The Korean Journal of Applied Statistics
2023, Vol. 36, No. 3, 225–243

DOI: https://doi.org/10.5351/KJAS.2023.36.3.225
Print ISSN1225-066X / Online ISSN 2383-5818

Bayesian marginalized two-part mixed effects model based on generalized gamma distribution

Yongtae Kwon[a], Keunbaik Lee[1,a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

Medical expenses, alcohol consumption, and rainfall are all examples of zero-inflated semicontinuous data. To analyze these data, two-part models have been proposed, consisting of a binary submodel that determines whether the data is zero or not, and a continuous submodel for data greater than zero. To analyze longitudinal zero-inflated semicontinuous data, the two-part models are extended to conditional two-part models and marginalized two-part models. We review two-part models and conditional/marginal two-part models in this paper. Then, in the marginalized two-part longitudinal models, we propose a Bayesian method for dealing with the frequentist problem of estimation failure and convergence. A simulation study is being carried out to compare our proposed model. We also use the Korea Health Panel Survey to examine the medical expenses of young people.

Keywords: two-part model, longitudinal semicontinuous data, generalized gamma distribution, Bayesian, marginalized two-part model

1. 서론

영과잉 반 연속 자료(zero-inflated semicontinuous data)는 0의 값이 많은 연속형 자료로 의료 비용, 음주량 그리고 강수량 등 사용자(user)와 비사용자(non-user)로 구분될 수 있는 자료에서 많이 띄는 형태이다. 해당 자료는 0에서부터 멀어질수록 점차 감소하며, 매우 치우친(skewed) 형태의 분포를 가진다. 이러한 자료들을 분석하는 방법으로는 자료들에 약간 상수 c를 더하고 로그 변환하여 전체 자료를 분석하는 방법과 0을 제외하고 0보다 큰 자료들로만 분석하는 방법 그리고 자료를 2개의 부모형으로 나눈 2-부분 모형(two-part model) 등이 있다 (Min과 Agresti, 2002). 자료에 상수를 더하여 분석하는 방법이나 0을 제외하고 0보다 큰 자료들만 분석하는 방법은 자료를 인위적으로 변화하고 자료 정보의 손실을 초래하게 된다. 이러한 손실을 피하기 위해 Duan 등 (1983)은 영과잉 반 연속 자료 분석을 위해 2-부분 모형을 제안하였다.

2-부분 모형은 2-부분 모형은 0인 자료들을 구분하는 이항 부모형(binary sub model)과 0보다 큰 자료들에 대한 연속형 부모형(continuous sub model)을 가지는 혼합 모형이다 (Duan 등, 1983). 연속형 부모형은 0보다 큰 자료들에 대해 조건부 연속형 분포를 가정하여 자료를 설명하였고, 그 조건부 연속형 분포로는 로그 정규 분포, 감마 분포 등 다양한 분포를 적용할 수 있다. 경시적 영과잉 반 연속 자료를 분석하기 위해서 조건부 2-부분 경시적 모형 (Olsen과 Schafer, 2001)과 주변화 2-부분 경시적 모형 (Smith 등, 2014)이 제안되었으며

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2C1002752). This paper was prepared by extracting part of Yongtae Kwon's thesis.

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: keunbaik@skku.edu

Published 30 June 2023 / journal homepage: http://kjas.or.kr

각각의 모형은 연속형 부모형에서 조건부 기댓값과 주변 기댓값을 이용한 모형이다. Smith 등 (2018)은 횡단면 자료에서 2-부분 모형과 동일하게 경시적 자료의 전반적인 해석이 가능한 주변화 2-부분 경시적 모형에서 반복 측정되는 자료들의 0의 비율이 적어도 10%이상이어야 한다고 하였다.

본 논문은 빈도주의 관점에서 주변화 2-부분 경시적 모형이 가지는 추정 및 수렴의 실패문제를 해결할 수 있으며 반복 측정된 자료에서 중도 절단된 자료를 유연하게 대처할 수 있는 베이지안 마코프-체인 몬테칼로 방법을 통한 모수 추정을 제안한다. 그리고 제안된 모형을 실제 자료인 한국의료패널 자료에 적합하여 분석하고자 한다.

본 논문의 구성은 다음과 같다. 우선 2절에서는 경시적 영과잉 반 연속 자료 분석을 위한 조건부 및 주변화 2-부분 경시적 모형을 살펴보고, 연속형 부모형에서 연속형 분포에 일반화 감마 분포를 적용한 주변화 2-부분 경시적 모형를 살펴본다. 3절에서는 일반화 감마 분포 기반 주변화 2-부분 경시적 모형의 베이지안 추정 방법을 제안한다. 4절에서는 3절에서 제시한 방법에 따른 모의실험을 진행하여 그 성능을 제시한다. 그리고 5절에서는 한국의료패널 자료에서 청년층 의료비에 대한 실증 분석, 결과를 종합한 6절 결론으로 구성되어 있다.

2. 2-부분 모형

2-부분 모형은 Duan 등 (1983)에 의해 제안되었으며 자료가 음이 아닌 값을 가지면서 0이 과도하게 많은 경우 사용하는 모형이다. 이 모형에서 0인지 아닌지를 구분하는 첫 번째 부모형과 0보다 큰 자료들에 대한 두 번째 부모형으로 구성되었다. 음이 아닌 확률 변수 $y_i(\geq 0)$ $(i = 1, \ldots, n)$에 대해서 횡단면 자료에 대한 2-부분 모형을 함수로 표현하면 다음과 같다.

$$f(y_i) = \begin{cases} 1 - \pi_i, & \text{if } y_i = 0; \\ \pi_i g(y_i), & \text{if } y_i > 0, \end{cases}$$

여기서 $\pi_i = P(Y_i > 0)$이고 $g(\cdot)$는 0보다 큰 값들의 분포이며 로그 정규 분포, 로그-왜 정규 분포 그리고 감마 분포 등 여러 분포를 적용시킬 수 있다. 2-부분 모형에서 π에 로짓 함수를 $g(\cdot)$에 로그 정규 분포를 가정하면, 첫 번째 부모형과 두 번째 부모형의 모수화를 다음과 같이 표현할 수 있다.

$$\begin{aligned} \text{이항 부모형} \ : \ & \text{logit}(\pi_i) = \text{logit}[P(Y_i > 0)] = \boldsymbol{x}_{1i}^\top \boldsymbol{\alpha}_1, \\ \text{연속형 부모형} \ : \ & \text{for } y_i > 0, \quad \text{또는} \\ & \mathrm{E}[\log(y_i) \mid y_i > 0] = \boldsymbol{x}_{2i}^\top \boldsymbol{\alpha}_2, \end{aligned}$$

여기서 $\boldsymbol{x}_{1i}$와 $\boldsymbol{x}_{2i}$는 각각 $p_1 \times 1$과 $p_2 \times 1$인 공변량 벡터이며, $\boldsymbol{\alpha}_1$과 $\boldsymbol{\alpha}_2$는 상응하는 모수 벡터이다. 두 번째 부모형에서의 로그 정규 분포는 자료가 0보다 큰 값들의 조건부 분포이다.

2.1. 경시적 자료를 위한 2-부분 모형

이 절에서는 횡단면 자료를 위한 2-부분 모형을 경시적 자료로 확장시킨 조건부 2-부분 경시적 모형과 주변화 2-부분 경시적 모형에 대해서 알아보고자 한다.

2.1.1. 조건부 2-부분 경시적 모형

경시적 반 연속 자료 분석을 위한 조건부 2-부분 경시적 모형(conditional two-part longitudinal model; CTP)은 Olsen와 Schafer (2001)에 의해 제안되었다. 음이 아닌 확률 변수 y_{ij} (≥ 0) $(i = 1, \ldots, n;\ j = 1, \ldots, n_i)$

에 대해서 경시적 자료에 대한 CTP 모형을 함수로 표현하면 다음과 같다.

$$f\left(y_{ij}\right) = \begin{cases} 1-\pi_{ij}, & \text{if } y_{ij}=0; \\ \pi_{ij} g\left(y_{ij}\right), & \text{if } y_{ij}>0, \end{cases} \tag{2.1}$$

여기서 $\pi_{ij} = \mathrm{P}(Y_{ij} > 0)$이고, $g(\cdot)$는 0보다 큰 값들의 분포이며, 앞에서 설명했던 2-부분 모형과 동일하게 $g(\cdot)$에 로그 정규 분포, 로그-왜 정규 분포 그리고 감마 분포 등의 분포를 적용시킬 수 있다.

앞서 2-부분 모형과 동일하게 첫 번째 부모형에는 로지스틱 회귀분석 그리고 두 번째 부모형에서 $g(\cdot)$를 로그 정규 분포를 가정한다. 그리고 각 부모형에 개체별 반복 측정의 특성을 고려한 임의효과(random effect)를 추가하여 모수화를 진행한다. 따라서 가정된 CTP 모형의 모수화는 다음과 같이 표현할 수 있다.

$$\begin{aligned} \operatorname{logit}\left(\pi_{ij}\right) &= \boldsymbol{x}_{1ij}^{\top}\boldsymbol{\gamma}_1 + \boldsymbol{z}_{1ij}^{\top}\boldsymbol{d}_{1i}, \\ E\left[\log\left(y_{ij}\right) \mid y_{ij}>0\right] &= \boldsymbol{x}_{2ij}^{\top}\boldsymbol{\gamma}_2 + \boldsymbol{z}_{2ij}^{\top}\boldsymbol{d}_{2i}, \end{aligned} \tag{2.2}$$

여기서 $\pi_{ij} = P(Y_{ij} > 0|\boldsymbol{d}_i)$이고 $\boldsymbol{x}_{1ij}$와 $\boldsymbol{x}_{2ij}$는 각각 $p_{x1} \times 1$과 $p_{x2} \times 1$인 공변량 벡터이며, $\boldsymbol{\gamma}_1$과 $\boldsymbol{\gamma}_2$는 상응하는 모수 벡터이다. 그리고 $\boldsymbol{d}_{1i}$과 $\boldsymbol{d}_{2i}$는 각각 $p_{z1} \times 1$과 $p_{z2} \times 1$인 임의효과 벡터이며, $\boldsymbol{z}_{1ij}$와 $\boldsymbol{z}_{2ij}$는 임의효과 벡터에 상응하는 공변량 벡터이다. 여기서 만약 $\boldsymbol{z}_{1ij}$와 $\boldsymbol{z}_{2ij}$가 모두 1인 경우 각각의 부모형은 임의절편 효과(random intercept effect)만 가지는 혼합효과 모형이 된다.

CTP 모형에서는 두 번째 부모형에서의 조건부 기댓값에 의해 0보다 큰 값들로만 모수를 추정하게 된다. 즉, 조건부 모수화로 인해 전체 자료에 대한 해석을 하지 못하고 0보다 큰 자료들에 대해 부분적인 해석만 가능하다. 예를 들어 CTP 모형으로 경시적 의료비 자료를 분석할 때, 의료비가 0원인 사람과 0원 보다 큰 사람들의 부분으로 나누어 각 부분에 대한 모수를 추정하기 때문에 의료비가 0원 보다 큰 사람들에 대해서만 의료비 해석이 가능하다. 그리고 위와 같이 조건부 모형에서 로그 변환된 자료값을 사용하게 되면 본래의 단위로 변환하여 해석할 수 없다는 한계점이 존재한다.

2.1.2. 주변화 2-부분 경시적 모형

앞서 제시한 조건부 모형의 문제를 개선하고자 Smith 등 (2014)은 CTP 모형의 두 번째 부모형에서 조건부 자료의 모형을 모수화 하는 것이 아닌 전체 자료에 대한 모형의 모수화 하는 주변화 2-부분 경시적 모형을 제안하였다.

주변화 2-부분 경시적 모형을 설명하기 앞서, 경시적 자료 분석에서 주변 모형(marginal model)이 가지는 의미를 우선 설명하고자 한다. 경시적 주변 모형은 개체간 공변량 효과를 파악하기 위해 주변 평균 $\mathrm{E}(Y_{ij})$를 통해 모형화를 한다. 즉, 주변 모형은 모집단 평균에 대한 공변량의 효과를 모형화 하는 모집단 평균 모형이며 함수는 다음과 같다.

$$\begin{aligned} \mathrm{E}\left(Y_{ij} \mid x_i\right) &= \mu_{ij}, \\ l\left(\mu_{ij}\right) &= x_{ij}^{\top}\beta, \end{aligned} \tag{2.3}$$

여기서 $l(\cdot)$는 연결 함수를 의미하며, 경시적 주변 모형은 연결 함수를 통해서 공변량에 의존한다.

다음으로 Smith 등 (2014)가 제안한 주변화 2-부분 경시적 모형(marginalized two-part longitudinal model; MTP)을 설명한다. MTP 모형은 주변 해석이 가능한 일반화 선형 모형과 CTP 모형의 구조적 특성이 합쳐진 모형이다. MTP 모형은 CTP 모형의 식 (2.1)과 같이 함수가 동일한 구조적 특성을 띄고 있지만, 두 번째 부모형에서 0보다 큰 자료들이 아니라 전체 자료에 대해서 모수화를 진행하여 조건부 해석이 아닌 주변 해석을 가능하게 하고 변환된 자료가 아닌 원자료 값을 사용한다는 점에서 CTP 모형과 차이가 있다.

MTP 모형의 첫 번째 부모형은 CTP 모형과 동일하게 임의효과를 추가한 로지스틱 혼합효과 모형을 가정하고, 두 번째 부모형은 로그선형 모형 즉, 임의효과를 추가한 혼합효과 모형을 가정한다. MTP 모형은 다음과 같이 표현된다.

$$\begin{aligned} \text{logit}\left(\pi_{ij}\right) &= \boldsymbol{x}_{1ij}^{\top}\boldsymbol{\beta}_1 + \boldsymbol{z}_{1ij}^{\top}\boldsymbol{b}_{1i}, \\ \log\left[E\left(Y_{ij} \mid \boldsymbol{b}_i\right)\right] &= \boldsymbol{x}_{2ij}^{\top}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2ij}^{\top}\boldsymbol{b}_{2i}, \quad \text{또는} \\ E\left(Y_{ij} \mid \boldsymbol{b}_i\right) &= \exp\left(\boldsymbol{x}_{2ij}^{\top}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2ij}^{\top}\boldsymbol{b}_{2i}\right), \end{aligned} \tag{2.4}$$

여기서 모수, 임의효과 및 공변량은 모두 CTP 모형 (2.2)에서 정의된 것과 동일하며, 두 번째 부모형은 주변모형 (2.1.2)의 로그연결 함수를 이용하여 공변량의 효과를 모형화하는 것과 동일하다.

위 모수화를 통해, CTP 모형과 MTP 모형의 두 번째 부모형의 추정할 모수의 의미는 CTP 모형의 고정효과 $\boldsymbol{\gamma}_2$와 임의효과 $\boldsymbol{d}_{2i}$는 0보다 큰 자료에 대한 효과를, MTP 모형의 고정효과 $\boldsymbol{\beta}_2$와 임의효과 $\boldsymbol{b}_{2i}$는 0을 포함한 전체 자료에 대한 효과를 의미한다.

MTP 모형에서는 CTP 모형과 동일하게 각 부모형에 임의효과가 존재하는데, 각각의 부모형의 임의효과 벡터가 서로 연관되어 있는 다변량 정규 분포를 가정한다. 임의효과에서는 각각의 부모형 내에서 반복 측정된 자료들의 상관관계 즉, 개체 내 상관관계와 첫 번째 부모형과 두 번째 부모형 간의 상관관계 즉, 부모형 간 상관관계 총 2가지의 상관관계를 고려한다. 임의효과에서 2가지 상관관계를 모두 고려한 이유는 개체 내 상관관계를 무시하게 되면 추정된 모수의 표준오차에 편향이 생기게 되고, 부모형 간의 상관관계를 무시하게 되면 모형의 계수 추정에 편향이 생기기 때문이다 (Su 등, 2009).

2.2. 주변화 2-부분 모형과 일반화 감마 분포

이 절에서는 두 번째 부모형에서 연속형 분포에 일반화 감마 분포를 적용한 주변화 2-부분 경시적 모형 (generalized gamma distribution based marginalized two-part model; MTP-GG)에 대해서 살펴보도록 하겠다.

2.2.1. 일반화 감마 분포

Manning 등 (2005)에 의해 제안된 일반화 감마 분포(generalized gamma distribution)는 위치 모수(location parameter), 척도 모수(scale parameter) 그리고 형태 모수(shape parameter)의 총 3개의 모수로 구성되어 있으며 확률 밀도 함수는 다음과 같다.

$$f\left(y;\kappa,\mu,\sigma\right) = \frac{\eta^{\eta}}{\sigma y\Gamma\left(\eta\right)\sqrt{\eta}}\exp\left[u\sqrt{\eta} - \eta\exp\left(|\kappa|u\right)\right], \tag{2.5}$$

여기서 $\Gamma(\cdot)$은 표준 감마 함수이고, $u = \text{sign}(\kappa)(\log(y)-\mu)/\sigma$이며 $\mu > 0$는 위치 모수, $\sigma > 0$는 척도 모수 그리고 $-\infty < \kappa < \infty$이며, $\eta = |\kappa|^{-2} > 0$이다.

일반화 감마 분포의 평균과 분산은 다음과 같다.

$$\begin{aligned} E\left(Y\right) &= \exp\left\{\mu + c\left(\sigma,\kappa\right)\right\}, \\ \text{Var}\left(Y\right) &= \left\{\exp\left(\mu\right)\kappa^{2\sigma}/\kappa\right\}^2 \times \left\{\frac{\Gamma\left(1/\kappa^2 + 2\sigma/\kappa\right)}{\Gamma\left(1/\kappa^2\right)} - \left[\frac{\Gamma\left(1/\kappa^2 + 2\sigma/\kappa\right)}{\Gamma\left(1/\kappa^2\right)}\right]^{-2}\right\}, \end{aligned}$$

여기서 $c(\sigma,\kappa) = (\sigma\log(\kappa^2))/\kappa + \log[\Gamma(1/\kappa^2 + \sigma/\kappa)] - \log[\Gamma(1/\kappa^2)]$이다.

일반화 감마 분포는 여러 분포로 변형이 가능한데, $\sigma = \kappa$인 경우는 감마 분포로, $\kappa = 1$인 경우는 와이블 (Weibull) 분포 그리고 $\kappa \to 0$로 가는 경우는 로그 정규 분포로 (극한 분포)변형이 가능하다. 그리고 Voronca

등 (2015)에 의하면 일반화 감마 분포를 적용하게 되면 다양한 형태들의 자료를 유연하게 설명할 수 있다고 하였는데, 이러한 점들을 통해 일반화 감마 분포가 영과잉 반 연속 자료에 대한 모형 적합이 용이할 것이라고 기대할 수 있다.

2.2.2. 일반화 감마 분포 기반 주변화 2-부분 모형

이 절에서는 2.1 절에서 제시한 경시적 영과잉 자료 분석을 위한 주변화 2-부분 모형에서 두 번째 부모형에 일반화 감마 분포를 적용한 모형 (Voronca 등, 2015)을 고찰한다. Voronca 등 (2015)의 모형에서 첫 번째 부모형은 로지스틱 회귀 모형에 임의효과를 추가한 혼합효과 모형이며, 두 번째 부모형은 앞에서 언급했던 일반화 감마 분포를 적용한 로그 선형 모형에 임의효과를 추가한 혼합효과 모형이다. 해당 모형을 식 (2.4)에서 제시한 모형이며, 두 부모형에 공통적으로 들어가는 임의효과 벡터들은 독립이 아닌 다변량 정규 분포를 가정한다.

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_{1i} \\ \boldsymbol{b}_{2i} \end{pmatrix} \sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22} \end{pmatrix}\right).$$

모수 추정을 위한 가능도 함수는 다음과 같다.

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\int \left(1-\pi_{ij}\right) I_{(y_{ij}=0)}\left[\pi_{ij} g\left(y_{ij};\kappa,\mu,\sigma,\boldsymbol{b}_{2i}\right)\right] I_{(y_{ij}>0)} \phi\left(\boldsymbol{b}_i;\boldsymbol{0},\boldsymbol{\Sigma}\right) d\boldsymbol{b}_i,$$

여기서 $\boldsymbol{\theta} = (\pi, \kappa, \mu, \sigma, \Sigma)$이고 $g(y_{ij};\kappa,\mu,\sigma,\boldsymbol{b}_{2i})$는 일반화 감마 분포 (2.5)이며, $\phi(\boldsymbol{b}_i;\boldsymbol{0},\boldsymbol{\Sigma})$는 평균 벡터가 $\boldsymbol{0}$이고 분산-공분산 행렬이 $\boldsymbol{\Sigma}$인 다변량 정규 분포이다.

일반화 감마 분포를 적용한 주변화 2-부분 경시적 모형에서 일반화 감마 분포의 평균과 로그 선형 모형인 두 번째 부모형과의 선형결합식을 통해 위치 모수 μ를 다음과 같이 재모수화할 수 있다.

$$\begin{aligned} & E\left(Y_{ij} \mid \boldsymbol{b}_i\right) = \Pr\left(Y_{ij} > 0 \mid \boldsymbol{b}_i\right) E\left(Y_{ij} \mid Y_{ij} > 0, \boldsymbol{b}_i\right), \\ \Leftrightarrow\ & \exp\left(\boldsymbol{x}_{2ij}^{\top}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2ij}^{\top}\boldsymbol{b}_{2i}\right) = \pi_{ij}\exp\left\{\mu_{ij} + c\left(\sigma,\kappa\right)\right\}, \\ \Leftrightarrow\ & \mu_{ij} = \boldsymbol{x}_{2ij}^{\top}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2ij}^{\top}\boldsymbol{b}_{2i} - \log\left(\pi_{ij}\right) - c\left(\sigma,\kappa\right), \end{aligned}$$

여기서 $c(\sigma,\kappa) = (\sigma\log(\kappa^2))/\kappa + \log[\Gamma(1/\kappa^2 + \sigma/\kappa)] - \log[\Gamma(1/\kappa^2)]$이다. 위치 모수 μ는 두 번째 부모형의 공변량 효과 $\boldsymbol{\beta}_2$, 임의효과 $\boldsymbol{b}_{2i}$, 첫 번째 부모형 그리고 일반화 감마 분포의 모수 (σ,κ)로 표현된다. μ의 재모수화를 통해 추정할 모수의 개수를 줄일 수 있고 이로 인해 모수 추정의 시간 단축 효과를 얻을 수 있다.

결과적으로 위치 모수 μ를 재모수화한 주변화 2-부분 경시적 모형은 다음과 같이 표현할 수 있다.

$$\begin{aligned} L\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\kappa,\sigma,\Sigma \mid \boldsymbol{y}\right) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\int & \left(1 - \frac{1}{1+\exp\left(-\boldsymbol{x}_{1ij}^{\top}\boldsymbol{\beta}_1 - \boldsymbol{z}_{1ij}^{\top}\boldsymbol{b}_{1i}\right)}\right) I_{(y_{ij}=0)} \\ & \times \left\{\frac{\kappa^{-2|\kappa|^{-2}}\exp\left\{Q(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\kappa,\sigma)\right\}}{\left(1+\exp\left(-\boldsymbol{x}_{1ij}^{\top}\boldsymbol{\beta}_1 - \boldsymbol{z}_{1ij}^{\top}\boldsymbol{b}_{1i}\right)\right)\sigma y_{ij}\Gamma\left(|\kappa|^{-2}\right)\sqrt{|\kappa|^{-2}}}\right\} I_{(y_{ij}>0)}\phi\left(\boldsymbol{b}_i;\boldsymbol{0},\boldsymbol{\Sigma}\right) d\boldsymbol{b}_i, \end{aligned}$$

여기서 Q 와 c는

$$\begin{aligned} Q\left(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\kappa,\sigma\right) = & \sqrt{|\kappa|^{-2}}\frac{\text{sign}\left(\kappa\right)\left(\log\left(y_{ij}\right) - \left(\boldsymbol{x}_{2ij}^{\top}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2ij}^{\top}\boldsymbol{b}_{2i} - \log\left(\text{expit}\left(\boldsymbol{x}_{1ij}^{\top}\boldsymbol{\beta}_1 + \boldsymbol{z}_{1ij}^{\top}\boldsymbol{b}_{1i}\right)\right) - c\left(\sigma,\kappa\right)\right)\right)}{\sigma} \\ & - |\kappa|^{-2}\exp\left(\kappa\left(\log\left(y_{ij}\right) - \left(\boldsymbol{x}_{2ij}^{\top}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2ij}^{\top}\boldsymbol{b}_{2i} - \log\left(\text{expit}\left(\boldsymbol{z}_{1ij}^{\top}\boldsymbol{\beta}_1 + \boldsymbol{x}_{1ij}^{\top}\boldsymbol{b}_{1i}\right)\right) - c\left(\sigma,\kappa\right)\right)\right)/\sigma\right) \end{aligned}$$

그리고 $c(\sigma,\kappa) = (\sigma\log(\kappa^2))/\kappa + \log[\Gamma(1/\kappa^2 + \sigma/\kappa)] - \log[\Gamma(1/\kappa^2)]$이다.

두 번째 부모형의 $g(\cdot)$에 일반화 감마 분포를 적용한 모형은 모수가 특정 값을 가질 때 앞서 언급한 일반적인 분포가 적용된 주변화 2-부분 모형으로 축소가 가능하다. 예를 들어 $c(\sigma,\kappa) = 0$이고 $\sigma = \kappa$이면 감마 분포가 적용된 주변화 2-부분 모형으로, $c(\sigma,\kappa) = \log[\Gamma(1+\sigma)]$이고 $\kappa = 1$이면 와이블 분포가 적용된 주변화 2-부분 모형으로, 그리고 $c(\sigma,\kappa) = \sigma^2/2$이고 $\kappa \to 1$이면 로그 정규 분포가 적용된 주변화 2-부분 모형으로 축소가 가능하다.

위 로그 가능도 함수를 통해 Voronca 등 (2015)는 빈도주의적 접근으로 1차 편미분한 값을 통해 최대화 하는 값을 모수를 추정하였으며, dual quasi Newton Raphson 알고리즘 기반 최적화를 통해 최대 가능도 추정량을 구하였다. 그리고 모형의 점근 표준 오차(asymptotic standard error)는 추정된 최대 가능도 추정량을 이용하여 피셔 정보(Fisher information)를 통해 계산된다. 또한 Jaffa 등 (2018)는 임의효과의 예측을 위해 adaptive Gaussian quadrature 접근을 이용한 경험적 베이즈 추정량을 사용하였다.

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\kappa}, \hat{\sigma}, \hat{\Sigma}\right) = \mathrm{diag}\left\{I^{-1}\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\kappa}, \hat{\sigma}, \hat{\Sigma}\right)\right\}.$$

하지만 여기서 Smith 등 (2017)은 임의효과의 구조가 고차원으로 갈수록 Gaussian quadrature 방법이 수렴에 실패한다고 하였다. 따라서 일반화 감마 분포를 적용한 주변화 2-부분 모형을 빈도주의적 방법이 아닌 베이지안 방법을 통해 접근해 보고자 한다.

3. 베이지안 추정 방법

이 절에서는 일반화 감마 분포 기반 주변화 2-부분 경시적 모형에 대해 베이지안 MCMC 추정 방법을 활용하여 무정보(non-informative) 사전 분포(prior distribution)와 가능도 함수로부터 사후 분포를 추정하고 해당 분포로부터 MCMC 표본을 추출하여 모수를 추정하는 방법을 제안해 보고자 한다.

3.1. 베이지안 MCMC

주변화 2-부분 경시적 모형의 모수인 $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \kappa, \sigma$ 그리고 Σ를 추정하기 위해서 Smith 등 (2017)에 의하면 빈도주의적 추정 방법인 가우시안 구적법(gaussian quadrature)으로 임의 기울기를 추가하여 모수를 추정하였다. 하지만 이 경우 모수 추정이 종종 수렴에 실패하게 된다. 따라서 임의효과에 대한 구조가 복잡해짐에 따른 모수 추정이 어려워지는 것이다. 이것을 극복하기 위해 우리는 베이지안 방법을 이용한 모수를 추정을 제안한다.

3.2. 사전 분포

주변화 2-부분 경시적 모형에 대한 모수들의 사전 분포는 다음과 같이 가정한다.

$$\begin{aligned}
\boldsymbol{\beta}_1 &\sim N(\beta_{10}, \Lambda_1), \\
\boldsymbol{\beta}_2 &\sim N(\beta_{20}, \Lambda_2), \\
\kappa &\sim U(a, b), \\
\sigma^2 &\sim \text{Inv.Gamma}(\omega_1, \omega_2), \\
\Sigma &\sim \text{Inv.Wishert}(\Omega, \omega_3),
\end{aligned}$$

여기서 대부분의 사전 분포는 무정보에 가까운 사전 분포를 가정한다. 사전 분포의 공변량 모수 $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$의 다변량 정규 분포 평균 벡터는 $\mathbf{0}$으로, 분산-공분산 행렬은 각각의 분산이 100인 대각 행렬을, κ의 균등 분포에서

a, b는 각각 −10, 10으로 σ^2의 역감마 분포는 ω_1, ω_2 모두 0.01, Σ의 역위샤트 분포에서 Ω는 임의효과 벡터의 차원만큼, ω_3는 임의효과 벡터의 개수만큼의 차원을 가진 단위 행렬의 값을 지정한다.

일반화 감마 분포 기반 주변화 2-부분 경시적 모형의 사후 분포에서 임의효과 벡터 $\boldsymbol{b}_i$에 대한 주변화가 용이하지 않기 때문에 $\boldsymbol{b}_i$에 대한 위계적 모형(hierarchical model)을 포함시켜 재추정한다. 재추정하는 사전 분포에 대한 모수들은 위에서 제시한 기존 모수의 분포와 동일하게 가정하며 임의효과 벡터 $\boldsymbol{b}_i|\Sigma$의 다변량 정규 분포에서 평균 벡터는 $\mathbf{0}$로 가정한다. 위와 같은 모수들을 통해 사후 분포를 추정하고 임의효과 벡터를 제외한 나머지 모수들에 대한 해석을 진행한다. 사후 분포에 대한 수식은 다음과 같다.

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \kappa, \sigma, \boldsymbol{b}, \Sigma \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \kappa, \sigma, \boldsymbol{b}, \Sigma) \times p(\boldsymbol{\beta}_1)\, p(\boldsymbol{\beta}_2)\, p(\kappa)\, p(\sigma)\, p(\boldsymbol{b} \mid \Sigma)\, p(\Sigma).$$

주변화 2-부분 경시적 모형에 대한 베이지안 추정은 통계 프로그램 R에 있는 rstan 패키지를 이용하여 모수 추정을 진행한다. rstan에서 사용하는 MCMC 추정 방법은 HMC기반인 NUTS를 사용하며, 추정된 사후 분포로부터 MCMC 표본을 추출 한다.

4. 모의실험

제안된 일반화 감마 분포 기반 주변화 2-부분 모형을 임의 추출한 표본으로부터 모수를 잘 추정하는지를 파악해 보고, 경쟁 분포인 로그-왜 정규 분포를 적용시켜 각기 다른 분포로부터 표본추출하여 일반화 감마 분포 기반 주변화 2-부분 경시적 모형과 로그-왜 정규 분포 기반 주변화 2-부분 경시적 모형(marginalized two-part longitudinal model based on log-skew normal distribution; MTP-LSN)에서의 모수 추정 결과를 비교해 보고자 한다.

첫 번째 모의실험은 일반화 감마 분포에서 임의추출한 표본을 이용하여 주변화 2-부분 모형의 모수를 베이지안 추정이 잘 작동함을 보여서 그 추정의 성능을 제시하고자 한다. 두 번째 모의실험은 주변화 2-부분 모형의 두 번째 부모형에서 기존에 Smith 등 (2017)에 의해 베이지안 방법으로 제안된 로그-왜 정규 분포와 본 논문에서 제안하는 일반화 감마 분포를 비교하고자 한다. 먼저 각각의 분포에서 표본 추출하여 각기 다른 분포 $g(\cdot)$에 적용하고 베이지안 방법으로 주변화 2-부분 모형의 모수를 추정하여 그 추정치들을 비교하고자 한다.

각각 모의실험에서는 난수발생을 위하여 식 (2.4)을 아래와 같이 고려하였다.

$$\begin{aligned}
\operatorname{logit}\left(\pi_{ij}\right) &= \beta_{10} + \beta_{11} t_{ij} + \beta_{12} G_i + b_{1i0}, \\
\log\left(E\left(Y_{ij} \mid \boldsymbol{b}_i\right)\right) &= \beta_{20} + \beta_{21} t_{ij} + \beta_{22} G_i + \beta_{23}\left(t_{ij} \times G_i\right) + b_{2i0},
\end{aligned}$$

여기서 $t_{ij} = j$ 그리고 G_i은 0 또는 1의 값을 가지고 각 그룹은 같은 크기의 표본을 가지며, 임의절편 벡터 $\boldsymbol{b}_{i0} = (b_{1i0}, b_{2i0})^\top$는 이변량 정규 분포 MVN($\mathbf{0}, \boldsymbol{\Sigma}$)에서부터 임의추출하였다. 그리고 두 번째 모의실험에 대한 연속형 분포 로그-왜 정규 분포 $g(\cdot)$은 다음과 같다.

$$\begin{aligned}
g\left(y_{ij}; \mu_{ij}, \omega, \kappa\right) &= \frac{2}{\omega y_{ij}} \phi\left(\frac{\log\left(y_{ij}\right) - \mu_{ij}}{\omega}\right) \Phi\left(\frac{\kappa}{\omega}\left(\log\left(y_{ij}\right) - \mu_{ij}\right)\right), \\
\mu_{ij} &= \boldsymbol{x}_{2ij}^\top \boldsymbol{\beta}_2 + \boldsymbol{z}_{2ij}^\top \boldsymbol{b}_{2i} - \log(2) - \log\left(\pi_{ij}\right) - \log\left[\Phi(\omega\delta)\right] - \frac{\omega^2}{2},
\end{aligned} \tag{4.1}$$

여기서 $\boldsymbol{x}_{2ij}$는 $p_{x2} \times 1$인 공변량 벡터이며, $\boldsymbol{\beta}_2$는 상응하는 모수 벡터이고, $\boldsymbol{b}_{2i}$는 $p_{z2} \times 1$인 임의효과 벡터이며, $\boldsymbol{z}_{2ij}$는 임의효과 벡터에 상응하는 공변량 벡터이다. 그리고 π_{ij}는 식 (2.4)의 것과 동일하며, $\omega > 0$는 로그 척도 모수, κ는 로그 형태 모수, $\delta = \kappa/(\sqrt{1+\kappa^2})$ 그리고 $\Phi(\cdot)$는 표준 정규 분포의 누적 분포 함수이다. 여기서도

Table 1: Average of probability zero in generated random number from each distribution

	MTP-GG			MTP-LSN
$P(Y_{ij}=0)$	$n=300$	$n=500$	$n=1000$	$n=500$
$t_{i1}=1$	35.4%	35.2%	35.1%	35.4%
$t_{i2}=2$	32.5%	32.4%	32.8%	32.7%
$t_{i3}=3$	29.7%	29.9%	30.1%	30.0%
$t_{i4}=4$	27.5%	27.2%	27.5%	27.6%

Table 2: Simulation results for parameter estimates of generalized gamma distribution based marginalized two-part model with sample sizes of 300, 500, and 1000

Parmater	$n=300$			$n=500$			$n=1000$		
(true value)	Mean	CP	PRB	Mean	CP	PRB	Mean	CP	PRB
β_{10} (0.70)	0.7022	97	0.31	0.6953	97	0.67	0.7087	97	1.24
β_{11} (0.15)	0.1516	95	1.07	0.1516	**94**	1.05	0.1462	**94**	2.56
β_{12} (−0.20)	−0.1962	**94**	1.91	−0.1796	97	**10.21**	−0.1937	97	3.15
β_{20} (−2.70)	−2.7083	97	0.31	−2.6795	98	0.76	−2.7085	99	0.31
β_{21} (0.50)	0.5055	**94**	1.09	0.4968	96	0.65	0.5018	96	0.36
β_{22} (−1.00)	−0.9932	**94**	0.68	−1.0079	98	0.79	−0.9865	97	1.35
β_{23} (0.25)	0.2455	96	1.80	0.2540	95	1.59	0.2466	**94**	1.34
σ_{11} (1.20)	1.1892	**94**	0.90	1.1843	96	1.31	1.2193	95	1.60
σ_{12} (1.00)	0.9951	**88**	0.49	0.9848	96	1.52	1.0054	**94**	0.54
σ_{22} (1.50)	1.5137	**94**	0.91	1.4841	**94**	1.06	1.5119	96	0.80
κ (1.50)	1.6132	97	7.55	1.5408	98	2.72	1.5187	**93**	1.25
σ^2 (1.00)	0.9724	95	2.76	0.9929	95	0.71	0.9917	96	0.83

The average of posterior means, coverage probability (CP), and percent relative bias (PRB) are displayed.

일반화 감마 분포 기반 주변화 2-부분 경시적 모형과 동일하게 Smith 등 (2017)이 제시한 위치 모수 μ를 식 (4.1)와 같이 재모수화한 로그-왜 정규 분포에서 난수발생하여 모의실험을 진행하였다.

각 부모형의 모수들의 참값은 다음과 같이 설정하였다. 첫 번째 부모형의 모수 $\beta_1 = (0.7, 0.15, -0.2)$, 두 번째 부모형의 모수 $\beta_2 = (-2.7, 0.5, -1, 0.25)$, 임의절편만 가정한 임의효과 벡터의 분산-공분산 행렬 $\Sigma = \left(\begin{smallmatrix} 1.2 & 1.0 \\ 1.0 & 1.5 \end{smallmatrix}\right)$, $\kappa = 1.5$ 그리고 $\sigma^2 = 1$ 모수를 가진 일반화 감마 분포 기반 주변화 2-부분 모형에서 300개, 500개 그리고 1,000개의 표본을 반복 수 4로 놓고 임의추출하여 모의실험을 진행하였다. 두 번째 모의실험에서는 첫 번째 모의실험에서의 모수와 동일하게 진행하였으며, 표본 수는 500개로 고정하고 로그-왜 정규 분포에서도 동일한 설정으로 임의추출하였다. 총 2번의 모의실험 모두 모수의 동일한 세팅으로부터 총 100번 시행 후, 각 시행별로 표본들의 평균값으로 모수 추정을 진행하였다. 각각의 모의실험에서는 0인 자료들의 비율을 25%에서 35%로 맞췄으며, 결과적으로 임의추출한 표본들의 평균적인 0값의 비율은 Table 1과 같다.

각각의 모의실험에 대한 평가지표는 최고 사후 밀도 구간(highest posterior density interval; HPDI)의 포함 확률(coverage probability; CP)과 상대 오차 백분율(percentage relative bias; PRB)을 사용하였다. 최고 사후 밀도 구간은 각 시행별 MCMC 표본들로부터 95%의 최고 밀도를 갖는 하한과 상한에 대한 구간이다. 최고 사후 밀도 구간의 포함 확률은 총 100번의 시행 중 몇 번의 시행이 최고 사후 밀도 구간 안에 모수가 포함되었는지를 확률로 표시한다. 최고 사후 밀도 구간의 포함 확률은 다음과 같다.

$$\mathrm{CP} = \frac{1}{100}\sum_{i}^{100} I_i^{\mathrm{HPDI}} \times 100\,(\%)\,,$$

여기서 I_i^{HPDI}는 최고 사후 밀도 구간 안에 모수가 포함되어 있으면 1, 아니면 0인 지시 함수이다.

다음으로 상대 오차 백분율은 모수에 대한 추정값에서 모수를 빼고 모수로 나눈 평가 지표이다. 해당 모의실험에서는 총 100번의 시행에 대한 모수 추정값들의 평균을 내어 해당 모수에서 빼고 모수로 나눈 값이며, 값이 작을수록 추정량이 모수를 잘 추정하였다고 할 수 있다. 상대 오차 백분율은 다음과 같다.

$$\text{PRB} = \left| \frac{\sum_i^{100} \hat{\theta}_i / 100 - \theta}{\theta} \right| \times 100\,(\%)\,,$$

여기서 θ는 주변화 2-부분 경시적 모형의 모수이다.

MCMC방법을 통해 추정된 사후 분포에서 각 시행별로 초기값(initial value)을 임의추출하였던 모형의 모수 값으로 설정하고 총 5,000번의 MCMC 표본을 추출하였다. 추출된 MCMC 표본 중에서 50% 즉, 처음부터 2,500개의 표본을 제외하고(burn-in) 나머지 2,500개 표본들에 대해서 각각 분포의 모형 적합에 대한 평가를 진행하였다.

4.1. 모의실험 : 일반화 감마 분포 적합

첫 번째 모의실험에서는 일반화 감마 분포 기반 주변화 2-부분 경시적 모형에서 앞에서 주어진 모수를 토대로 임의 표본 추출을 하여, 베이지안 MCMC 방법으로 사후 분포를 추정하여 MCMC 표본을 추출한 후, 평균으로 모수를 추정하였다. Table 2는 100번의 모의 시행에 대한 모수 추정값의 평균과 최고 사후 밀도 구간의 포함 확률 그리고 상대 오차 백분율의 결과이다.

여기서 β_{ij}는 i 번째 부모형의 j 번째 회귀 계수이다. 첫 번째 모의실험 결과에 Table 2에 따르면, 표본의 수가 높아질수록 최고사후밀도구간의 포함 확률이 즉, 모수를 포함하는 확률이 높아짐을 알 수 있으며, 추정값과 모수를 비교하는 상대 오차 백분율도 표본 수 500개일 때 β_{12}가 10.21%임을 제외하고 표본 수가 높아질수록 낮아지게 되는 것으로 나타났다. 결과적으로 각 표본들에 대한 모의실험 전체 모수의 최고사후밀도구간의 포함 확률 평균과 상대 오차 백분율의 평균은 각각 300개에서 94.6%과 2.0%, 500개에서는 96.2%와 2.3% 그리고 1,000개에서는 95.7%과 1.5%로 표본 수 500개의 모의실험에서 최고사후밀도구간의 포함 확률이 가장 높았지만, 상대 오차 백분율 또한 가장 높은 것으로 나타났다. 이는 베이지안 MCMC 추정에서 임의 표본 추출의 개수가 높아질수록 일반화 감마 분포 기반 주변화 2-부분 경시적 모형의 모수 추정이 정확해진다고 할 수 있다.

4.2. 모의실험 : 일반화 감마 분포와 로그-왜 정규 분포 적합

두 번째 모의실험은 먼저 일반화 감마 분포 또는 로그-왜 정규 분포에서 각각 표본을 추출하고 그 자료를 일반화 감마 분포와 로그-왜 정규 분포를 가지는 2-부분 모형에 베이지안 MCMC 방법을 통해 모수를 추정하여 그 결과를 비교하고자 한다. 해당 모의실험은 일반화 감마 분포가 적용된 주변화 2-부분 모형이 경쟁 분포인 로그-왜 정규 분포와 비교해서 서로 다른 분포에서 추출한 자료에 대해서도 모수 추정의 성능이 좋은지 파악해보려 한다.

로그-왜 정규 분포 기반 주변화 2-부분 모형을 베이지안 모수 추정을 위해서 사전 분포가 필요하며 이는 Smith 등 (2017)이 제시한 분포를 가정하였다. 사전 분포에 사용되는 모수는 3절에서 제시한 일반화 감마 분포 기반 주변화 2-부분 모형의 사전 분포 모수와 동일하게 지정하였고 임의효과 벡터 $\boldsymbol{b}_i$에 대한 위계적 모형 또한 동일하게 가정하였다. 본 모의실험에서 사용될 로그-왜 정규 분포와 일반화 감마 분포 기반 주변화 2-부분

Table 3: Comparison of results from two models

Parmater	MTP-GG			MTP-LSN		
(true value)	Mean	CP	PRB	Mean	CP	PRB
β_{10} (0.70)	0.6953	97	0.67	0.7001	**94**	0.02
β_{11} (0.15)	0.1516	**94**	1.05	0.1511	**94**	0.76
β_{12} (−0.20)	−0.1796	97	**10.21**	−0.2212	98	**10.62**
β_{20} (−2.70)	−2.6795	98	0.76	−2.6909	**94**	0.34
β_{21} (0.50)	0.4968	96	0.65	0.5031	**92**	0.62
β_{22} (−1.00)	−1.0079	98	0.79	−0.9960	95	0.40
β_{23} (0.25)	0.254	95	1.59	0.2499	**94**	0.04
σ_{11} (1.20)	1.1843	96	1.31	1.2062	**91**	0.52
σ_{12} (1.00)	0.9848	96	1.52	0.9643	**92**	3.57
σ_{22} (1.50)	1.4841	**94**	1.06	1.3863	**85**	7.58

Generate random number from generalized gamma distribution based marginalized two-part model with sample size of 500. The average of posterior means, coverage probability (CP), and percent relative bias (PRB) are displayed.

모형의 사전 분포는 다음과 같다.

$$\begin{aligned}
\boldsymbol{\beta}_1 &\sim N\left(\mathbf{0}, 100 \times \boldsymbol{I}_{p_{x1}}\right), \\
\boldsymbol{\beta}_2 &\sim N\left(\mathbf{0}, 100 \times \boldsymbol{I}_{p_{x2}}\right), \\
\kappa &\sim U\left(-10, 10\right), \\
\sigma^2 &\sim \text{Inv.Gamma}\left(0.01, 0.01\right), \\
\boldsymbol{b}_i \,|\, \Sigma &\sim \text{MVN}\left(\mathbf{0}, \Sigma\right), \\
\Sigma &\sim \text{Inv.Wishert}\left(2, \boldsymbol{I_2}\right).
\end{aligned}$$

4.2.1. 일반화 감마 분포 추출

첫 번째 경우는 첫 번째 모의실험의 동일한 모수 설정으로 일반화 감마 분포 기반 주변화 2-부분 경시적 모형으로부터 임의 표본 500개를 추출하여 그 자료를 각각의 분포 기반 주변화 2-부분 경시적 모형에 적합시켰다. 총 100번의 시행에 대한 결과는 Table 3과 같다.

우리가 예상할 수 있듯이 일반화 감마 분포를 이용한 주변화 2-부분 모형은 자료를 생성한 모형이므로 상대 오차 백분율이 낮고, 최고사후밀도구간의 포함 확률이 94% 이상임을 알 수 있다. 로그-왜 정규 분포를 이용한 주변화 2-부분 모형도 상대 오차 백분율이 작음을 알 수 있다. 하지만 각각 모형의 모의실험 전체 모수의 최고사후밀도구간의 포함 확률 평균과 상대 오차 백분율 평균은 MTP-GG에서 각각 96.1%과 1.96% 그리고 MTP-LSN에서는 92.9%와 2.45%로 나타났다. 일반화 감마 분포가 로그-왜 정규 분포보다 최고사후밀도구간의 포함 확률이 평균적으로 높으며, 상대 오차 백분율은 평균적으로 더 낮음을 알 수 있다. 이는 앞서 제시한 것같이 자료를 생성한 모형이 일반화 감마 분포이기에 기인한 것으로 생각이 된다.

4.2.2. 로그-왜 정규 분포 추출

두 번째 경우는 로그-왜 정규 분포 기반 주변화 2-부분 모형으로부터 500개의 표본을 추출하여 각각의 분포 기반 주변화 2-부분 모형에 적합 시켰다. Table 4는 총 100번의 시행에 대한 결과를 제시하고 있다.

로그-왜 정규 분포을 이용한 모형 적합의 경우 해당 분포로부터 표본을 추출한 모형이므로 상대 오차 백분율이 작고, 최고사후밀도구간의 포함 확률이 충분히 큼을 알 수 있다. 그와 비슷하게 일반화 감마 분포

Table 4: Comparison of results from two models

Parameter	MTP-GG			MTP-LSN		
(true value)	Mean	CP	PRB	Mean	CP	PRB
β_{10} (0.70)	0.6866	95	1.91	0.6894	**93**	0.22
β_{11} (0.15)	0.1511	95	0.72	0.1473	95	1.79
β_{12} (−0.20)	−0.1956	95	2.21	−0.1902	**94**	4.90
β_{20} (−2.70)	−2.674	95	0.96	−2.6728	97	1.01
β_{21} (0.50)	0.4970	100	0.61	0.4961	98	0.78
β_{22} (−1.00)	−1.0068	96	0.68	−1.0032	96	0.32
β_{23} (0.25)	0.2491	97	0.34	0.2502	97	0.07
σ_{11} (1.20)	1.2284	98	2.37	1.1832	**91**	1.40
σ_{12} (1.00)	0.9810	**93**	1.90	0.9753	**93**	2.47
σ_{22} (1.50)	1.4695	**93**	2.04	1.4713	95	1.91

Generate random number from log-skew normal distribution based marginalized two-part model with sample size of 500. The average of posterior means, coverage probability (CP), and percent relative bias (PRB) are displayed.

를 이용한 모형 또한 상대 오차 백분율이 작음을 알 수 있고, 최고사후밀도구간의 포함 확률은 로그-왜 정규 분포의 결과 보다 큼을 알 수 있다. 이는 상대적으로 사후 표준 편차가 커져서 나타나는 현상이다. 그리고 앞에서 비교한 각각 모형의 모의실험 전체 모수의 최고사후밀도구간의 포함 확률 평균과 상대 오차 백분율 평균은 MTP-GG에서 각각 95.7%과 1.37% 그리고 MTP-LSN에서는 94.9%와 1.49%로 나타났다. 이 결과를 보아 일반화 감마 분포의 베이지안 MCMC 추정은 자료의 생성한 분포에 상관없이 잘 적용됨을 알 수 있다.

4절에서는 일반화 감마 분포 기반 주변화 2-부분 경시적 모형에 대해 3절에서 제시한 베이지안 방법을 통해 모의실험을 진행하였다. 총 2가지 경우의 모의실험을 진행하였으며, 첫 번째 모의실험에서는 임의추출한 표본의 개수가 많아질수록 일반화 감마 분포 기반 주변화 2-부분 경시적 모형의 모수 추정에 대한 2가지 평가 지표 최고 사후 밀도 구간의 포함 확률이 높아지고 상대 오차 백분율이 낮게 나타나 모수 추정이 잘 되었음을 확인할 수 있었다. 두 번째 모의실험, 경쟁 분포인 로그-왜 정규 분포와의 비교에서는 일반화 감마 분포가 로그-왜 정규 분포를 적용했을 때 보다 2개의 평가 지표 최고 사후 밀도 구간의 포함 확률이 전반적으로 높게 나타났고, 상대 오차 백분율이 낮게 나타나 경쟁 분포와 비교에서도 좋은 결과가 나타났다.

결과적으로 3절에서 제안한 일반화 감마 분포 기반 주변화 2-부분 경시적 모형의 베이지안 접근이 2가지 경우의 모의실험에서 좋은 결과가 나타났다. 다음 5절에서는 실증 분석 즉, 베이지안 접근을 이용해 일반화 감마 분포 기반 주변화 2-부분 경시적 모형을 한국의료패널의 경시적 의료비 자료에 적용해 보고자 한다.

5. 실증 분석

이 절에서는 3절에서 제안한 일반화 감마 분포 기반 주변화 2-부분 경시적 모형의 베이지안 방법으로 한국의료패널 경시적 의료비 자료에 대한 실증 분석을 진행한다. 본 분석은 실제 자료에서도 제안한 방법이 잘 적용되는지 확인해보고, 더 높은 차원의 임의효과 즉, 기울기를 추가한 임의효과 구조에서도 잘 추정되는지 검증해본다.

5.1. 한국의료패널자료 설명

한국의료패널은 한국보건사회연구원과 국민건강보험공단이 공동으로 보건 의료비용과 의료비 지출 수준의 변화를 파악하는 패널조사이며, 의료이용 형태와 의료비 지출 규모에 대한 정보뿐만 아니라 의료이용 및

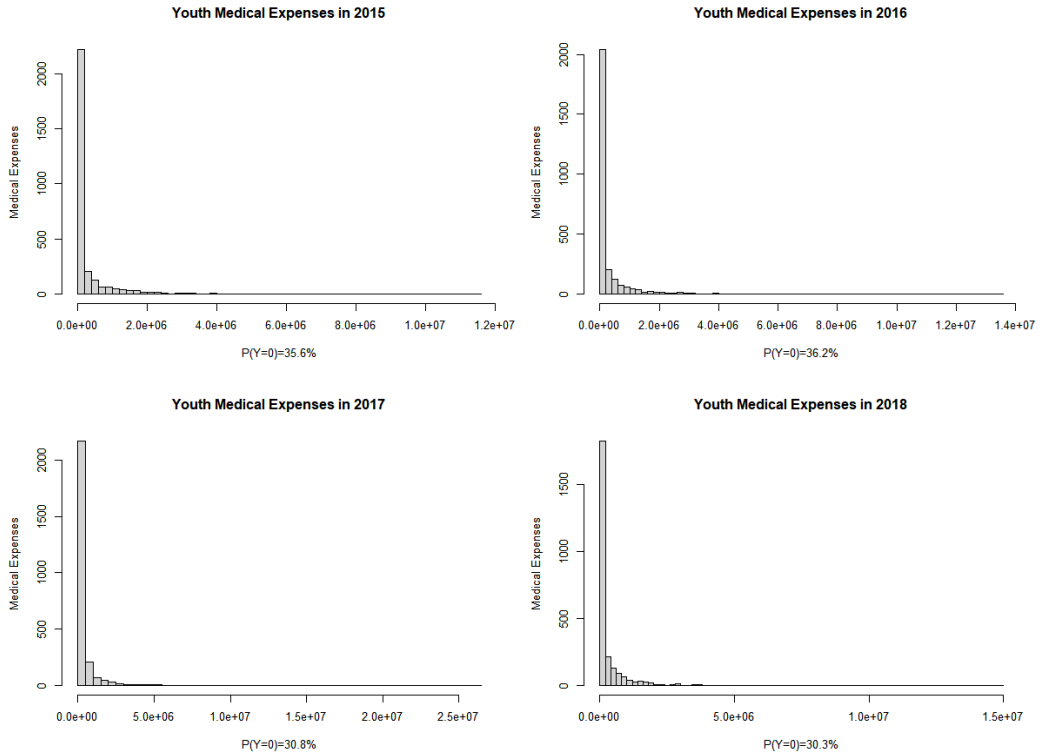

Figure 1: *Histogram of medical expenses for subjects aged 19 to 34.*

의료비 지출에 영향을 미치는 요인들을 포괄하여 심층적으로 분석할 수 있는 패널 자료를 구축하는데 목적을 둔다. 실증분석에서는 한국의료패널 자료에서 2015년부터 2018년도까지 4년간 2015년도 기준 만 19세부터 만 34세까지 청년층 의료비에 대한 경시적 자료분석을 하고자 한다. 또한 중장년층의 의료비를 분석한 논문 '세대별 의료비 지출에 영향을 미치는 요인분석' (황연희, 2011)에서 사용된 변수를 참고하여 해당 변수들이 청년층 의료비 지출에 어떠한 영향을 미치는지 일반화 감마 분포 기반 주변화 2-부분 경시적 모형과 베이지안 방법을 통해 파악해 보고자 한다. 본 연구는 한국보건사회연구원과 국민건강보험공단이 공동으로 주관하는 한국의료패널 2008년~2018년 연간 데이터(Version 1.7.2)를 활용하였다.

본 논문의 분석에서 사용되는 반응 변수는 한국의료패널 자료에서 2015년도부터 2018년도까지 개인지출 의료비(I_MEDICAL_EXP1)을 사용하며 해당 의료비는 보건 의료 서비스에서 응급의료비, 입원의료비 그리고 외래의료비와 의약품에서 응급처방 약 값, 입원처방 약 값 그리고 외래처방 약 값을 모두 합친 금액이다. 앞서 설명한 분석 목표에 맞게 분석 대상인 청년층을 가리기 위해서 2015년도 조사일 기준 만 19세부터 만 34세까지 분석대상을 선정하였다. 결과적으로 선정된 청년층의 수는 2015년도 기준 2,936명이고 해당 청년들을 2016, 2017 그리고 2018년도까지 추적조사하여 2018년도 기준 약 86.2%의 표본이 유지되었다. 추적조사 중에 중단된 자료들은 임의결측(missing at random; MAR)이며, 중단되기 전까지 자료들을 포함하여 분석을 진행하였다.

Figure 1은 청년층 의료비에 대한 연도별 히스토그램이다. 히스토그램에서는 2015년도부터 2018년도까지 의료비가 0원인 값이 각 연도별 자료에서 30%대로 존재하며, 오른쪽으로 심하게 치우쳐진 분포를 띄고

Table 5: Summary statistics of medical expenses

Medical expenses	Year			
	2015	2016	2017	2018
$P(Y = 0)$	35.6%	36.2%	30.8%	30.3%
$E(Y)$	286,476	305,979	324,023	322,278
$\sigma(Y)$	718,961	855,053	1,407,228	853,606
Med(Y)	28,450	29,700	43,200	48,900
Max(Y)	11,519,410	13,459,860	26,263,290	14,906,696

있다. 청년층 의료비를 더 수치적으로 파악하기 위해 Table 5로 요약통계량을 나타내었다. 그리고 Table 5는 청년층 의료비에 대한 연도별 요약통계량 표이다. 청년층의 의료비에서 0원의 비율은 2015년 기준 35.6%에서 2018년 30.3%까지 연도가 지날수록 감소하는 경향이 있어 청년층 의료비용에 대한 발생이 늘어나는 것으로 볼 수 있으며, 의료비의 평균값 280,000원대부터 320,000원대까지 그리고 중앙값은 28,000원대부터 48,000원대까지 상승하는 것을 확인할 수 있다. 또한 2017년에 의료비의 최댓값이 26,263,290원으로 다른 연도에 비해 2배 이상이 나와 표준 편차의 값이 2017년에 가장 큰 것으로 나타났다. 청년층 의료비 자료에서 의료비용 0원의 비율이 높은 것과 분포에서 치우쳐진 정도 그리고 추적 조사 중 절단된 자료들의 특징을 보았을 때, 베이지안 방법을 활용한 일반화 감마 분포 기반 주변화 2-부분 경시적 모형 적합이 타당해 보인다.

다음으로 설명 변수는 성별(gender), 경제활동 여부(work), 교육수준(edu), 거주 지역(region), 담배 여부(cig), 음주 정도(drink), 만성질환 개수(number of diseases), 나이(age) 그리고 총 가구 소득(total income)으로 분석에 사용되는 모든 설명 변수들에 대한 결측치는 해당 객체를 삭제하였다. 설명 변수에서 연속형 변수인 총 가구 소득은 로그 변환을 통해 분포의 대칭성을 맞춰 주었고 과도하게 많은 범주를 가지고 있는 범주형 변수인 교육수준, 거주 지역 그리고 음주 정도를 재범주화하여 범주를 축소시켰다. 교육수준은 9개의 범주에서 무학부터 중학교 3학년까지 중졸 이하(middle school), 고등학교 1학년에서 3학년까지는 고졸(high school), 대학교 1학년부터 박사까지는 대졸 이상(over-college)으로 3개의 범주로, 거주 지역에서는 17개의 범주에서 서울, 경기도 거주는 수도권(capital) 그리고 그 외는 비수도권(non-capital) 총 2개로 범주로, 마지막으로 음주 정도는 8개의 범주에서 마시지 않음(non-drinker), 월 3회 이하(under 3/month) 그리고 3회 초과(over 3/month)로 총 3개로 범주를 축소하였다.

5.2. MTP-GG모형 적합

이 절에서는 위에서 주어진 2015년도부터 2018년도까지 반응 변수 한국의료패널의 청년층 의료비에 대해 9개의 설명 변수 나이, 로그 변환한 총소득, 만성질환 개수, 성별, 경제활동 여부, 교육수준, 거주 지역, 담배 여부, 음주 정도를 베이지안 방법으로 일반화 감마 분포 기반 주변화 2-부분 경시적 모형에 적합하여 설명하고자 한다. 먼저, 경시적 청년층 의료비 자료를 위한 일반화 감마 분포 기반 주변화 2-부분 경시적 모형에서의 변수 선택을 위해 임의효과를 추가한 가장 간단한 혼합효과 모형 즉, 각각의 부모형에 임의절편만 가정한 모형으로 베이지안 모수 추정을 진행하여 베이지안 관점에서 유의한 변수를 추려낸다. 그 다음으로 각각의 부모형에서 베이지안 관점에서 유의한 변수들로 임의효과 구조를 다양하게 하여 모형을 추정한 후 WAIC (Watanabe-Akaike information criteria) 통계량으로 최종 모형을 선정한다. 마지막으로 선정된 최종 모형에 대한 MCMC 알고리즘의 추적 그림(trace plot)을 통해 수렴 여부를 확인한다. 최종적으로 분석에 필요한 베이지안 MCMC은 4개의 체인에 각 5,000번을 반복 후 처음 2,500개를 제외한 MCMC 표본들 즉, 총 10,000개의 표본에 대해서 모수 추정을 진행하였다.

Table 6: Data analysis result table of MTP-GG with random intercepts for medical expenses

Random intercepts model	1^{st} sub model	2^{nd} sub model
Intercept	−3.12 (0.82) [−4.71,−1.50]*	8.42 (0.52) [7.41,9.42]*
Gender (female)	1.37 (0.10) [1.17,1.57]*	0.77 (0.07) [0.64,0.90]*
Age	0.07 (0.01) [0.05,0.09]*	0.06 (0.01) [0.05,0.07]*
Log (total income)	0.09 (0.06) [−0.03,0.21]	0.12 (0.04) [0.04,0.20]*
Region (non-capital)	−0.27 (0.09) [−0.45,−0.09]*	−0.12 (0.06) [−0.23,−0.00]*
Number of diseases	1.30 (0.36) [0.61,1.99]*	0.69 (0.15) [0.40,1.00]*
Edu (high)	0.75 (0.56) [−0.34,1.84]	0.67 (0.35) [−0.03,1.36]
Edu (over-college)	0.68 (0.54) [−0.35,1.79]	0.60 (0.35) [−0.11,1.26]
Work (jobless)	0.19 (0.08) [0.04,0.35]*	0.10 (0.05) [0.00,0.20]*
Drink (under 3/month)	0.39 (0.08) [0.23,0.55]*	0.10 (0.05) [−0.01.−0.20]
Drink (over 3/month)	0.47 (0.10) [0.28,0.67]*	0.05 (0.06) [−0.07,0.18]
Cig (smoker)	0.18 (0.11) [−0.03,0.39]	0.07 (0.07) [−0.09,0.21]
σ_{11}	3.73 (0.24) [3.26,4.18]*	
σ_{12}	1.95 (0.11) [1.74,2.17]*	
σ_{22}	1.4 (0.08) [1.25,1.55]*	
κ	−0.29 (0.04) [−0.36,−0.20]*	
σ^2	1.97 (0.04) [1.88,2.06]*	

* indicates the 95% credible interval does not include 0.

5.2.1. 사전 분포

한국의료패널의 청년층 의료비에 대한 일반화 감마 분포 기반 주변화 2-부분 모형의 사후 분포를 추정하기 위한 사전 분포는 다음과 같이 가정한다.

$$\begin{aligned}
\boldsymbol{\beta}_1 &\sim N\left(\mathbf{0}, 100 \times \boldsymbol{I}_{p_{x1}}\right), \\
\boldsymbol{\beta}_2 &\sim N\left(\mathbf{0}, 100 \times \boldsymbol{I}_{p_{x2}}\right), \\
\kappa &\sim U\left(-10, 10\right), \\
\sigma^2 &\sim \text{Inv.Gamma}\left(0.01, 0.01\right), \\
\boldsymbol{b}_i \mid \boldsymbol{\Sigma} &\sim \text{MVN}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \\
\boldsymbol{\Sigma} &\sim \text{Inv.Wishert}\left(\dim\left(Z\right), I_{p_Z}\right),
\end{aligned}$$

여기서 $Z = [z_{1ij} \vdots z_{2ij}]$이고 I_{p_Z}는 Z의 차원만큼 가지는 단위 행렬이다.

5.2.2. 변수 선택

Table 6는 각각의 부모형에 임의절편만 가정한 일반화 감마 분포 기반 주변화 2-부분 경시적 모형을 베이지안 MCMC 방법으로 모수를 추정한 결과이다.

여기서 $*$는 최고사후밀도구간에 0이 포함되어 있지 않는 것을 의미하며, $(\cdot)$은 사후 분포의 표준 편차 그리고 $[\cdot]$은 95% 최고사후확률밀도 구간이다. 베이지안 방법을 통한 모수 추정 결과에 의하면 첫 번째 부모형에서는 성별, 나이, 거주 지역, 만성질환 개수, 경제활동 여부 그리고 음주 정도가 베이지안 관점에서 유의하다고 판단되었고 두 번째 부모형에서는 성별, 나이, 로그 변환한 총 가구 소득, 거주 지역, 만성질환 개수 그리고 경제활동 여부 변수가 유의하게 나타났다. 또한 임의절편 벡터의 분산-공분산과 일반화 감마

Table 7: Random effects structure

Model comparison		1^{st} sub model	2^{nd} sub model	WAIC
Random intercepts	model 1.	~ 1	~ 1	205739.7
Random intercepts & slope	model 2.	~ 1	~ 1 + gender	205747.0
	model 3.	~ 1	~ 1 + work	**205736.6**
	model 4.	~ 1	~ 1 + region	205740.9

Table 8: Final model parameter estimation table

Final model	1^{st} sub model	2^{nd} sub model
Intercept	−1.42 (0.30) [−1.99,−0.85]*	9.39 (0.36) [8.65,10.07]*
Gender (female)	1.43 (0.09) [1.25,1.61]*	0.79 (0.06) [0.68,0.90]*
Age	0.06 (0.01) [0.04,0.08]*	0.06 (0.01) [0.05,0.07]*
Log (total income)		0.09 (0.03) [0.02,0.15]*
Region (non-capital)	−0.28 (0.09) [−0.46,−0.10]*	−0.13 (0.06) [−0.25,−0.02]*
Number of diseases	1.28 (0.35) [0.61,1.99]*	0.69 (0.15) [0.39,0.99]*
Work (jobless)	0.17 (0.08) [0.02,0.34]*	0.09 (0.06) [−0.02,0.21]
Drink (under 3/month)	0.30 (0.07) [0.16,0.44]*	
Drink (over 3/month)	0.39 (0.09) [0.22,0.56]*	
σ_{11}	3.73 (0.24) [3.25,4.21]*	
σ_{12}	1.95 (0.13) [1.70,2.19]*	
σ_{13}	0.01 (0.14) [−0.26,0.28]	
σ_{22}	1.39 (0.10) [1.19,1.58]*	
σ_{23}	−0.08 (0.08) [−0.24,0.08]	
σ_{33}	0.22 (0.09) [0.07,0.39]*	
κ	−0.28 (0.04) [−0.37,−0.20]*	
σ^2	1.95 (0.05) [1.86,2.04]*	

* indicates the 95% credible interval does not include 0.

분포의 모수들 또한 유의하게 나타났다. 베이지안 관점에서 유의하다고 판단되는 첫 번째 부모형과 두 번째 부모형의 변수들로 다음 절에서는 임의효과 구조를 다양하게 하여 모형 적합을 시도해 보고자 한다.

5.2.3. 최종 모형

위 모형 적합 결과를 토대로 다양한 임의효과 구조를 적합 시켜 WAIC를 통해 최종 모형 선택을 진행한다. 통계량 WAIC는 Watanabe와 Opper (2010)에 의해 제안되었으며, 다음과 같다.

$$\text{WAIC} = -2\left(\text{lppd} - p_{\text{WAIC}}\right),$$

여기서 $\text{lppd} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \log E_{\boldsymbol{\theta}}[p(y_{ij}|\boldsymbol{\theta})]$이고 $p_{\text{WAIC}} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \text{Var}_{\boldsymbol{\theta}}[\log(p(y_{ij}|\boldsymbol{\theta}))]$이며, $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{b}_i, \boldsymbol{\Sigma}, \kappa, \sigma)$이다.

결과적으로 비교할 모형은 임의절편만 가정한 모형, 임의절편과 성별 기울기를 가정한 모형, 임의절편과 경제활동 여부 기울기를 가정한 모형 그리고 임의효과 절편에 거주 지역 기울기를 가정한 모형으로, 총 4가지 모형을 적합한 WAIC 결과는 Table 7과 같다.

WAIC를 통해 모형 선택 결과, 첫 번째 부모형에 임의절편과 두 번째 부모형에 임의절편 및 경제활동 여부의 기울기를 가정한 모형이 4개의 모형 중에서 WAIC가 205736.6으로 가장 낮게 나타났다. 최종적으로 선택된 모형의 모수 추정 결과는 Table 8와 같다.

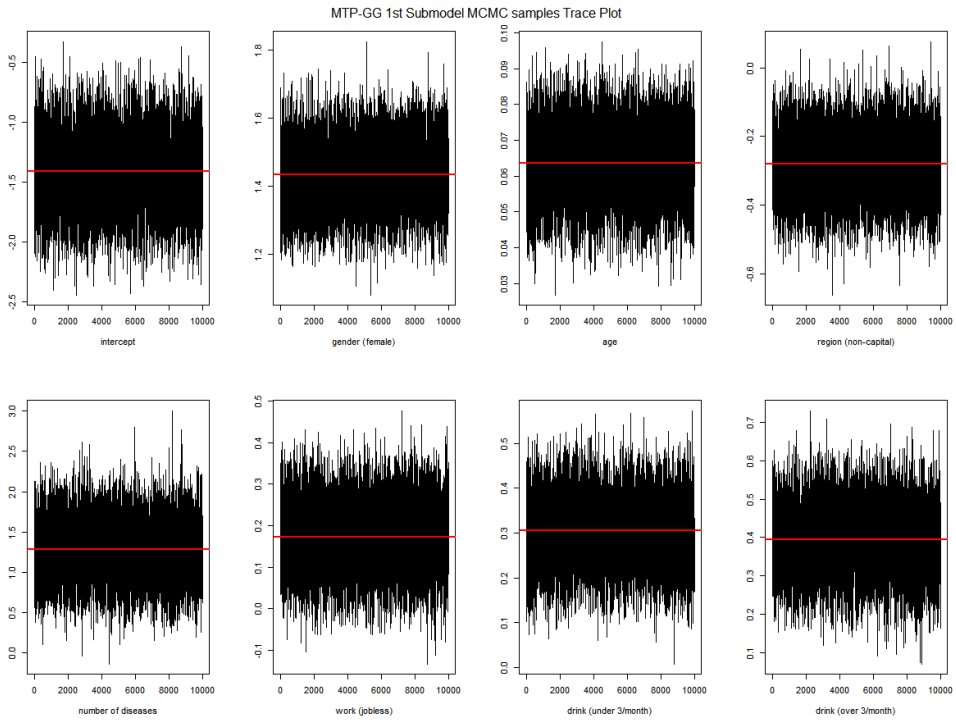

Figure 2: *Trace plot of 1st sub model parameters for checking MCMC algorithm convergence.*

첫 번째 부모형에 임의절편와 두 번째 부모형에 임의절편 및 경제활동 여부의 기울기를 추가한 최종 모형의 모수를 추정한 결과, 두 번째 부모형에서 경제활동 여부 그리고 임의효과 분산-공분산에서 σ_{13}와 σ_{23}이 95% 사후확률밀도구간에 0을 포함하여 베이지안 관점에서 유의하지 않게 나타났다.

5.2.4. 수렴 진단

최종 모형인 첫 번째 부모형에 임의절편과 두 번째 부모형에 임의절편 및 경제활동 여부의 기울기를 가정한 모형의 베이지안 MCMC 알고리즘에 대한 수렴(convergence) 여부는 추적 그림(trace plot)으로 확인한다. 추적 그림에서 MCMC 표본들이 특정한 패턴을 보이지 않고 잘 섞여 있으면, MCMC 알고리즘이 잘 수렴한다고 판단할 수 있다. 최종 모형에서의 추적 그림은 Figure 2와 같으며, 빨간 선은 각 모수에 대한 사후 분포의 평균을 의미한다.

추적 Figure 2은 최종 모형에서 첫 번째 부모형의 모수의 베이지안 MCMC 표본들이 평균을 중심으로 특정 패턴을 보이지 않고 표본들이 잘 섞여 있음을 확인할 수 있어, MCMC 알고리즘이 잘 수렴했다고 할 수 있다.

5절에서는 일반화 감마 분포 기반 주변화 2-부분 경시적 모형과 3절에서 제안한 베이지안 방법을 통해 2015년도부터 2018년도까지 한국의료패널에서 만 19세부터 만 34세 청년층 의료비에 대한 실증분석을 진행하였다. 많은 범주를 가지는 범주형 변수 교육수준, 거주 지역 그리고 음주 정도를 재범주화를 통해 범주를 축소시키고, 총 가구 소득에 대해 로그 변환을 하는 전처리 과정을 통해 총 9개의 설명 변수로 모형 적합을 실시하였다. 각각의 부모형에 임의절편만 가정한 모형 적합 결과, 첫 번째 부모형에서는 성별, 나이, 거주 지

역, 만성질환 개수, 경제활동 여부 그리고 음주 정도가 두 번째 부모형에서는 성별, 나이, 로그 변환한 총 가구 소득, 거주 지역, 만성질환 개수 그리고 경제활동 여부가 베이지안 관점에서 유의한 변수로 나타났다. 최종적으로 선정된 변수들을 통해 다양한 임의효과 구조를 가지는 4가지 모형 간의 비교 결과, 첫 번째 부모형에 임의절편과 두 번째 부모형에 임의절편 및 경제활동 여부의 기울기를 가정한 모형이 WAIC가 가장 낮게 나타나 한국의료패널의 청년층 의료비에 가장 적합한 모형으로 선택되었고, MCMC 표본들에 대한 추적 그림을 통해 해당 모형의 MCMC 알고리즘이 잘 수렴된 것을 확인할 수 있었다.

결과적으로 실제 자료인 한국의료패널 청년층 의료비 자료에 대해 3절에서 제안한 일반화 감마 분포 기반 주변화 2-부분 경시적 모형의 베이지안 모수 추정 방법이 각각의 부모형에 임의절편만 가정한 가장 간단한 혼합효과 모형부터 두 번째 부모형에 임의기울기를 추가한 3차원의 임의효과 구조를 가진 혼합효과 모형까지 모수 추정과 수렴이 잘 되었음을 확인할 수 있었다.

6. 결론

본 논문에서는 영과잉 반 연속 자료들을 직관적으로 분석할 수 있는 2-부분 모형을 소개하였다. 더 나아가 2-부분 모형을 경시적 자료로 확장시킨 조건부, 주변화 2-부분 경시적 모형에 대한 문헌 연구를 하였다. 그 중 자료에 대한 주변 해석이 가능한 주변화 2-부분 모형에서 자료를 포괄적으로 설명할 수 있는 일반화 감마 분포를 연속형 부모형에 적용하여 베이지안 MCMC 방법을 통한 모수 추정을 제안하였다. 제안한 베이지안 방법은 표본크기가 커짐에 따라 그 추정이 잘 작동함을 모의실험을 통하여 보였다. 그리고 또 다른 모의실험에서 일반화 감마 분포와 로그-왜 정규 분포를 연속형 부모형에 적용하여 그 추정치들의 비교를 통해 일반화 감마 분포를 이용한 주변화 2-부분 모형이 로그-왜 정규 분포와 비교해서도 그 성능이 좋음을 알 수 있었다.

제안된 베이지안 일반화 감마 분포 기반 주변화 2-부분 모형을 이용하여 한국의료패널 청년층 의료비 자료에 대한 실증분석을 진행하였다. 각 부모형에 임의효과 절편만 가정한 모형 추정 결과, 베이지안 관점으로 첫 번째 부모형에서는 성별, 나이, 수도권 거주 여부, 만성질환 개수, 경제활동 여부 그리고 음주 정도 변수가 유의하게 나타났고 두 번째 부모형에서는 성별, 나이, 로그 변환 한 총 가구 소득, 수도권 거주 여부, 만성질환 개수 그리고 경제활동 여부 변수가 유의하게 나타났다. 해당 변수들을 통해 여러 가지 임의효과 구조를 가정하여 모형 적합을 시도한 결과, 첫 번째 부모형에 임의절편을 가정하고 두 번째 부모형에 임의절편과 경제활동 여부의 기울기를 가정한 모형의 WAIC가 가장 적게 나타나 한국의료패널 청년층 의료비에 가장 적합한 모형으로 나타났다. 결과적으로 제안한 베이지안의 추정 방법이 임의효과의 차원이 높아짐에 따라 모수 추정에 실패하는 빈도주의적 방법과는 다르게 3차원의 임의효과의 모수 추정 및 수렴이 잘 되었음을 알 수 있었다.

References

Duan N, Manning WG, Morris CN, and Newhouse JP (1983). A comparison of alternative models for the demand for medical care, *Journal of Business & Economic Statistics*, **1**, 115–126.

Jaffa MA, Gebregziabher M, Garrett SM, Luttrell DK, Lipson KE, Luttrell LM, and Jaffa AA (2018). Analysis of longitudinal semicontinuous data using marginalized two-part model, *Journal of Translational Medicine*, **16**, 1–15.

Manning WG, Basu A, and Mullahy J (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data, *Journal of Health Economics*, **24**, 465–488.

Min Y and Agresti A (2002). Modeling nonnegative data with clumping at zero: A survey.

Olsen MK and Schafer JL (2001). A two-part random-effects model for semicontinuous longitudinal data, *Journal of the American Statistical Association*, **96**, 730–745.

Smith VA, Maciejewski ML, and Olsen MK (2018). Modeling semicontinuous longitudinal expenditures: A practical guide, *Health Services Research*, **53**, 3125–3147.

Smith VA, Neelon B, Preisser JS, and Maciejewski ML (2017). A marginalized two-part model for longitudinal semicontinuous data, *Statistical Methods in Medical Research*, **26**, 1949–1968.

Smith VA, Preisser JS, Neelon B, and Maciejewski ML (2014). A marginalized two-part model for semicontinuous data, *Statistics in Medicine*, **33**, 4891–4903.

Su L, Tom BD, and Farewell VT (2009). Bias in 2-part mixed models for longitudinal semicontinuous data, *Biostatistics*, **10**, 374–389.

Voronca DC, Gebregziabher M, Durkalski VL, Liu L, and Egede LE (2015). Marginalized two part models for generalized gamma family of distributions, Cornell University Library, Available from: https://*arXiv preprint arXiv:1511.05629*

Watanabe S and Opper M (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research*, **11**, 3571–3594.

황연희 (2011). 세대별 의료비 지출에 영향을 미치는 요인 분석.

Received December 28, 2022; Revised February 24, 2023; Accepted March 8, 2023

일반화 감마 분포 기반 베이지안 주변화 2-부분 혼합 모형

권용태[a], 이근백[1,a]

[a]성균관대학교 통계학과

요 약

영과잉 반 연속 자료는 0의 값이 많은 연속형 자료를 의미하며, 그 예로는 의료 비용, 음주량 그리고 강수량 등이 있다. 이러한 자료를 분석하기 위한 2-부분 모형이 있으며, 이 모형은 자료를 0인지 아닌지를 판단하는 이항 부모형과 0보다 큰 자료들에 대한 연속형 부모형으로 구성되어 있다. 경시적 영과잉 반 연속 자료 분석을 위한 모형은 2-부분 모형을 확장한 조건부 2-부분 모형과 주변화 2-부분 모형이 있다. 본 논문에서는 영과잉 반 연속 자료 분석을 위한 2-부분 모형과 경시적 영과잉 반 연속 자료 분석을 위한 주변화 2-부분 모형을 고찰한다. 그리고 빈도주의 관점에서 주변화 2-부분 모형의 모수 추정 및 수렴실패의 문제를 해결하기 위하여 베이지안 모수추정을 제안한다. 제안된 방법의 성능을 비교하기 위하여 모의실험을 수행하고, 실제 자료인 한국의료패널의 청년층 의료비 분석을 위하여 제안된 추정 방법을 이용한다.

주요용어: 2-부분 모형, 경시적 영과잉 반 연속 자료, 일반화 감마 분포, 베이지안, 주변화 2-부분 경시적 모형

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2022R1A2C1002752). 이 논문은 권용태의 석사논문의 일부를 발췌하였음.

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: keunbaik@skku.edu