The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.2.123
33(2), 123–136

Joint analysis of binary and continuous data using skewed logit model in developmental toxicity studies

Yeong-hwa Kim[a] · Beom Seuk Hwang[a,1]

[a]Department of Applied Statistics, Chung-Ang University

(Received September 17, 2019; Revised October 30, 2019; Accepted November 5, 2019)

Abstract

It is common to encounter correlated multiple outcomes measured on the same subject in various research fields. In developmental toxicity studies, presence of malformed pups and fetal weight are measured on the pregnant dams exposed to different levels of a toxic substance. Joint analysis of such two outcomes can result in more efficient inferences than separate models for each outcome. Most methods for joint modeling assume a normal distribution as random effects. However, in developmental toxicity studies, the response distributions may change irregularly in location and shape as the level of toxic substance changes, which may not be captured by a normal random effects model. Motivated by applications in developmental toxicity studies, we propose a Bayesian joint model for binary and continuous outcomes. In our model, we incorporate a skewed logit model for the binary outcome to allow the response distributions to have flexibly in both symmetric and asymmetric shapes on the toxic levels. We apply our proposed method to data from a developmental toxicity study of diethylhexyl phthalate.

Keywords: Bayesian inference, diethylhexyl phthalate, joint modeling, Markov chain Monte Carlo, skewed logit model

1. 서론

하나의 개체에서 두 개 이상의 측정치가 동시에 관찰되는 경우는 의학, 공학, 사회과학, 자연과학 등 여러 다양한 분야에서 흔히 나타난다. 특히 생명과학 분야에서는 어떤 특정 처리(treatment)의 수준에 따라 서로 연관된 두 개 이상의 변수가 동시에 영향을 받는 경우를 종종 고려한다. 예를 들어, 발달 독성학(developmental toxicity studies)에서는 어떤 독성 물질의 각기 다른 수준에 노출된 임신한 어미 쥐에 대해 쥐의 몸무게, 간의 무게, 자궁의 무게(gravid uterine weight), 죽은 태아(pup)의 숫자, 태아의 기형(malformation) 여부, 태아의 무게(fetal weight) 등 여러가지 다양한 변수들이 어떻게 영향을 받는지를 연구한다. 이때 각 변수들을 독립적인 개별모형으로 분석하게 되면 편향된(biased) 결과를 얻을 수 있기 때문에 두 개 이상의 변수들의 상관성을 고려한 결합모형을 사용해야 한다 (Catalano와 Ryan, 1992; Regan과 Catalano, 1999; Dunson, 2000; McCulloch, 2008). 이때 변수들의 형태는 이산형,

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2019R1C1C1011710).
[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bshwang@cau.ac.kr

명목형, 연속형 등 다양한 형태를 띨 수 있는데, 이와 같이 각기 다른 형태의 자료들을 결합해서 분석할 때 잠재변수(latent variable)를 활용하는 결합모형이 주로 사용되어왔다 (Catalano와 Ryan, 1992; Regan과 Catalano, 1999). Catalano와 Ryan (1992)은 서로 상관관계가 있는 이진수 변수와 연속형 변수에 대한 결합모형을 제시하였는데 이진수 변수에 내포되어 있는 연속형의 잠재변수를 가정하고 그 잠재변수와 연속형 변수가 정규분포를 따르는 랜덤효과(random effect)를 공유함으로써 두 변수의 상관관계를 고려하는 방식으로 두 변수를 분석하였다. Dunson (2000)은 지수족(exponential family) 분포를 가정한 혼합형 자료에 대하여 베이지안 방법을 통해서 일반화된 결합모형을 제시하였다.

잠재변수를 가정한 랜덤효과 방법은 형태가 다른 두 변수에 대한 결합모형의 대표적인 방법으로 사용되어왔지만, 랜덤효과를 주로 정규분포로 가정함으로써 그 한계점을 드러내었다. 연속형 변수가 비대칭의 분포 형태를 가지거나 이진수 자료가 극단적인 결과값(극단적으로 많은 0의 수 또는 1의 수)을 가지는 경우 정규분포 랜덤효과 모형은 자료의 비대칭성을 제대로 설명해주지 못하는 단점을 가지게 된다. 특정 분포를 가정하지 않고 불규칙적인 분포의 형태를 모형화하는 방법으로서 비모수 베이지안(nonparametric Bayesian) 방법이 여러 분야에서 다양하게 사용되어왔다 (Sethuraman, 1994; MacEachern, 1999; De Iorio 등, 2004). 특히 Dunson 등 (2003), Li 등 (2004), Hwang과 Pennell (2014), Hwang과 Pennell (2018)은 비모수 베이지안 방법을 형태가 다른 두 변수들의 결합모형에 적용하여 정규분포의 랜덤효과 모형의 단점을 해결하였다.

비대칭적인 이진수 자료를 모형화하는데 사용되는 또 다른 방법으로 Chen 등 (1999)은 비대칭 연결함수 모형(skewed link model)을 제시하였다. 비대칭 연결함수 모형은 사건이 일어날 확률이 0으로 근접하는 속도와 1로 근접하는 속도를 서로 다르게 설정하여 대표적인 대칭 모형인 로지스틱 회귀모형(logistic regression model)이나 프로빗 회귀모형(probit regression model)이 설명하지 못하는 자료의 불균형성을 설명하였다. Chen 등 (2001)이나 Kim과 Hwang (2019)은 비대칭 로짓 모형(skewed logit model)을 베이지안 추론 방법에 기반하여 모형 분석을 하였다.

본 논문에서는 서로 상관관계가 존재하는 이진수 자료와 연속형 자료의 결합모형에 비대칭 로짓 모형을 도입하여 불균형적인 형태의 두 변수를 분석하려고 한다. 특히 발달 독성학에서 흔히 발생하는 독성 물질의 각기 다른 용량 수준에 따른 두 변수의 변화를 보기 위해 미국 국립 독성학 프로그램(National Toxicology Program; NTP)에서 수행한 실험용 쥐에 대한 독성물질 diethylhexyl phthalate (DEHP)에 대한 데이터를 사용하여 모형의 적합성을 측정하고자 한다. 2장에서는 이진수(binary) 자료와 연속형(continuous) 자료에 대한 기본적인 모형 설정을 설명하고 비대칭 로짓 모형을 도입한 결합모형을 소개한다. 3장에서는 제시된 모형에 대해 베이지안 추론 방법을 소개한다. 자료의 가능도 함수, 변수들의 사전분포(prior distribution)와 사후분포(posterior distribution)를 차례로 소개하고, 이를 구체적으로 계산하기 위한 Markov chain Monte Carlo (MCMC) 방법을 소개한다. 4장에서는 독성물질 DEHP가 임신한 쥐에 어떤 영향을 미치는 지에 대한 연구를 소개하며 비대칭 로짓 모형을 사용한 결합모형이 이 데이터에 어떻게 적용되어 결과를 도출하고 있는지 분석한다. 마지막으로 5장에서는 본 논문을 요약 정리하고 향후 후속 연구의 방향에 대해 논의한다.

2. 이진수 자료와 연속형 자료에 대한 결합모형

2.1. 결합모형의 구성

x_i는 독성 물질의 용량 수준을 나타내며 d개의 이산형의 값을 가진다고 가정한다. $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2})^T$는 독성 물질의 i번째 용량 수준에 노출된 j번째 개체에서 측정된 2×1 벡터로 이루어진 반응변수를 나타낸다 ($i = 1, \ldots, d$, $j = 1, \ldots, n_i$). 이때 Y_{ij1}은 이진수 형태의 변수이고, Y_{ij2}는 연속형 형태의 변수

이며, 두 변수는 서로 연관되어 있다고 가정한다. 예를 들어, 발달 독성학 연구에서 독성 물질의 상이한 수준에 노출된 임신한 어미 쥐가 가진 태아의 기형 여부와 태아의 몸무게를 두 변수로 고려할 수 있을 것이다. 이때 두 변수는 서로 음의 상관관계가 존재한다고 알려져 있다. 이진수 변수를 모형화하기 위해 Albert와 Chib (1993)이 제시한 잠재변수 방법을 이용하면 다음과 같이 표현할 수 있다.

$$y_{ij1} = \begin{cases} 1, & \text{if } y^*_{ij1} \geq 0, \\ 0, & \text{if } y^*_{ij1} < 0. \end{cases}$$

이때 두 변수 사이의 상관관계를 고려하기 위해 y_{ij1}에 대한 잠재변수 y^*_{ij1}과 y_{ij2}는 다음과 같은 결합모형을 구성한다.

$$y^*_{ij1} = \beta_{10} + \beta_{11}x_i + \delta_{i1}z_{ij} + \epsilon_{ij1}, \tag{2.1}$$

$$y_{ij2} = \beta_{20} + \beta_{21}x_i + \delta_{i2}z_{ij} + \epsilon_{ij2}. \tag{2.2}$$

이때 $\boldsymbol{\beta} = (\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21})^T$는 독성 물질 용량 x_i의 고정효과(fixed effect)의 절편과 기울기를 나타내고, z_{ij}는 개체 특정(subject-specific) 랜덤효과로 $z_{ij} \sim G$를 따르며, $(\delta_{i1}, \delta_{i2})$는 랜덤효과에 대한 계수를 나타낸다. 랜덤효과 z_{ij}는 y^*_{ij1}와 y_{ij2}에 공유됨으로써 두 변수 사이의 상관관계를 설명해준다. 그리고 오차항에 대해서는 $\epsilon_{ij1} \sim F_1$, $\epsilon_{ij2} \sim F_2$라고 가정한다 ($i = 1, \ldots, d$, $j = 1, \ldots, n_i$).

2.2. 비대칭 연결 모형의 도입

이진수 변수의 잠재변수와 연속형 변수 사이에 공유되는 랜덤효과 z_{ij}는 일반적으로 정규분포를 따른다고 가정한다 (Dunson 등, 2003). 하지만 독성 물질 DEHP의 경우에서처럼 한쪽으로 치우쳐서 발생하는 불균형적인 이진수 자료의 경우 대칭적인 형태를 띠고 있는 정규분포로는 그 특징을 제대로 반영하지 못한다. 이러한 경우 비대칭 연결함수 모형(asymmetric link function model)을 사용하면 사건이 발생할 확률이 0으로 접근하는 속도와 1로 접근하는 속도를 서로 다르게 설정하여 이진수 자료의 비대칭성을 설명해준다. Chen 등 (1999)이 사용한 비대칭 로짓 모형을 위의 결합모형에 도입하면 z_{ij}의 분포 G는 비대칭 분포의 누적분포함수, 오차항에 대한 F_1과 F_2는 대칭 분포의 누적분포함수를 따른다고 가정할 수 있다. 이때 식 (2.1)과 (2.2)에 있는 δ_{i1}과 δ_{i2}는 i번째 용량 수준에서 측정된 이진수 변수의 잠재변수와 연속형 변수의 비대칭 모수(skewness parameter)를 각각 나타낸다($-\infty < \delta_{ij} < \infty$, $i = 1, \ldots, 5$, $j = 1, 2$).

이진수 변수를 모형화할 때 비대칭 연결 모형을 사용하면 분포 G와 F_1, 그리고 비대칭 모수 δ_{i1}의 값에 따라 여러가지 다른 형태의 모형을 포함하게 된다. 예를 들어, 비대칭 모수(δ_{i1})가 0이거나 G가 퇴화분포(degenerate distribution)를 따른다면, 비대칭 연결 모형은 대칭 연결함수 모형과 같은 형태를 가지게 된다. 즉, F_1을 표준정규분포로 가정하면 프로빗 회귀모형과 같은 형태가 되고, 표준 로지스틱 분포로 가정하면 로지스틱 회귀모형의 형태를 갖게 된다. 실제 데이터 분석에서 분포 G는 다음과 같은 절반 표준정규분포(half-standard normal distribution)를 가정하기도 하고,

$$g(z) = \frac{2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z > 0.$$

다음과 같이 지수분포(exponential distribution)를 가정하여 사용하는 경우도 종종 있다.

$$g(z) = e^{-z}, \quad z > 0.$$

비대칭 모수인 δ_{i1}이 0이 아닌 경우에 F_1 분포가 표준정규분포를 따르면 비대칭 연결 모형은 비대칭 프로빗 모형(skewed probit model)이 되고, F_1 분포가 표준 로지스틱 분포를 따르면 비대칭 로짓 모형이 된다.

이진수 자료에서 가정한 비대칭 분포 G를 식 (2.1)과 (2.2)에서처럼 연속형 자료와 랜덤효과를 통해서 공유하게 되면 연속형 자료의 비대칭적인 특징 역시 설명할 수 있다. 이때 비대칭 모수 δ_{i1}과 δ_{i2}는 모형의 비대칭 정도를 설명해준다. 예를 들어, 비대칭 모수가 0이면 대칭 모형이 되기 때문에 사건이 일어날 확률 (p_{ij})이 0으로 접근하는 속도와 1로 접근하는 속도가 같게 된다. 하지만, 비대칭 모수가 0보다 크면 y_{ij1}의 주변분포(marginal distribution)는 오른쪽으로 치우치게 되고(skewed to the right), 이에 따라 p_{ij}는 1로 접근할 때의 속도가 0일때 보다 더 빨라지게 된다. 반대의 경우에는 y_{ij1}의 주변분포가 왼쪽으로 치우친 분포(skewed to the left)가 되고, p_{ij}는 0으로 접근할 때의 속도가 1일때 보다 더 빠른 성질을 가지게 된다. 이때 모든 i에 대하여 $\delta_{i1} = \delta_1$가 성립하면 분포의 비대칭 정도는 각 용량 수준에 상관없이 동일하다는 것을 의미한다. 하지만, 발달 독성학에서는 독성 물질의 용량 수준이 증가함에 따라 비대칭의 방향이 반대로 변화하는 경우를 흔히 볼 수 있다. 따라서 모형에서 δ_1 대신에 δ_{i1}을 사용함으로써 용량 수준에 따른 상이한 비대칭성을 설명해줄 수 있다. 이는 δ_{i2}에 대해서도 마찬가지이고, 또한 $\delta_{i1} \neq \delta_{i2}$을 가정함으로써 보다 유연한 결합모형을 고려하게 된다.

2.3. 비대칭 로짓 모형을 사용한 결합모형

본 논문에서는 비대칭 연결 모형의 가장 대표적인 형태인 비대칭 로짓 모형을 결합모형에 도입하여 이진수 자료와 연속형 자료의 연관성을 분석하고자 한다. 즉, 다음과 같이 식 (2.1)과 (2.2)에서 랜덤효과 z_{ij}는 절반표준정규분포를 따르고, ϵ_{ij1}는 표준 로지스틱 분포, ϵ_{ij2}는 정규분포를 따른다고 가정한다.

$$g(z_{ij}) = \frac{2}{\sqrt{2\pi}} e^{-\frac{z_{ij}^2}{2}}, \qquad F_1(\epsilon_{ij1}) = \frac{e^{\epsilon_{ij1}}}{1+e^{\epsilon_{ij1}}}, \qquad F_2(\epsilon_{ij2}) = \Phi\left(\frac{\epsilon_{ij2}}{\sigma}\right).$$

이때 $z > 0$이고, $g(\cdot)$은 절반표준정규분포의 확률밀도함수, $F_1(\cdot)$과 $F_2(\cdot)$는 각각 표준 로지스틱 분포와 정규분포의 누적분포함수를 나타낸다. 또한, Φ는 표준정규분포의 누적분포함수이다. 결합모형의 관찰된 데이터 가능도함수(observed data likelihood function)는 다음과 같이 구할 수 있다.

$$\begin{aligned} L\left(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}\right) = \prod_{i=1}^{d} \prod_{j=1}^{n_i} \int_{-\infty}^{\infty} & \left[F_1(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij})\right]^{y_{ij1}} \left[1 - F_1(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij})\right]^{1-y_{ij1}} \\ & \times N\left(y_{ij2}; \beta_{20} + \beta_{21} x_i + \delta_{i2} z_{ij}, \sigma^2\right) g(z_{ij}) dz_{ij}, \end{aligned}$$

여기에서 $N(\mu, \sigma^2)$는 평균 μ와 분산 σ^2을 가지는 정규분포를 나타낸다. 실제 데이터 분석에서는 계산을 용이하게 하기 위해서 잠재변수 z를 고려한 모수 $(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2)$의 완전데이터 가능도함수(complete data likelihood function)를 다음과 같이 구하여 이용한다.

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}) &= \prod_{i=1}^{d} \prod_{j=1}^{n_i} \left[F_1(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij})\right]^{y_{ij1}} \left[1 - F_1(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij})\right]^{1-y_{ij1}} \\ &\quad \times N(y_{ij2}; \beta_{20} + \beta_{21} x_i + \delta_{i2} z_{ij}, \sigma^2) g(z_{ij}) \\ &= \prod_{i=1}^{d} \prod_{j=1}^{n_i} \left[\frac{\exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)}{1 + \exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)}\right]^{y_{ij1}} \left[\frac{1}{1 + \exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)}\right]^{1-y_{ij1}} \\ &\quad \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\left\{y_{ij2} - (\beta_{20} + \beta_{21} x_i + \delta_{i2} z_{ij})\right\}^2\right] \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2} z_{ij}^2}. \end{aligned} \tag{2.3}$$

3. 결합모형에 대한 베이지안 추론

3.1. 사전분포와 사후분포

일반적인 베이지안 추론 방법을 시행하기 위해 미지의 모수들에 대한 사전분포를 선택한 후 가능도함수를 바탕으로 결합사후분포(joint posterior distribution)를 구하고자 한다. Chen 등 (1999)과 Kim과 Hwang (2019)은 비대칭 로짓 모형의 모수 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$와 δ에 대해서 무정보(noninformative) 사전분포를 다음과 같이 고려하였다.

$$p(\boldsymbol{\beta}, \delta) \propto p(\delta).$$

이때 $p(\delta)$는 적절한(proper) 사전분포 또는 부적절한 균등(improper uniform) 사전분포를 가정할 수 있는데, Chen 등 (1999)은 그 사후분포가 항상 적절성(propriety) 이 보장됨을 증명하였다. 본 논문에서는 계산의 용이함과 사후분포의 적절성 등을 고려하여 세 모수 $(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma)$에 대하여 각각 다음과 같이 독립적인 사전분포를 고려하였다.

$$\boldsymbol{\beta} \sim \text{MVN}\left(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta}\right), \qquad \boldsymbol{\delta} \sim \text{MVN}\left(\boldsymbol{\mu_\delta}, \boldsymbol{\Sigma_\delta}\right), \qquad \sigma^2 \sim \text{IG}(a, b). \tag{3.1}$$

여기서 $\text{MVN}(\boldsymbol{\mu}, \Sigma)$는 평균 $\boldsymbol{\mu}$와 공분산행렬 Σ를 가지는 다변량정규분포를 나타내고, $\boldsymbol{\mu_\beta}$와 $\boldsymbol{\mu_\delta}$의 차원은 각각 4×1과 $2d \times 1$이고, $\boldsymbol{\Sigma_\beta}$와 $\boldsymbol{\Sigma_\delta}$의 차원은 각각 4×4과 $2d \times 2d$이다. $\text{IG}(a, b)$는 형태모수(shape parameter) a와 척도모수(scale parameter) b를 가지는 역감마 분포(inverse gamma distribution)를 나타내며, 연속형 변수의 오차항이 취하는 정규분포의 분산에 대한 공액사전분포(conjugate prior distribution)로서 사용되었다. 식 (2.3)에 정의된 완전데이터 가능도함수와 식 (3.1)의 모수에 대한 사전분포를 이용하여 다음과 같이 결합사후분포를 구한다.

$$\begin{aligned}
p\left(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{z} | \boldsymbol{Y}, \boldsymbol{X}\right) &\propto L\left(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}\right) p(\boldsymbol{\beta}) p(\boldsymbol{\delta}) p(\sigma^2) \\
&\propto \prod_{i=1}^{d} \prod_{j=1}^{n_i} \left\{ \left[\frac{\exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)}{1 + \exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)} \right]^{y_{ij1}} \left[\frac{1}{1 + \exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)} \right]^{1 - y_{ij1}} \right. \\
&\quad \left. \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2} \left\{ y_{ij2} - (\beta_{20} + \beta_{21} x_i + \delta_{i2} z_{ij}) \right\}^2 \right] \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2} z_{ij}^2} \right\} \\
&\quad \times \text{MVN}\left(\boldsymbol{\beta}; \boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta}\right) \times \text{MVN}\left(\boldsymbol{\delta}; \boldsymbol{\mu_\delta}, \boldsymbol{\Sigma_\delta}\right) \times \text{IG}(\sigma^2; a, b).
\end{aligned} \tag{3.2}$$

식 (3.2)의 결합사후분포로부터 각 모수와 잠재변수 z에 대한 조건부 사후분포(conditional posterior distribution)를 계산하면 다음과 같다.

$$\begin{aligned}
p(\boldsymbol{\beta}|-) \propto \prod_{i=1}^{d} \prod_{j=1}^{n_i} &\left\{ \left[\frac{\exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)}{1 + \exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)} \right]^{y_{ij1}} \left[\frac{1}{1 + \exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)} \right]^{1 - y_{ij1}} \right. \\
&\left. \times \exp\left[-\frac{1}{2\sigma^2} \left\{ y_{ij2} - (\beta_{20} + \beta_{21} x_i + \delta_{i2} z_{ij}) \right\}^2 \right] \right\} \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu_\beta})^T \Sigma_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu_\beta}) \right\}
\end{aligned} \tag{3.3}$$

$$\begin{aligned}
p(\boldsymbol{\delta}|-) \propto \prod_{i=1}^{d} \prod_{j=1}^{n_i} &\left\{ \left[\frac{\exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)}{1 + \exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)} \right]^{y_{ij1}} \left[\frac{1}{1 + \exp\left(\beta_{10} + \beta_{11} x_i + \delta_{i1} z_{ij}\right)} \right]^{1 - y_{ij1}} \right. \\
&\left. \times \exp\left[-\frac{1}{2\sigma^2} \left\{ y_{ij2} - (\beta_{20} + \beta_{21} x_i + \delta_{i2} z_{ij}) \right\}^2 \right] \right\} \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\delta} - \boldsymbol{\mu_\delta})^T \Sigma_\delta^{-1} (\boldsymbol{\delta} - \boldsymbol{\mu_\delta}) \right\}
\end{aligned} \tag{3.4}$$

$$p(\sigma^2|-) \sim \text{IG}\left(a + \frac{1}{2}\prod_{i=1}^{d} n_i, b + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{n_i}\{y_{ij2} - (\beta_{20} + \beta_{21}x_i + \delta_{i2}z_{ij})\}^2\right) \tag{3.5}$$

$$p(\boldsymbol{z}|-) \propto \prod_{i=1}^{d}\prod_{j=1}^{n_i}\left\{\left[\frac{\exp(\beta_{10} + \beta_{11}x_i + \delta_{i1}z_{ij})}{1 + \exp(\beta_{10} + \beta_{11}x_i + \delta_{i1}z_{ij})}\right]^{y_{ij1}}\left[\frac{1}{1 + \exp(\beta_{10} + \beta_{11}x_i + \delta_{i1}z_{ij})}\right]^{1-y_{ij1}}\right.$$
$$\left.\times \exp\left[-\frac{1}{2\sigma^2}\{y_{ij2} - (\beta_{20} + \beta_{21}x_i + \delta_{i2}z_{ij})\}^2\right]\exp\left(-\frac{1}{2}z_{ij}^2\right)\right\}. \tag{3.6}$$

3.2. Markov chain Monte Carlo

미지의 모수를 추론하기 위해 3.1장에서 계산된 조건부 사후분포를 바탕으로 MCMC 방법을 사용하고자 한다. 식 (3.5)와 같이 특정 분포의 형태를 가지는 조건부 사후분포의 경우 깁스 샘플러(Gibbs sampler) (Casella와 George, 1992) 방법을 사용할 수 있으나, 특정 분포의 형태를 띠고 있지 않은 경우는 깁스 샘플러 방법을 사용할 수가 없다. 이를 해결하기 위해 Chen 등 (1999)은 결합 사후분포가 로그 오목함(log-concave)을 보인 후 적응 기각 알고리즘(adaptive rejection algorithm)을 사용하였고, Kim과 Hwang (2019)은 결합 사후분포의 오목성에 상관없이 사용할 수 있는 보다 일반적인 메트로폴리스-해스팅스(Metropolis-Hastings; M-H) 알고리즘 (Chib과 Greenberg, 1995)을 적용하였다. 본 논문에서는 모수 $\boldsymbol{\beta}$와 $\boldsymbol{\delta}$에 대해서 Kim과 Hwang (2019)에서 사용한 분산조정 메트로폴리스(adaptive Metropolis) 알고리즘 (Haario 등, 2005)을 적용하였다. 즉, 샘플의 채택 비율(acceptance rate)을 조정하여 수렴 상태를 개선하기 위해 메트로폴리스 알고리즘 각 단계에서 사용하는 제안분포(proposal distribution)의 분산을 샘플들의 경험적(empirical) 분산과 조정계수를 이용하여 변화시켰다. 예를 들어, $t+1$ 시점의 모수 $\theta^{(t+1)}$을 업데이트하기 위해서 다음과 같이 정규분포에서 샘플링된 후보값(candidate value) θ^*를 고려한다.

$$\theta^* \sim N\left(\theta^{(t)}, V^{(t)}\right),$$
$$V^{(t)} = \begin{cases} V^{(0)}, & \text{if } t \le t_0, \\ s\text{Var}\left(\theta^{(0)}, \ldots, \theta^{(t)}\right) + s\varepsilon, & \text{if } t > t_0, \end{cases}$$

여기서 $V^{(0)}$는 모수 θ에 대한 제안분포의 초기 분산값이고, ε은 분산이 0이 되는 것을 막기 위한 아주 작은 상수값을 나타낸다. 또한, s는 후보값의 채택 비율을 조정해주는 조정 계수를 나타내는데, Gelman 등 (2014)은 d-차원에 대해 $2.4/\sqrt{d}$의 조정계수를 사용했을 때 최적의 채택 비율을 얻을 수 있다고 제시하였다. 이때 대략적인 최적의 채택 비율은 $d = 1$일때 0.44이고, 고차원($d > 5$)으로 갈수록 그 비율은 줄어들어 대략 0.23 정도일때가 최적이 된다고 제안하였다. 또한 잠재변수인 $\boldsymbol{z}$를 업데이트하기 위해서는 M-H 방법 중 특별한 경우인 독립 샘플러(independence sampler)를 사용한다. 즉 샘플의 수렴 여부를 개선하기 위해 메트로폴리스 알고리즘의 매 단계마다 제안분포로서 절반표준정규분포를 독립적으로 사용하였다. 식 (3.3)–(3.6)에 나타난 조건부 사후분포를 바탕으로 다음과 같은 단계를 거쳐 깁스 샘플러 내의 메트로폴리스-해스팅스 알고리즘(Metropolis-Hastings-within-Gibbs sampler)을 시행한다.

Step 1: 모수의 초기값 $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\delta}^{(0)}, \sigma^{2(0)})$를 설정하고, 잠재변수의 초기값 $z_{ij}^{(0)}$ $(i = 1, \ldots, d,\ j = 1, \ldots, n_i)$을 절반표준정규분포에서 생성한다.

Step 2: t시점의 값이 $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\delta}^{(t)}, \sigma^{2(t)}, \boldsymbol{z}^{(t)})$로 주어졌을 때, $t+1$ 시점의 값을 식 (3.3)–(3.6)을 이용하여 다음과 같이 순차적으로 업데이트한다.

- 식 (3.3)로부터 분산조정 M-H 방법을 사용하여 $\boldsymbol{\beta}$를 업데이트한다:
  $p(\boldsymbol{\beta}^{(t+1)}|\boldsymbol{\delta}^{(t)}, \sigma^{2(t)}, \boldsymbol{z}^{(t)}, \boldsymbol{Y}, \boldsymbol{X})$
- 식 (3.4)로부터 분산조정 M-H 방법을 사용하여 $\boldsymbol{\delta}$를 업데이트한다:
  $p(\boldsymbol{\delta}^{(t+1)}|\boldsymbol{\beta}^{(t+1)}, \sigma^{2(t)}, \boldsymbol{z}^{(t)}, \boldsymbol{Y}, \boldsymbol{X})$
- 식 (3.5)의 역감마 분포로부터 깁스 샘플러 방법을 사용하여 σ^2를 업데이트한다:
  $p(\sigma^{2(t+1)}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \boldsymbol{z}^{(t)}, \boldsymbol{Y}, \boldsymbol{X})$
- 식 (3.6)로부터 독립 샘플러 방법을 사용하여 잠재변수 $\boldsymbol{z}$를 업데이트한다:
  $p(\boldsymbol{z}^{(t+1)}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}, \sigma^{2(t+1)}, \boldsymbol{Y}, \boldsymbol{X})$

Step 3: Step 2로 돌아가서 수렴할 때까지 반복한다.

여러 다양한 값들을 초기값으로 설정해서 위의 단계를 거쳐 MCMC를 실행한 후에 베이지안 진단법(Bayesian diagnostics)을 사용하여 알고리즘의 수렴 여부를 확인한다. 본 논문에서는 가장 일반적인 방법인 trace plot과 Gelman과 Rubin의 Potential Scale Reduction Factor, $\hat{R}$ (Gelman 등, 2014)을 사용하여 수렴이 잘 이루어졌는지 확인하였다.

3.3. 사후 예측분포

본 논문에서 제시한 결합모형에서 사후 예측분포(posterior predictive distribution)는 다음과 같이 구할 수 있다. 먼저 이진수 자료에 대해서 $y_{ij} = 1$일 사후 예측확률을 계산하면 다음과 같다.

$$\begin{aligned} p(y_{ij1} = 1|\boldsymbol{y}) &= \iiiint F_1(\beta_{10} + \beta_{11}x_i + \delta_{i1}z_{ij})p(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, z_{ij}|\boldsymbol{y})dz_{ij}d\boldsymbol{\beta}d\boldsymbol{\delta}d\sigma^2 \\ &= \iiiint \frac{\exp(\beta_{10} + \beta_{11}x_i + \delta_{i1}z_{ij})}{1 + \exp(\beta_{10} + \beta_{11}x_i + \delta_{i1}z_{ij})} p(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, z_{ij}|\boldsymbol{y})dz_{ij}d\boldsymbol{\beta}d\boldsymbol{\delta}d\sigma^2. \end{aligned} \tag{3.7}$$

연속형 자료에 대한 사후 예측분포는 다음과 같이 계산할 수 있다.

$$\begin{aligned} p(y_{ij2}|\boldsymbol{y}) &= \iiiint p(y_{ij2}|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, z_{ij})p(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, z_{ij}|\boldsymbol{y})dz_{ij}d\boldsymbol{\beta}d\boldsymbol{\delta}d\sigma^2 \\ &= \iiiint \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\{y_{ij2} - (\beta_{20} + \beta_{21}x_i + \delta_{i2}z_{ij})\}^2\right] \\ &\quad \times p(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, z_{ij}|\boldsymbol{y})dz_{ij}d\boldsymbol{\beta}d\boldsymbol{\delta}d\sigma^2. \end{aligned} \tag{3.8}$$

식 (3.7)과 (3.8)에서 얻어진 이진수 자료와 연속형 자료에 대한 사후 예측분포는 직접 적분을 사용하여 계산하기가 쉽지 않다. 따라서 여기서는 몬테카를로 적분(Monte Carlo integration) 방법을 활용하여 근사적인 값을 계산해낸다. 즉, 3.2절의 MCMC 과정에 사후 예측분포의 몬테카를로 적분 계산을 포함시켜 그 과정을 수행한다.

4. 자료의 분석

4.1. 자료의 탐색

이 장에서는 2장과 3장에서 제안한 비대칭 로짓 모형을 사용한 베이지안 결합모형을 실제 발달 독성학 데이터에 적용하여 분석하고자 한다. 본 연구에서 사용한 자료는 미국 NTP에서 수행한 실험용 쥐

Table 4.1. Summary of the diethylhexyl phthalate study data

Dose (ppm)	Dams	Pups	Litter size Median (Range)	Malformations No. (%)	Fetal weight(grams) Mean (SD)
0	28	301	11 (3–16)	9 (3.0%)	0.94 (0.11)
250	26	291	11 (3–16)	5 (1.7%)	0.97 (0.11)
500	26	281	11 (3–15)	38 (13.5%)	0.92 (0.12)
1000	18	147	8 (3–13)	53 (36.1%)	0.90 (0.15)
1500	10	53	5 (1–10)	44 (83.0%)	0.79 (0.15)

에 대한 독성물질 DEHP의 영향에 대한 연구(NTP study: TER84064)로부터 나온 자료이다. 이 연구에서는 임신한 어미 쥐(dam)에 독성 물질 DEHP의 용량 수준을 0ppm, 250ppm, 500ppm 1000ppm, 1500ppm으로 증가시키면서 어미 쥐의 몸무게, 간의 무게, 자궁의 무게, 태아의 무게, 태아의 기형 여부 등을 측정하였다. 본 논문에서는 DEHP의 각기 다른 용량 수준에 따른 어미 쥐가 기형인 태아를 최소한 하나 이상을 가지고 있는지에 대한 여부와 어미 쥐가 가지고 있는 태아의 평균 무게에 대한 변수를 사용하고자 한다. Table 4.1은 본 논문에 사용된 자료를 간략히 요약한 것인데, DEHP의 용량 수준이 증가함에 따라 어미 쥐가 갖는 기형인 태아의 수가 점차로 증가하는 추세를 보이다가 1000ppm과 1500ppm의 용량에서 급격히 증가함을 알 수 있다. 또한, 태아의 무게 평균값도 점차로 감소하다가 1500ppm의 용량에서 급격히 감소하고 있음을 알 수 있다. 어미 쥐가 갖는 태아의 수(litter size)는 용량 수준이 증가함에 따라 그 평균값이 감소하는 추세를 보여주고 있다.

Figure 4.1은 DEHP의 용량 수준에 따라 태아의 무게가 어떤 형태의 분포를 가지는지를 보여준다. 용량 수준이 작을 때는 태아의 무게 분포가 오른쪽으로 치우친 형태를 지니고 있지만, 용량 수준이 증가할수록 분포의 모양이 왼쪽으로 치우친 형태를 가지게 된다. 이런 특징은 태아의 기형 여부를 나타낸 변수에서도 나타난다. 즉, 기형인 태아의 비율이 용량 수준이 작을 때는 3% 미만으로 굉장히 작지만, 용량 수준이 1500ppm일때는 83%로 극단적으로 큰 값을 가지고 있다. 이는 DEHP의 용량 수준에 따라 변수의 비대칭성의 방향이 서로 뒤바뀌는 특성을 잘 설명해주고 있다. 사전 분석으로서 태아의 기형 여부와 태아의 무게에 대한 로지스틱 회귀분석 결과를 보면 두 변수 사이에 강한 음의 상관관계가 존재함을 알 수 있다 (로그 오즈비 $= -4.4$, p-value $= 0.032$). 따라서 DEHP 데이터는 이진수 자료와 연속형 자료에 대한 비대칭 로짓 모형을 사용한 결합모형에 적용하기에 적합할 것이다.

4.2. 베이지안 모형

x_i는 독성 물질 DEHP의 용량 수준을 나타내며 5개의 서로 다른 값을 가지는데($i = 1, \ldots, 5$), 분석의 편의상 단위를 조정하기 위해 ppm 대신 쥐 사료에 들어간 DEHP의 비율(%)을 사용한다 (0%, 0.025%, 0.05%, 0.1%, 0.15%). $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2})^T$는 DEHP의 i번째 용량 수준에 노출된 j번째 개체에서 측정된 2×1 벡터로 이루어진 반응변수를 나타낸다 ($i = 1, \ldots, 5$, $j = 1, \ldots, n_i$). 이때 Y_{ij1}은 이진수 형태의 변수인 기형인 태아가 하나 이상 있는지 여부를 나타내며, Y_{ij2}는 연속형 형태의 변수로서 태아의 무게를 나타낸다. 4.1에서 언급한대로 두 변수는 강한 음의 상관관계를 가지고 있기 때문에 식 (2.1)과 (2.2)에 적용해서 결합모형을 구성할 수 있다.

베이지안 추론 방법을 사용하기 위해 세 모수 $(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma)$에 대하여 각각 다음과 같이 무정보적인 사전분포를 독립적으로 고려하였다.

$$\boldsymbol{\beta} \sim \text{MVN}\left(\mathbf{0}, 100 I_4\right), \qquad \boldsymbol{\delta} \sim \text{MVN}\left(\mathbf{0}, 100 I_{10}\right), \qquad \sigma^2 \sim \text{IG}(1, 10),$$

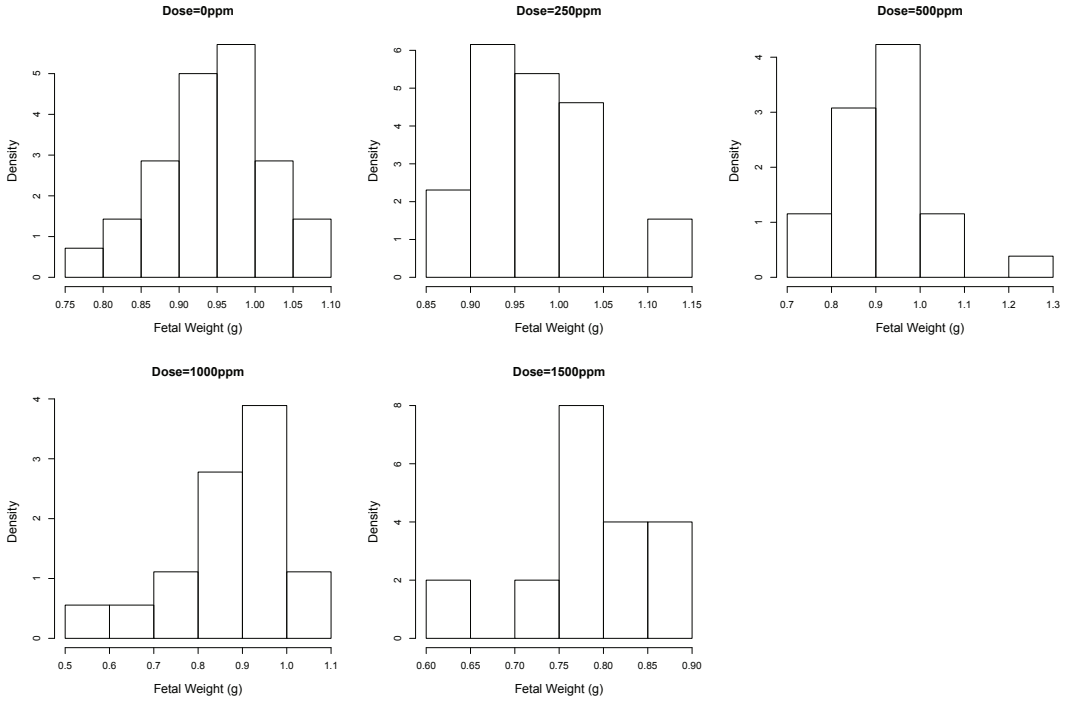

Figure 4.1. Histograms of fetal weight at each dose level of diethylhexyl phthalate.

여기에서 정보를 가진(informative) 사전분포를 가정하여 추론해도 그 결과는 크게 바뀌지 않는 것을 확인하여 사전분포의 효과를 최소로 하기 위해 무정보적인 사전분포를 가정한 모형을 제시하였다. 모수 추정방법으로는 3.1과 3.2에서 소개한 MCMC 방법 중 하나인 깁스 샘플러 내의 메트로폴리스-해스팅스 알고리즘을 사용하였다. 50,000번의 반복시행과 25,000번의 제거(burn-in)를 통해 얻은 표본을 바탕으로 추정치를 계산하였고, 각 모수 추정치의 수렴여부를 판단하기 위해 trace plot과 Gelman과 Rubin의 $\hat{R}$을 사용하여 확인하였다.

제안된 모형의 성능을 보여주기 위해서 일반적인 대칭 로짓 모형을 사용한 결합모형과의 비교를 고려하였다. 대칭 결합모형은 식 (2.1)과 (2.2)에서 $\delta_{i1} = \delta_1$, $\delta_{i2} = \delta_2$ $(i = 1, \ldots, 5)$로 놓고, z_{ij}의 분포 G를 정규분포 $N(0, \sigma_z^2)$로 가정함으로써 얻을 수 있다. 두 모형의 비교를 위해서 베이지안 분석에서 흔히 사용하는 모형 비교 통계량인 deviance information criterion (DIC)를 사용하였다 (Spiegelhalter 등, 2002). DIC는 모수의 편차의 사후평균과 모형의 복잡한 정도에 대한 페널티의 합으로 구성되어 있기 때문에, 더 작은 DIC를 가지는 모형이 데이터를 더 잘 적합시킨다는 것을 의미한다. 본 논문에서는 랜덤효과를 가지는 모형에 직접적으로 적용할 수 있는 수정된 DIC를 사용하여 모형 비교를 하였다 (Celeux 등, 2006).

4.3. 분석결과

본 논문에서 제시한 비대칭 로짓 모형을 사용한 결합모형에서 MCMC 알고리즘을 통해 추출된 주요 모수들의 표본값의 trace plot은 Figure 4.2에서 확인할 수 있다. 대체적으로 수렴이 잘 이루어지고 있음

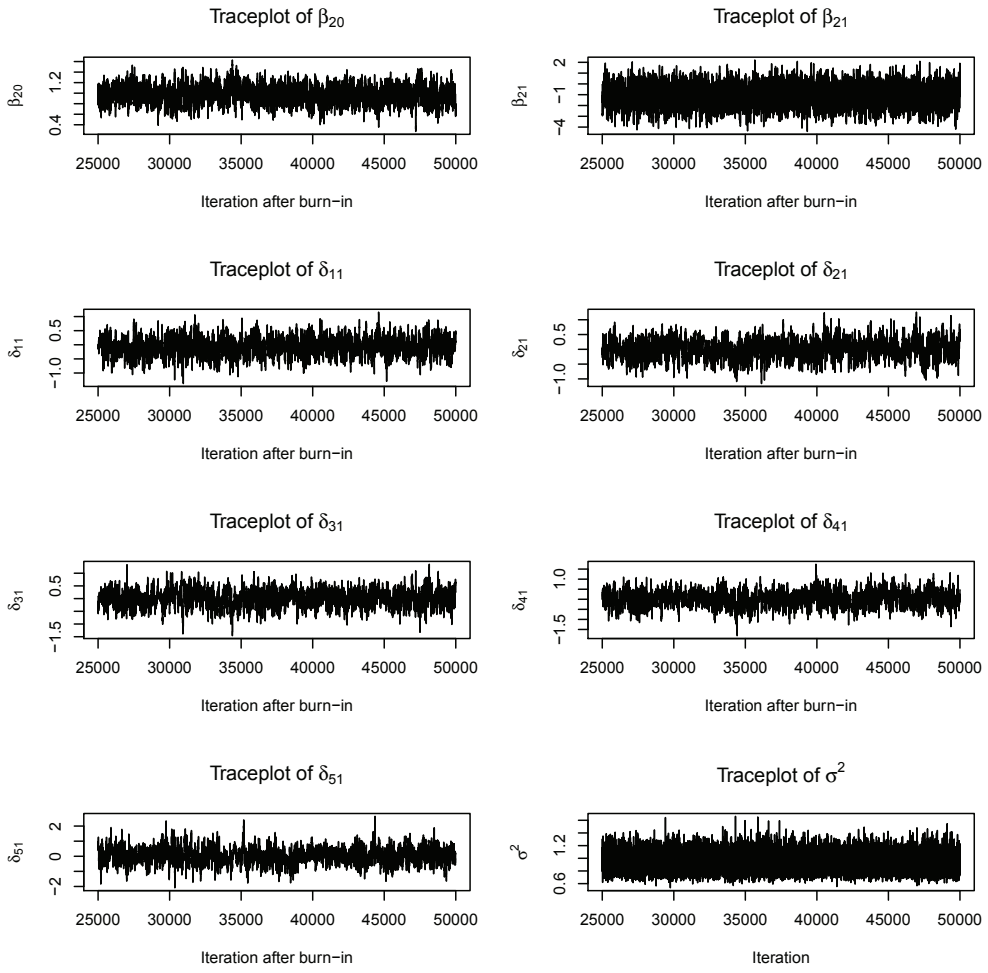

Figure 4.2. Traceplots of parameters, $\beta_{20}, \beta_{21}, \delta_{12}, \delta_{22}, \delta_{32}, \delta_{42}, \delta_{52}, \sigma^2$ in the diethylhexyl phthalate data analysis.

을 알 수 있고, Gelman과 Rubin의 $\hat{R}$ 역시 1에 근접한 값을 가지고 있음을 확인하였다.

DIC를 기반으로 모형의 적합도를 비교해 보면, 비대칭 로짓 모형의 DIC(= 364.2)가 대칭 로짓 모형의 DIC(= 371.5)보다 약간 더 작은 값을 가짐으로써 비대칭 로짓 모형을 사용한 결합모형이 DEHP 데이터에 더 적합하다고 결론내릴 수 있다.

Table 4.2는 비대칭 로짓 모형을 사용한 결합모형을 DEHP 데이터에 적합시켜 나온 결과로서 모수의 사후평균(posterior mean)과 95% 신용구간(credible interval; C.I.)을 보여주고 있다. β_{11}과 β_{21}은 독성 물질 DEHP의 각기 다른 용량 수준에 따른 고정효과를 나타낸다. β_{11}이 양의 값을 가지는 것은 DEHP의 수준이 증가할수록 기형인 태아가 최소한 한마리 이상 있을 확률이 증가함을 보여주는 것이고, 반면에 β_{21}은 음의 값을 가짐으로써 DEHP의 수준이 증가할수록 태아의 무게가 점차 감소한다는 사실을 반영해주고 있다. 또한 태아의 기형 여부와 태아의 무게는 서로 음의 상관관계를 나타내고 있음

Table 4.2. Posterior means and 95% credible intervals for parameters in the diethylhexyl phthalate data analysis

	Parameter	Estimate	95% C.I.
Malformation model	β_{10}	0.362	(−0.620, 1.319)
	β_{11}	1.056	(0.090, 2.128)
	δ_{11}	−4.205	(−8.067, −0.986)
	δ_{21}	−4.224	(−8.271, −1.110)
	δ_{31}	0.170	(−2.086, 2.763)
	δ_{41}	3.091	(−0.207, 7.345)
	δ_{51}	3.854	(0.189, 8.692)
Weight model	β_{20}	0.975	(0.655, 1.289)
	β_{21}	−1.092	(−2.294, −0.072)
	δ_{12}	−0.069	(−0.696, 0.548)
	δ_{22}	−0.001	(−0.640, 0.624)
	δ_{32}	0.002	(−0.646, 0.645)
	δ_{42}	0.025	(−0.717, 0.751)
	δ_{52}	0.023	(−0.975, 1.114)
	σ^2	0.927	(0.710, 1.211)

을 알 수 있다. 이러한 결과는 Table 4.1에서 보여준 데이터의 기본 내용과 동일한 패턴을 지니고 있음을 확인할 수 있다.

비대칭 모수인 δ의 추정치를 통해 우리 모형의 특징을 살펴보면, DEHP의 용량 수준이 증가함에 따라 이진수 변수 모형에서의 δ값은 음의 값에서 점차 증가하여 양의 값을 가지는 패턴을 취하고 있다. 즉, 용량 수준이 0ppm과 250ppm일 때는 음의 δ값을 가짐으로써($\delta_{11} = -4.205, \delta_{21} = -4.224$) 이진수 변수에 내포된 잠재변수의 주변분포가 왼쪽으로 치우친 형태를 띠게 되어, 태아의 기형 발생 확률이 0으로 접근하는 속도가 1일때의 속도보다 더 빠른 비대칭적 성질을 가지게 된다. 반면에 용량 수준이 1000ppm과 1500ppm 등으로 높아질 때는 양의 δ값을 가지게 되어($\delta_{41} = 3.091, \delta_{51} = 3.854$) 잠재변수가 오른쪽으로 치우친 주변분포를 가지게 되며 태아의 기형 발생 확률이 1로 접근하는 속도가 더 빠른 성질을 보여주게 된다. 이는 Table 4.1에서 보여주는 데이터의 기본 성질과 동일한 결과로서 독성 물질의 용량 수준이 증가함에 따라 자료의 비대칭성의 방향이 뒤바뀌는 특성을 잘 나타내주고 있다. 하지만 이러한 특성은 태아의 무게 변수에서는 관찰되지 않는다. 즉, $\boldsymbol{\delta_2} = (\delta_{12}, \delta_{22}, \delta_{32}, \delta_{42}, \delta_{52})$는 0에 가까운 값을 가지고 있기 때문에 분포의 비대칭성을 띠고 있지 않음을 알 수 있다. 이는 Figure 4.1의 히스토그램에 나와 있는 것처럼 DEHP의 수준에 따른 태아의 무게 분포의 형태가 심한 비대칭성을 지니고 있지 않기 때문에 나온 결과라고 유추할 수 있다.

Figure 4.3는 DEHP 데이터 분석 결과를 그래프로 나타내고 있는데, 3.3절에서 계산한 사후 예측분포를 토대로 독성 물질에 노출된 어미 쥐가 최소한 한 마리의 기형인 태아를 가질 확률과 어미 쥐가 가지는 태아의 무게의 변화를 보여주고 있다. 독성 물질의 수준이 증가할수록 기형인 태아를 가질 확률은 전반적으로 증가하는 패턴을 보여주고 있다. 또한 95% 신용구간을 보면 독성 물질의 용량이 0ppm과 250ppm일 때는 왼쪽으로 치우친 형태를 가지고 있고, 용량이 1000ppm과 1500ppm으로 상당히 클 경우에는 오른쪽으로 치우친 형태를 띠고 있음을 알 수 있다. 이는 위에서 δ값이 갖는 패턴과 동일한 의미를 가지고 있음을 나타내준다. 또한 독성 물질의 용량 수준이 증가함에 따라 태아의 무게의 분포는 왼쪽으로 변화(shift)하는 모습을 보여주고 있고 이런 결과 역시 Table 4.1과 Figure 4.1에서 보여준 데이터의 특성을 반영하는 결과임을 알려 준다.

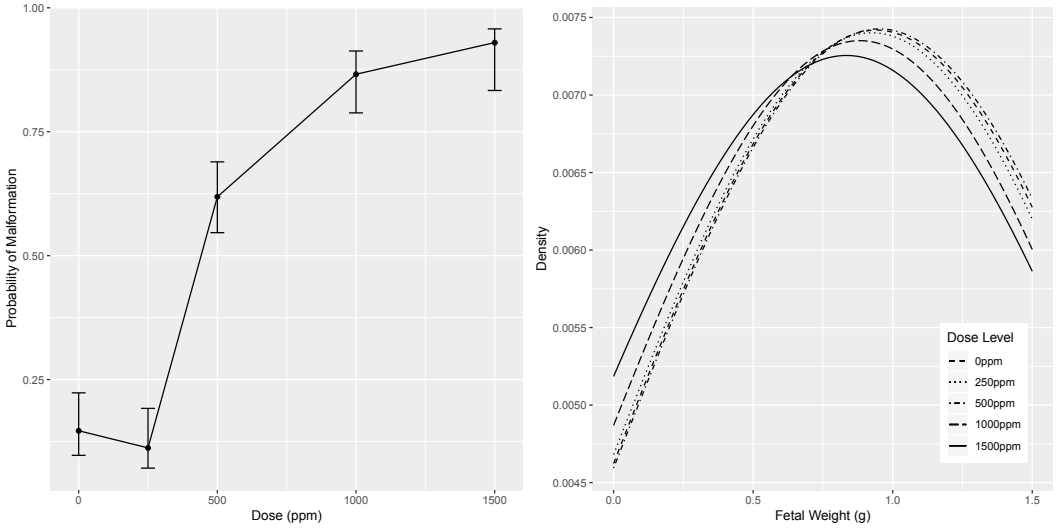

Figure 4.3. DEHP data results: posterior predictive probabilities of having at least one malformed pups (means and 95% credible intervals) at each dose level (left panel), and posterior predictive densities for fetal weight at each dose level (right panel).

5. 결론 및 논의

발달 독성학에서 이진수 자료와 연속형 자료를 결합모형으로 분석할 때에는 Catalano와 Ryan (1992)이 제시한 잠재변수에 기초한 모형을 주로 사용하게 된다. 이런 모형은 기본적으로 정규분포를 따르는 랜덤효과를 가정하게 되는데, 비대칭적인 형태를 가지는 이진수 자료가 포함되는 경우 그 변수의 특징을 잡아내지 못하는 단점이 있다. 이를 해결하기 위한 방법 중 하나로 Chen 등 (1999)이 제안한 비대칭 로짓 모형을 결합모형에 적용하여 우리의 모형을 제시하였다. 독성 물질 DEHP의 용량 수준이 증가함에 따라 기형인 태아의 존재 여부와 태아의 무게가 어떻게 변하는지에 대해 우리의 결합모형을 적용했을 때, 비대칭 로짓 모형의 장점이 극명하게 드러남을 알 수 있었다. 즉, 비대칭 모수인 δ의 추정치가 DEHP의 수준이 증가함에 따라 음의 값에서 차츰 양의 값으로 변하는 패턴을 보여줌으로써 용량 수준이 낮을 때와 높을 때 서로 반대 방향의 비대칭성을 가진다는 것을 반영하였다. 또한, DEHP의 용량 수준의 변화에 따른 기형인 태아의 존재 여부와 태아의 무게의 변화가 서로 음의 상관관계를 가지게 되어 기본 자료의 탐색 결과와 일치하는 분석 결과를 보여주었다.

본 연구에서 도입한 비대칭 로짓 모형은 용량 수준의 변화에 따른 분포의 비대칭성을 잘 반영하고 있지만 다봉분포(multimodal distribution)와 같은 형태의 불규칙성은 나타내지 못하는 한계점이 있다. 좀 더 복잡한 분포의 변화를 반영하기 위해서 혼합모형(mixture model)이나 비모수 베이지안(nonparametric Bayesian) 방법을 이진수 변수의 잠재변수에 가정한 후 결합모형에 도입하여 분석한다면 보다 일반화된 모형이 될 것이다. 이때 여러가지 다양한 모형들을 비교하고 적합성을 평가하기 위해서 베이지안 분석 방법에서 흔히 사용하는 CPO (conditional predictive ordinates) 통계량이나 DIC (deviance information criterion) 측정치를 사용한다면 보다 유용한 분석이 될 수 있을 것이다. Table 4.1에 나와 있는 것처럼 독성 물질 DEHP의 용량이 높은 수준(1000ppm 또는 1500ppm)에서는 어미 쥐의 개체 수가 현저히 적은 특징을 볼 수 있는데, 이러한 현상이 결과에 미치는 영향을 보기 위해 Dunson 등 (2003)이 제시한 군집의 크기(litter size)와 변수들의 결합모형을 고려해볼 필요도 있을 것이다. 또

한 Hwang과 Pennell (2014)에서 위험 평가(risk assessment) 분야에 응용한 것처럼 본 논문의 결합모형을 기준용량(benchmark dose) 분석에 적용하여 다른 모형들과 그 결과들을 비교해보는 것도 좋은 후속 연구가 될 수 있을 것이다.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, **88**, 669–679.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler, *The American Statistician*, **46**, 167–174.

Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, **87**, 651–658.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criterion for missing data models, *Bayesian Analysis*, **1**, 651–674.

Chen, M. H., Dey, D. K., and Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data, *Journal of the American Statistical Association*, **94**, 1172–1186.

Chen, M. H., Dey, D. K., and Shao, Q. M. (2001). Bayesian analysis of binary data using skewed logit models, *Calcutta Statistical Association Bulletin*, **51**, 12–30.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, *The American Statistician*, **49**, 327–335.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An ANOVA model for dependent random measures, *Journal of the American Statistical Association*, **99**, 205–215.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes, *Journal of the Royal Statistical Society: Series B*, **62**, 355–366.

Dunson, D. B., Chen, Z., and Harry, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes, *Biometrics*, **59**, 521–530.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, CRC Press, New York.

Haario, H., Saksman, E., and Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC, *Computational Statistics*, **20**, 265–273.

Hwang, B. S. and Pennell, M. L. (2014). Semiparametric Bayesian joint modeling of a binary and continuous outcome with applications in toxicological risk assessment, *Statistics in Medicine*, **33**, 1162–1175.

Hwang, B. S. and Pennell, M. L. (2018). Semiparametric Bayesian joint modeling of clustered binary and continuous outcomes with informative cluster size in developmental toxicity assessment, *Environmetrics*, **29**, e2526, 1–15.

Kim, S. B. and Hwang, B. S. (2019). A Bayesian skewed logit model for high-risk drinking data, *Journal of the Korean Data & Information Science Society*, **30**, 335–348.

Li, E., Zhang, D., and Davidian, M. (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements, *Biometrics*, **60**, 1–7.

MacEachern, S. N. (1999). Dependent nonparametric process. In *ASA Proceeding of the Section on Bayesian Statistical Science*, American Statistical Association: Alexandria, VA.

McCulloch, C. (2008). Joint modelling of mixed outcome types using latent variables, *Statistical Methods in Medical Research*, **17**, 53–73.

Regan, M. M. and Catalano, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology, *Biometrics*, **55**, 760–768.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica*, **4**, 639–650.

Spiegelhalter, D. J., Best, N. G., Carline, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B*, **64**, 583–639.

발달 독성학에서 비대칭 로짓 모형을 사용한 이진수 자료와 연속형 자료에 대한 결합분석

김영화[a] · 황범석[a,1]

[a]중앙대학교 응용통계학과

(2019년 9월 17일 접수, 2019년 10월 30일 수정, 2019년 11월 5일 채택)

요 약

하나의 개체에서 여러가지 측정치가 동시에 관찰되는 경우는 다양한 연구 분야에서 흔히 나타난다. 발달 독성학 연구에서는 특정 독성 물질의 각기 다른 수준에 노출된 임신한 어미 쥐에 대해 기형인 태아의 존재와 태아의 무게가 동시에 측정된다. 이런 두 변수를 결합하여 모형화하는 것은 각기 독립적인 두 모형으로 분석하는 것보다 더 효율적인 결과를 낸다고 알려져 있다. 대부분의 결합 모형은 정규분포를 랜덤효과로 가정하여 분석한다. 그러나 발달 독성학 연구에서처럼 반응변수들의 분포가 독성 물질이 변함에 따라 불규칙하게 변하는 경우 정규분포의 가정으로는 그 특징을 잡아낼 수 없게 된다. 본 논문에서는 이진수 자료와 연속형 자료에 대해 비대칭 로짓 모형을 사용한 베이지안 결합모형을 제시한다. 본 모형은 비대칭 로짓 모형을 사용함으로써 반응변수의 분포의 형태가 독성 물질의 수준에 따라 대칭/비대칭의 형태를 자유롭게 띨 수 있는 장점을 가지고 있다. 모형의 적합성을 살펴보기 위해 발달 독성학 연구에서 독성 물질 DEHP에 적용하여 그 결과를 확인해본다.

주요용어: 결합모형, 독성 물질, 마코프체인 몬테카를로, 베이지안 추론, 비대칭 로짓 모형

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2019R1C1C1011710).

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: bshwang@cau.ac.kr