The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.6.923
27(6), 923–932

Hurdle Model for Longitudinal Zero-Inflated Count Data Analysis

Iktae Jin[a] · Keunbaik Lee[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received October 13, 2014; Revised October 30, 2014; Accepted November 5, 2014)

Abstract

The Hurdle model can to analyze zero-inflated count data. This model is a mixed model of the logit model for a binary component and a truncated Poisson model of a truncated count component. We propose a new hurdle model with a general heterogeneous random effects covariance matrix to analyze longitudinal zero-inflated count data using modified Cholesky decomposition. This decomposition factors the random effects covariance matrix into generalized autoregressive parameters and innovation variance. The parameters are modeled using (generalized) linear models and estimated with a Bayesian method. We use these methods to carefully analyze a real dataset.

Keywords: Random effects covariance matrix, generalized linear model, modified Cholesky decomposition, truncated Poisson model.

1. 서론

관찰 값들이 셀 수 있는 자료일 때 계수 자료(count data)라고 한다. 계수 자료는 여러 분야에서 이용되고 있는데 여기서 영(zero)의 값이 가정한 분포보다 과잉(inflation) 관측 되는 경우가 있다. 예를 들면 어떠한 특정지역에서 발생하는 교통사고 건수를 조사한다고 하자. 여기서 교통사고 건수는 포아송 분포를 따른다고 가정 할 수 있다. 특정지역에서는 교통사고가 잘 일어나지 않기 때문에 영이 많이 관측된다. 이를 포아송 모형으로 분석을 한다면 정보가 손실되고 추정값이 편향되는 경우가 발생한다. 이러한 경우를 위해 영이 과잉 관측된 부분을 설명하는 모형이 필요하였고, 제로팽창포아송(zero inflated Poisson; ZIP)모형 (Lambert, 1992)과 허들모형(hurdle model) (Mullahy, 1986)이 제안되었다.

ZIP모형은 영과잉 자료를 분석하기 위한 일반적으로 많이 사용되고 있는 모형이다. 이 모형은 영과 영 이외의 값의 구분을 위한 로지스틱 회귀모형과 포아송 모형으로 이루어진 혼합모형이다. 따라서 ZIP모형은 영이 과잉 관측 됐을때만 사용 할 수 있고, 영이 거의 관측이 되지 않았을(deflation) 때는 확률을 표현하는 로지스틱 회귀모형에서의 영향력 추정값이 무한대로 추정되는 문제가 발생한다 (Min과

This project was supported by Basic Science Research Program through the National Research Foundation of Korea(KRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1A1004002, NRF-2014R1A1A2054997).

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-Gu, Seoul 110-745, Korea. E-mail: keunbaik@skku.edu

Agresti, 2005). 이러한 문제를 해결하기 위해서 허들모형이 제안되었다. 허들모형은 이항모형(binary model)과 절삭된 포아송모형(truncated Poisson model)으로 이루어진 혼합모형이다. 이 모형의 장점은 영이 과잉이나, 과소일 때 모두 사용할 수 있다. 이 논문에서는 경시적 영과잉 자료분석을 위한 허들모형을 제안할 것이다.

경시적 범주형자료(longitudinal categorical data)는 일정기간 동안 반복 측정된 자료이다. 따라서 반복 측정된 개체에서 나온 결과치들은 서로 상관관계가 있고, 이 상관관계를 고려하여 모수추정을 하여야 한다. 이러한 자료의 분석을 위한 모형으로 일반화선형혼합모형(generalized linear mixed models; GLMMs)이 자주 사용되어 왔다 (Breslow와 Clayton, 1993). GLMMs에서 변량효과(random effects)의 공분산행렬(covariance matrix)이 반복 측정된 결과치들의 상관관계를 설명한다. 그리고 개체들 간의 변동 및 시간에 따른 변동을 설명한다. 그러나 공분산행렬은 주로 고차원(high dimension)이고, 양정치(positive definite)를 만족해야 하기 때문에 추정이 쉽지가 않다. 그래서 AR(1)과 같은 간단한 형태의 공분산행렬 구조를 가정한다. 그리고 등분산성을 가지는 행렬을 가정한다. 하지만 이러한 가정은 매우 강한 가정이기 때문에 고정된 효과(fixed effects) 추정에 편의(bias)가 발생할 수 있다 (Heagerty와 Kurland, 2001). 이 논문에서 우리는 수정된 콜레스키분해(modified Cholesky decomposition; Pourahmadi, 1999) 방법을 통하여 더 높은 차수를 가질 수 있는 AR(p) 형태의 공분산행렬을 제안한다. 그리고 공변량(covariate)에 의존하는 이분산성(heteroscedasticity)의 공분산행렬을 제안한다.

수정된 콜레스키 방법은 공분산행렬을 일반화자기회귀모수(generalized autoregressive parameters; GARPs)와 혁신분산(innovation variances; IVs)모수로 분해하여 이분산성을 가진 높은 차원의 AR구조로 쉽게 확장된다 (Pourahmadi, 1999). 일반화자기회귀모수는 한 시점의 변량효과를 그 이전의 변량효과들이 얼마나 영향을 주는지를 나타내는 계수이고, 혁신분산은 변량효과의 분산모수이다. 이러한 방법은 선형모형에서 제안되어 선형혼합모형(linear mixed model)으로 확장되어 사용되었다 (Daniels와 Pourahmadi, 2002; Daniels와 Zhao, 2003; Pan와 Mackenzie, 2003, 2006). 논문에서 수정된 콜레스키분해를 통해 일반적인 형태의 이분산성을 가지는 변량효과 공분산행렬을 가지는 허들모형을 제안한다.

일반화자기회귀모수와 혁신모수의 추정을 위해서 Pourahmadi (1999)는 각각 선형회귀 및 로그선형모형(log linear model)을 제안하여 뉴턴-랍슨의 방법을 이용하였다. 그 이후 선형혼합모형에서 변량효과의 공분산행렬의 일반화자기회귀모수와 혁신모수의 추정시에도 뉴턴-랍슨의 방법이 이용되었다 (Pan과 Mackenzie, 2003, 2006). 베이지안 방법에 의한 모수추정 또한 Daniels와 Pourahmadi (2002)과 Daniels와 Zhao (2003)에 의해서 제안되었고, 일반화선형혼합모형에도 확장되어 사용되었다 (Lee, 2013).

본 논문의 구성은 다음과 같다. 2장에서는 영 과잉 자료에 대한 허들모형과 변량효과의 공분산 행렬을 수정된 콜레스키 분해로 표현한 것을 소개한다. 3장에서는 모수를 베이지안 방법으로 추정하는 방법과 모델을 선택하는 방법에 대해 소개를 한다. 4장에서는 실제 자료인 New Orleans Murder Rate 자료를 분석한다. 마지막으로 5장에서는 결론을 제시한다.

2. 수정된 콜레스키분해를 사용한 허들모형

이 장에서는 먼저 허들모형을 소개하고 다음으로 변량효과의 공분산행렬을 수정된 콜레스키분해로 표현한 것을 소개한다.

2.1. 허들모형 (Hurdle model)

Y_{it}는 개체 $i(i = 1, \ldots, N)$의 $t(t = 1, \ldots, n_i)$번째의 가산형 반응변수이다. 그리고 Y_{it}는 변량효과 b_{it}가 주어 졌을 때 조건부독립을 가정한다. $\boldsymbol{x}_{it}$는 Y_{it}에 상응하는 공변량이다. 허들모형은 이항모형과

절삭된 포아송모형으로 이루어진 혼합모형이고, 다음과 같이 표현된다.

$$\mathrm{P}(y_{it}|x_{it}) = \begin{cases} 1-p_{it} & \text{if } y_{it}=0, \\ p_{it}\dfrac{e^{-\lambda_{it}}\lambda_{it}^{y_{it}}}{y_{it}!(1-e^{-\lambda_{it}})} & \text{if } y_{it}=1,2,\ldots, \end{cases} \tag{2.1}$$

여기서

$$\mathrm{logitp}_{it} = \boldsymbol{x}_{it}^T\boldsymbol{\beta}_1 + b_{it}, \tag{2.2}$$

$$\log\lambda_{it} = \boldsymbol{x}_{it}^T\boldsymbol{\beta}_2 + \delta b_{it}, \tag{2.3}$$

$$\boldsymbol{b}_i \sim \text{i.i.d. } N(0,\boldsymbol{\Sigma}_i), \tag{2.4}$$

$\boldsymbol{\beta}_1$과 $\boldsymbol{\beta}_2$는 회귀계수를 나타내는 $r\times 1$ 벡터이고, $\boldsymbol{x}_{it}$는 시간 t에서 개체 i에 대한 $r\times 1$ 공변량 벡터이다. 변량효과 $\boldsymbol{b}_i=(b_{i1},\ldots,b_{in_i})^T$이고 $\boldsymbol{\Sigma}_i$는 $n_i\times n_i$인 행렬이다. $1-p_{it}$가 $e^{-\lambda_{it}}$보다 크면 자료에서 0이 과잉 관측된 것이고 작으면 0이 과소 관측된 것이다 (Neelon 등, 2010). 모수 δ는 영이 아님과 가산자료의 평균의 관계를 나타내고 있다. δ가 0인 경우는 이 둘의 관계는 없다는 의미이고, 양수인 경우는 양의 관계임을 알 수 있다.

2.2. 수정된 콜레스키분해 (Modified Cholesky Decomposition)

변량효과의 공분산행렬 $\boldsymbol{\Sigma}_i$는 고차원이고 양정치를 만족해야 한다. 이를 위해 수정된 콜레스키분해를 사용하여 해결한다. 변량효과 $\boldsymbol{b}_i$는 b_{it}가 이전의 $b_{i1},\ldots,b_{i,t-1}$로 이루어진 회귀식으로 표현된다는 가정을 한다. 식으로 표현하면 다음과 같다.

$$b_{i1} = e_{i1}, \tag{2.5}$$

$$b_{it} = \sum_{j=1}^{t-1}\phi_{itj}b_{ij} + e_{it}, \quad \text{for t=2},\ldots,n_i, \tag{2.6}$$

$\boldsymbol{e}_i=(e_{i1},\ldots,e_{in_i})^T$는 평균이 0, 분산이 $\boldsymbol{D}_i$인 정규분포이고 $\boldsymbol{D}_i = diag(\sigma_{i1}^2,\ldots,\sigma_{in_i}^2)$이다. 식 (2.5)와 식 (2.6)를 행렬로 표현하면 다음과 같다.

$$\boldsymbol{T}_i\boldsymbol{b}_i = \boldsymbol{e}_i, \tag{2.7}$$

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_{i21} & 1 & 0 & \cdots & 0 \\ -\phi_{i31} & -\phi_{i32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\phi_{in_i1} & -\phi_{in_i2} & -\phi_{in_i3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \\ \vdots \\ b_{in_i} \end{pmatrix} = \begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \\ \vdots \\ e_{in_i} \end{pmatrix},$$

여기서 $\boldsymbol{T}_i$는 대각이 1이고 (t,j)번째가 $-\phi_{itj}$인 하삼각행렬이다. 식 (2.7)의 양 변에 분산을 취하면 다음과 같다.

$$\boldsymbol{T}_i\boldsymbol{\Sigma}_i{\boldsymbol{T}_i}^T = \boldsymbol{D}_i, \tag{2.8}$$

따라서 $\boldsymbol{\Sigma}_i$는 ϕ_{itj}의 일반화자기회귀모수(generalized autoregressive parameters; GARPs)과 σ_{it}^2의 혁신분산(innovation variances; IVs)모수로 표현된다. 일반화자기회귀모수는 어느 시점의 변량효과에 시

점 이전의 변량효과들이 얼마나 영향을 주는지를 나타내는 계수이고, 혁신분산은 변량효과의 분산모수이다. 일반화자기회귀모수와 혁신분산은 다음과 같은 회귀식으로 표현된다.

$$\phi_{itj} = \omega_{itj}^T \boldsymbol{\gamma}, \quad \log\left(\sigma_{it}^2\right) = h_{it}^T \boldsymbol{\eta}, \tag{2.9}$$

ω_{itj}는 시간에 차이를 나타내는 지시함수(indicator function)이고 h_{it}는 성별이나 지역같은 개인의 특성을 나타내는 공변량이고 $\boldsymbol{\gamma}$와 $\boldsymbol{\eta}$는 벡터로 추정해야 하는 모수이다 (Pourahmadi, 2000; Pourahmadi와 Daniels, 2002; Daniels와 zhao, 2003; Lee 등 2011; Lee, 2013). 식 (2.8)과 같이 모형을 하면 제한적인 부분이 없기 때문에 공변량으로 변량효과 공분산행렬을 모형화 할 수 있고 공분산행렬이 이 분산성을 가질 수 있는 장점이 있으며 일반화선형모형으로 일반화자기회귀모수와 혁신분산의 모형화가 간편하다는 장점이 있다.

3. 베이지안 방법

베이지안 방법은 모수의 사후분포(posterior distribution)를 찾아 모수를 추정하는 방법이다. 사후분포는 우도함수(likelihood function)와 사전분포(prior distribution)를 이용하여 구할 수 있다. 이 장에서는 먼저 베이지안 모형화를 소개하고 다음으로 모형 선택(Model Selection) 방법인 DIC(Deviance Information Criterion)대해 소개한다.

3.1. 베이지안 모형화

우선 모수들의 사전분포를 정한다. 모수인 $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, $\boldsymbol{\delta}$의 사전분포는 다음과 같다.

$$\boldsymbol{\beta}_1 \sim N\left(\mathbf{0}, \sigma_{\beta_1}^2 \boldsymbol{I}\right), \tag{3.1}$$

$$\boldsymbol{\beta}_2 \sim N\left(\mathbf{0}, \sigma_{\beta_2}^2 \boldsymbol{I}\right), \tag{3.2}$$

$$\boldsymbol{\gamma} \sim N\left(\mathbf{0}, \sigma_{\gamma}^2 \boldsymbol{I}\right), \tag{3.3}$$

$$\boldsymbol{\eta} \sim N\left(\mathbf{0}, \sigma_{\eta}^2 \boldsymbol{I}\right), \tag{3.4}$$

$$\boldsymbol{\delta} \sim N\left(\mathbf{0}, \sigma_{\delta}^2 \boldsymbol{I}\right), \tag{3.5}$$

여기서 $\sigma_{\beta_1}^2$, $\sigma_{\beta_2}^2$, σ_{γ}^2, σ_{η}^2, 그리고 σ_{δ}^2는 큰 값으로 하여 무정보적 사전분포(noninformative prior)를 고려한다. $\boldsymbol{I}$는 각 모수에 차원에 맞춘 단위행렬이다.

식 (2.1)과 식 (3.1)–식 (3.5)에 각각 제시된 표본분포와 사전분포를 이용하여 나타낸 결합사후분포는 다음과 같다.

$$\begin{aligned} f(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{b} | \boldsymbol{y}) \propto\ & f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\delta}) \\ & \times \phi(\boldsymbol{b}|\mathbf{0}, \boldsymbol{\gamma}, \boldsymbol{\eta}) \phi(\boldsymbol{\beta}_1|\mathbf{0}, \sigma_{\beta_1}^2) \phi\left(\boldsymbol{\beta}_2|\mathbf{0}, \sigma_{\beta_2}^2\right) \phi\left(\boldsymbol{\gamma}|\mathbf{0}, \sigma_{\gamma}^2\right) \phi\left(\boldsymbol{\eta}|\mathbf{0}, \sigma_{\eta}^2\right) \phi\left(\boldsymbol{\delta}|\mathbf{0}, \sigma_{\delta}^2\right) \\ =\ & \prod_{i=1}^{N} \prod_{t=1}^{n_i} (1 - p_{it})^{I(y_{it}=0)} \left(p_{it} \frac{e^{-\lambda_{it}} \lambda_{it}^{y_{it}}}{y_{it}!(1 - e^{-\lambda_{it}})} \right)^{1 - I(y_{it}=0)} \\ & \times \phi(\boldsymbol{b}|\mathbf{0}, \boldsymbol{\gamma}, \boldsymbol{\eta}) \phi(\boldsymbol{\beta}_1|\mathbf{0}, \sigma_{\beta_1}^2) \phi\left(\boldsymbol{\beta}_2|\mathbf{0}, \sigma_{\beta_2}^2\right) \phi\left(\boldsymbol{\gamma}|\mathbf{0}, \sigma_{\gamma}^2\right) \phi\left(\boldsymbol{\eta}|\mathbf{0}, \sigma_{\eta}^2\right) \phi\left(\boldsymbol{\delta}|\mathbf{0}, \sigma_{\delta}^2\right), \end{aligned} \tag{3.6}$$

여기서 $f(y|)$는 허들모형을 위한 우도함수이고, $\phi(\cdot)$는 (다변량)정규분포((multivariate) normal probability distribution)를 나타낸다. 모수들의 결합사후분포로부터의 베이지안 추론을 위한 깁스 샘플러(Gibbs sampler)을 이용한다. 하지만 대부분의 조건부 확률분포가 모르는 분포이므로 메트로폴리

Table 4.1. Mean and Variance of independent variables

		1940	1950	1960	1970	1980	1990	2000
Problack	Mean	0.333	0.277	0.422	0.490	0.439	0.542	0.575
	Variance	0.052	0.055	0.159	0.142	0.130	0.123	0.133
Pro15to24	Mean	0.096	0.094	0.063	0.084	0.181	0.064	0.076
	Variance	0.002	0.002	0.0002	0.0005	0.001	0.0005	0.001
IQ	Mean	0.327	0.341	0.316	0.318	0.436	0.352	0.341
	Variance	0.003	0.003	0.002	0.004	0.014	0.013	0.014

스-해스팅스(Metropolis-Hastings) 알고리즘을 이용해야 한다. 하지만 실제 분석에서는 WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml)를 이용하여 이러한 과정을 수행한다.

3.2. 모형선택

베이지안 모형 선택 방법에는 사후예측손해(posterior predictive loss; Gelfand와 Ghosh, 1998)와 편차정보기준(DIC; deviance information criterion)이 있다. 이 두 방법은 모두 모형의 복잡성과 적합성을 각각 고려한 기준치이다 (Daniels과 Hogan, 2008). 본 논문에서는 DIC를 이용하여 모형선택을 한다. DIC는 AIC(Akaike Information Criterion)와 유사하며 편차들의 평균에서 모수의 사후평균을 대입한 편차를 빼서 계산된다. θ를 모든 모수들을 모아 놓은 벡터라고 하면, DIC는 아래와 같이 정의 한다.

$$\text{DIC} = 2\overline{\text{Dev}}(\theta) - D(\tilde{\theta}),$$

여기서 $\text{Dev}(\theta) = -2\log L(\theta|y)$이고 $D(\tilde{\theta}) = \text{Dev}(E_\theta(\theta|y))$이다. DIC는 $\overline{\text{Dev}}(\theta)$와 $D(\tilde{\theta})$를 추정 또는 근사 방법에 따라 여러가지 방식으로 표현된다. DIC는 위의 모형에 적합하지 않기 때문에 모형에 적합한 DIC를 제안한다 (Celeux 등, 2006).

$$\text{DIC}_3 = -4E[\log L(\theta|y)|y] + 2\log \widehat{L}(y), \tag{3.7}$$

여기서

$$\widehat{L}(y) = E_\theta[f(y|\theta)|y], \tag{3.8}$$

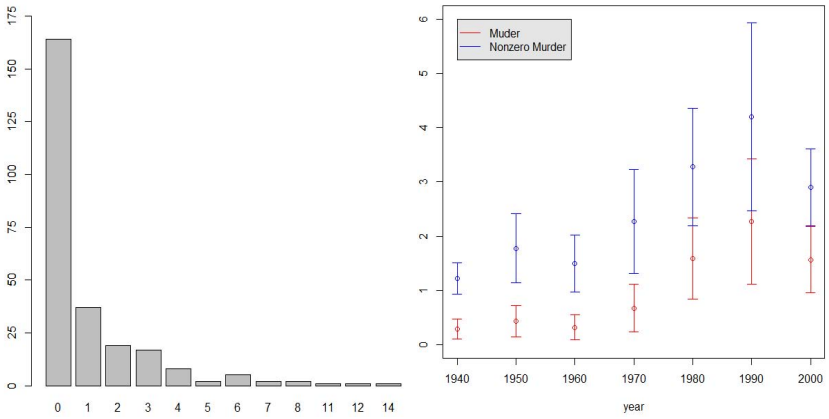
식 (3.7)에서 $E[\log L(\theta|y)|y]$와 $\widehat{L}(y)$는 마코프 체인 몬테카를로(Markov chain Monte Carlo; MCMC) 방법을 사용하여 근사적으로 구한다.

4. 뉴올리언즈 (New Orleans) 살인율 자료

본 절에서는 뉴올리언즈 살인율 자료를 살펴 보고, 앞의 절에서 제시된 통계적 방법을 이용하여 분석한다.

4.1. 자료기술

이 자료는 Lee 등 (2011)에 처음 분석된 자료로서 뉴올리언즈 지역의 37개의 인구조사 표준지역(census tract) 및 경찰관할지역(police zone)에서 1940년부터 2000년까지 10년 단위로 살인사건수를 조사한 자료이다. 따라서 60년간 조사된 자료로서 259개의 관찰값을 가진다. 이 연구의 목적은 60년에 걸쳐



(a) Frequency (b) Confidence Interval

Figure 4.1. Barplot and Confidence Interval plot of response variable

Table 4.2. Description of models in terms of $\log(\sigma_{it}^2)$ and ϕ_{itj}

Model	$\log(\sigma_{it}^2)$	ϕ_{itj}
Model 1	$\log(\sigma_{it}^2) = \eta_0$	$\phi_{itj} = \gamma_0 I_{(\lvert t-j\rvert=1)}$
Model 2	$\log(\sigma_{it}^2) = \eta_0 + \eta_1 Problack$	$\phi_{itj} = \gamma_0 I_{(\lvert t-j\rvert=1)}$
Model 3	$\log(\sigma_{it}^2) = \eta_0 + \eta_1 Pro15to24$	$\phi_{itj} = \gamma_0 I_{(\lvert t-j\rvert=1)}$
Model 4	$\log(\sigma_{it}^2) = \eta_0$	$\phi_{itj} = \gamma_0 I_{(\lvert t-j\rvert=1)} + \gamma_1 I_{(\lvert t-j\rvert=2)}$
Model 5	$\log(\sigma_{it}^2) = \eta_0 + \eta_1 Problack$	$\phi_{itj} = \gamma_0 I_{(\lvert t-j\rvert=1)} + \gamma_1 I_{(\lvert t-j\rvert=2)}$
Model 6	$\log(\sigma_{it}^2) = \eta_0 + \eta_1 Pro15to24$	$\phi_{itj} = \gamma_0 I_{(\lvert t-j\rvert=1)} + \gamma_1 I_{(\lvert t-j\rvert=2)}$

서 살인율의 변화에 영향을 미치는 인구통계학적인 요인(demographic factor)을 조사하는 것이 목적이다. 고려한 요인들은 흑인의 비율(Problack), 소득의 분위수(IQ), 15 ∼ 24세의 비율(Pro15to24)이다. Table 4.1.은 연도별로 이 요인들의 평균과 분산을 나타내고 있다. 표를 보면 흑인의 비율은 1940년에서 2000년으로 갈수록 증가하는 것을 볼 수 있고, 15 ∼ 24세 비율과 소득분위는 변화가 없다는 것을 볼 수 있다.

총 259개의 관찰값들 중에 약 63%의 값들이 0이므로 우리는 영과잉 가산자료로 생각할 수 있다. Figure 4.1의 (a)는 살인사건 수의 빈도를 나타낸 것인데 여기서 영의 빈도가 많음을 볼 수 있다. Figure 4.1의 (b)는 연도별 살인사건 수와 영을 제외한 살인사건 수의 평균과 95% 신뢰구간을 그림으로 나타낸 것이다. 빨간색 선은 살인사건 수이고 파란색 선은 영을 제외한 살인사건 수이다. 영을 제외한 평균들의 신뢰구간이 영을 포함한 구간들과 겹치지 않음을 알 수 있다. 그림을 보면 1990년도까지 살인의 수는 평균적으로 증가하는 것을 볼 수 있다.

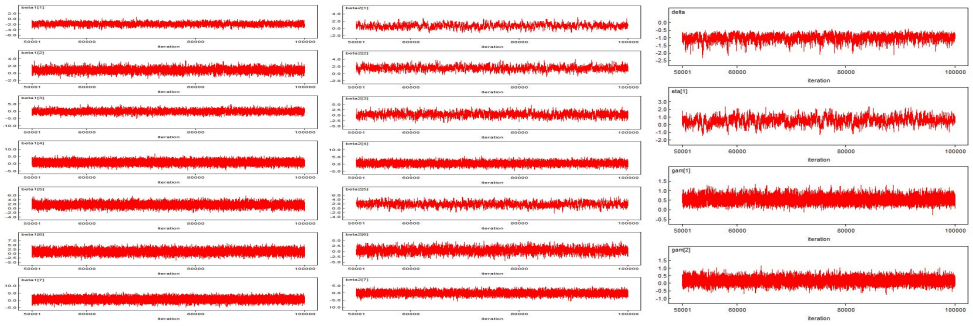
4.2. 자료분석

분석의 단순화를 위해서 구역간 살인사건의 수는 독립이라고 가정하였다. 응답변수인 살인사건 수는 인구에 비례해서 커짐으로 지역의 인구를 offset로 고려하여 분석하였다. 독립변수는 앞에서 제시된 흑인의 비율(Problack), 소득의 분위수(IQ), 15 ∼ 24세의 비율(Pro15to24), 그리고 연도(year)와 그것의 제곱(year^2) 그리고 세제곱(year^3)을 사용하였다.

Table 4.3. Posterior means(95% confidence intervals in the parentheses)

	모형					
	1	2	3	4	5	6
Fixed parameters: β						
$\beta 1$[1](Intercept)	−1.863* (−2.972,−0.792)	−1.858* (−2.946,−0.824)	−1.863* (−2.957,−0.826)	−1.865* (−2.982,−0.798)	−1.862* (−2.941,−0.807)	−1.863* (−2.943,−0.796)
$\beta 1$[2](Probalck)	0.914 (−0.211, 2.038)	0.919 (−0.197,2.026)	0.896 (−0.248,2.039)	0.879 (−0.264,2.023)	0.895 (−0.229,1.996)	0.895 (−0.226,2.025)
$\beta 1$[3](IQ)	−0.091 (−2.335, 2.137)	−0.140 (−2.305,2.024)	−0.126 (−2.322,2.054)	−0.102 (−2.339,2.116)	−0.115 (−2.307,2.085)	−0.099 (−2.306,2.070)
$\beta 1$[4](Pro15to24)	0.989 (−1.541, 3.550)	1.003 (−1.555,3.568)	0.986 (−1.577,3.564)	0.962 (−1.603,3.540)	0.971 (−1.602,3.527)	0.964 (−1.586,3.540)
$\beta 1$[5](Year)	1.619 (−0.304, 3.545)	1.639 (−0.272,3.567)	1.627 (−0.276,3.564)	1.643 (−0.300,3.587)	1.631 (−0.323,3.575)	1.639 (−0.283,3.603)
$\beta 1$[6]($Year^2$)	1.168 (−1.299,3.644)	1.162 (−1.277,3.620)	1.169 (−1.285,3.634)	1.185 (−1.274,3.632)	1.187 (−1.249,3.654)	1.181 (−1.269,3.648)
$\beta 1$[7]($Year^3$)	0.587 (−2.050,3.208)	0.597 (−2.038,3.208)	0.599 (−2.042,3.240)	0.622 (−2.003,3.278)	0.599 (−2.031,3.245)	0.598 (−2.069,3.255)
$\beta 2$[1](Intercept)	0.721 (−0.422,1.925)	0.813 (−0.396,2.025)	0.802 (−0.310,1.957)	0.708 (−0.432,1.877)	0.697 (−0.508,1.870)	0.699 (−0.479,1.877)
$\beta 2$[2](Probalck)	1.519* (0.410,2.678)	1.452* (0.241,2.675)	1.556* (0.421,2.728)	1.536* (0.395,2.717)	1.479* (0.318,2.684)	1.532* (0.398,2.757)
$\beta 2$[3](IQ)	0.303 (−1.792,2.391)	0.211 (−1.847,2.303)	0.239 (−1.821,2.274)	0.366 (−1.711,2.423)	0.358 (−1.724,2.398)	0.325 (−1.701,2.427)
$\beta 2$[4](Pro15to24)	0.536 (−1.819,2.965)	0.531 (−1.858,2.905)	0.506 (−1.892,2.898)	0.549 (−1.889,2.955)	0.561 (−1.849,2.891)	0.543 (−1.841,2.901)
$\beta 2$[5](Year)	1.849 (−0.186,3.890)	1.753 (−0.209,3.742)	1.741 (−0.206,3.775)	1.850 (−0.203,3.821)	1.926 (−0.046,3.904)	1.867 (−0.127,3.924)
$\beta 2$[6]($Year^2$)	0.308 (−2.179,2.788)	0.378 (−2.068,2.854)	0.372 (−2.058,2.776)	0.392 (−1.967,2.786)	0.349 (−2.089,2.766)	0.409 (−2.041,2.856)
$\beta 2$[7]($Year^3$)	−0.190 (−2.826,2.392)	−0.182 (−2.818,2.469)	−0.168 (−2.789,2.408)	−0.114 (−2.686,2.454)	−0.154 (−2.714,2.440)	−0.130 (−2.746,2.463)
GARP: γ						
γ[1](AR1)	0.645* (0.360,0.889)	0.641* (0.359,0.881)	0.647* (0.369,0.884)	0.542* (0.207,0.937)	0.553* (0.214,0.948)	0.552* (0.212,0.944)
γ[2](AR2)				0.171 (-0.255,0.601)	0.156 (-0.277,0.593)	0.167 (-0.264,0.610)
IV parameters: η						
η[1]	0.436 (-0.677,1.462)	0.560 (-0.535,1.592)	0.520 (-0.458,1.517)	0.538 (-0.553,1.547)	0.546 (-0.611,1.662)	0.464 (-0.555,1.504)
η[2]		-0.303 (-1.374,0.769)	0.058 (-1.277,1.404)		-0.232 (-1.273,0.831)	0.045 (-1.293,1.367)
δ	-1.032* (-1.583,-0.664)	-1.061* (-1.614,-0.666)	-0.999* (-1.474,-0.652)	-1.001* (-1.502,-0.659)	-1.05* (-1.600,-0.653)	-1.024* (-1.512,-0.665)
DIC	2211.031	2188.559	2167.143	2127.397	2181.899	2238.325

수정된 콜레스키분해의 일반화 자기회귀모수와 혁신분산에 따라서 6개의 모형을 Table 4.2에 제시하였다. 모형1, 2, 그리고 3은 모두 AR(1)구조를 가지며 모형1은 등분산성을, 모형2는 Problack에 의존하는 이분산성을, 그리고 모형3은 Pro15to24에 의존하는 이분산성을 가지는 변량효과 공분산행렬을 가정한다. 모형4, 5, 그리고 6은 모두 AR(2)구조를 가지며, 모형4는 등분산성을, 모형5는 Problack에 의존하는 이분산성을, 그리고 모형6은 Pro15to24에 의존하는 이분산성을 가지는 변량효과 공분산 행렬을 가정한 모형들이다.



(a) beta1 (b) beta2 (c) delta, eta, gamma

Figure 4.2. trace plot

이러한 6개의 모형을 베이지안 방법에 의한 모수들의 추정을 위하여 WinBUGS와 R 패키지 R2WinBUGS의 BUGS 함수를 사용하였다. 총 500,000개의 난수를 5의 간격(thin=5)으로 추출하여 100,000개을 가지고 앞의 50,000개는 버리고 나머지 50,000개의 난수로 분석에 사용하였다.

Table 4.3은 Table 4.2에 제시된 6개의 모형을 추정한 결과와 모형선택의 기준치인 DIC의 결과들이다. 모형1-6의 DIC결과로 모형 4가 가장 적합한 모형임을 알 수 있다(모형1: 2211.031, 모형2: 2188.559, 모형3: 2167.143, 모형4: 2127.397, 모형5: 2181.899, 모형6: 2238.325). 모형 4는 AR(2)구조를 가지면서 등분산성을 만족하는 변량효과의 공분산 행렬을 가지는 모형이다. 모든 모형에서 회귀계수들의 추정치들이 비슷한 사후평균값들을 가짐을 알 수 있다. 살인율에서 흑인의 비율이 95% 신용구간에 0을 포함하지 않으므로 유의한 변수임을 알 수 있다. 그러므로 흑인의 비율이 증가하면 살인율이 증가함을 알 수 있다. 일반자기회귀모수 중 AR(1)의 모수가 유의하고, 혁신분산은 유의하지 않음을 알 수 있다. 그리고 δ 또한 유의미하고 그 값이 음의 값을 나타내고 있다.

4.3. 수렴진단

MCMC 알고리즘의 수렴성을 진단하기 위한 여러가지 방법이 있다. 몬테카를로 오차 확인, MCMC 알고리즘으로부터 추출된 모수에 대한 난수값을 R에서 진단하는 방법, 사후밀도함수그림 확인, 자기상관그림 확인, 시도표 확인, 분위수그림을 확인 그리고 겔만-루빈 통계량을 이용하는 방법 등이 있다. 이 중에서 본 논문은 시도표를 통한 수렴성을 확인하는 방법을 이용한다. 시도표로 수렴여부를 진단하는 방법은 그림에서 특정한 패턴을 보이지 않고, 난수들이 잘 섞여 있을 때에 MCMC 알고리즘이 수렴 한다고 할 수 있다. Figure 4.2는 모형4의 모수들에 대한 시도표를 나타낸 그림이다. 그림을 보면 β_1, β_2, δ, η, γ의 시도표는 특정한 패턴을 보이지 않는 것으로, 모두 수렴한다고 할 수 있다.

5. 결론

본 논문에서는 경시적 영과잉 가산자료 분석을 위해 수정된 콜레스키 분해를 통해 일반적인 형태의 이분산성을 가지는 변량효과 공분산행렬을 가지는 허들모형을 제안했다. 공분산행렬은 공변량에 의존하는 일반화자기회귀모수와 혁신분산에 의해 표현된다. 일반화자기회귀모수는 시간에 차이를 나타내는 지시함수로 회귀식이 표현되고 일반적인 AR구조를 설명할 수 있다. 혁신분산은 개인의 특성을 나타내는 공변량으로 회귀식을 표현하여 등분산성을 만족하게 할 수 있다. 그리고 혁신분산에 loglinear모형을 가정함으로써 공분산행렬의 양정치성을 항상 만족하게 제안하였다. 모수들의 추정은 베이지안 방법을 이용

한 WinBUGS 프로그램을 사용해서 추정하였다. 본 논문은 영과잉 가산자료 분석에서 허들모형이 어떻게 쓰이고 추정되는지 살펴보았다.

본 논문에서 사용된 뉴올리언즈 살인율 자료분석에서 공분산행렬이 등분산성을 가지는 AR(2) 모형이 가장 적합한 모형이었다. 일반 자기회귀모수 중 AR(1)의 모수는 유의하고 혁신분산 모수는 유의하지 않다. 나머지 모수들에서는 살인율에서의 흑인의 비율이 유의하므로 흑인의 비율이 증가하면 살인율이 증가한다. 수렴진단 결과는 모수들의 시도표에서 특정한 패턴이 나타나지 않았기 때문에 모두 수렴한다.

References

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 125–134.

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006). Deviance Information Criteria for Missing Data Models, *Bayesian Analysis*, **1**, 651–674.

Daniels, J. M. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data, *Biometrika*, **89**, 553–566.

Daniels, J. M. and Zhao, Y. D. (2003). Modelling the random effects covariance matrix in longitudinal data, *Statistics in Medicine*, **22**, 1631–1647.

Daniels, M. J. and Hogan, J. W.(2008). Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis, *Chapman & Hall/CRC*.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach, *Biometrika*, **85**, 1–13.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models, *Biometrika*, **88**, 973–985.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Lee, K., Joo, Y., Song, J. J. and Harper, D. W. (2011). Analysis of zero-inflated clustered count data: A marginalized model approach, *Computational Statistics & Data Analysis*, **55**, 824–837.

Lee, K. (2013). Bayesian modeling of random effects covariance matrix for generalized linear mixed models, *Communications for Statistical Applications and Methods*, **20**, 235–240.

Mullahy, J. (1986). Specification and testing of some modified count data models, *Journal of Econometics*, **33**, 341–365.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data, *Statistical Modelling*, **5**, 1–19.

Neelon, B. H., O'Malley, A. J. and Normand, S. T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use, *Statistical Modelling*, **10**, 421–439.

Pan, J. X. and Mackenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies, *Biometrika*, **90**, 239–244.

Pan, J. X. and MacKenzie, G. (2006). Regression models for covariance structures in longitudinal studies, *Statistical Modelling*, **6**, 43–57.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, *Biometrika*, **86**, 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix, *Biometrika*, **87**, 425–435.

Pourahmadi, M. and Daniels, M. J. (2002). Dynamic conditionally linear mixed models for longitudinal data, *Biometrika*, **58**, 225–231.

영과잉 경시적 가산자료 분석을 위한 허들모형

진익태[a] · 이근백[a,1]

[a]성균관대학교 통계학과

(2014년 10월 13일 접수, 2014년 10월 30일 수정, 2014년 11월 5일 채택)

요 약

허들모형은 영이 과잉 가산자료를 분석하기 위해서 사용되어 왔다. 이 모형은 이산부분을 위한 로짓모형과 절삭된 가산부분을 위한 절삭된 포아송모형의 혼합모형이다. 이 논문에서 우리는 경시적 영과잉 가산자료를 분석하기 위해서 수정된 콜레스키 분해을 이용하여 일반적인 이분산성을 가지는 변량효과 공분산행렬을 제안한다. 수정된 콜레스키 분해는 변량효과 공분산행렬을 일반화자기상관 모수와 혁신분산모수로 분리되면, 이러한 모수들은 베이지안 일반화 선형모형을 통해 추정된다. 그리고 실제 자료분석을 통하여 설명한다.

주요용어: 변량효과 공분산행렬, 일반화선형혼합모형, 수정된 콜레스키 분해, 절삭된 포아송모형.

이 논문은 2012, 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2012R1A1A1004002, NRF-2014R1A1A2054997).

[1]교신저자: (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: keunbaik@skku.edu