The Korean Communications in Statistics
Vol. 13 No. 1, 2006, pp. 177-190

Bayesian Modeling of Mortality Rates for Colon Cancer[1)]

Hyunjoong Kim[2)]

Abstract

The aim of this study is to propose a Bayesian model for fitting mortality rate of colon cancer. For the analysis of mortality rate of a disease, factors such as age classes of population and spatial characteristics of the location are very important. The model proposed in this study allows the age class to be a random effect in addition to its conventional role as the covariate of a linear regression, while the spatial factor being a random effect. The model is fitted using Metropolis-Hastings algorithm. Posterior expected predictive deviances, standardized residuals, and residual plots are used for comparison of models. It is found that the proposed model has smaller residuals and better predictive accuracy. Lastly, we described patterns in disease maps for colon cancer.

Keywords : Mortality rate; Bayesian modeling; Metropolis-Hastings algorithm; Posterior distribution; Disease mapping.

1. 연구 배경과 목적

소지역 추정의 문제에서 특정 질병에 의한 사망률의 모형화는 여러면에서 중요한 의미를 갖는다. 첫째, 특정질병에 대한 위험요인을 밝힌다는 것이다. 둘째, 특정질병이 발생하는 공간적 특성을 찾을 수 있다는 것이다. 셋째, 위험요인과 연결되는 공간적 특성을 규명하여 "요주의 지역"을 찾게 된다는 것이다. 추가적으로 사망률의 모형화가 갖는 실용적 의미중 하나는 그것이 질병지도(disease mapping)의 작성에 필요한 단계가 된다는 것에 있다. 예를 들어 Pickle, Mungiole, Jones & White (1996)는 1988년과 1992년 사이의 미국 주요 질병 사망률에 대한 모형화를 통하여 미국전역의 질병지도를 작성한 바 있다. Nandram, Sedransk & Pickel (1999)은 이전 모형을 개선하여 암 사망률에 대한 질병지도를 제공하였다. 또한 Nandram, Sedransk & Pickel (2000)은 COPD(만성폐쇄성폐질환)에 대한 질병지도를 제공하였다. 이러한 질병지도는 요주의 지역을 밝히고 궁극적으로 지역적 특성과 원인을 밝히게 된다는 면에서 공중보건을 위한 매우 중요한 도구가 된다. 이러한 관점들에서 사망률의 모형화는 의미있는 연구가 될 수 있다.

1) 이 논문은 연세대학교 상경대학 기초학문분야 육성기금의 지원에 의하여 이루어진 것임.
2) Assistant Professor, Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea. E-mail : hkim@yonsei.ac.kr

본 논문의 주된 목적은 대장암 사망률에 대한 모형들을 제안하고 평가하며, 궁극적으로 질병지도를 비교하는데 있다. 특히 연령적 특성과 공간지역적 특성을 모형에 변인으로 포함하여 각 연령층별 대장암 사망률을 베이지안 방법인 Metropolis-Hastings 알고리즘을 통하여 모형화하고자 한다. 분석을 위하여 본 논문은 Nandram, Sedransk & Pickel (1999)에 사용된 연령층과 지역단위를 그대로 사용하기로 한다.

2. 자료 구조

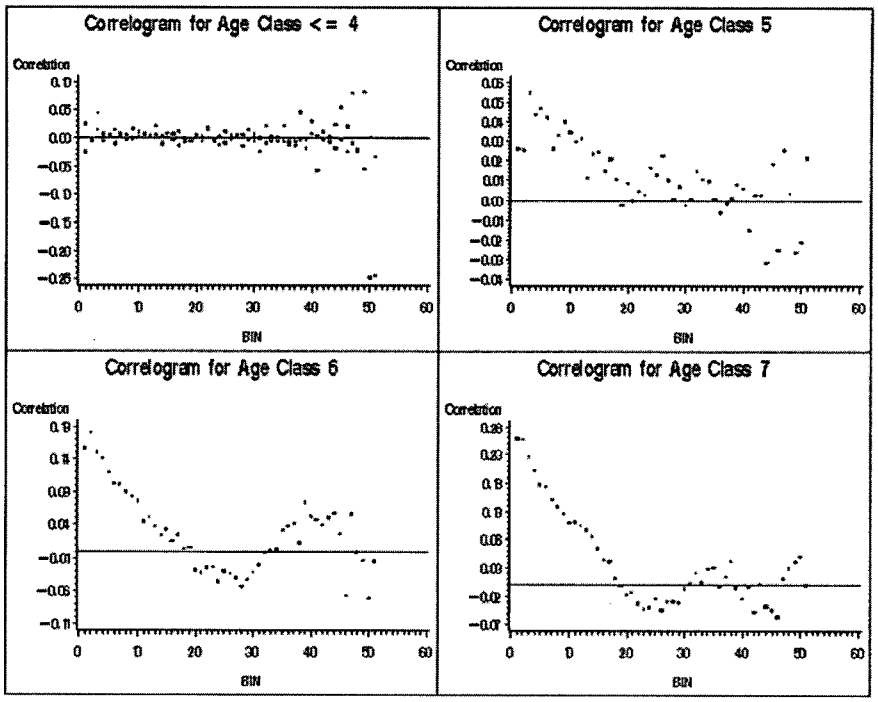
2.1 데이터 설명

본 논문에서 사용된 데이터는 1988년과 1992년 사이에 미국에서 발부된 사망증명서의 기록을 정리한 데이터베이스에서 가져온 것이다. 데이터베이스는 구체적으로 연령, 인종, 성별, 거주지, 사망원인별로 사망자의 수를 기록하였는데, 이는 미국 National Center for Health Statistics (NCHS)에 보고된 사망증명서에 기초한 것이다. 관측된 사망률은 1988년과 1992년 사이의 5년간 누적사망률을 계산한 것이다.

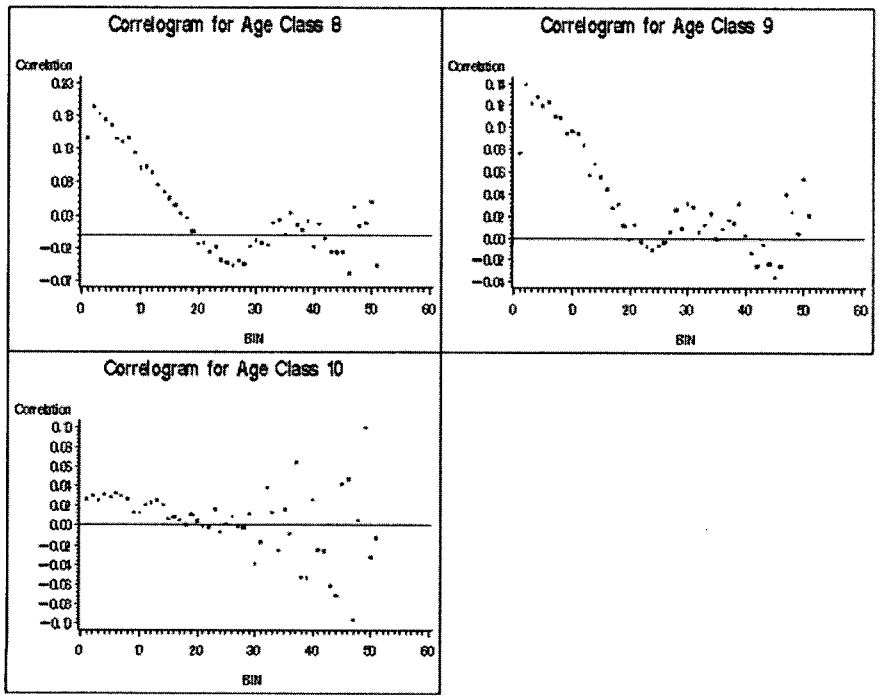
모형에 사용된 연령층은 0-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85 이상으로 나누었고, 각각 1, 2, 3, ⋯, 9, 10의 숫자로 표현한다 (Nandram, Sedransk & Pickel 1999). 공간지역은 Health Service Area (HSA)라는 작은 단위로 표현하였다. HSA는 NCHS가 설정한 공중 의료 지역단위로서 유사한 의료환경에 속한 county를 묶은 것이다. 미국에는 총 780개의 HSA가 있는데 각 HSA에 속한 county의 수는 1에서 20개까지 분포하고 있고 중위수는 2개이다. 뉴욕시를 제외하고는 한 HSA는 최소 250 평방마일의 면적을 갖고 있다. 여러개의 HSA를 묶어 미국 전역을 12개의 대지역으로 나누었으며 Nandram, Sedransk & Pickel (1999), 그리고 Kim & Nandram (2002)과 마찬가지로 본 논문에서 베이지안 모형은 각 대지역별로 따로 수행하였다.

2.2 데이터의 Correlograms

<그림 1>과 <그림 2>는 각 연령층별로 예측된 사망률에 대한 correlogram이다. 한 개의 "bin" 사이즈는 50마일로써 연령층에 관계없이 근접한 지역의 사망률끼리는 상관관계가 높다는 것을 알 수 있다. 750마일 이내의 이웃지역과의 상관계수는 0.1과 0.25의 범위내에 있으며 높은 연령층에서 상관계수가 더 높다는 것을 알 수 있다. 여기서 연령층 0-4, 5-14, 15-24, 25-34 는 대장암이 거의 발생하지 않는 연령층인 관계로 0-34의 연령층으로 통합되었다. 마지막으로 연령층 45-54, 55-64, 65-74, 75-84의 correlogram은 서로 비슷한 형태를 띠는 것을 관찰할 수 있다.



<그림 1> 연령층 0-34, 35-44, 45-54, 55-64에 대한 correlogram



<그림 2> 연령층 65-74, 75-84, 85 이상에 대한 correlogram

3. 대장암 사망률 분석

먼저 d_{ij}를 HSA i와 연령층 j의 사망자의 수라 하자 ($i=1,\cdots,N$; $j=1,\cdots,a$). 마찬가지로 n_{ij}는 HSA i와 연령층 j의 인구수이다. Brillinger (1996)와 Pickle, Mungiole, Jones & White (1996)의 논문에서 사용한 것과 마찬가지로 사망자 수에 대한 분포를 다음과 같이 사용한다.

$$d_{ij}\Big|\lambda_{ij} \overset{\text{ind}}{\sim} \text{Poisson}(n_{ij}\lambda_{ij}),\ i=1,\cdots,N, j=1,\cdots,a \quad . \tag{3.1}$$

여기서 고정효과 λ_{ij}는 각 연령층별 사망률을 의미한다. 참고로 연령층을 통합한 사망률은 다음과 같은 가중평균을 이용하여 구할 수 있다. 여기서 w_j는 미국 인구의 각 연령층별 빈도 비율이다.

$$R_i = \sum_{j=1}^{a} w_j \lambda_{ij} \quad .$$

Correlogram과 마찬가지로 연령층 0-4, 5-14, 15-24, 25-34는 대장암이 거의 발생하지 않는 연령층인 관계로 0-34의 연령층으로 통합되어 모든 분석이 수행되었다.

3.1 오프셋 모형 (Offset Model)

본 절에서는 Nandram, Sedransk & Pickle (1999)이 제안한 모형중 하나를 먼저 고려한다. 이 모형은 각 대지역별로 연령층에 대한 선형모형과 HSA에 대한 임의효과 모형을 가정한 것이다. 즉 일반화 선형모형의 연결함수로 다음과 같은 함수를 사용한다.

$$\log\lambda_{ij} = \boldsymbol{x}_j^T\underline{\beta} + \nu_i \quad \text{그리고} \quad \nu_i|\eta^2 \sim iid\ N(0,\eta^2), i=1,\cdots,N$$

여기서 독립변수 $\boldsymbol{x}_j$는 연령이 높을수록 큰 값을 갖는 연령층 변수이다. 사망률을 모형화 하는데 있어서 연령의 효과를 선형관계로 설명하고자 하며 ν_i는 지역에 따른 사망률의 변동을 설명하고자 함이다. 연령의 효과는 높은 연령층에서는 감소하므로 knot이 포함된 3차의 고정효과를 가정한다. 즉 연령층 벡터는 연령층 j에 대하여 $\underline{x}_j^t = (1, (j-1), (j-1)^2, (j-1)^3, \max\{0,(j-7)^3\})$으로 코딩되어 사용된다 ($j=4,\cdots,10$).

선형모형의 회귀계수 $\underline{\beta}$와 분산 η^2는 다음과 같은 사전 확률분포함수를 가정한다.

$$p(\underline{\beta}) = 1 \quad \text{그리고} \quad \eta^2 \sim \Gamma(\frac{b}{2}, \frac{c}{2}), \quad \text{여기서} \quad b=c=0.002.$$

이 모형은 소지역 추정에서 많이 사용되는 것으로 오프셋 모형이라 불리운다. Nandram, Sedransk & Pickle (1999)와 Nandram, Sedransk & Pickel (2000)의 논문과 마찬가지로 $\underline{\beta}$와 η^2의 분포로 부적절(improper) 사전분포를 사용하였다. 이 모형을 적합시키기 위하여 본 논문에서는 Metropolis-Hastings 알고리즘을 수행한다. 이를 위하여 Product of Kernels Principle (Chib & Greenberg 1995)을 이용하면 각각의 조건

부 사후확률분포로부터 표본을 번갈아가며 추출할 수 있는 장점이 있다. Metropolis 단계에서 proposal 밀도함수를 구하는 방법으로는 각 조건부 사후확률분포를 중앙값을 기준으로 2차 테일러 확장으로 근사시키는 방법을 이용한다 (Gelfand, Sahu & Carlin 1995).

연령층의 개수를 a, 그리고 HSA의 개수를 N으로 했을때 (본 데이터의 경우에는 $a=7$ and $N=780$), 결합확률밀도함수는 다음과 같다.

$$P(\underline{\beta},\underline{\nu},\eta^2 \mid \underline{d}) \propto \prod_{i=1}^{N}\prod_{j=1}^{a} \exp[(\underline{x}_j'\underline{\beta}+\nu_i)d_{ij} - n_{ij}e^{(\underline{x}_j'\underline{\beta}+\nu_i)}]$$
$$\times \prod_{i=1}^{N}(\frac{1}{\eta^2})^{\frac{1}{2}} e^{-(\frac{1}{2\eta^2})\nu_i^2} (\frac{1}{\eta^2})^{\frac{b}{2}+1} e^{-\frac{c}{2\eta^2}} \tag{3.2}$$

먼저, $\underline{\beta},\eta^2,\underline{d}$ 가 주어진 상태에서 ν_i의 조건부 사후확률분포함수를 고려해보자. 결합확률분포함수에 로그함수를 취한 것을 $\Delta(\nu_i)$이라 하면, 다음과 같은 식을 구할 수 있다. 이 식에서 ν_i와 관련없는 부분은 삭제하였다.

$$\Delta(\nu_i) = \nu_i \sum_{j=1}^{a} d_{ij} - e^{\nu_i}\sum_{j=1}^{a} n_{ij}e^{\underline{x}_j^t\underline{\beta}} - \frac{\nu_i^2}{2\eta^2} \quad .$$

이를 1차 미분하면

$$\Delta'(\nu_i) = \sum_{j=1}^{a} d_{ij} - e^{\nu_i}\sum_{j=1}^{a} n_{ij}e^{\underline{x}_j^t\underline{\beta}} - \frac{\nu_i}{\eta^2}$$

이 되고 2차 미분하여 다음과 같은 식을 얻는다.

$$\Delta''(\nu_i) = -[\frac{1}{\eta^2} + e^{\nu_i}\sum_{j=1}^{a} n_{ij}e^{\underline{x}_j^t\underline{\beta}}] \quad .$$

여기서 조건부 사후확률밀도함수는 로그 오목함수이다. 따라서 $\Delta'(\nu)=0$ 그리고 $\nu_i/\eta^2=0$이라 하면 ν_i의 추정치를 구할 수 있게 된다 (Gilks & Wild 1992).

$$\hat{\nu}_i = \log[\sum_{j=1}^{a} d_{ij} / \sum_{j=1}^{a} n_{ij}e^{\underline{x}_j^t\underline{\beta}}] \quad .$$

ν_i의 조건부 확률밀도함수를 $\hat{\nu}_i$을 중심으로 2차 테일러 확장을 수행하는 과정에서 다음과 같은 평균과 분산을 갖는 정규분포로 근사함을 알 수 있다.

$$Mean(\nu_i \mid \underline{\beta},\eta^2,\underline{d}) \approx \hat{\nu}_i - [\sum_{j=1}^{a} d_{ij} - e^{\hat{\nu}_i}\sum_{j=1}^{a} n_{ij}e^{\underline{x}_j^t\underline{\beta}} - \frac{\hat{\nu}_i}{\eta^2}][\frac{1}{\eta^2} + e^{\hat{\nu}_i}\sum_{j=1}^{a} n_{ij}e^{\underline{x}_j'\underline{\beta}_i+\hat{\nu}_i}]^{-1}, \tag{3.3}$$

$$Var(\nu_i \mid \underline{\beta},\eta^2,\underline{d}) \approx [\frac{1}{\eta^2} + e^{\hat{\nu}_i}\sum_{j=1}^{a} n_{ij}e^{\underline{x}_j'\underline{\beta}_i+\hat{\nu}_i}]^{-1}. \tag{3.4}$$

이를 이용하여 Metropolis 단계에서 다음과 같은 proposal 밀도함수를 이용한다.

$$\nu_i \mid \underline{\beta},\eta^2,\underline{d} \sim N(\mu,\Sigma) \ ,$$

여기서 μ와 Σ는 각각 (3.3)과 (3.4)와 같다. η의 조건부 확률밀도함수는 결합 확률밀도함수로 부터 다음과 같이 구할 수 있다.

$$\eta^{-2} \mid \underline{\beta}, \underline{\nu}, \underline{d} \sim \Gamma(\frac{N+b}{2}, \frac{c+\sum_{i=1}^{N} \nu_i^2}{2}) \quad .$$

이제 $\underline{\beta}$의 표본을 추출하기 위한 조건부 사후확률분포함수를 근사하고자 한다. 먼저 결합 확률분포함수에 로그를 취하고 $\underline{\nu}, \eta^2, \underline{d}$가 주어진 상태에서 β의 조건부 사후확률분포함수를 고려해 보자. 결합 확률분포함수에 로그함수를 취한 것을 $\Delta(\underline{\beta})$라고 하면 (상수항은 제외),

$$\Delta(\underline{\beta}) = (\sum_{i=1}^{N}\sum_{j=1}^{a} \underline{x}_j' d_{ij})\underline{\beta} - \sum_{i=1}^{N}\sum_{j=1}^{a} n_{ij} e^{\underline{x}_j'\underline{\beta}_i + \hat{\nu}_i}$$

이 되고 1차와 2차 미분의 결과식을 다음과 같이 쓴다.

$$\Delta'(\underline{\beta}) = \sum_{i=1}^{N}\sum_{j=1}^{a} (d_{ij} - n_{ij} e^{\underline{x}_j'\underline{\beta}_i + \hat{\nu}_i})\underline{x}_j \quad ,$$

$$\Delta''(\underline{\beta}) = -\sum_{i=1}^{N}\sum_{j=1}^{a} n_{ij} e^{\underline{x}_j'\underline{\beta}_i + \hat{\nu}_i} \underline{x}_j \underline{x}_j' \quad .$$

ν_i와 마찬가지로 $\underline{\beta}$의 조건부 확률밀도함수를 $\hat{\underline{\beta}}$을 중심으로 2차 테일러 확장을 수행하려 하므로 먼저 $\hat{\underline{\beta}}$을 계산할 필요가 있다. 본 논문에서는 $\hat{\underline{\beta}}$을 가중회귀분석을 사용하여 구하기로 한다. 반응변수 y_j를 다음과 같이 정의하고

$$y_j = \log(\sum_{i=1}^{N} d_{ij} / \sum_{i=1}^{N} n_{ij}), j = 1, \ldots, a \quad ,$$

Pickle et al. (1996)과 유사하게 가중선형회귀모형을 아래와 같이 가정한다.

$$y_j = \underline{x}_j' \underline{\beta} + e_j \quad ,$$

여기서 $e_j \sim N(0, \gamma^2 / d_{\cdot j})$이며 $d_{\cdot j} = \sum_{i=1}^{N} d_{ij}$이다. 가중선형회귀모형에 의한 $\underline{\beta}$의 추정치는 다음과 같다.

$$\hat{\underline{\beta}} = (\sum_{j=1}^{a} d_{\cdot j} \underline{x}_j \underline{x}_j')^{-1} \sum_{j=1}^{a} d_{\cdot j} y_j \underline{x}_j \quad .$$

$\underline{\beta}$의 조건부 확률밀도함수를 $\hat{\underline{\beta}}$을 중심으로 2차 테일러 확장을 수행하는 과정에서 다음과 같은 평균과 분산을 갖는 정규분포로 근사함을 알 수 있고 이를 Metropolis 단계에서 proposal 밀도함수로 이용한다.

$$\underline{\beta} \mid \underline{d}, \eta^2, \underline{d}, \tau^2 \sim N(\underline{\mu}, \Sigma) \quad ,$$

여기서

$$\underline{\mu} = \hat{\underline{\beta}} + (\sum_{i=1}^{N}\sum_{j=1}^{a} n_{ij} e^{\underline{x}_j'\hat{\underline{\beta}} + \delta_i} \underline{x}_j \underline{x}_j')^{-1} \sum_{i=1}^{N}\sum_{j=1}^{a} (d_{ij} - n_{ij} e^{\underline{x}_j'\hat{\underline{\beta}} + \delta_i})\underline{x}_j \, ,$$

$$\Sigma = (\sum_{i=1}^{N}\sum_{j=1}^{a} n_{ij} e^{\underline{x}_j'\hat{\underline{\beta}} + \delta_i} \underline{x}_j \underline{x}_j')^{-1} \, .$$

이상의 proposal 밀도함수들을 이용하여 Metropolis-Hastings 알고리즘을 수행하여

$\underline{\beta}$, ν, σ^2의 조건부 사후분포함수로부터 교대로 확률표본을 추출하고 수렴할 때까지 이 과정을 반복한다.

3.2 연령적 특성 임의효과 모형

3.1절의 오프셋 모형은 연령적 특성이 사망률에 미치는 효과는 선형모형에 의한 고정효과로 가정하였다. 이에 본 논문에서는 오프셋 모형을 확장하여 연령적 특성에 해당하는 임의효과를 모형에 추가하고자 한다. 그 이유는 사망률에 대한 모형이 연령별 고정효과에 추가하여 연령별 임의효과에 의해 더 잘 적합되리라 기대하기 때문이다. 이러한 방법은 Nandram, Sedransk and Pickle (2000)의 논문에서 이미 사용된 바 있다. 구체적으로 일반화 선형모형의 연결함수로 다음과 같은 함수를 사용한다.

$$\log(\lambda_{ij}) = \underline{x}_j'\underline{\beta} + \nu_i + \delta_j ,$$

여기서 사전분포함수는 다음과 같다.

$$\nu_i \mid \sigma_1^2 \sim N(0,\sigma_1^2), \quad \delta_j \mid \sigma_2^2 \sim N(0,\sigma_2^2),$$

그리고

$$P(\underline{\beta}) = 1 \quad , \ \sigma_1^{-2}, \sigma_2^{-2} \sim \Gamma(\frac{b}{2}, \frac{c}{2}), \ b = c = 0.002.$$

따라서 연령층에 해당하는 δ_j가 연령의 변동성을 설명할 것으로 기대한다. 3.1절과 유사한 절차를 거쳐 Metropolis-Hastings 알고리즘을 수행하고자 한다. 먼저 모든 모수를 포함한 결합 확률밀도함수를 구하면 다음과 같다.

$$P(\underline{\beta},\underline{\nu},\underline{\delta},\sigma_1^2,\sigma_1^2 \mid \underline{d}) \propto \prod_i^N \prod_j^a \exp[(\underline{x}_j'\underline{\beta} + \nu_i + \delta_j)d_{ij} - n_{ij}e^{(\underline{x}_j'\underline{\beta}+\nu_i+\delta_j)}] \times \prod_i^N (\frac{1}{\sigma_1^2})^{\frac{1}{2}} e^{-\frac{1}{2\sigma_1^2}\nu_i^2}$$

$$\times \prod_j^a (\frac{1}{\sigma_2^2})^{\frac{1}{2}} e^{-\frac{1}{2\sigma_2^2}\delta_j^2} \times (\frac{1}{\sigma_1^2})^{\frac{b}{2}+1} e^{-\frac{c}{2\sigma_1^2}} \times (\frac{1}{\sigma_2^2})^{\frac{b}{2}+1} e^{-\frac{c}{2\sigma_2^2}} .$$

단순화를 위하여 연령층에 해당하는 변인을 다음과 같이 변수변환을 고려하도록 한다.

$$\phi_j = \underline{x}_j'\underline{\beta} + \delta_j .$$

그 결과로 결합 확률밀도함수는 다음과 같이 변하게 된다.

$$P(\underline{\beta},\underline{\nu},\underline{\phi},\sigma_1^2,\sigma_2^2 \mid \underline{d}) \propto \prod_i^N \prod_j^a \exp[(\nu_i + \phi_j)d_{ij} - n_{ij}e^{(\nu_i+\delta_j)}] \times \prod_i^N (\frac{1}{\sigma_1^2})^{\frac{1}{2}} e^{-\frac{1}{2\sigma_1^2}\nu_i^2}$$

$$\times \prod_j^a (\frac{1}{\sigma_2^2})^{\frac{1}{2}} e^{-\frac{1}{2\sigma_2^2}(\phi_j - \underline{x}_j'\underline{\beta})^2} \times (\frac{1}{\sigma_1^2})^{\frac{b}{2}+1} e^{-\frac{c}{2\sigma_1^2}} \times (\frac{1}{\sigma_2^2})^{\frac{b}{2}+1} e^{-\frac{c}{2\sigma_2^2}} . \qquad (3.5)$$

모수들 중에서 $\sigma_1^{-2}, \sigma_2^{-2}, \beta$의 조건부 사후확률분포함수는 (3.5)식으로부터 쉽게 구할 수 있다. 즉,

$$P(\sigma_1^{-2} \mid \underline{\beta},\underline{\phi},\underline{\nu},\sigma_2^2,\underline{d}) \propto \Gamma(\frac{N+b}{2}, \frac{c+\sum_i^N \nu_i^2}{2}) \quad ,$$

$$P(\sigma_2^{-2} \mid \underline{\beta},\underline{\phi},\underline{\nu},\sigma_1^2,\underline{d}) \propto \Gamma(\frac{a+b}{2}, \frac{c+\sum_j^a (\phi_j - \underline{x}_j'\underline{\beta})^2}{2}) \quad ,$$

$$\underline{\beta} \mid \underline{\nu},\underline{\phi},\sigma_1^2,\sigma_2^2,\underline{d} \sim N(\underline{\mu},\sigma_2^2\Sigma),$$

여기서

$$\underline{\mu} = (\sum_j \underline{x}_j'\underline{x}_j)^{-1}(\sum_j \phi_j \underline{x}_j)^{-1}, \ \Sigma = (\sum \underline{x}_j \underline{x}_j')^{-1}$$

이다. 모수 ϕ_j에 대한 조건부 사후확률분포함수인

$$P(\phi_j \mid \underline{\beta},\underline{\nu},\sigma_1^2,\sigma_2^2,\underline{d}) \propto \prod_i^N \exp[(\nu_i+\phi_j)d_{ij} - n_{ij}e^{\nu_i+\delta_j}] \times e^{-\frac{1}{2\sigma_2^2}(\phi_j - \underline{x}_j'\underline{\beta})^2}$$

으로부터 랜덤표본을 추출하기는 쉽지 않다. 따라서 3.1절에 기술된 바와 같이 2차 테일러 확장을 이용한 방법으로 전개하여 조건부 사후확률분포함수를 근사시킨 후 Metropolis-Hastings 알고리즘을 사용하여 표본을 추출하려 한다.

ϕ_j의 조건부 사후확률분포함수에 로그를 취한 값을 $\Delta(\phi_j)$이라 하면,

$$\Delta(\phi_j) = A(\phi_j) - \frac{1}{2\sigma_2^2}(\phi_j - \underline{x}_j'\underline{\beta})^2,$$

여기서

$$A(\phi_j) = \sum_{i=1}^N (\nu_i+\phi_j)d_{ij} - n_{ij}e^{(\nu_i+\phi_j)}$$

와 같다. $A(\phi_j)$에 1차 및 2차 미분을 수행한 결과를 각각 구하면

$$\frac{\partial A(\phi_j)}{d\phi_j} = \sum_i (d_{ij} - n_{ij}e^{(\nu_i+\phi_j)}) \ ,$$

$$\frac{\partial^2 A(\phi_j)}{d\phi_j^2} = -\sum_i n_{ij}e^{(\nu_i+\phi_j)}$$

과 같다. $\hat{\phi}_j$를 중심으로한 2차 테일러 확장의 응용으로 ϕ_j의 조건부 사후확률분포함수를 다음과 같은 분포로 근사시킬 수 있으므로 이를 proposal 밀도함수로 사용한다.

$$\phi_j \mid \underline{\beta},\underline{\nu},\sigma_1^2,\sigma_2^2,\underline{d} \sim N\left(\frac{\hat{\phi}_j \sum_{i=1}^N n_{ij}e^{\nu_i+\hat{\phi}_j}}{\sum_{i=1}^N n_{ij}e^{\nu_i+\hat{\phi}_j} + \frac{1}{\sigma_2^2}}, \frac{1}{\sum_{i=1}^N n_{ij}e^{\nu_i+\hat{\phi}_j} + \frac{1}{\sigma_2^2}}\right),$$

여기서

$$\hat{\phi}_j = \log(\sum_j d_{ij} / \sum_i n_{ij}e^{\nu_i})$$

이다.

모수 ν_i에 대한 조건부 사후확률분포함수 역시 랜덤표본을 추출하기는 쉽지 않다. 같은 절차에 의하여 ν_i의 조건부 사후확률분포함수에 로그를 취한 값을 $\Delta(\nu_i)$라 하기로 한다.

$$\Delta(\nu_i) = A(\nu_i) - \frac{1}{2\sigma_1^2}\nu_i^2,$$

여기서

$$A(\nu_i) = \sum_j (\nu_i + \phi_j)d_{ij} - n_{ij}e^{(\nu_i+\phi_j)}.$$

다음과 같이 1차 및 2차 미분을 수행한다.

$$\frac{\partial A(\nu_i)}{\partial \nu_i} = \sum_j (d_{ij} - n_{ij}e^{(\nu_i+\phi_j)})$$

그리고

$$\frac{\partial^2 A(\nu_i)}{\partial \nu_i^2} = -\sum_j n_{ij}e^{(\nu_i+\phi_j)}.$$

ν_i의 추정치는

$$\hat{\nu}_i = \log\left(\frac{\sum_j d_{ij}}{\sum_j n_{ij}e^{\phi_j}}\right)$$

와 같고, 테일러 확장을 이용하여 ν_i의 proposal 밀도함수를 다음과 같이 구한다.

$$\nu_i|\underline{\beta},\underline{\phi},\sigma_1^2,\sigma_2^2,\underline{d} \sim N\left(\frac{\hat{\nu}_i\sum n_{ij}e^{\hat{\nu}_i+\phi_j}}{\sum n_{ij}e^{\hat{\nu}_i+\phi_j}+\frac{1}{\sigma_1^2}}, \frac{1}{\sum n_{ij}e^{\hat{\nu}_i+\phi_j}+\frac{1}{\sigma_1^2}}\right).$$

최종적으로 3.2절에서 유도된 모수 $\underline{\beta},\underline{\nu},\underline{\phi},\sigma_1^2,\sigma_2^2$의 조건부 사후확률분포들로부터 Metropolis-Hastings 알고리즘에 의해 교대로 랜덤표본이 추출된다.

4. 모형의 평가

본 장에서는 세가지의 모형평가방법을 사용하여 3.1절과 3.2절에 소개된 모형의 비교분석을 하고자 한다. 그리고 질병지도를 이용하여 관찰된 대장암 사망률과 모형에 의해 예측된 평균사망률의 패턴을 비교하고자 한다.

4.1 사후 기대예측차 (Posterior Expected Predictive Deviance)

먼저 사후확률분포로부터 예측된 사망자의 수와 관찰된 실제 자료와의 차이를 측정하는 사후 기대예측차,

$$E\{P(\underline{d}^{obs}, \underline{d}^{\neq w})|\underline{d}^{obs}\},$$

를 고려하자. $P(\underline{d}^{obs}|\underline{d}^{\neq w})$는 예측값과 관찰값과의 차이를 측정하는 함수이다. 만약 제안된 모형이 적절하다면 이 함수는 작은 값을 갖게 된다. 여기서 예측값 $\underline{d}^{\neq w}$는 확률분포함수

$$f(\underline{d}^{\neq w}|\underline{d}^{obs}) = \int g(\underline{d}^{\neq w}|\underline{\lambda})h(\underline{\lambda}|\underline{d}^{obs})d\underline{\lambda}$$

를 따르는 확률벡터인데, $h(\underline{\lambda}|\underline{d}^{obs})$ 모수 $\underline{\lambda}$의 사후확률분포함수이고 $g(\underline{d}^{\neq w}|\underline{\lambda})$는 식 (3.1)에 주어진 확률질량함수이다.

차이함수 $P(\underline{d}^{obs}|\underline{d}^{\neq w})$는 다음과 같은 세가지 종류가 흔히 사용된다.

1. 카이제곱 방법

$$P(\underline{d}^{obs},\underline{d}^{\neq w}) = \sum_i\sum_j (d_{ij}^{obs} - d_{ij}^{\neq w})^2/(d_{ij}^{\neq w}+0.5).$$

2. 순위 방법

$$P(\underline{d}^{obs},\underline{d}^{\neq w}) = \sqrt{12}\sum_i\sum_j \{c_{ij}/(a+1)-0.5\}(d_{ij}^{obs} - d_{ij}^{\neq w}),$$

여기서 a는 연령층의 개수이고 $(a=7)$, $c_{ij} = rank(d_{ij}^{obs} - d_{ij}^{\neq w})$이다.

3. 포아송 방법

$$P(\underline{d}^{obs},\underline{d}^{\neq w}) = 2\sum_i\sum_j \left\{ (d_{ij}^{obs}+0.5)\ln\left(\frac{d_{ij}^{obs}+0.5}{d_{ij}^{\neq w}+0.5}\right) - (d_{ij}^{obs} - d_{ij}^{\neq w}) \right\}.$$

카이제곱 방법이 가장 널리 사용되는 것이고 순위방법은 Hettmansperger (1984, Chapter 5)이 제안한 것으로 윌콕슨 점수에 기초하여 구성되었다. 포아송방법은 포아송 표본분포의 가정하에 Waller, Carlin, Xia and Gelfand (1997) 가 제안한 것이다. 카이제곱 방법과 포아송 방법은 대표본 이론상으로 유사하다고 알려져 있다. 이상의 세가지 측도를 사용하여 두 모형의 성과를 비교한 결과는 <표 1>에 나와 있다. 여기서 모형1은 오프셋 모형이고 모형2는 연령적 특성 임의효과 모형을 의미한다. <표 1>의 결과는 연령적 특성 임의효과 모형이 모든 대지역에서 오프셋 모형보다 우수하였음을 명확히 보이고 있다.

4.2 표준화 잔차 (Standardized Residuals)

두 번째 모형평가방법으로 다음과 같이 정의되는 표준화 잔차를 사용하고자 한다 (Gelman, Carlin, Stern, Rubin 1995). 이 방법을 이용하여 모형들을 비교하고자 할 때에는 과적합의 문제를 피하기 위하여 교차타당성 (cross-validation) 방법을 사용하여야 한다. 먼저 $P(\underline{d}^{obs},\underline{d}^{\neq w}) = 2\sum_i\sum_j \left\{ (d_{ij}^{obs}+0.5)\ln\left(\frac{d_{ij}^{obs}+0.5}{d_{ij}^{\neq w}+0.5}\right) - (d_{ij}^{obs} - d_{ij}^{\neq w}) \right\}$를 (ij)에 위치한 관찰치만을 제외한 총 관찰치의 집합이라 하자. 또한 사망률 r_{ij}는 $r_{ij} = d_{ij}/n_{ij}$ 라고 정의하자.

<표 1> 사후기대예측차를 이용한 두 모형의 성과 비교

대지역	카이제곱 방법		순위 방법		포아송 방법	
	모형 1	모형 2	모형 1	모형 2	모형 1	모형 2
1	567.23	339.24	2275.02	1725.67	475.34	309.90
2	1342.39	823.78	5475.54	4281.73	1145.97	760.47
3	832.67	624.41	2592.96	2237.64	709.15	551.84
4	2116.61	1569.18	6885.92	5976.54	1859.90	1452.58
5	1568.84	1399.86	4252.68	4058.79	1336.82	1211.71
6	2422.14	1822.71	8011.16	6700.34	2133.38	1631.87
7	639.08	615.51	1447.00	1376.72	554.52	516.95
8	1516.79	1428.26	3487.29	3297.82	1304.56	1222.31
9	2143.97	1828.79	5733.44	5123.67	1819.89	1565.42
10	719.08	575.10	1884.89	1645.96	619.70	513.89
11	512.29	530.95	840.55	812.47	444.31	424.21
12	1148.23	715.43	4539.05	3412.99	1010.22	661.48

교차타당성 잔차는 $a_{ij} = r_{ij} - E(r_{ij}|\underline{d}_{(ij)})$이라 정의할 수 있고, 그에 따른 표준화 잔차는

$$DRES_{ij} = \frac{r_{ij} - E(r_{ij}|\underline{d}_{(ij)})}{SD(r_{ij}|\underline{d}_{ij})}$$

점추정값인 $E(r_{ij}|\underline{d}_{(ij)})$와 비교된다. 이러한 표준화 잔차를 이용하여 제안된 모형이 실제 자료에 얼마나 잘 적합되는지를 평가할 수 있다.

만약 (ij)번째 관찰치가 $|DRES_{ij}| \geq q$를 만족한다면 이 관찰값은 모형이 잘 적합시키지 못하는 이상점에 해당한다고 할 수 있다. 여기서 q는 주로 3 혹은 4를 사용한다. 본 논문에서는 $|DRES_{ij}| \geq q$에 해당하는 관찰값의 수를 측정함으로써 두 모형의 적합도를 비교하고자 한다. 이를 위해 각 연령층별 이상점의 개수와 각 대지역별 이상점의 개수를 비교한다. <표 2>와 <표 3>은 이에 대한 결과를 보이고 있다. 여기서 <표 2>는 각 연령층별로 <표 3>은 각 대지역별로 이상점에 해당하는 HSA의 개수를 기록한 것이다. 모형 2를 사용하면 표준화 잔차가 매우 큰 HSA의 개수가 감소하므로 모형 1보다 더 좋은 모형이라 결론지을 수 있다.

4.3 잔차분석

각 모형의 잔차와 잔차의 표준편차를 시각화함으로써 모형의 성과를 비교해 볼 수도 있다. <그림 3>은 잔차와 잔차의 표준편차를 ± 2의 밴드를 이용하여 그린 것이다.

<표 2> 연령층별 표준화 잔차를 이용한 두 모형의 성과 비교

대지역	$\|DRES_{ij}\| \geq 3$		$\|DRES_{ij}\| \geq 4$	
	모형 1	모형 2	모형 1	모형 2
1	11	2	8	0
2	29	6	16	2
3	14	5	5	1
4	46	17	23	5
5	21	8	9	1
6	46	11	23	1
7	6	4	3	1
8	15	6	7	3
9	29	10	11	2
10	10	2	6	1
11	2	5	2	2
12	28	3	16	0

<표 3> 대지역별 표준화 잔차를 이용한 두 모형의 성과 비교

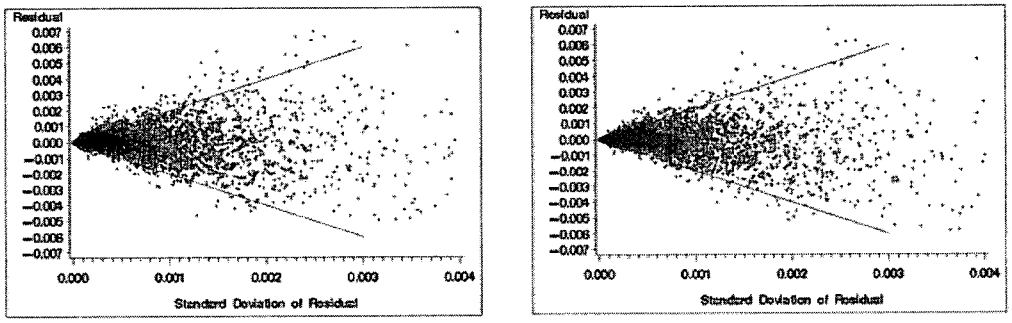
연령층	$\|DRES_{ij}\| \geq 3$		$\|DRES_{ij}\| \geq 4$	
	모형 1	모형 2	모형 1	모형 2
≤ 4	110	21	77	8
5	32	36	11	21
6	8	11	0	0
7	52	9	22	1
8	16	9	7	2
9	12	7	3	0
10	27	12	9	5

포아송 분포의 평균과 분산은 같다는 특성상 사망률이 높은 HSA의 잔차 표준편차는 더 커지게 마련이므로 ± 2의 밴드는 평행하지 않고 깔때기 모양을 하게 된다. <그림 3>에 의한 결과는 모형 2가 모형 1보다 약간 좋은 결과를 보이고 있다.

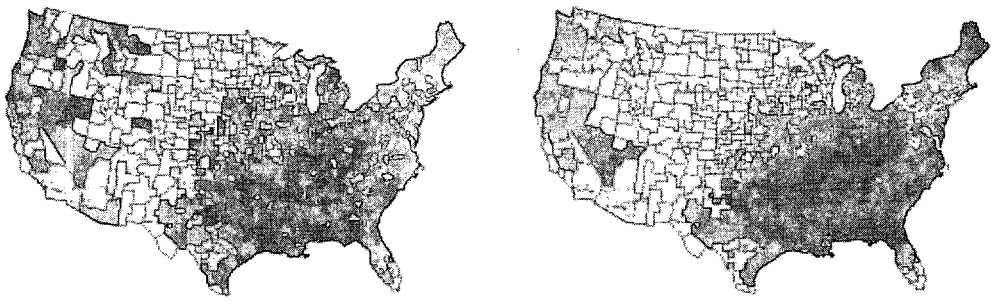
4.4 질병지도

<그림 4(a)>는 75-84세 연령층 대장암 사망률의 관측값을 이용하여 질병지도를 그린 것이다. 또한 같은 연령층에서 모형 2의 모수추정값 (사후분포의 평균값)을 이용한 사망률의 예측치를 나타낸 질병지도는 <그림 4(b)>에 표현되었다. 사망률은 5가지 색상으로 표현되었으며 인구 10만명당 대장암 사망자수 0-490명, 491-560명, 561-620명, 621-690명, 691명 이상의 순서대로 진한 색을 띤다. 모형 1의 모수추정값을 이용한 사망률 질병지도는 모형 2의 지도와 유사하여 생략하였으며 다른 연령층에 대한 질병지도 역시 생략하였다. 결론적으로 모형 2에 의한 사망률의 질병지도는 공간효과의 영

향으로 실제 관찰값을 기준으로 한 질병지도보다 더 매끄러운 예측값을 보이고 있으나 전체적인 패턴에 있어서는 큰 차이를 보이지 않는다. 질병지도에 의하면 아팔라치안 산맥의 주변지역과 동남부 지역의 사망률이 높은 것을 알 수 있다.



<그림 3> 잔차와 잔차의 표준편차



(a) (b)

<그림 4> 관찰된 사망률과 사후평균사망률을 이용한 대장암의 질병지도 - 연령층 75-84세 기준

5. 결론

본 논문에서는 대장암 사망률에 대한 베이지안 모형을 제안하였다. 대장암을 포함한 모든 질병에 의한 사망률은 연령적 특성과 지역적 특성이 함께 고려되어야 더 정확한 분석을 할 수 있다. 기존의 대장암 사망률 분석은 공간지역적 특성을 임의효과로 놓고 연령적 특성을 선형회귀식에 포함한 고정효과로 모형을 적합해 왔다. 본 논문에서는 대장암의 경우에 있어서 연령적 특성이 임의효과의 역할을 하는 모형을 세우고 이 모형의 성과를 평가하였다. 본 논문에서 고려된 모형들은 베이지안 방법인 Metropolis-Hastings 알고리즘을 이용하여 적합하였고, 이를 위한 이론적 전개과정을 구체적으로 보였다. 적합된 모형을 비교하기 위해서 사후 기대예측차와 표준화 잔차 그리고 잔차분석을 활용하였으며, 결론적으로 연령적 특성이 임의효과로 추가된 두 번째 모형이 더 좋은 예측력을 갖는 것을 알 수 있었다. 마지막으로 추정값을 이용한

질병지도를 작성하여 관찰된 사망률에 근거한 질병지도와 비교하였으며, 공간효과로 인하여 모형 추정값에 의한 질병지도가 더 매끄러운 형태를 가지며 적합된 모형이 관찰값을 잘 예측함을 보였다.

참고문헌

[1] Brillinger, D.R. (1996). The natural variability of vital rates and associated statistics. *Biometrics*, Vol. 42, 693-734.

[2] Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, Vol. 49, 327-335.

[3] Gelfand, A.E., Sahu, S.K., and Carlin, B.P. (1995). Efficient Parameterizations for Normal Linear Mixed Models. *Biometrika*, Vol. 82, 479-488.

[4] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

[5] Gilks, W.R. and Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, Vol. 41, 337-348.

[6] Hettmansperger, T. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

[7] Nandram, B. and Kim, H. (2002). Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical Computation and Simulation*, Vol. 72, 319-340.

[8] Nandram, B., Sedransk, J., and Pickle, L. (1999). Bayesian Analysis of Mortality Rates for U.S., Health Service Areas. *Sankhya*, Series B, Vol. 61, 145-165.

[9] Nandram, B., Sedransk, J. and Pickle, L. (2000). Bayesian analysis and mapping of mortality rates for chronic obstructive pulmonary disease. *Journal of the American Statistical Association* Vol. 95, 1110-1118.

[10] Pickle, L.W., Mungiole, M., Jones, G.K., and White, R. C. (1996). *Atlas of U.S. Mortality*. National Center for Health Statistics, Hyattsville, MD.

[11] Waller, L., Carlin, B., Xia, H. and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association,* Vol. 92, 607-617.

[Received October 2005, Accepted April 2006]