The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 515-520

Empirical Bayes Estimate for Mixed Model with Time Effect

Yong-Chul Kim[1)]

Abstract

In general, we use the hierarchical Poisson-gamma model for the Poisson data in generalized linear model. Time effect will be emphasized for the analysis of the observed data to be collected annually for the time period. An extended model with time effect for estimating the effect is proposed. In particularly, we discuss the Quasi likelihood function which is used to numerical approximation for the likelihood function of the parameter.

Keywords : empirical Bayes estimate, hierarchical model, time effect, Quasi likelihood function.

1. 서론

분석하고자 하는 자료에 적합한 모형을 만들기 위하여 많은 연구가 되어져 왔다. 특히 사망률 분석에 대한 경우에 자료의 형태가 어느 한정된 기간에 발생되어지는 자료로 해석한다면 포아송 분포를 따르므로 포아송 자료를 설명할 수 있는 모형을 이용 할 수 있다.

Manton, Woodbury 그리고 Stallard(1981)는 사망률 분석에 범주형과 연속형의 혼합 모형을 제안하였다. 이 모형은 같은 수준의 위험도를 가지고 있는 암 사망자 수는 이산형인 포아송 분포를 따르고 개인별 위험도는 연속형인 감마분포를 따른다고 가정하였다. 따라서 주어진 모집단의 자료의 모수 대하여 암 사망자 수는 음 이항 분포를 따른다. 또한 감마분포의 모수 (α, β)를 각각 인구통계와 지역변수로써 고정효과를 갖는다고 가정하여 분석하였다. Tsutakawa(1988)는 계층적 혼합 모형으로써 포아송-감마 모형을 채택하여 암 사망률을 추정하였다. 이 모형은 인구통계는 고정적 요인으로 하고 지역변수는 임의의 요인으로 가정하였다. 본 논문에서 사용될 모형은 계층적 혼합 모형에 시간의 요인을 고려하여 인구통계 부분을 시간에 대하여 선형화한 확장 모형이다. 횡단적 조사(longitudinal study)에서는 시점이 다르게 나타나므로 시간의 효과는 존재한다고 가정하였다.

혼합 모형에서 모수의 추정에 사용될 우도함수는 종종 비분석적 형태의 함수로 표현되어지며 모수를 구하는 방법에 있어서도 근사적 접근 방법으로 수치적 해법을 이용하기도 한다. Quasi 우도함수의 사용은 일반적 우도함수의 근사적 대용방법이나 비분석적인 경우에 사용하고 있다. 특히

1) Associate professor, School of Computer and Information Yongin University, Yongin city Kyunggido, 449-714, E-mail: yckim@eve.yongin.ac.kr

관찰치의 분포함수의 평균과 분산의 비례 관계에 의하여 Quasi 우도함수의 형태가 결정되어지므로 우도 함수의 대체 방법인 Quasi 우도함수 방법은 추론에 있어서도 우도함수가 갖는 성질을 그대로 유지시키며 모수를 추정할 수 있다.

본 논문에서는 시간을 고려한 혼합 모형을 제시하고 모수 추정은 경험적 베이즈 추정치를 사용하였다. 또한 평균과 분산구조의 형태에 따라서 결정되는 Quasi 우도함수의 이용 가능성에 대하여 언급하였다. 다음 장에서는 Quasi 우도함수와 확장된 Quasi 우도함수의 사전결과에 대하여 논하고, 제 3장에서는 시간을 고려한 혼합 모형에 대하여 설명하고 우도함수와 Quasi 우도함수의 적용에 대하여 논의하였다. 그리고 제 4장에서는 관련된 예와 결론을 논의하였다.

2. 사전결과

2.1. Quasi 우도함수의 정의

통계적 모형에 의하여 관찰되어지는 관찰치 y_i(i=1,2,...,n)를 각각이 서로 독립적이고 기대값은 μ_i이고 분산은 μ_i의 함수 $h(\mu_i)$라고 하자. 또한 평균 μ_i는 다른 모수 $\alpha_1, \alpha_2, \ldots, \alpha_k$의 함수라 하자. 그러면 각각의 관찰치에 대하여 Quasi 우도함수 $q(y_i, \mu_i)$은 다음으로 정의되어진다.

$$\frac{\partial q(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{h(\mu_i)} \tag{2.1}$$

또는

$$q(y_i, \mu_i) = \int_{\mu_i} \frac{y_i - \beta_i}{f(\beta_i)} d\beta_i + y_i \tag{2.2}$$

의 함수이다.

2.2. 확장된 Quasi 우도함수

근래에는 이전의 Quasi 우도함수를 변형하여 확장된 Quasi 우도함수를 Nelder와 Pegibon(1987)에 의하여 제시되었다. 확장된 Quasi 우도함수는 분산의 구조식이 평균 μ와 이산 모수 ϕ에 대하여 $var(y) = \phi V_\theta(\mu)$일 때 확장된 Quasi 우도함수는 다음과 같이 주어진다.

$$Q_\theta(y; \mu) = -\frac{1}{2} \log\{2\pi\phi V_\theta(y)\} - \frac{1}{2} \phi^{-1} D_\theta(y; \mu) \tag{2.3}$$

여기서 $D_\theta(y; \mu) = -2\int_y^\mu \frac{y-u}{V_\theta(u)} du$ 이다.

특히, 분산구조식

$$V_\theta(\mu) = \mu^\theta \ , \ \theta = 1, 2 \tag{2.4}$$

일 때 일반적으로 확장된 Quasi 우도함수를 사용한다. 평균과 분산이 선형관계에 있는 경우에

$$D_\theta(y; \mu) = 2\{y\log(y/\mu) - (y-\mu)\} \tag{2.5}$$

이다.

3. 시간을 고려한 혼합 모형과 Quasi 우도함수

이장에서는 시간을 고려한 혼합 모형을 기술하고 모형에 관련된 모수의 우도함수 대용으로 평균과 분산구조의 형태에 따라서 결정되는 Quasi 우도함수의 사용을 제시하고자 한다.

3.1. 모형과 모수의 정의

인구 크기 n_{ijk}와 독립적인 관찰치 y_{ijk}는 k 시점에서 i 번째 지역에 속하고 j 번째 인구 그룹에 속하고 평균 p_{ijk}인 포아송 분포를 따른다고 가정하자. 그리고 i 번째 지역의 효과 z_i는 확률변수이고 사전분포로써 평균이 1이되고 모수 γ를 갖는 역 감마함수를 따른다고 가정하자. 계층적 모형에서 p_{ijk}는 양의 값을 가져야 하므로 지수변환을 하지 않고 사용한다면 공액 사전분포로써 감마분포를 이용 할 수 있다. 임의 값 p_{ijk} 는 형태 모수 α 이고 범위 모수는 β 인 감마분포를 따른다. 모수의 형태에 따라서 다양한 모형을 가정할 수 있다. 모수 $\beta = \alpha z_i \theta_{1j} e^{\theta_{2j} t_k}$라하고 j 번째 그룹에 대한 초 모수 $(\alpha, \theta_{1j}, \theta_{2j})$는 상수이고 t_k는 k번째 시점 값이라 가정하면 p_{ijk}의 기대값과 분산을 다음과 같다.

$$E(p_{ijk}|\theta_{1j}, \theta_{2j}, \alpha, z_i) = z_i \theta_{1j} e^{\theta_{2j} t_k} \tag{3.1}$$

이고

$$Var(p_{ijk}|\theta_{1j}, \theta_{2j}, \alpha, z_i) = \alpha (z_i \theta_{1j} e^{\theta_{2j} t_k})^2 \tag{3.2}$$

이다.

위 (3.1)과 (3.2)와 같이 정의되어진다면 모형은 일반적인 로그 선형 모형의 근사형으로 표현 할 수 있다. 즉,

$$\log p_{ijk} = \log z_i + \log \theta_{1j} + \theta_{2j} t_k + \varepsilon_{ijk}, \tag{3.3}$$

ε_{ijk}는 일반적인 선형 모형의 오차이다.

위 (3.3)의 모형에서 상수 모수 $(\theta_{1j}, \theta_{2j})$는 j 번째 그룹의 고정 효과를 나타낸다.

3.2. 평균 p_{ijk}의 추정

3.1에서 언급한 모형을 이용하면 사후 추정치인 p_{ijk}의 추정은 경험적 베이즈 추정치로 가능하다. 즉,

$$E(p_{ijk}|y_{ijk}, \theta_{1j}, \theta_{2j}, \alpha, z_i) = \frac{y_{ijk} \widehat{p_{ijk}} + \widehat{p_{ijk}}/\alpha}{n_{ijk} \widehat{p_{ijk}} + \alpha^{-1}}.$$

위의 식에서 $\widehat{p_{ijk}} = z_i \theta_{1j} e^{\theta_{2j} t_k}$ 이다.

3.3. 초 모수 $(\alpha, \theta_{1j}, \theta_{2j}, \gamma)$의 추정

초 모수의 추정은 기본적인 우도함수를 이용한 최우 추정 방법과 우도함수의 평균과 분산 구조

식을 이용한 Quasi 우도함수의 최우 추정 방법에 관하여 제시하고자 한다.

3.3.1 초 모수 $(\alpha, \theta_{1j}, \theta_{2j}, \gamma)$의 우도함수

위의 모형을 가정으로 한다면 우도함수는 다음과 같다.

$$q(\alpha, \theta_{1j}, \theta_{2j}, \gamma|y) = \prod_{i=1} \int \prod_{j=1} \prod_{k=1} f(y_{ijk}|\alpha, n_{ijk}\alpha^{-1}\theta_{1j}e^{\theta_{2j}t_k}z_i)h(z_i|\gamma)dz_i.$$

여기에서 f는 음 이항분포이고 h는 역 감마 분포함수이다.

log를 취하면,

$$\log q(\alpha, \theta_{1j}, \theta_{2j}, \gamma|y) = \sum_{i=1} \int \prod_{j=1} \prod_{k=1} f(y_{ijk}|\alpha, n_{ijk}\alpha^{-1}\theta_{1j}e^{\theta_{2j}t_k}z_i)h(z_i|\gamma)dz_i$$

이다.

3.3.2 초 모수 $(\alpha, \theta_{1j}, \theta_{2j}, \gamma)$의 Quasi 우도함수

Quasi 우도함수는 분포함수의 형태보다도 평균과 분산 구조식을 알고 있으면 구 할 수 있다. 위에서 언급한 모형 중에 음 이항분포 함수 f의 평균과 분산구조는 다음과 같다. 평균과 분산은

$$E(y_{ijk}|\theta_{1j}, \theta_{2j}, \alpha^{-1}, z_i) = n_{ijk}z_i\theta_{1j}e^{\theta_{2j}t_k} = \mu'$$

이고,

$$Var(p_{ijk}|\theta_{1j}, \theta_{2j}, \alpha^{-1}, z_i) = \mu' + \alpha\mu'^2$$

이다.

분산의 구조식이 μ'에 대하여 이차식으로 표현되어진다. 복잡한 수식으로 표현되는 경우에 평균과 분산의 구조식을 이용하여 우도함수의 대용으로 Quasi 우도함수 식 (2.1)의 사용이 가능 할 것이다. 특히 모형의 분산 구조식이 μ'에 대하여 일차식으로 표현되어진다면 확장된 Quasi 우도함수를 식 (2.3)에 적용할 수 있다.

그러나 모형의 Quasi 우도함수 역시 선형 구조식이 아니므로 우도함수 만큼 복잡한 형태를 나타낸다. 이러한 점을 고려하여 예제는 우도함수를 이용하겠다.

4. 예제 및 결론

예제는 1973년부터 1984년까지 미국 미주리주에서 발생한 45세에서 64세, 65세에서 74세 그리고 75세이상으로 그룹을 나눈 여성 암 발생 자료를 이용하겠다. 시간을 고려한 혼합 모형의 가정 하에서 시간을 1974년, 1977년, 1980년, 1983년으로 나누고 우도 함수를 이용한 초 모수 $(\alpha, \theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \gamma)$의 추정치는 다음과 같다.

$$\hat{\alpha} = 838.85 \qquad \hat{\theta}_{11} = 0.000758$$

$$\hat{\theta}_{12} = 0.001128 \qquad \hat{\theta}_{13} = 0.001168$$

$$\hat{\theta}_{21} = 0.183109 \qquad \hat{\theta}_{22} = 0.238777$$

$$\hat{\theta}_{23} = 0.215682 \qquad \hat{\gamma} = 22.0.$$

실제 발생자료에 대하여 백만 명당 각 그룹에 대한 발생건수와 시간을 고려한 혼합모형을 이용한 발생건수를 년도별로 표시하면 <그림 1>부터 <그림 3>이다. 년도에 따라서 각 나이 그룹은 증가를 보이고 있다. 65세부터 75세 그룹이 가장 큰 증가를 나타내는 것을 알 수 있으며 각 나이 그룹별로 증가율이 다르므로 시간의 요인을 고려하는 것을 필요로 한다. 또한 본 논문의 시간을 고려한 혼합 모형은 주어질 시간에 대하여 예측이 가능 할 것이다. 그리고 추정값이 과소추정이 된 이유는 많은 인구수를 갖고 있는 발생 지역의 발생비율이 적은 인구를 가지고 있는 발생 지역의 발생비율을 지배하기 때문인 것 같다.

혼합 모형의 우도함수는 감마 분포의 모수 (α, β)에 의해서 모형의 평균과 분산의 구조식이 결정이 된다. 구조식을 재 정의하여 만약 우도함수가 평균과 분산의 구조가 선형인 경우이면

$$E(y|\alpha, \beta) = \mu'$$

이고

$$Var(y|\alpha, \beta) = \phi\mu'$$

이다.

Quasi 우도함수는 평균과 분산의 관계식에 대하여 결정되므로 식 (2.3)을 이용하면 다음과 같다.

$$Q(\alpha, \beta|y) \propto \phi^{-\frac{1}{2}} \left(\frac{y}{\mu'}\right)^{\frac{y}{\phi}} \exp\left(\frac{y-\mu'}{\phi}\right). \tag{4.1}$$

식 (4.1)과 같이 Quasi 우도함수는 평균과 분산의 관계식을 이용하며 분포의 함수 식과 관계없이 근사적으로 이용 할 수 있다.

혼합 모형에서 Quasi 우도함수의 특성을 활용한다면 복잡한 구조식을 갖는 우도함수를 간단한 함수 식으로 조정이 가능하다. 혼합 모형은 모수의 구조식에 따라 다양한 형태의 모형을 가정할 수 있다. 어떤 모형이 적합한 가에 대하여 모형의 선택 방법에 대하여 보다 연구가 필요하다.

참고문헌

[1] Manton, K.G., Woodbury, M.A., and Stallard, E.(1981). A Variance Components Approach to Categorical Data Models With Heterogeneous Mortality Rates in North Carolina Counties, Biometrics, 37, 259-269.

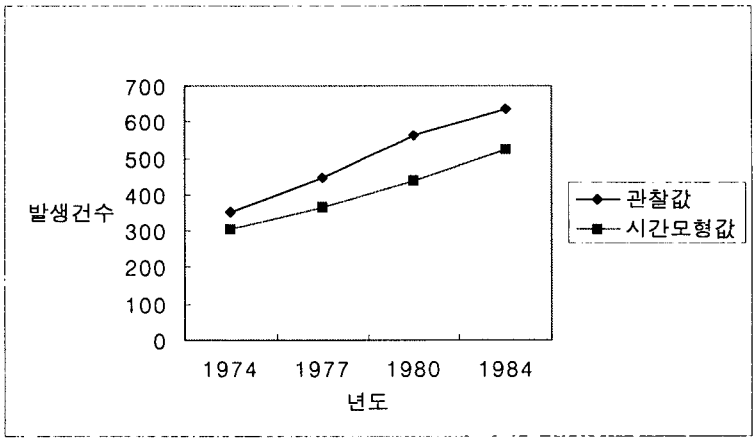
[2] Nelder, J.A. and Pregibon, D.(1987), "An Extended Quasi-Likelihood Function," *Biometrika*, 74, 221-232.

[3] Nelder, J.A. and Wedderburn, R.W.M.(1972), "Generalized Linear Models," *Journal of the Royal Statisty, tics Socie*Series A, 135, 370-384

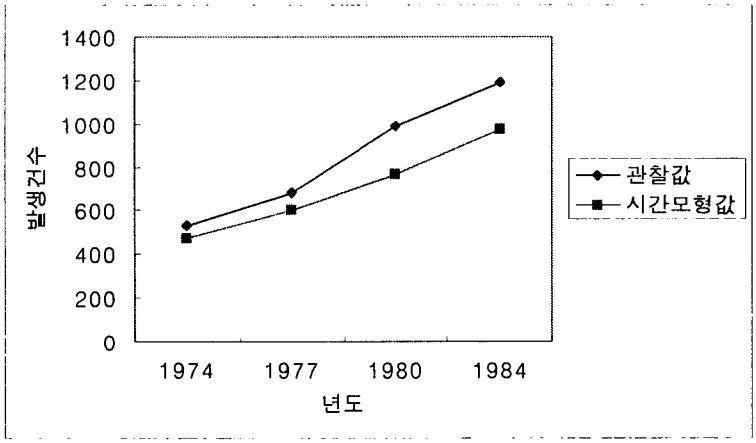
[4] Tsutakawa, R.K.(1988). Mixed Model for Analyzing Geographic Variability in Mortality Rates, Journal of American Statistical Association, 83, 37-42.

[5] Wedderburn, R.W.M.(1974), "Quasi-Likelihood Function," *The annals of Statistics*, Vol. 11,No. 1, 59-67.

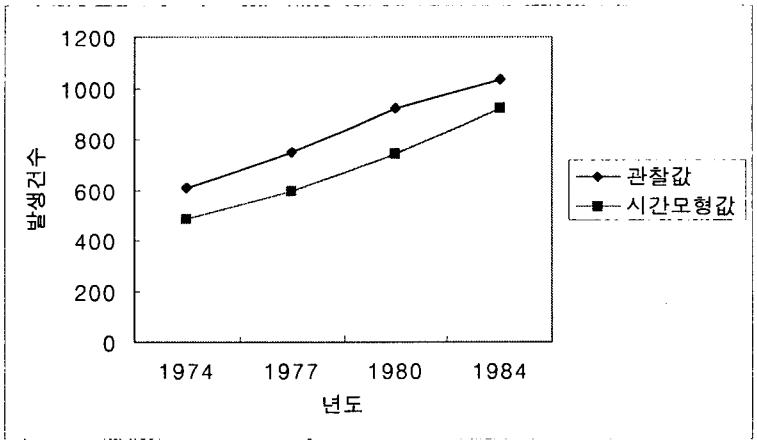
[2002년 4월 접수, 2002년 8월 채택]



<그림 1> 45세부터 65세미만까지의 관찰값과 시간모형의 추정값



<그림 2> 65세부터 75세미만까지의 관찰값과 시간모형의 추정값



<그림 3> 75세이상의 관찰값과 시간모형의 추정값