응용통계연구 제 17권 3호, 2004년, pp. 431-443

HGLM과 EB 추정법을 이용한 질병지도의 작성 *

김영원 [1] 조나경 [2]

요 약

본 연구에서는 질병지도작성(disease mapping)을 위해 인접지역의 정보를 효과적으로 활용할 수 있는 EB(empirical Bayes) 추정법과 HGLM(hierarchial generalized linear model)을 기초로 한 추정법을 다룬다. 사례연구로 이 추정방법들을 이용하여 2000년 사망원인통계자료를 이용해 경상도 및 전라도의 112개 시·군·구 단위 행정자치구역별 45세 이상 폐암 사망률을 산출하고, 경상도 및 전라도 지역 폐암 사망률 지도를 작성한다. 아울러 제시된 방법들에 위해 얻어진 추정치들의 변동과 3년간 평균 사망률을 기준으로 구한 MSD(mean square deviation)를 이용하여 추정방법들의 특성을 비교 분석한다.

주요용어: EB추정법, HGLM, 사망률, 소지역추정, 질병지도,

1. 서론

우리나라에서도 지역별 통계 산출 필요성이 증대됨에 따라 각종 정부 공식통계에 있어서 전국 또는 도 단위의 통계뿐만 아니라 시·군·구 등의 소지역 통계에 대한 요구가 날로 높아지고 있다. 또한 컴퓨터의 발달로 GIS개념을 도입하여 각 지역별 통계를 한눈으로 보기 쉽게 지도에 나타내어 지역별 특성을 파악하는 연구가 활발히 진행되고 있다. 특히 해당지역의 경제 환경 요인이 질병에 따른 사망률 또는 이환율에 많은 영향을 줄 수 있기 때문에 최근 지역별 환경 요인에 따른 사망률 및 이환율 분석이 우리나라에서도 급격히 관심 대상으로 부각되고 있다.

그러나 날로 증가하고 있는 소지역별 사망률 또는 이환율에 대한 관심과는 달리 보건관련 통계를 주로 다루는 통계청 또는 보건사회연구원 등에서는 아직 광역시 또는 도 단위 통계에 중점을 두고 있다. 따라서 시·군·구 등과 같은 소지역별 보건통계를 파악하는 것이 어려운 것이 현실이고, 동시에 지역별 차이를 한 눈에 볼 수 있도록 사망률이나 이환율을 지도에 표현하는 질병지도에 대한 연구가 전혀 이루어지고 있지 않다. 그러므로 우리나라에서도 적절한 보건통계 작성 및 분석을 위해 질병지도 관련 연구에 관심을 갖는 것이 절실히 요구된다.

질병의 지리학적 변동을 한 눈에 파악하기 위해 소지역별 사망률이나 이환율을 지도에 나타내는 질병지도작성(disease mapping)은 역학 연구를 위한 도구로 그 활용이 점차 증대되고 있다. Howe(1970)에 의하면 이러한 시도는 이미 18세기에 시작되었으며, 컴퓨터와 컴

* 본 연구는 숙명여자대학교 2003년도 교내연구비 지원에 의해 수행되었음

1) (140-742) 서울 용산구 청파동 효창원길 52, 숙명여자대학교 통계학과, 교수,
E-mail: ywkim@sookmyung.ac.kr
2) (140-742) 서울 용산구 청파동 효창원길 52, 숙명여자대학교 통계학과 대학원

퓨터 그래픽의 발전으로 좀 더 쉽게 질병지도를 작성하는 것이 가능해졌다. 하지만 질병지도 작성을 위한 추정기법에 대한 연구는 1980년대 후반부터 본격적으로 관심을 갖게 되었다(Marshall, 1991).

질병지도의 추정문제를 본격적으로 다루기 시작한 연구는 Clayton과 Kaldor(1987)라고 볼 수 있으며, 이들은 포아송-감마 모형을 기반으로 하는 EB(empirical Bayes)추정량을 제안하였다. 또한 Marshall(1991)은 이들이 제시한 EB추정방법을 변형한 선형최적 EB 추정량 (linear best empirical Bayes estimator)을 질병지도 작성에 활용하였다. 최근에 Ghosh 등(1998)은 이산형 자료에 대한 소지역추정기법에 해당하는 GLM(generalized linear model)을 기반으로 한 HB(hierarchical Bayes) 추정법에 대한 이론을 제시하였으며, Rao(2003)는 최근까지의 많은 연구 결과들을 토대로 질병지도에 활용 가능한 소지역 추정기법들을 체계적으로 정리하고 있다.

특히, 본 연구에서는 Ghosh 등(1998)이 제시한 GLM을 기초로 한 HB추정법의 대안으로 상당히 효과적일 것으로 판단되는 Lee와 Nelder(1996, 2001)가 제시한 HGLM(hierarchical generalized linear model)을 질병지도 작성을 위해 활용하는 방안을 새로이 제안한다. HGLM은 질병지도 작성을 위해 적합한 고정 및 랜덤 효과를 포함한 혼합일반화선형모형에 적용될 수 있으며, HB추정법과는 달리 모형의 모수에 대한 사전분포를 가정하는 대신, 전통적인 통계적 추론입장에서 확장된 개념의 우도함수에 해당하는 h-likelihood(Lee와 Nelder, 1996)를 기반으로 한 통계적 추론방법이다.

사례 연구로 제시된 추정방법들을 적용하여 우리나라 시·군·구 질병지도 작성문제를 다루고자 한다. 이를 위해 2000년 통계청 사망원인 통계자료를 토대로 경상도 및 전라도지역 내의 시·군·구에 대한 45세 이상 폐암 사망률 추정 및 질병지도 작성방안을 제시한다. 참고로 이 지역의 2000년 45세 이상 폐암 사망자수는 4,997명이고, 2000년 45세 이상 주민등록 인구는 5,338,159명이다. 경상도 및 전라도 지역에는 4개 광역시에 31개 구가 있고, 도 지역은 51개의 군과 31개의 시로 구성되어 있는데, 그 중 울릉군을 제외한 112개의 시·군·구 지역의 45세 이상 폐암 사망률을 연구대상으로 한다.

본 연구에서는 2000년 통계청의 사망원인통계자료 및 주민등록인구자료를 바탕으로, 우선 기존의 연구결과들을 기초로 한 이들 지역의 폐암 사망률에 대한 직접추정방식에 의한 표준사망률(standardized mortality ratio), 전지역(total) 적률추정 및 지역별(local) 적률추정을 이용한 EB추정결과를 산출하는 동시에 새로 제시한 HGLM을 적용해 구한 추정결과를 바탕으로 각 추정기법들에 따른 질병지도를 작성하여 그 결과를 비교 분석한다. 실제 우리나라 폐암 사망률 사례분석 결과를 토대로 각 추정방법의 효율성을 직접 비교할 수는 없다. 따라서 여기서는 각 방법에 따른 추정결과의 비교를 통해 제시된 추정법들의 특성을 비교하는데 중점을 두고 있다.

2. 질병지도 작성을 위한 소지역 추정

2.1. 경험적 베이즈 추정

우선 2000년도 경상도 및 전라도 지역의 112개 시·군·구(울릉군제외)의 45세 이상 인구

를 대상으로 한 폐암 표준사망률 산출방법을 살펴보도록 한다. Clayton과 Kalder(1987) 등이 제시한 것과 같이 소지역 i에 대한 45세 이상 폐암 표준사망률을 추정하기 위한 직접추정량 $\hat{\theta}_i$은 다음과 같다.

$$\hat{\theta}_i = y_i/e_i = \frac{y_i}{n_i\Big(\sum_i y_i/\sum_i n_i\Big)}, \quad i = 1, ..., 112 \tag{2.1}$$

여기서 y_i는 소지역 i에서 45세이상 폐암 사망자수이고 n_i는 45세이상 주민등록인구이다.

한편 Clayton과 Kalder(1987)는 직접추정의 경우 인접지역간에도 차이가 너무 크게 나타나는 문제를 완화하고, 인접지역간의 유사성을 반영하기 위해 포아송-감마 모형을 가정한 EB추정법을 적용하여 직접추정값을 평활(smoothing)시켜 질병지도를 작성하는 방법을 제시했다. Marshall(1991)은 이를 변형한 선형최적 EB추정량(best linear empirical Bayes estimator)을 질병지도 작성에 활용하였다. 특히 Marshall은 EB추정에서 베이즈 모수 추정을 위해 전체 지역을 대상으로 한 적률추정법과 인접 지역만을 이용하는 적률추정법을 고려하고 있다.

질병지도 작성을 위해 Marshall(1991)이 제시한 선형최적 EB추정량은 다음과 같다. θ_i를 i번째 소지역의 연간 사건 발생비율, y_i는 포아송분포를 따르는 연간 발생 건수라고 가정하면, θ_i의 MLE인 $x_i = y_i/n_i$이고, θ_i가 주어진 경우, x_i의 평균과 분산은 각각 $E(x_i|\theta_i) = \theta_i$, $var(x_i|\theta_i) = \theta_i/n_i$ 이다.

선형최적 베이즈 추정량을 고려하기 위해, θ_i의 구체적인 분포를 가정하지 않고 θ_i의 평균을 m_i, 분산을 A_i라고 가정하면, 결합분포에서 x_i의 평균은 $E_\theta\{E(x_i|\theta_i)\} = m_i$이고 분산은 다음과 같다.

$$var(x_i) = var_\theta\{E(x_i|\theta_i)\} + E_\theta\{var(x_i|\theta_i)\} = A_i + m_i/n_i$$

따라서 θ_i 의 최량선형 베이즈추정량은 다음과 같은 수축추정량 형태로 표현된다.

$$\hat{\theta}_i = m_i + C_i(x_i - m_i) \tag{2.2}$$

여기서, $C_i = A_i/(A_i + m_i/n_i)$이다. Marshall(1991)은 모형의 단순화를 위해 모든 i, 다시 말해 전체 지역에 대해 $A_i = A$, $m_i = m$으로 가정하였다. 식 (2.2)에서 m과 A를 전체 지역의 자료에서 구한 적률추정량 $\tilde{m}$과 $\tilde{A}$로 대체하게 되면 Marshall(1991)이 제안한 전지역(total) 적률추정 EB추정량(EB_T)이다. 이를 표준 사망률 형식으로 표현하면 다음과 같다.

$$\tilde{\theta}_i = \frac{\tilde{m} + \tilde{C}_i(x_i - \tilde{m})}{\sum_i y_i/\sum_i n_i} \tag{2.3}$$

여기서 $\tilde{C}_i = \frac{\tilde{A}}{\tilde{A}+\tilde{m}/n_i} = \frac{s^2 - \tilde{m}/\bar{n}}{s^2 - \tilde{m}/\bar{n} + \tilde{m}/n_i}$ 이고, 가중표본분산 $s^2 = \frac{1}{n}\sum_i n_i(x_i - \tilde{m})^2$을 이용한 추정량으로 $\tilde{A} = s^2 - \tilde{m}/\bar{n}$ 를 사용한 것이고, 만약 이것이 음수인 경우, 즉 $s^2 < \tilde{m}/\bar{n}$일때는, $\tilde{A} = 0$으로 처리한다.

한편 지리적으로 인접한 지역은 환경 및 경제적인 요인 때문에 상당부분 유사성을 갖는 것이 일반적이다. 따라서 Marshall(1991)은 비슷한 이환율을 가질 가능성이 높은 서로 인접

한 소지역을 하나의 지역으로 묶어 이 지역에 동일한 사전 모수를 가정하는, 다시 말해 지역(local) 적률추정량을 사용한 EB추정량(EB_L)을 아울러 제시하였다. 이 경우 θ_i의 추정은 직접추정치 x_i가 인접지역 평균으로 끌려가는 형태로 나타나게 된다. 이를 정리하면 식 (2.3)에서 $\tilde{C}_i$, $\tilde{m}$, s^2, n 을 각각 i지역의 인접지역만을 대상으로 계산하고, 이를 $\tilde{C}_{(i)}$, $\tilde{m}_{(i)}$, $s^2_{(i)}$, $\tilde{n}_{(i)}$ 으로 표현하면, 인접지역 개념을 도입한 EB추정량은 다음과 같이 나타낼 수 있다.

$$\tilde{\theta}_{(i)} = \frac{\tilde{m}_{(i)} + \tilde{C}_{(i)}(x_i - \tilde{m}_{(i)})}{\sum_i y_i / \sum_i n_i} \tag{2.4}$$

식 (2.4)의 인접지역 개념을 도입한 EB_L 추정량을 구하기 위해서는 각 소지역에 대한 인접지역을 어떻게 정의할 것인지를 결정해야 한다. Clayton과 Kaldor(1987)은 인접지역 개념을 표현하기 위해 다음과 같은 요소들로 구성된 인접 행렬(W)을 사용하고 있다.

$$W_{ij} = \begin{cases} 1 & i\text{와 } j\text{는 인접지역} \\ 0 & \text{그렇지 않을 때} \end{cases}$$

본 연구에서는 두 가지 방법으로 인접지역을 정의하고자 한다. 경상도와 전라도 전체 시·군·구를 광역시와 도지역으로 분류하고, 광역시의 경우 구 단위를 기준으로 그 밖의 도지역은 시·군 단위를 기준으로 인접행렬을 정의한다. 구체적으로 인접지역은 다음과 같이 두 가지 방법으로 정의한다. 첫 번째는 지역적으로 경계를 같이 하는지 여부를 기준으로 인접지역을 정의한다. 이는 지리적으로 가까운 위치에 있는 소지역들이 지리 환경적인 측면에서 유사성을 가질 수 있다는 것을 반영한 것이다. 이런 방식으로 정의된 인접지역 개념을 반영한 EB추정량을 $EB_L(1)$으로 표시하기로 한다. 두 번째는 동일한 광역시의 경우 광역시내의 모든 구를, 시·군의 경우 각 도내에서 시는 시끼리 군은 군끼리 인접지역으로 정의한다. 이는 광역시와 도지역을 구분하고 동시에 도지역에서 시지역과 군지역을 구분한 것으로, 도시화율에 따른 지역간의 유사성을 반영하기 위한 것이다. 이런 방식으로 정의된 인접지역 개념을 반영한 EB추정량은 $EB_L(2)$로 표시하기로 한다.

2.2. HGLM에 의한 추정

이산형 자료에 대한 소지역 추정기법으로 Ghosh 등(1998)이 제시한 GLM을 기초로 한 베이즈 접근법의 대안으로 Lee와 Nelder(1996, 2001)가 제시한 HGLM을 활용하는 방안을 고려해 볼 수 있다. HGLM은 고정 및 랜덤 효과를 포함한 혼합일반화선형모형에 적용될 수 있다. HB추정을 위해서는 Gibbs Sampling과 같은 복잡한 계산과정을 거쳐야 하고, 아직 이런 복잡한 수행과정을 손쉽게 처리할 수 있는 소프트웨어를 이용할 수 없다는 현실적인 문제를 갖고 있다. 하지만 HGLM은 이런 복잡한 과정을 수행할 필요가 없고, 특히 통계처리 소프트웨어인 GenStat에서 손쉽게 수행될 수 있다는 것도 실제 활용에 있어서 큰 장점이 될 수 있다.

Lee와 Nelder(1996)가 제시한 HGLM은 다음과 같다. $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ 는 관측 자료이고, 랜덤효과를 나타내는 관측이 안 되는 확률변수는 $\mathbf{u} = (u_1, u_2, \cdots, u_m)^T$ 로 나타내

며, 랜덤효과 $\mathbf{u}$가 주어진 경우, 확률변수 $\mathbf{y}$ 는 다음과 같은 GLM 로그우도함수를 갖는다.

$$l_0(\theta(\mu_0), \phi; y|u) = \sum([y_i\theta(\mu_{0i}) - b\{\theta(\mu_{0i})\}]/\phi_i + k(y_i, \phi_i))$$

여기서, $\theta(\mu_{0i})$는 정준모수(canonical parameter), ϕ_i는 산포모수(dispersion parameter)를 나타낸다. 선형 예측량(predictor)은 다음의 형식을 취한다.

$$\eta_0 = g(\mu_0) = \mathbf{X}\underline{\beta} + \mathbf{Z}v$$

여기서, $\mu_0 = (\mu_{01}, \mu_{02}, \cdots, \mu_{0n})^T$, $g()$ 는 연계함수(link function), $\mathbf{X}$는 고정효과 $\underline{\beta}$에 대한 모형 행렬이고, $\mathbf{Z}$는 랜덤효과 $v = g_1(u)$ 에 대한 모형 행렬이다. 여기서, $v_i = g_1(u_i)$ 는 단조증가함수이다. 한편 랜덤효과 u_i 에 대해 적절한 분포를 가정한다.

이와 같은 HGLM모형에 대한 추론은 Lee와 Nelder(1996)가 제시한 h-likelihood를 기반으로 가능하다. HGLM을 기반으로 한 통계적 추론은 HB추정에 대한 대안으로 매우 효과적인 것으로 알려져 있으며, HB추정법과 달리 사전분포에 대한 가정이 필요 없고, 베이즈 추론과는 달리 고전적인 통계추론 개념인 우도함수를 기반으로 한다는 특징을 갖는다. 참고로 HB추정법의 경우 계산상 많은 어려움이 있는 데 반해 HGLM은 확장된 형태의 우도함수인 h-likelihood를 기반으로 산출된 추정방정식을 이용하여 모수와 랜덤효과에 대한 MHLE(maximum h-likelihood estimation)을 구하게 된다. HGLM와 관련된 실제 자료 분석은 통계처리 패키지인 GenStat에 의해 수행이 가능하다.

본 연구에서는 HGLM을 활용하여 질병지도 작성을 위한 소지역 추정을 위해 구체적으로 다음과 같은 모형을 고려한다. 사망률처럼 관측자료가 포아송분포에 따른다는 가정이 적합한 경우 소지역 추정을 위한 다음과 같은 HGLM모형을 고려할 수 있다.

$$y_i \sim Poisson(\theta_i) \quad \& \quad log(\theta_i) = \mu_i = x_i^T\underline{\beta} + u_i$$

여기서, x_i는 i번째 소지역의 지역특성 등을 나타내는 보조변수, u_i는 i번째 소지역의 지역특성을 나타내는 랜덤효과이다. 실제 폐암 사망률 분석을 위해 $Y_{ik}|\theta_{ik} \sim ind\ Poisson(\theta_{ik})$, v_i 는 평균이 0이고 분산이 σ^2 이라고 가정한다. Y_{ik}는 i번째 소지역의 k번째 그룹(성별; 남, 여, 나이; 45-54세, 54-64세, 65세이상)에서의 폐암 사망자수이고 $\theta_{ik} = E_{ik}e^{\mu_{ik}}$으로 예측 사망자수를 나타낸다. 이 때, $E_{ik} = n_{ik}\left(\sum_{i,k} y_{ik}/\sum_{i,k} n_{ik}\right)$ 이고, $\mu_{ik} = log\left(\frac{\theta_{ik}}{E_{ik}}\right)$는 로그상대위험이다.

본 연구에서는 HGLM을 활용하기 위해 두 가지 경우를 고려한다. 첫 번째는 지역만을 고려한 모형 HGLM(1), $\mu_{ik} = \tau + u_i + \epsilon_{ik}$ 이고, 두 번째는 성별, 연령, 지역을 모두 고려한 모형 HGLM(2), $\mu_{ik} = x_{ik}^T\underline{\beta} + u_i + \epsilon_{ik}$ 를 연구대상으로 한다. 그룹 범주 요인을 좀 더 명확히 표현하기위해 성별-연령별 범주를 나타내는 첨자 k를 성별요인 s, 연령요인 a 로 표시하면 그룹 요인은 다음과 같이 표현될 수 있다.

$$x_{as}^T\beta = v_s\alpha + z_a\gamma + v_s z_a\xi$$

여기서, $v_s = \begin{cases} 0 & \text{if } s = 1 \ \text{(남자)} \\ 1 & \text{if } s = 2 \ \text{(여자)} \end{cases}$ 이고 $z_a = \begin{cases} -1 & \text{if } a = 1 \ \text{(45세} \sim \text{54세)} \\ 0 & \text{if } a = 2 \ \text{(55세} \sim \text{64세)} \\ 1 & \text{if } a = 3 \ \text{(65세이상)} \end{cases}$ 이다.

3. 시·군·구별 사망률 추정결과

경상도 및 전라도 지역의 45세 이상 폐암 사망률을 앞에서 제시한 다양한 형식의 EB 추정방법으로 산출하고, 이들 추정결과를 기초로 SPSS/Maps로 질병지도를 작성한 결과는 그림 3.1-그림 3.4와 같다.

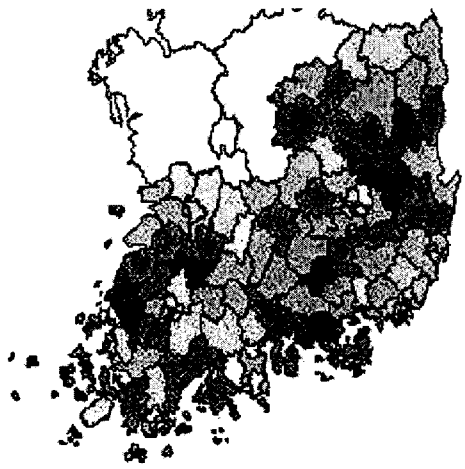


그림 3.1: 직접추정

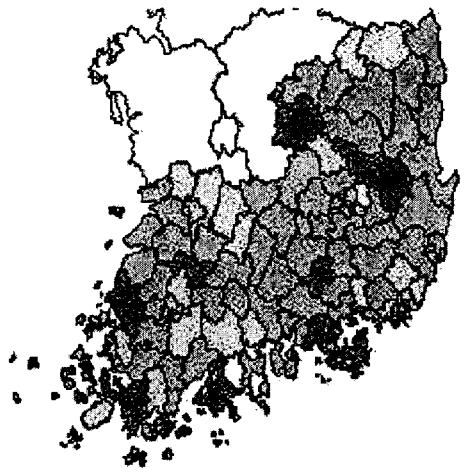
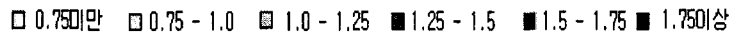


그림 3.2: EB_T 추정



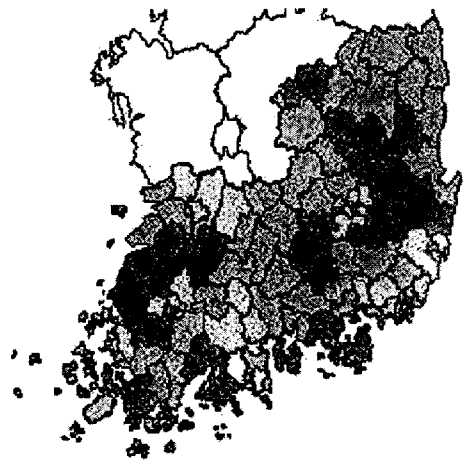


그림 3.3: $EB_L(1)$ 추정

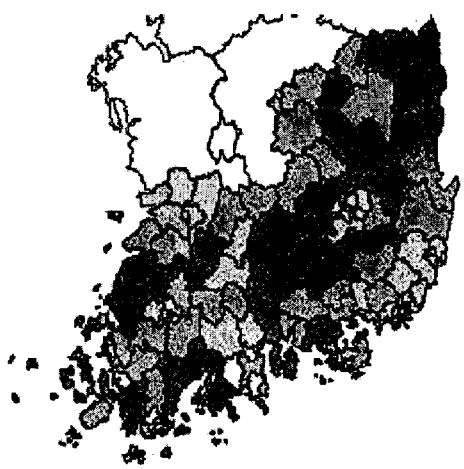


그림 3.4: $EB_L(2)$ 추정

그림 3.1은 자료에서 직접 구한 표준사망률 지도로 표준사망률은 0.60과 2.20사이의 값을 가지며, 경남 의령군, 전남 함평군, 전북 순창군의 45세 이상 폐암 사망률이 다른 지역보다 특히 높게 나타났고, 폐암 사망률이 0.75미만인 11개 지역 중에서는 광역시의 구가 8개 지역인 것으로 나타났다. 그림 3.2의 EB_T는 0.72와 1.50사이의 값을 가지며, 표준사망률에 대한 직접추정결과와 비교해 보면, 폐암 사망률이 높거나 낮은 소지역이 얼마나 효과적으로 전체 평균 쪽으로 끌려갔는지 알 수 있다.

이제 지역별 적률추정을 이용한 EB추정결과에 대해 살펴보자. 지역별 적률추정을 이용한 EB추정결과는 각 소지역의 인접지역 정의에 의해 차이가 있다. 우선, 경계선 공유여부를 기준으로 인접지역을 정의한 $EB_L(1)$의 경우, 가장 적은 경우 2개, 가장 많은 경우 9개의 소지역을 인접지역으로 갖으며, 평균적으로 5.16개 소지역을 인접지역으로 갖는다. 한편 $EB_L(2)$의 경우 가장 적은 경우 5개, 가장 많은 경우 17개의 소지역을 인접지역으로 갖으며, 평균적으로 10.96개 소지역을 인접지역으로 갖는다. 그림 3.3의 $EB_L(1)$과 그림 3.4의 $EB_L(2)$를 살펴보면, $EB_L(1)$은 0.74와 1.66사이의 값을 가지며 $EB_L(2)$는 0.71과 1.58사이의 값을 갖는다. 또한 두 경우 모두 EB_L은 대체적으로 직접추정과 EB_T를 적절히 잘 보정하고 있다는 것을 볼 수 있다.

마지막으로 HGLM을 활용하는 경우, 실제 자료 분석을 GenStat에 의해 수행하여 얻은 잔차분석 결과는 그림 3.5 및 그림 3.6과 같다.

그림 3.5: HGLM(1)의 잔차 정규확률그림과 히스토그램

그림 3.6: HGLM(2)의 잔차 정규확률그림과 히스토그램

모형의 적합성을 파악하기 위한 잔차분석 결과를 살펴보면, 그림 3.5에 있는 HGLM(1)의 경우는 정규성 가정에서 상당히 벗어나는 것으로 보이는 반면 그림 3.6에 있는 HGLM(2)의 경우는 정규성 가정이 상당히 적절한 것으로 보인다. 따라서 HGLM(2) 모형이 HGLM(1)보다 사망률 자료에 더 적합한 모형이라는 것을 알 수 있으며, 이는 HGLM 모형에 성별 및 나

이 변수를 반영하는 것이 효과적이라는 것을 입증해 주고 있다.

표 3.1: HGLM(2)의 GenStat결과

	estimate	s.e.	t
Constant	-1.2022	0.0644	-18.67
나이 2	1.5844	0.0652	24.32
나이 3	2.7211	0.0612	44.49
성별 2	-1.0811	0.1144	-9.45
나이 2 .성별 2	-0.5928	0.1357	-4.37
나이 3 .성별 2	-0.4324	0.1208	-3.58

또한 표 3.1을 보면 성별, 연령, 지역 모두를 고려한 HGLM(2)의 경우, 제시된 $|t|$를 보면 통계적으로 모든 변수가 유의한 것으로 나타났다. 따라서 HGLM모형에서는 지역효과만을 고려한 HGLM(1)보다는 성별, 연령, 지역 모두를 고려한 HGLM(2)가 더 적합하다는 결론을 얻을 수 있다.

한편 HGLM(1)과 HGLM(2)의 경우 앞에 제시된 EB추정법과 달리 모형을 통해 각 소지역의 성별, 연령 범주에 따른 예측 사망자수를 얻게 된다. 앞에 제시된 EB추정결과와의 비교를 위해서는 HGLM 결과를 토대로 성별, 연령별 범주를 통합한 각 소지역별 45세 이상 사망률 추정이 필요하다. 따라서 각 소지역의 성별, 연령별 범주에 따른 예측 사망자수($\hat{\theta}_{ik}^H$)를 합쳐서 i 소지역 예측 사망자수($\hat{\theta}_i^H$)를 산출하고, 즉 $\hat{\theta}_i^H = \sum_k \hat{\theta}_{ik}^H$, 이를 토대로 i 소지역 사망률 추정치를 구한다.

앞에서 본 것처럼 사망률 자료에 대한 적합도가 높은 것으로 판단되는 HGLM(2)를 기초로 산출된 시·군·구별 사망률 추정결과를 지도에 나타내면 그림 3.7과 같다. 이 그림을 보면, 앞의 EB_L과는 다른 관점에서 직접 추정결과를 보정해 주고 있는데, 여기서는 각 지역별 성별, 연령별 차이가 추정결과에 반영되고 있다.

Clayton과 Kaldor(1987), 그리고 Marshall(1991) 등이 지적한 것과 같이 직접추정량을 사용하는 경우 경제, 환경적인 요소가 매우 유사한 인접지역에서도 논리적인 설명이 곤란할 정도로 사망률 격차가 상당히 크게 발생할 수 있다. 하지만 인접지역의 정보를 종합적으로 반영하는 EB추정량을 사용하면 인접지역의 사망률 격차를 완화할 수 있다. 한편 이런 EB추정량의 경우 사망률 분석에 있어 매우 중요한 요인인 성별 또는 연령 등이 사망률에 미치는 영향을 분석하기에는 적합한 방법은 아니다. 따라서 본 연구에서 새로 제안한 HGLM을 활용한 분석기법은 성별 또는 연령이 사망률에 미치는 영향을 통계적 모형을 통해 정확히 분석할 수 있다는 점에서 상당히 효과적인 분석기법이 될 수 있다.

참고로 $\hat{\theta}_i^H$는 Marshall(1991) 등이 제시한 EB추정법에서 반영하지 못한 각 소지역에 대해 성별 및 연령 요인에 따른 사망률 변동을 추정식에 반영하고 있기 때문에 상당히 효율적인 추정결과를 제공할 수 있다고 예상된다. $\hat{\theta}_i^H$ 의 효율성에 대한 연구는 실용적인 측면에서 뿐만 아니라 통계 이론적 측면에서도 매우 중요한 연구 과제라고 생각된다.

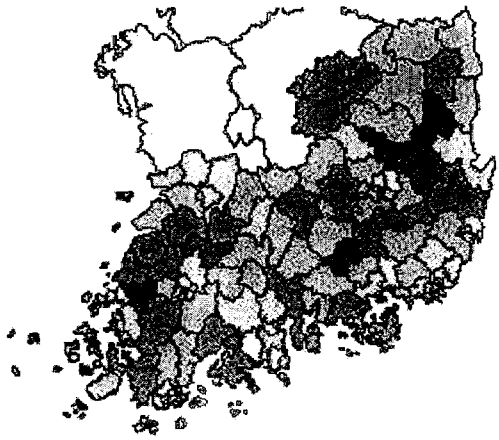


그림 3.7: HGLM(2) 추정결과

4. 추정방법들의 특성 비교

질병지도 작성을 위해 제시된 추정방법들의 경우, 적절한 평활(smothing)을 통해 직접 추정에서 나타나는 인접지역간의 극단적인 변동을 제거하여 경제·환경적인 요인으로 지역별 사망률을 설명하는 것을 목적으로 한다. 따라서 이들 추정방법의 오차는 실제적으로 별 의미가 없으며, 이보다는 추정량이 어느 정도 지역간 특성을 반영하면서 극단적인 추정치를 도출하지 않고 안정적인 추정결과를 제공하느냐 하는 것이 주관심대상이 된다. 따라서 여기서는 제시된 추정방법들의 안정성(stability)에 초점을 두고 추정방법들의 특성을 비교해 보고자 한다. 우선 각 추정방법에 따른 안정성을 보기 위해 각 추정방법에 의해 얻어진

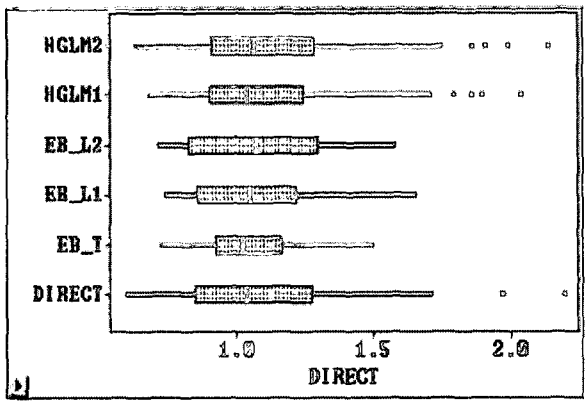


그림 4.1: 소지역별 폐암 사망률 추정결과들의 상자그림

112개 시·군·구에 대한 추정결과를 상자그림으로 나타내면 그림 4.1과 같다. 그림 4.1을 살펴보면 모든 경우 추정결과의 중앙값이 대략 1.0과 1.1사이인 것으로 나타났고, 대체로 큰 값 방향으로 꼬리가 긴 비대칭의 모양을 보인다. 직접추정, HGLM(1), HGLM(2)의 경우 다른 추정법보다 변동이 크고 이상값이 나타남을 볼 수 있다. 특히 HGLM모형의 경우 EB_L과는 달리 각 지역별 특성을 별개로 나타내는 랜덤효과가 반영되어 있고, 인접지역간의 유사성이 모형에 반영되어 있지 않기 때문에 지역간 변동이 크게 나타나고 있다. 이에 반해 EB추정은 변동이 크지 않고 이상값도 발견되고 있지 않다. 따라서 직접추정치에 비해 EB추정치가 상당부분 안정성을 보이고 있다는 것을 알 수 있다.

한편 제시된 추정방법의 타당성 및 특성을 검토하기 위해 2000년 사망원인통계자료를 이용하여 얻어진 각 추정방법에 따른 추정결과를 연구대상 연도인 2000년을 중심으로 한 3개년(1999, 2000, 2001)도 각 소지역별 45세 이상 폐암 평균사망률과 비교해 추정기법에 따른 차이를 비교해 보고자 한다. 여기서 3개년도 평균 폐암 사망률은 통계청의 3개년도 사망원인통계자료에서 구한 것이다. 2000년 자료를 기초로 한 각 시·군·구 폐암 사망률 추정값과 3개년 평균 폐암 사망률과의 비교를 위해 다음과 같이 정의된 MSD(mean square deviation)를 사용한다.

$$MSD = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta_i)^2$$

여기서, θ_i는 i 소지역 3개년 45세 이상 평균 폐암 사망률이고, $\hat{\theta}_i$ 은 각 추정법으로 2000년 자료를 토대로 구한 소지역 추정결과이다. 여기서 m 은 추정대상 소지역 수를 나타낸다. 추정방법에 따라 구한 지역별 MSD를 정리하면 표 4.1과 같다.

표 4.1: 지역별 폐암 사망률 추정치들의 MSD 비교

MSD	직접추정량	EB_T	$EB_L(1)$	$EB_L(2)$	HGLM(1)	HGLM(2)
부산광역시	0.01578	0.01386	0.00872	0.00936	0.01410	0.01143
대구광역시	0.01418	0.01392	0.02376	0.01714	0.01398	0.01220
광주광역시	0.00403	0.00751	0.01502	0.01664	0.00531	0.00561
울산광역시	0.00399	0.01515	0.01878	0.01842	0.00911	0.00355
전라북도	0.04129	0.04175	0.02517	0.02351	0.02617	0.01228
전라남도	0.04470	0.01600	0.01599	0.00916	0.02670	0.02342
경상북도	0.02114	0.02694	0.01913	0.01730	0.04006	0.05361
경상남도	0.04426	0.03948	0.02385	0.01583	0.03167	0.02544
전 체	0.02962	0.02469	0.01876	0.01509	0.02570	0.02412

각 추정법의 MSD는 지역별로 차이를 보이고 있지만 대체적으로 도의 시군지역의 경우 $EB_L(2)$가 가장 안정적인 것으로 판단된다. 반면에 광역시의 경우 HGLM(2)가 대체적으로 안정적인 것으로 보인다. 경상도와 전라도의 모든 시·군·구 지역을 종합적으로 보면(표 4.1의 전체 참조) $EB_L(1)$과 $EB_L(2)$이 상대적으로 안정적인 것으로 판단된다. 이는 질병

지도 작성을 위한 통계적 추정기법을 적용함에 있어 인접지역 개념을 적절히 활용하는 것이 매우 중요하다는 것을 시사하는 것으로 받아들일 수 있다.

참고로 HGLM의 경우 EB_L과 상당히 다른 추정결과를 보여주고 있다. 이는 본 연구에서 사용된 HGLM에서 지역효과를 나타내는 u_i를 서로 독립이라고 가정하고 있기 때문에 EB_L에서와 같이 인접지역 개념을 도입한 분석이 이루어지고 있지 못하다는 것에서 그 원인을 찾을 수 있다. 따라서 향후 연구에서 HGLM에서 지역효과를 나타내는 랜덤성분 u_i에 대해 Besag(1974)이 제시한 CAR(conditional autoregression) 등과 같이 공간 상관관계(spatial correlation)를 반영할 수 있는 가정을 추가한다면 좀 더 개선된 결과를 얻을 수 있을 것으로 판단된다.

본 연구에서는 HGLM은 성별, 연령별 효과를 공변량으로 반영하고 있지만, 공간 상관관계는 반영하고 있지 않다. 반면에 EB 추정에서는 성별, 연령별 효과가 반영되어 있지 않고, 대신 인접 지역간의 공간 상관관계가 반영되어 있다. 따라서 이들 방법들의 효율성을 직접 비교하는 것은 의미가 없다. 아울러 HGLM을 적용한 경우 여기서 사용된 추정량의 MSE를 산출할 수 있는 이론이 아직 개발되어 있지 않다. 따라서 본 연구는 실제 중요한 연구 주제인 표본오차 계산 문제를 다루고 있지 못하다는 한계를 갖고 있다. 참고로 Ghosh 등(1998)이 사용한 HB 추정에 있어서도 오차를 베이즈 사후분포의 분산으로 설명하고 있다는 점에 유의하기 바란다. 따라서 향후 관련 연구에서 설계기반 MSE를 추정하는 문제에 대한 이론 개발이 요구된다.

5. 결 론

본 연구에서는 질병지도 작성을 통한 통계분석에 사용될 수 있는 소지역 추정기법을 정리하고, 사례분석을 위해 2000년 사망원인통계 자료를 이용하여 경상도 및 전라도의 112개 시·군·구 45세 이상 폐암 사망률의 표준사망률에 대한 직접추정량, 전지역에 대한 적률추정량 및 지역별 적률추정량을 활용한 EB추정량, HGLM를 활용한 추정법을 검토하고, 이를 토대로 지역별 사망률 특성을 한 눈에 파악할 수 있는 폐암 사망률 지도를 작성하였다. 아울러 지역별 3년간 평균 사망률과 추정치와의 편차를 종합해 주는 MSD를 기준으로 이 추정량들의 안정성을 비교 분석한 결과 표준사망률, EB_T보다는 EB_L 또는 HGLM을 기초로 한 추정법이 좀 더 안정적인 추정방법이라는 것을 알 수 있었다. 따라서 사망원인 통계조사에서 지역별 사망률을 추정할 때 EB_L이나 HGLM을 적용시키면 시·군·구 소지역 통계의 안정성이 상당히 향상될 수 있을 것으로 판단된다.

특히 본 연구에서는 질병지도 작성을 위해 특히 HGLM을 활용한 소지역 추정법을 새로 제시하고 있는 데, 제시된 추정법의 효율성을 이론적으로 검증할 수 있는 연구가 앞으로 필요하다. 아울러 HGLM에서 지역효과를 나타내는 랜덤성분 u_i에 공간 상관관계를 반영하는 모형에 대한 이론 및 활용에 대한 추가적인 연구가 효과적인 질병지도 작성기법의 개발에 크게 기여할 수 있을 것으로 판단된다. 한편 지역별 폐암발생에 영향을 주는 지역별 환경지수 등과 같은 보조변수를 추가로 모형에 활용하는 추정방법에 대한 연구도 고려해 볼 필요가 있을 것 같다.

참고문헌

통계청 (2000). 사망원인통계연보.

Beasg, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society B*, **36**, 192-236.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, **43**, 671-681.

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. (1998). Generalized linear models for small-area estimation, *Journal of the American Statistical Association*, **93**, 273-282.

Howe, G. M. (1970). *National Atlas of disease Mortality in the United Kingdom*, revised and enlarged edn. Nelson, London.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society A*, **161**, 121-60.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions, *Biometrika*, **88**, 987-1006.

Marshall, R. J. (1991). Mapping disease and mortality rates using empirical Bayes estimators, *Applied Statistics*, **40**, 283-294.

Rao, J. N. K (2003). *Small Area Estimation*, John Wiley and Sons, New York.

[2004년 4월 접수, 2004년 8월 채택]

HGLM and EB Estimation Methods for Disease Mapping *

Young-won Kim [1] Na-Kyung Cho [2]

ABSTRACT

For the purpose of disease mapping, we consider the four small area estimation techniques to estimate the mortality rate of small areas; direct, Empirical estimation with total moment estimator and local moment estimator, Estimation based on hierarchial generalized linear model. The estimators are compared by empirical study based on lung cancer mortality data from 2000 Annual Reports on the Cause of Death Statistics in Gyeongsang-Do and Jeonla-Do published by Korean National Statistical Office. Also the stability and efficiency of these estimators are investigated in terms of mean square deviation as well as variation of estimates.

Keywords: Empirical Bayes estimation, Hierarchical generalized linear model, Small area estimation, Disease mapping

* This Research was supported by the Sookmyung Women's University Research Grants 2003

1) Professor, Department of Statistics, Sookmyung Women's University, Hypchangwon-gil 52, Chungpa-dong, Yongsan-Gu, Seoul, 140-742, Korea
E-mail: ywkim@sookmyung.ac.kr

2) Graduate Student, Department of Statistics, Sookmyung Women's University, Hypchangwon-gil 52, Chungpa-dong, Yongsan-Gu, Seoul, 140-742, Korea