(\frac{(m-3)+v_\gamma}{2}, \frac{v_\gamma\lambda_\gamma+\gamma'\Sigma_\gamma^{-1}\gamma}{2}\right) \end{aligned}$$

- α의 평균 θ_α에 대한 사후확률분포는 다음과 같이 정규분포를 따른다:

$$p(\theta_\alpha|\cdot) \propto \exp\left\{-\frac{1}{2}\sum_{i=2}^{m}\frac{(\alpha_i-\theta_\alpha)^2}{\sigma_\alpha^2}\right\}\times 1$$

$$\therefore \theta_\alpha|\cdot \sim N\left(\bar{\alpha}, \frac{\sigma_\alpha^2}{(m-1)}\right)$$

- 절대연도 효과를 나타내는 모수 γ들의 상관계수 ρ에 대한 사후확률분포는 다음과 같은 비례관계를 따른다:

$$p(\rho|\cdot) \propto |\Sigma_\gamma^{-1}|^{1/2}\exp\left\{-\frac{1}{2}\gamma'\Sigma_\gamma^{-1}\gamma\right\}\cdot I(0,1)$$

참고문헌

황형태, 이성임, 방미진 (2005). 이혼율에 대한 새로운 지표의 개발 및 적용: 1990-2003년도의 우리나라 이혼률 분석. <통계연구>, **10**, 23-37.

Barnett, G. and Zehnwirth, B. (2000). Best estimates for reserves. *Proceedings of the Casualty Actuarial Society* , **LXXXVII**, 245-303.

Brosius, E. (1992). Loss development using credibility. *Casualty Actuarial Society Part 7 Exam Study Kit.*

De Vylder, F. and Goovaerts, M. J. (1979). *Proceedings of the first meeting of the contact group "Actuarial Science",* KU Leuven, Belgium.

England, P. D and Verrall, R. J. (2002). Stochastic claims reserving in general insurance (with discussion). *British Actuarial Journal,* **8,** III.

Kim, Y. (2006). A Comparative study for several bayesian estimators under balanced loss function. *Journal of the Korean Data & Information Science Society,* **17**, 291-300.

Kremer, E. (1982). IBNR-Claims and the two-way model of ANOVA. *Scandinavian Actuarial Journal*, **1**, 47-55.

Lee, S. (2007). Population projections for local governments in Korea: based on Hamilton-Perry & Auto regression. *Journal of the Korean Data & Information Science Society,* **18**, 955-961.

Mack, T. (1993). Distribution -free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, **23**, 213-225.

Mack, T. (1994). Which stochastic model is underlying the chain ladder method?. *Insurance: Mathematics and Economics*, **15**, 133-138.

Murphy (1994). Unbiased loss development factors. *Proceedings of the Casualty Actuarial Society*, **LXXXI**, 154-222.

Verrall, R. J. (1990). Bayes and empirical bayes estimation for the chain ladder model. *Astin Bulletin*, **20,** 217-243.

Zehnwirth, B. (1994). Probabilistic development factor models with applications to loss reserve variability, prediction intervals and risk based capital. *Casualty Actuarial Society Forum*, Spring 1994, **2**.

Journal of the Korean Data & Information Science Society
2009, **20**(2), 411–423

한국데이터정보과학회지

Prediction in run-off triangle using Bayesian linear model

Ju-Mi Lee[1] · Johan Lim[2] · Kyu S. Hahn[3] · Kyeong Eun Lee[4]

[1]Clinical Trial Center, Kyungpook National University Hospital
[2]Department of Statistics, Seoul National University
[3]Underwood International College, Yonsei University
[4]Department of Statistics, Kyungpook National University

Received 22 January 2009, revised 19 March 2009, accepted 23 March 2009

Abstract

In the current paper, by extending Verall (1990)'s work, we propose a new Bayesian model for analyzing run-off triangle data. While Verall's (1990) work only accounts for the calendar year and evolvement time effects, our model further accounts for the "absolute time" effects. We also suggest a Markov Chain Monte Carlo method that can be used for estimating the proposed model. We apply our proposed method to analyzing three empirical examples. The results demonstrate that our method significantly reduces prediction error when compared with the existing methods.

Keywords: Bayesian linear model, Markov chain Monte Carlo method, run-off triangle data.

[1] Researcher, 200 Dongduk-ro, Jung-gu, Daegu 700-721, Korea.
[2] Associate Professor, 599 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea.
[3] Assistant Professor, 262 Seongsanno, Seodaemun-gu, Seoul 120-712, Korea.
[4] Corresponding author: Assistant Professor, 1370 Sankyuk-dong, Buk-gu, Daegu 702-701, Korea. E-mail: artlee@knu.ac.kr