

환경오염물질이 여성 불임에 미치는 영향에 대한 베이지안 잠재계층모형 분석[†]

최윤경¹ · 황범석²

^{1,2}중앙대학교 응용통계학과

접수 2018년 8월 20일, 수정 2018년 9월 6일, 게재확정 2018년 9월 10일

요약

대표적인 환경오염물질인 PCB가 여성의 불임에 미치는 영향에 대해 조사한 LIFE 연구를 기반으로 해서 그 관계를 분석한 잠재계층모형을 제시하였다. 본 모형에서는 피험자의 잠재적인 위험군이 존재한다고 가정하고 잠재계층 변수를 도입하여 PCB와 여성의 불임에 대한 로짓 모형을 연결시켜주었다. 또한 반연속적인 형태를 띠는 PCB 값을 다루기 위해 혼합 분포 모형을 적용시켰다. 구체적인 분석 방법으로 MCMC에 기반을 둔 베이지안 접근법을 사용하였고, 모형의 비교를 위해서 DIC를 계산하여 최적의 모형을 찾아내려고 하였다. 분석 결과 피험자들이 속한 잠재적인 위험군에 따라 불임에 대한 확률이 영향을 받고 있음을 확인할 수 있었다.

주요용어: 마코프체인 몬테카를로, 메트로폴리스 알고리즘, 반연속적인 자료, 잠재계층 변수, LIFE 연구.

1. 서론

환경오염물질에 의해 발생하는 질병의 종류 및 형태는 매우 다양하다. 특히 지속적인 환경오염물질에 의 노출은 여성의 생식 능력의 감소 또는 불임에 큰 영향을 미칠 수 있다. 예를 들어, 폴리염화바이페닐 (polychlorinated biphenyls, PCBs)은 생물체 내에서 농축 현상을 나타내는 대표적인 환경오염물질로 농축량에 따라 여성의 생식 기능의 이상을 일으킨다고 알려져 있다 (Meeker 등, 2011). 미국의 국립 보건원 산하 국립아동보건인간개발연구소 (*Eunice Kennedy Shriver National Institute of Child Health and Human Development, NICHD*)에서 시행된 LIFE 연구 (*Longitudinal Investigation of Fertility and the Environment Study*)에서는 36가지 종류의 PCB 농축량과 가임기 여성의 임신 여부에 대한 연구를 진행하였다 (Buck Louis 등, 2011).

임신 여부를 반응변수로 가정할 경우 일반적으로 로지스틱 회귀분석을 사용하여 분석할 수 있으나, LIFE 연구에서의 PCB 농축량은 과도하게 많은 0의 값을 가지는 반연속 자료 (semicontinuous data)의 성격을 띠고 있어서 단순 로지스틱 회귀분석은 과도하게 많은 0에 대한 설명이 어려울 수 있다 (Zhang 등, 2012; Hwang 등, 2018). 이러한 문제를 해결하기 위해 화학물질 노출 및 질병 데이터를 분

[†] 이 논문은 2016년도 중앙대학교 CAU GRS 지원에 의하여 작성되었고, 2016년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2016R1D1A1B03933334).

¹ (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과, 석사과정.

² 고신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과, 조교수.

E-mail: bshwang@cau.ac.kr

석하는 데 유용한 방법인 잠재계층모형 (latent class model)이 종종 사용되어왔다 (Lee 등, 2018). 사람마다 같은 환경오염물질에 노출되더라도 반응여부나 반응속도에 차이가 있을 수 있기 때문에 피험자들을 잠재된 위험군으로 나눠서 환경오염물질과 그 반응변수간의 관계를 알아보는데 적용하는 것이다. Lin 등 (2002)은 경시적으로 측정된 전립선 특이항원 (prostate-specific antigen)과 전립선암의 관계를 알아보기 위해 잠재계층모형을 가정하고 분석하였다. Neelon 등 (2011)은 산모의 출산경력이 정신건강에 미치는 영향에 대해 베이지안 방법에 기초한 잠재계층모형을 제안하였다. Liu 등 (2015)은 경시적 자료와 생존 자료에 대한 결합 모형을 사용할 때 잠재계층변수를 고려하였고, Zhang 등 (2012)은 반연속적인 형태를 띠는 환경오염물질의 결합체와 여성의 자궁내막증 (endometriosis)의 발생률에 대해서 잠재계층모형을 가정하고 분석하였다. 또한 Hwang 등 (2018)은 LIFE 자료를 사용해서 커플들의 위험군을 잠재계층변수로 가정하고 여성의 불임 여부와 화학물질 간의 관계를 분석하였다.

이와 같은 잠재계층모형은 주로 최대가능도 방법 (maximum likelihood method)을 통해서 분석되어 왔고, 그에 따르는 복잡한 계산을 해결하기 위해서 다양한 계산법들이 제안되어 왔다. Lin 등 (2002)은 최대가능도추정량 (maximum likelihood estimator, MLE)을 계산하기 위해 EM 알고리즘을 사용하였고, Zhang 등 (2012)은 MCEM (Monte Carlo EM) 알고리즘을 활용하여 복잡한 계산을 해결하였다. 또한 Liu 등 (2015)은 가우스 구적 접근법 (Gaussian quadrature approach)을 적용하여 MLE를 계산하였다. Neelon 등 (2011)과 Hwang 등 (2018)은 MCMC (Markov chain Monte Carlo) 방법을 토대로 베이지안 접근법을 사용하였다.

본 논문에서는 Zhang 등 (2012)과 Hwang 등 (2018)에서 사용된 잠재계층모형을 가정하여 환경오염물질인 PCB와 여성의 불임 가능 사이의 연관성을 베이지안 접근법을 토대로 분석하려고 한다. 2장에서는 LIFE 연구에 적용될 수 있는 잠재계층모형을 설명한다. 3장에서는 베이지안 추론을 위해서 사전분포 (prior distribution)를 가정하고 그에 따르는 사후분포 (posterior distribution)를 계산한 후에 MCMC 방법의 단계를 설명한다. 또한 최적의 잠재계층의 개수를 결정하기 위해 모형 비교 방법인 DIC (Deviance Information Criterion)를 간략히 설명한다. 4장에서는 실제 LIFE 연구 자료를 토대로 PCB와 여성의 불임 간의 관계에 대해 분석한다. 끝으로 5장에서는 본 논문의 결론을 제시하고 향후 연구의 방향에 대해 논의한다.

2. 잠재계층모형

반응변수 Y_i ($i = 1, \dots, n$)는 i 번째 여성의 불임 상태를 나타내는 이항변수로 임신에 실패했을 때 (infertility) 1의 값을 가진다고 가정한다. 설명변수 X_i 는 i 번째 여성에게서 측정된 환경오염물질의 농도를 나타낸다. 이때, Y_i 와 X_i 의 관계를 설명해주기 위해 여성의 불임에 대한 위험도를 나타내는 잠재계층 변수 (latent class variable) L_i 를 모형에 도입한다. 즉, i 번째 여성의 k 번째 위험 그룹 ($k = 1, \dots, K$)에 속한다면, $L_i = k$ 가 되고, 이때 k 가 증가함에 따라 불임에 대한 위험도가 증가한다고 가정한다. 예를 들어, $K = 3$ 인 잠재계층모형은 저위험 ($L_i = 1$), 중간위험 ($L_i = 2$), 고위험 ($L_i = 3$) 계층으로 구성된다. 또한, 잠재계층 k 에 속할 확률을 $\pi_k = P(L_i = k)$ 로 정의한다. 일반적으로 잠재계층모형에서는 두 변수 Y_i 와 X_i 의 조건부 독립 (conditionally independent)을 가정한다. 즉, 두 변수의 관계는 잠재계층 변수 L_i 에 의해서만 설명이 되고, 동일한 잠재계층 내에서 두 변수는 아무런 관련을 가지지 않는다 (McCutcheon, 1987). 잠재계층 변수가 주어졌을 때, i 번째 여성의 불임이 될 확률은 다음과 같이 나타낼 수 있다.

$$\text{logit}P(Y_i = 1 | \mathbf{R}_i) = \beta_0 + \sum_{k=1}^{K-1} \beta_k R_{ki}, \quad (2.1)$$

여기서 R_{ki} 는 잠재계층 변수를 나타내기 위한 지시변수로서 다음과 같이 정의된다.

$$R_{ki} = \begin{cases} 1, & \text{if } L_i = k, \quad k = 1, \dots, K - 1, \\ 0, & \text{otherwise.} \end{cases} \quad (2.2)$$

이러한 모형에 잠재계층 변수 L_i 를 순서형 변수로 가정하여 식 (2.1)의 오른쪽 항을 $\beta_0 + \beta_1 L_i$ 와 같이 표현할 수도 있으나, 이럴 경우 잠재계층 간의 변화가 계층에 상관없이 일정해진다는 단점이 생긴다. 따라서 본 논문에서는 보다 일반적인 방법인 지시변수를 도입하여 모형을 정의한다. 이때, K 개의 잠재계층이 존재한다면 $K - 1$ 개의 지시변수가 생성됨을 알 수 있다. 또한, $\beta = (\beta_0, \dots, \beta_{K-1})^T$ 는 잠재계층 변수에 상응하는 계수 벡터를 나타낸다.

LIFE 연구 (Longitudinal Investigation of Fertility and the Environment Study)에서는 36개의 각기 다른 PCB를 다루고 있는데, 본 논문에서는 그 중의 하나인 PCB177에 초점을 맞춰 분석하려고 한다. PCB177은 약 26% 가량이 농도가 0이거나 측정 불가일 정도로 너무 작아서 0으로 취급되고 있다. 이러한 반연속적인 (semicontinuous) 데이터는 0의 값이 발생하는 점획률분포와 0보다 큰 값에 대한 연속형 분포의 혼합 분포의 형태를 통해서 모형화할 필요가 있다 (Liu 등, 2010; Neelon 등, 2011; Liu 등, 2012; Kim 등, 2017). Figure 2.1에서 보는 바와 같이 PCB177의 분포는 과도하게 많은 0의 값을 가지고 있고, 0보다 큰 값에 대해 로그변환 시킨 후의 분포는 정규분포를 근사적으로 따르고 있다.

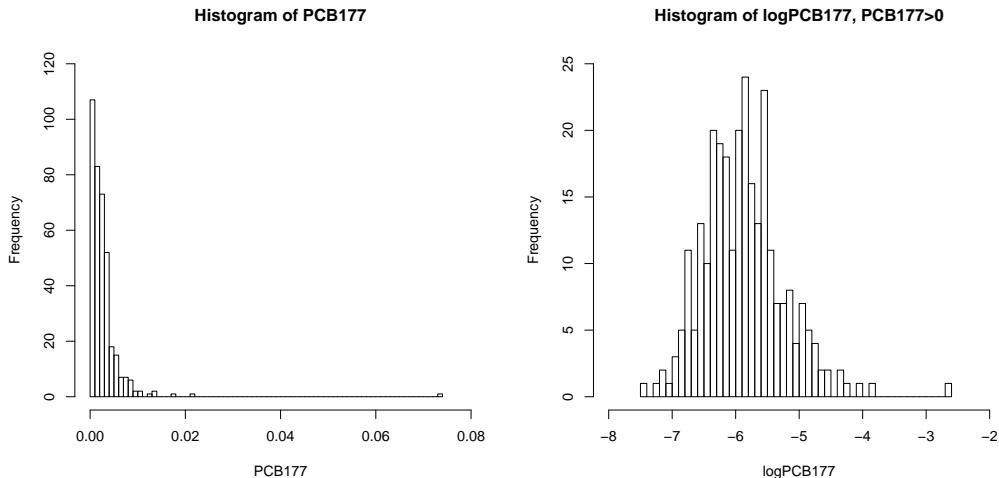


Figure 2.1 Histograms of PCB177 and logPCB177

PCB177의 이러한 특성을 바탕으로 X_i 를 두개의 잠재변수 U_i 와 V_i 를 사용하여 다음과 같이 나타낼 수 있다.

$$U_i = \begin{cases} 1, & \text{if } X_i \neq 0 \\ 0, & \text{if } X_i = 0 \end{cases} \quad \text{and} \quad V_i = \begin{cases} X_i, & \text{if } X_i \neq 0 \\ \text{irrelevant}, & \text{if } X_i = 0, \end{cases} \quad (2.3)$$

여기서 U_i 는 X_i 의 0의 여부를 나타내는 지시변수이고, V_i 는 0이 아닌 X_i 값을 나타내는 연속형 변수이다. Figure 2.1을 바탕으로 V_i 는 다음과 같이 로그정규분포를 따른다고 가정한다.

$$V_i | \mathbf{R}_i \sim \log N(\mu_i(R_i), \sigma^2), \quad (2.4)$$

여기서 $\mu_i(\mathbf{R}_i)$ 와 σ^2 는 각각 $\log V_i$ 의 평균과 분산을 나타낸다. $\log V_i$ 의 평균은 잠재계층 변수 R_i 에 따라 달라진다고 가정하면, 다음과 같은 관계를 가지게 된다.

$$\mu_i(\mathbf{R}_i) = \alpha_0 + \sum_{k=1}^{K-1} \alpha_k R_{ki}, \quad (2.5)$$

여기서 R_{ki} 는 식 (2.2)에서 정의된 바와 같고, $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_{K-1})^T$ 는 거기에 상응하는 계수 벡터를 나타낸다. 마지막으로 X_i 가 0이 아닐 확률은 0이 아닌 X_i 값의 평균과 관련이 있다고 가정하고 다음과 같은 로짓모형을 설정한다.

$$\text{logit}P(U_i = 1 | \mathbf{R}_i) = \eta_0 + \eta_1 \mu_i(\mathbf{R}_i), \quad (2.6)$$

식 (2.1)에 있는 여성의 불임에 대한 로짓모형과 식 (2.3)-(2.6)으로 구성된 PCB177 농도에 관한 모형이 잠재계층 변수 \mathbf{R}_i 에 의해 연결되어 있음을 알 수 있다. 이와 같은 결합 모형을 바탕으로 한 완전데이터 가능도함수 (complete data likelihood function)는 다음과 같다.

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2, \boldsymbol{\pi} | \mathbf{Y}, \mathbf{U}, \mathbf{V}, \mathbf{R}) &= \prod_{i=1}^n \left(\frac{e^{\Lambda_i}}{1+e^{\Lambda_i}} \right)^{y_i} \left(\frac{1}{1+e^{\Lambda_i}} \right)^{1-y_i} \\ &\times \left(\frac{e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}}{1+e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{u_i} \left(\frac{1}{1+e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{1-u_i} \\ &\times [\log N(V_i; \mu_i(\mathbf{R}_i), \sigma^2)]^{u_i} \times \pi_{L_i}, \end{aligned} \quad (2.7)$$

여기서 $\Lambda_i = \beta_0 + \sum_{k=1}^{K-1} \beta_k R_{ki}$ 이고, $\pi_{L_i} = P(L_i = k)$ 로 정의된다.

3. 베이지안 추론

3.1. 사전분포의 가정

잠재계층모형의 베이지안 분석을 위해 식 (2.1), (2.5), (2.6)에 있는 $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}$ 에 대해 각각 다음과 같이 독립적인 사전분포를 고려한다.

$$\begin{aligned} \boldsymbol{\alpha} &\sim MVN(\boldsymbol{\mu}_{\alpha}, \boldsymbol{\Sigma}_{\alpha}), \\ \boldsymbol{\beta} &\sim MVN(\boldsymbol{\mu}_{\beta}, \boldsymbol{\Sigma}_{\beta}), \\ \boldsymbol{\eta} &\sim MVN(\boldsymbol{\mu}_{\eta}, \boldsymbol{\Sigma}_{\eta}), \end{aligned} \quad (3.1)$$

여기서 $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 은 평균 $\boldsymbol{\mu}$ 와 공분산행렬 $\boldsymbol{\Sigma}$ 를 가지는 다변량정규분포를 나타낸다. 사전분포의 평균인 $\boldsymbol{\mu}_{\alpha}$ 와 $\boldsymbol{\mu}_{\beta}$ 의 차원은 $K \times 1$ 이고, 공분산인 $\boldsymbol{\Sigma}_{\alpha}$ 와 $\boldsymbol{\Sigma}_{\beta}$ 의 차원은 $K \times K$ 이다. $\boldsymbol{\mu}_{\eta}$ 와 $\boldsymbol{\Sigma}_{\eta}$ 의 차원은 각각 2×1 과 2×2 이다. 또한, PCB177 농도의 양의 값 (V_i)에 대한 로그정규분포의 척도모수 (scale

parameter)에 대해서는 공액사전분포 (conjugate prior distribution)인 역감마 분포 (inverse Gamma distribution)를 취한다: $\sigma^2 \sim IG(a_\sigma, b_\sigma)$. 각 잠재계층에 속할 확률에 대해서는 공액사전분포인 디리슈레 분포 (Dirichlet distribution)를 따른다고 가정한다.

$$(\pi_1, \dots, \pi_K) \sim Dirichlet(e_1, \dots, e_K),$$

마지막으로 모형에서 잠재계층의 총개수 K 는 알려져 있다고 가정하고, 모형 비교 방법을 통해 최적의 K 를 구한다. 모형 비교 방법에 대해서는 3.4장에서 자세히 다룬다.

3.2. 사후분포의 계산

식 (2.7)에서 정의된 완전데이터 가능도함수와 3.1장에서 가정한 사전분포를 이용하여 다음과 같이 결합사후분포를 계산한다.

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2, \boldsymbol{\pi} | \mathbf{Y}, \mathbf{U}, \mathbf{V}, \mathbf{R}) &\propto p(\mathbf{Y}, \mathbf{U}, \mathbf{V}, \mathbf{R} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2, \boldsymbol{\pi}) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\boldsymbol{\eta}) p(\sigma^2) p(\boldsymbol{\pi}) \\ &\propto \prod_{i=1}^n \left[\left(\frac{e^{\Lambda_i}}{1 + e^{\Lambda_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\Lambda_i}} \right)^{1-y_i} \right. \\ &\quad \times \left(\frac{e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}}{1 + e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{u_i} \left(\frac{1}{1 + e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{1-u_i} \\ &\quad \times \left. \{ \log N(V_i; \mu_i(\mathbf{R}_i), \sigma^2) \}^{u_i} \times \pi_{L_i} \right] \\ &\quad \times MVN(\boldsymbol{\alpha}; \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}) \times MVN(\boldsymbol{\beta}; \boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \times MVN(\boldsymbol{\eta}; \boldsymbol{\mu}_{\boldsymbol{\eta}}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) \\ &\quad \times IG(\sigma^2; a_\sigma, b_\sigma) \times Dirichlet(\boldsymbol{\pi}; e_1, \dots, e_K). \end{aligned} \quad (3.2)$$

식 (3.2)에서 구한 결합사후분포로부터 각 모수에 대한 조건부 사후분포를 계산하면 다음과 같다.

$$\begin{aligned} p(\boldsymbol{\alpha} | -) &\propto \prod_{i=1}^n \left[\left(\frac{e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}}{1 + e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{u_i} \left(\frac{1}{1 + e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{1-u_i} \right. \\ &\quad \times \left. \{ \log N(V_i; \mu_i(\mathbf{R}_i), \sigma^2) \}^{u_i} \right] \times \exp \left\{ -\frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\mu}_{\boldsymbol{\alpha}})^T \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} (\boldsymbol{\alpha} - \boldsymbol{\mu}_{\boldsymbol{\alpha}}) \right\}, \end{aligned} \quad (3.3)$$

$$p(\boldsymbol{\beta} | -) \propto \prod_{i=1}^n \left[\left(\frac{e^{\Lambda_i}}{1 + e^{\Lambda_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\Lambda_i}} \right)^{1-y_i} \right] \times \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}}) \right\}, \quad (3.4)$$

$$\begin{aligned} p(\boldsymbol{\eta} | -) &\propto \prod_{i=1}^n \left[\left(\frac{e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}}{1 + e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{u_i} \left(\frac{1}{1 + e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{1-u_i} \right] \\ &\quad \times \exp \left\{ -\frac{1}{2} (\boldsymbol{\eta} - \boldsymbol{\mu}_{\boldsymbol{\eta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} (\boldsymbol{\eta} - \boldsymbol{\mu}_{\boldsymbol{\eta}}) \right\}, \end{aligned} \quad (3.5)$$

$$p(\sigma^2 | -) \sim IG \left(a_\sigma + \frac{1}{2} \sum_{i=1}^n u_i, b_\sigma + \frac{1}{2} \sum_{i=1}^n u_i (\log V_i - \mu_i(\mathbf{R}_i)^2) \right), \quad (3.6)$$

$$p(\boldsymbol{\pi} | -) \sim Dirichlet \left(\sum_{i=1}^n I(L_i = 1) + e_1, \dots, \sum_{i=1}^n I(L_i = K) + e_K \right), \quad (3.7)$$

$$\begin{aligned} p(L_i = k | -) &\propto \left(\frac{e^{\Lambda_i}}{1 + e^{\Lambda_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\Lambda_i}} \right)^{1-y_i} \times \left(\frac{e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}}{1 + e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{u_i} \left(\frac{1}{1 + e^{\eta_0 + \eta_1 \mu_i(\mathbf{R}_i)}} \right)^{1-u_i} \\ &\times \{ \log N(V_i; \mu_i(\mathbf{R}_i), \sigma^2) \}^{u_i} \times \pi_{L_i}. \end{aligned} \quad (3.8)$$

3.3. Markov Chain Monte Carlo

식 (3.3)-(3.8)에 나와 있는 각 모수의 조건부 사후분포를 바탕으로 갑스 샘플러 (Gibbs sampler)와 메트로폴리스-해스팅스 (Metropolis-Hastings(M-H)) 알고리즘을 결합한 Markov chain Monte Carlo (MCMC) 알고리즘을 사용해서 모수를 추정하려고 한다. 다음과 같은 단계를 거쳐 MCMC 알고리즘은 진행된다.

Step 1: 모수의 초기값 $(\boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\eta}^{(0)}, \sigma^{2(0)}, \boldsymbol{\pi}^{(0)})$ 을 설정하고, 잠재계층 변수 L_i ($i = 1, \dots, n$)를 $\boldsymbol{\pi}^{(0)}$ 를 토대로 생성한다.

Step 2: t 시점의 값이 $(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\eta}^{(t)}, \sigma^{2(t)}, \boldsymbol{\pi}^{(t)})$ 로 주어졌을 때, $t+1$ 시점의 값을 식 (3.3)-(3.7)을 이용하여 다음과 같이 순차적으로 업데이트한다.

- 식 (3.3)을 바탕으로 M-H 방법을 사용하여 $\boldsymbol{\alpha}$ 를 샘플링한다:

$$p(\boldsymbol{\alpha}^{(t+1)} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\eta}^{(t)}, \sigma^{2(t)}, \boldsymbol{\pi}^{(t)}, -)$$

- 식 (3.4)를 바탕으로 M-H 방법을 사용하여 $\boldsymbol{\beta}$ 를 샘플링한다:

$$p(\boldsymbol{\beta}^{(t+1)} | \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\eta}^{(t)}, \sigma^{2(t)}, \boldsymbol{\pi}^{(t)}, -)$$

- 식 (3.5)를 바탕으로 M-H 방법을 사용하여 $\boldsymbol{\eta}$ 를 샘플링한다:

$$p(\boldsymbol{\eta}^{(t+1)} | \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \sigma^{2(t)}, \boldsymbol{\pi}^{(t)}, -)$$

- 식 (3.6)을 바탕으로 Gibbs sampler 방법을 사용하여 σ^2 를 샘플링한다:

$$p(\sigma^{2(t+1)} | \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)}, \boldsymbol{\pi}^{(t)}, -)$$

- 식 (3.7)를 바탕으로 Gibbs sampler 방법을 사용하여 $\boldsymbol{\pi}$ 를 샘플링한다:

$$p(\boldsymbol{\pi}^{(t+1)} | \boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\eta}^{(t+1)}, \sigma^{2(t+1)}, -)$$

Step 3: 잠재계층 변수 L_i ($i = 1, \dots, n$)를 식 (3.8)을 바탕으로 M-H 방법을 사용하여 업데이트 한다.

Step 4: Step 2로 돌아가서 수렴할 때까지 반복한다.

일반적으로 메트로폴리스-해스팅스 알고리즘을 사용할 때, 추정값이 잘 수렴하지 않는 경우가 발생하기도 한다. 이를 해결하기 위해 본 논문에서는 분산조정 메트로폴리스 (adaptive Metropolis) 알고리즘을 사용하였다 (Haario 등, 2005). 즉, 제안분포 (proposal distribution)의 분산을 결정할 때, 경험적 (empirical) 분산과 조정계수를 사용하여 매 단계 조정된 제안분포를 통해서 후보값 (candidate value)을 고려한다. 예를 들어, $t+1$ 시점의 모수 $\theta^{(t+1)}$ 을 업데이트하기 위해서 후보값 θ^* 을 다음과 같이 결정한다.

$$\begin{aligned}\theta^* &\sim N(\theta^{(t)}, V^{(t)}), \\ V^{(t)} &= \begin{cases} V^{(0)}, & \text{if } t \leq t_0 \\ s\text{Var}(\theta^{(0)}, \dots, \theta^{(t-1)}) + s\epsilon, & \text{if } t > t_0, \end{cases}\end{aligned}$$

여기서 $V^{(0)}$ 는 모수 θ 에 대한 제안분포의 초기 분산값이고, s 는 후보값의 채택 비율 (acceptance rate)을 조정해주는 조정계수, 그리고 ϵ 은 아주 작은 상수값을 나타낸다. 알고리즘의 수렴 여부를 확인하기 위해서 trace plot과 더불어 Gelman과 Rubin의 potential scale reduction factor, \hat{R} 을 사용한다 (Gelman 등, 2014).

3.4. 모형의 비교

잠재계층모형에서 최적의 K 를 결정하기 위해서 모형 비교 분석을 시행한다. 베이지안 분석에서 주로 쓰이는 모형 비교 방법 중 하나인 DIC (deviance information criterion)는 다음과 같이 정의된다 (Spiegelhalter 등, 2002).

$$DIC = \overline{D(\theta)} + p_D,$$

여기서 $D(\theta)$ 는 모수의 편차 (Deviance)로서 $D(\theta) = -2\log f(y|\theta) + 2\log h(y)$ 로 정의되고, $\overline{D(\theta)}$ 는 편차의 사후평균을 나타낸다. 그리고 $p_D = \overline{D(\theta)} - D(\hat{\theta}) = E[D(\theta)|y] - D(E[\theta|y])$ 는 모형에서 사용된 모수의 수 또는 모형의 복잡한 정도에 대한 폐널티를 의미한다. 그러므로 여러 모형들 중에서 더 작은 DIC를 가지는 모형이 상대적으로 데이터를 더 잘 적합하고 있다고 말할 수 있다. 하지만, 잠재변수가 포함되어 있는 모형에서는 모수 θ 가 항상 식별 가능한 것은 아니기 때문에 위에서 정의된 DIC가 올바른 값을 제공하지 못하는 경우가 발생한다. 이를 보완하기 위해서 Celeux 등 (2006)은 잠재변수가 포함된 모형에 대해서 사용할 수 있는 수정된 DIC를 다음과 같이 제안하였다.

$$DIC_4 = -4E_{\theta, Z}[\log f(y, Z|\theta)|y] + 2E_Z[\log f(y, Z|E_{\theta}[\theta|y, Z])|y], \quad (3.9)$$

여기서 Z 는 직접 관찰되지 않는 잠재변수를 나타내고, 수정된 DIC (DIC_4)는 Z 를 포함한 완전데이터 가능도함수 (complete data likelihood function)를 바탕으로 계산된다. 본 논문에서는 잠재계층의 총 개수 K 를 다양한 값으로 고정시켜서 모형화한 후, 식 (3.9)에서 정의된 DIC (DIC_4)를 계산하여 비교 함으로써 최적의 K 를 결정한다.

4. LIFE 연구 자료의 분석

4.1. 자료의 탐색

이 장에서는 2장과 3장에서 제안된 잠재계층모형에 대한 베이지안 추론방법을 실제 자료에 적용하여 그 결과를 살펴보고자 한다. 이 장에서 사용할 LIFE 연구 (Longitudinal Investigation of Fertility and the Environment Study)에서는 임신을 시도하고 있는 총 501쌍의 커플들을 12개월 간 추적 조사하여 여성의 불임 여부를 조사하였다. 실제 결과를 얻어서 분석에 사용된 커플들의 수는 378쌍이고, 그 중에서 56명의 여성의 불임이라는 결과를 얻게 되었다. 또한 환경오염물질인 PCB177의 농축량은 99명 (약 26%)의 여성의 0또는 측정 불가일 정도로 너무 작아서 0으로 취급되는 값을 가지고 있다. Figure 2.1에서 반연속적인 PCB177 자료의 특성을 확인할 수 있다.

4.2. 베이지안 모형

베이지안 추론을 위해 각 모수에 대한 다음과 같은 무정보적인 (noninformative) 사전분포를 사용하여 사후 추정을 진행하였다. 일반적으로 무정보적이지만 적절한 (proper) 공액사전분포를 사용하면 적절한 사후분포를 얻을 수 있다.

$$\begin{aligned}\boldsymbol{\alpha} &\sim MVN(\mathbf{0}, 100I_K), \quad \boldsymbol{\beta} \sim MVN(\mathbf{0}, 100I_K), \quad \boldsymbol{\eta} \sim MVN(\mathbf{0}, 100I_2), \\ \sigma^2 &\sim IG(1, 1), \quad \boldsymbol{\pi} \sim Dirichlet(1, \dots, 1),\end{aligned}$$

여기서 K 는 잠재계층의 총개수를 나타낸다. 최적의 K 를 찾아내기 위해 $K = 2, 3, 4, 5$ 를 가정한 모형을 각각 적합시킨 후 모형 비교 방법인 DIC를 비교하였다. 모수 추정방법으로는 3.3장에서 소개한 대로 김스 샘플러와 분산조정 메트로폴리스 알고리즘을 결합한 MCMC 방법을 사용하였는데, 서로 다른 5개의 chain을 가정한 후 50,000번의 반복시행과 25,000번의 제거 (burn-in)를 통해 얻은 표본을 바탕으로 추정치를 계산하였다. 각 모수 추정치의 수렴여부를 판단하기 위해 trace plot과 Gelman과 Rubin의 \hat{R} 을 사용하였다.

4.3. 분석결과

MCMC 알고리즘을 통해 추출된 각 모수의 표본값들은 trace plot을 통해 확인한 결과 대체적으로 잘 수렴되고 있음을 알 수 있었다. 특히, 메트로폴리스 알고리즘을 사용할 때에는 수렴을 개선하기 위해 제안분포의 후보값의 채택비율 (acceptance rate)을 0.44로 조정한 분산조정 메트로폴리스 알고리즘을 사용하였다 (Gelman 등, 2014). Gelman과 Rubin의 \hat{R} 역시 1에 가깝게 나와서 수렴에 큰 문제가 없음을 확인할 수 있었다.

Table 4.1은 잠재계층의 총개수에 따른 각기 다른 네 개의 모형을 LIFE 자료에 적합시킨 결과로서 각 모수의 추정치와 95% 신뢰구간 및 DIC (Deviance Information Criterion)를 나타내고 있다. 먼저 DIC를 기준으로 모형을 비교해보면 3개의 잠재계층으로 이루어진 모형 (3-class model)이 가장 작은 값 (-1872.5)을 가지고 있어서 LIFE 자료를 가장 잘 설명해주고 있다는 것을 알 수 있다. 잠재계층의 수가 4개, 5개로 늘어날수록 DIC 값은 증가하고 있지만, 6개 이상의 모형들도 그런 패턴을 가진다고 확신할 수는 없다. 하지만, Table 4.1에 있는 잠재계층에 속할 확률 (π)의 추정값을 비교해보면 고위험군으로 갈수록 그 계층에 속할 확률이 0에 가까워지는 것을 확인할 수 있다. 예를 들어, 5개의 잠재계층 모형에서 $\pi_4 = 0.08, \pi_5 = 0.02$ 로서 고위험군에 속할 확률은 0에 가깝게 나타난다. 따라서 위험 계층을 더 세부적으로 나누는 것은 큰 의미가 없고, 또한 모형의 절약 (parsimonious model) 측면에서 봤을 때도 6개 이상의 잠재계층 모형을 고려할 필요없이 3개의 잠재계층 모형이 최적의 모형이라고 결론내릴 수 있다.

최적의 모형으로 선택된 3개의 잠재계층으로 이루어진 모형에 대한 구체적인 해석은 다음과 같다. 해석의 편의상 3개의 잠재계층을 저위험군, 중위험군, 고위험군으로 정한다면, 먼저 불임에 대한 오즈 (odds)는 저위험군보다 중위험군에서 약 19% ($=\exp(0.17)$) 정도 더 높게 나타났다. 반면에 고위험군에서는 저위험군에서보다 약 두배 가량 ($=\exp(0.69)$)이 더 높게 나타났다. 즉, 환경오염물질에 대한 잠재적인 위험도가 높아질수록 불임에 대한 오즈가 증가하고, 특히 중위험군에서 고위험군으로 바뀌면서 더 큰 증가량을 보이고 있음을 알 수 있다. 이는 Table 4.2에 나온 결과로도 뒷받침된다. Table 4.2는 각기 다른 네 개의 모형에서 각 잠재계층에서 발생하는 불임의 확률을 추정한 값을 보여주고 있다. 3개의 잠재계층으로 이루어진 모형에서 추정된 불임의 확률은 고위험군으로 갈수록 점점 높아지고 있고, 특히 고위험군으로 넘어가면서 훨씬 더 큰 증가량을 보여주고 있다. 이는 앞서 오즈의 해석과 같은 패턴

을 보여준다. 다른 모형에서도 비슷한 패턴을 보여주고 있다. 예를 들어, 4개의 잠재계층으로 이루어진 모형에서도 전체적으로 위험도가 높아질수록 불임의 확률이 증가하고, 특히 두 번째 위험군에서 세 번째 위험군으로 바뀔 때 더 높은 증가량을 보여주고 있으며, 네 번째 위험군으로 바뀔 때는 증가량이 감소하고 있음을 알 수 있다. 5개의 잠재계층으로 이루어진 모형도 같은 패턴을 보여주고 있다. Figure 4.1은 이러한 패턴을 더 명확하게 보여주고 있다. 위험도가 증가할수록 불임의 확률이 같은 비율로 증가하는 것이 아니라 어느 단계에서 급격한 증가를 보여주는 것이다. 즉 잠재적인 분계점 (threshold)이 존재한다고 추측할 수 있다.

또한 3개의 잠재계층으로 이루어진 모형에서 α 추정치를 살펴보면, 위험도가 증가할수록 0이 아닌 PCB 농축량의 로그 평균값은 증가하고 있다. 특히 저위험군에서 중위험군으로 바뀔 때보다 (-7.19→-5.70) 중위험군에서 고위험군으로 바뀔 때 (-5.70→-4.09) 더 많이 증가함을 알 수 있다. 2개, 4개, 5개의 잠재계층으로 이루어진 모형들도 비슷한 패턴을 가지고 있음을 알 수 있다. 그리고, η_1 값이 양의 추정치를 갖는다는 것은 PCB 농축량이 양의 값을 가질 확률 또한 고위험군으로 갈수록 증가한다는 것을 의미한다.

Table 4.1 Posterior means and 95% credible intervals for parameters in LIFE data analysis

Parameter	Estimate (C.I.)			
	2-class model	3-class model	4-class model	5-class model
α_0	-5.39(-5.62,-4.90)	-4.09(-4.17,-4.01)	-1.27(-1.31,-1.20)	-0.85(-0.91,-0.78)
α_1	0.48(-0.02,1.25)	-3.10(-3.13,-3.07)	-3.53(-3.57,-3.49)	-0.23(-0.27,-0.18)
α_2	-	-1.61(-1.65,-1.56)	-4.60(-4.66,-4.52)	0.16(-0.08,0.42)
α_3	-	-	-0.04(-0.09,-0.002)	-3.12(-3.51,-2.73)
α_4	-	-	-	-3.87(-3.94,-3.79)
β_0	-0.97(-1.02,-0.93)	-1.15(-1.31,-1.07)	-1.62(-1.69,-1.52)	-1.93(-2.02,-1.83)
β_1	-0.73(-0.78,-0.62)	-0.86(-0.93,-0.77)	-0.46(-0.49,-0.35)	-0.63(-0.74,-0.55)
β_2	-	-0.69(-0.76,-0.60)	-0.39(-0.46,-0.31)	-0.58(-0.65,-0.51)
β_3	-	-	-0.11(-0.18,-0.03)	-0.11(-0.21,0.02)
β_4	-	-	-	-0.05(-0.14,0.01)
η_0	-0.07(-1.92,1.74)	-0.05(-1.07,1.14)	-0.25(-1.73,1.33)	-0.92(-1.50,-0.31)
η_1	0.08(-0.07,0.24)	0.12(0.02,0.21)	0.34(0.13,0.55)	0.56(-1.07,0.18)
σ^2	0.32(0.25,0.40)	0.33(0.26,0.41)	0.30(0.22,0.39)	0.31(0.22,0.40)
π_1	0.04(0.005,0.11)	0.007(0.0002,0.01)	0.09(0.01,0.19)	0.01(0.001,0.07)
π_2	0.96(0.89,0.99)	0.96(0.92,0.99)	0.88(0.78,0.97)	0.02(0.0004,0.11)
π_3	-	0.03(0.007,0.08)	0.01(0.0003,0.05)	0.87(0.73,0.95)
π_4	-	-	0.01(0.0003,0.05)	0.08(0.01,0.16)
π_5	-	-	-	0.02(0.0004,0.09)
DIC	-1849.1	-1872.5	-1784.2	-1671.6

Table 4.2 Fitted probability of infertility by latent class models in LIFE data analysis

Class	Fitted Probability of Infertility			
	2-class model	3-class model	4-class model	5-class model
1	0.154	0.118	0.111	0.072
2	0.275	0.137	0.118	0.075
3	-	0.240	0.151	0.115
4	-	-	0.165	0.121
5	-	-	-	0.127

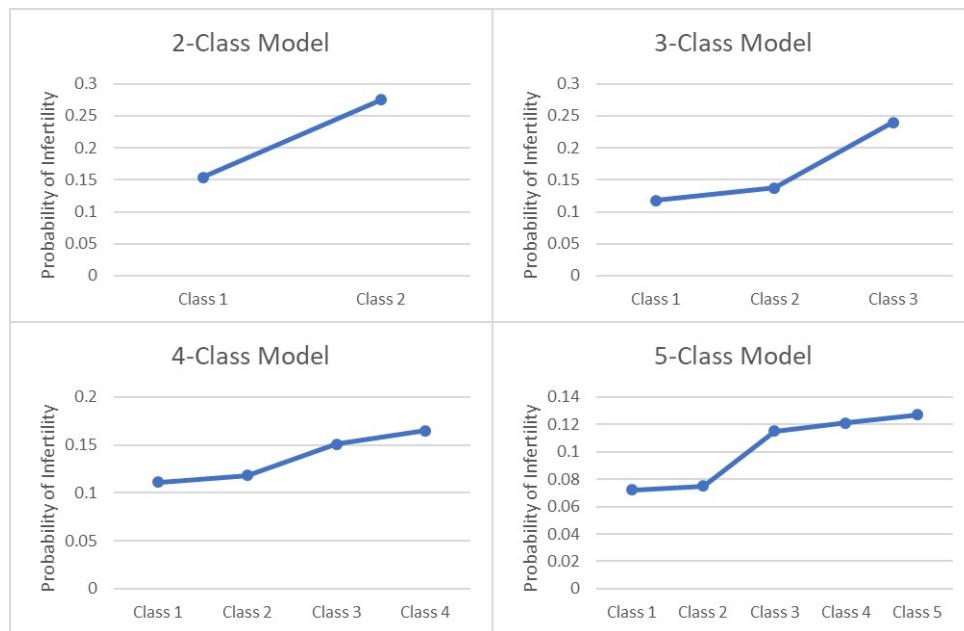


Figure 4.1 Fitted probability of infertility by latent class models in LIFE data analysis

5. 결론

본 논문에서는 LIFE 연구 (Longitudinal Investigation of Fertility and the Environment Study)에서 나온 자료를 토대로 환경오염물질인 PCB177과 여성의 불임 여부에 대해서 분석하였다. PCB177에 노출된 양과 여성 불임의 확률을 잠재계층 변수를 통해 연결시킨 후 베이지안 접근법을 사용하여 결과를 도출하였다. 잠재계층모형을 통해서 각 피해자가 환경오염물질에 반응하는 정도가 다르다는 가정하에 각기 다른 잠재적인 위험군에 속한 여성들의 불임의 정도가 서로 다른 특성을 지니고 있음을 확인하려고 하였다. 이때 PCB 농축량이 반연속적인 형태를 띠고 있어서 0의 값이 발생하는 부분과 0보다 큰 값에 대해서 각기 다른 분포를 가정한 혼합 분포 모형을 적용하였다. 또한 이러한 결합모형을 분석하는데 있어서 생기는 복잡한 계산을 베이지안 접근법을 사용하여 해결하려고 하였다.

분석 결과 환경오염물질에 대한 잠재계층이 바뀔 때 그 위험도가 높아질수록 불임에 대한 오즈가 증가하고 있음을 알 수 있었다. 하지만 증가하는 정도가 선형의 관계를 가지고 있지는 않았고, 어느 특정 계층에서 급격한 증가를 보여주었다. 로지스틱 함수를 통해서 계산된 각 잠재계층에서의 불임의 확률도 같은 패턴을 가지며 어느 특정 계층에서 급격한 증가를 보여줘서 잠재적인 분계점의 존재를 추측할 수 있게 해주었다. 마찬가지로 잠재계층의 위험도가 증가할수록 PCB 농축량이 양의 값을 가질 확률과 0이 아닌 PCB 농축량의 로그 평균값은 증가하며 그 증가량 또한 특정 계층을 지나면서 급격히 증가함을 확인할 수 있었다.

본 논문에서는 PCB177의 농축량과 불임 여부를 통해서 잠재계층을 구성하는 모형을 제시하였다. 하지만 두 개의 변수들 관계만으로 잠재계층을 구성하면 정보의 부족으로 모형이 제한적일 수 밖에 없다. 이때 피해자들의 공변량 (covariates)을 모형에 포함시킨다면 보다 많은 정보를 활용해서 잠재계층을 구성하는데 도움을 줄 수 있을 것이다. 즉, 여성의 불임에 영향을 줄 수 있는 나이, BMI, 흡연여부 등을 식 (2.1)이나 (2.5)에 공변량으로 모형에 추가하면 보다 정확한 분석이 될 수 있을 것이다. LIFE 연구

에서는 총 36개의 PCB를 고려하였지만 본 논문에서는 PCB177이라는 환경오염물질 하나만 모형에 적용하였다. PCB는 하나만으로도 인체에 영향을 미칠 수 있지만, 여러개의 PCB가 복합적인 상호작용에 의해서 더 큰 영향을 미칠 수도 있다. 따라서 두 개 이상의 PCB가 여성의 임신에 미치는 영향에 대해서 분석해 보는 것도 향후 과제 중 하나가 될 수 있을 것이다. 또한 잠재계층의 총개수가 늘어나면 지시변수의 수가 증가하기 때문에 모형이 복잡해져서 계산의 어려움을 느낄 수 있다. 이때 잠재계층을 순서형 자료 (ordinal data)로 인식하여 분석을 하면 계산의 복잡성을 감소시킬 수 있다. 다만 이런 경우 계층 간의 변화가 계층에 상관없이 일정해진다는 단점이 생겨서 이를 보완하기 위한 방법이 고안되어야 한다.

References

- Buck Louis, G. M., Schisterman, E. F., Sweeney, A. M., Wilcosky, T. C., Goce-Langton, R. E., Lynch, C. D., Barr, D. B., Schrader, S. M., Kim, S., Chen, Z. and Sundaram, R. (2011). Designing prospective cohort studies for assessing reproductive and developmental toxicity during sensitive windows of human reproduction and development - the LIFE Study. *Paediatric and Perinatal Epidemiology*, **25**, 413-424.
- Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006). Deviance information criterion for missing data models. *Bayesian Analysis*, **1**, 651-674.
- 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., Chen, Z., Buck Louis, G. M. and Albert, P. S. (2018). A Bayesian multi-dimensional couple-based latent risk model with an application to infertility, *Unpublished manuscript*.
- Kim, J., Bang, S. and Kwon, O. (2017). Analysis of scientific military training data using zero-inflated and Hurdle regression. *Journal of the Korean Data & Information Science Society*, **28**, 1511-1520.
- Lee, Y., Ha, Y. K., Cho, Y-S. and Kim, H. (2018). The latent classes and factors affecting quality of life among South Koreans with metabolic syndrome. *Journal of the Korean Data & Information Science Society*, **29**, 313-326.
- Lin, H., McCulloch, C. H., Turnbull, B. W. and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, **97**, 53-65.
- Liu, Y., Liu, L. and Zhou, J. (2015). Joint latent class model of survival and longitudinal data: An application to cprca study. *Computational Statistics and Data Analysis*, **91**, 40-50.
- McCutcheon, A. L. (1987). *Latent class analysis*, Sage Publications, California.
- Meeker, J. D., Maity, A., Missmer, S. A., Williams, P. L., Mahalingaiah, S., Ehrlich, S., Berry, K. F., Altshul, L., Perry, M. J., Cramer, D. W. and Hauser, R. (2011). Serum concentrations of polychlorinated biphenyls in relation to in vitro fertilization outcomes. *Environmental Health Perspectives*, **119**, 1010-1016.
- Neelon, B., O'Malley, A. J. and Normand, S.-L. T. (2011). A Bayesian two-part latent class model for longitudinal medical expenditure data: Assessing the impact of mental health and substance abuse parity. *Biometrics*, **67**, 280-289.
- 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.
- Zhang, B., Chen, Z. and Albert, P. S. (2012). Latent class models for joint analysis of disease prevalence and high-dimensional semicontinuous biomarker data. *Biostatistics*, **13**, 74-88.

A Bayesian latent class model for effect of environmental pollutants on female infertility[†]

Yoon Kyung Choi¹ · Beom Seuk Hwang²

^{1,2}Department of Applied Statistics, Chung-Ang University

Received 20 August 2018, revised 6 September 2018, accepted 10 September 2018

Abstract

We proposed a latent class model to examine the association between an environmental pollutant (PCB) and female infertility in the LIFE study. We assumed there exist latent risk groups of subjects and linked the PCB exposure and logit model for female infertility through the latent class variable. Also, semicontinuous PCB exposure was analyzed through a mixture of a degenerate distribution at zero and a continuous distribution for nonzero values. We took a Bayesian perspective to inference and used Markov chain Monte Carlo algorithms to obtain posterior estimates of model parameters. We calculated and compared DICs for all comparable models to find the most appropriate model for LIFE study data. As a result, we found that the risk of infertility was affected by latent risk groups of PCB exposure.

Keywords: Latent class model, LIFE study, Markov chain Monte Carlo, Metropolis algorithm, semicontinuous data.

[†] This research was supported by the Chung-Ang University Graduate Research Scholarship in 2016, and supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B03933334).

¹ Graduate student, Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea.

² Corresponding Author: Assistant professor, Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bshwang@cau.ac.kr