The Korean Journal of Applied Statistics
2022, Vol. 35, No. 6, 703–724

DOI: https://doi.org/10.5351/KJAS.2022.35.6.703
Print ISSN1225-066X / Online ISSN 2383-5818

KCYP data analysis using Bayesian multivariate linear model

Insun Lee[a], Keunbaik Lee[1,a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

Although longitudinal studies mainly produce multivariate longitudinal data, most of existing statistical models analyze univariate longitudinal data and there is a limitation to explain complex correlations properly. Therefore, this paper describes various methods of modeling the covariance matrix to explain the complex correlations. Among them, modified Cholesky decomposition, modified Cholesky block decomposition, and hypersphere decomposition are reviewed. In this paper, we review these methods and analyze Korean children and youth panel (KCYP) data are analyzed using the Bayesian method. The KCYP data are multivariate longitudinal data that have response variables: School adaptation, academic achievement, and dependence on mobile phones. Assuming that the correlation structure and the innovation standard deviation structure are different, several models are compared. For the most suitable model, all explanatory variables are significant for school adaptation, and academic achievement and only household income appears as insignificant variables when cell phone dependence is a response variable.

Keywords: Bayesian variable selection, KCYPS data, modified Cholesky decomposition, multivariate longitudinal data

1. 서론

경시적 자료란 한 개체에서 반복 측정된 자료이다. 따라서 이러한 형태의 자료를 분석하려면 반복 측정된 자료에 존재하는 상관관계를 올바르게 추정하면서 설명 변수의 효과를 분석해야 한다. 일반적으로 경시적 연구에서는 다변량 경시적 자료가 주로 생성된다. 하지만 기존의 통계적 모형은 대부분 다변량 경시적 자료를 단변량 경시적 자료로 나눠서 각각의 속성별로 분석하였고, 그 결과 다변량 경시적 자료에 존재하는 복잡한 상관관계를 제대로 설명하지 못하게 된다. 여기서 복잡한 상관관계란 다변량 경시적 자료는 여러 개의 속성이 반복 측정되는 자료이므로 하나의 반응 변수의 반복에서 일어나는 상관관계보다 더 복잡한 관계를 가진다는 의미이다. 그 복잡한 상관관계는 다음과 같다. 1) 다른 시점에서의 동일한 속성들간 상관관계, 2) 다른 시점에서 다른 속성들 간 상관관계, 3) 같은 시점에서의 서로 다른 속성들 간 상관관계이다. 따라서 다변량 경시적 자료를 단변량 경시적 자료로 나눠서 분석하는 것은 다른 시점에서의 동일한 속성들간 상관관계만을 설명하는 모형을 가지고 분석하는 것으므로 설명 변수의 반응 변수에 대한 효과의 추정에서 편향이 발생할 수 있다 (Kim과 Zimmerman, 2012). 이러한 편향을 피하기 위하여 앞서 제시한 복잡한 상관관계를 모두 고려할 수 있는 통계적 모형이 필요하다.

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 Insun Lee'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 31 December 2022 / journal homepage: http://kjas.or.kr

다변량 경시적 자료 분석에서 앞서 제시한 세 가지의 상관관계를 설명하는 공분산 행렬이 필요하며, 이 공분산 행렬을 모형화하기 위한 통계적 모형들이 제안되어 왔다. 우선, 단변량 경시적 자료 분석에서 공분산 행렬을 모형화하는 방법인 수정된 콜레스키 분해(modified Cholesky decomposition; MCD) (Pourahmadi, 1999)를 다변량 선형 모형으로 확장한 수정된 콜레스키 블록 분해(modified Cholesky block decomposition; MCBD)가 제안되었다 (Kim과 Zimmerman, 2012). 이는 두 번의 수정된 콜레스키 분해를 통해 공분산 행렬의 모수를 일반화 자기회귀모수(generalized autoregressive parameters; GARP)와 혁신 공분산 행렬(innovation covariance matrix; ICM)로 분해하며, 공분산 행렬의 이분산성(heteroscedasticity)과 양정치성(positive-definiteness)을 만족하게 만든다. 하지만 이 방법은 동일 시점의 반응 변수들 간의 공분산 행렬인 ICM에 대해 순서가 존재하지 않음에도 불구하고, 임의의 순서를 부여하는 자연스럽지 않은 가정이 필요하다. 이를 해결하기 위해 Lee 등 (2020)은 다변량 정규 분포를 이용한 선형 모형(multivariate normal linear modle; MNLM)을 제안하였다. 이 방법은 Kim과 Zimmerman (2012)과 동일하게 공분산 행렬을 일반화 자기회귀모수 행렬과 혁신 공분산 행렬로 분해하고, 혁신 공분산 행렬을 다시 혁신 표준편차 행렬과 상관계수 행렬로 분해한다. 이 상관계수 행렬을 초구분해(hypersphere decomposition; HD)를 적용하여 모형화하였다. 이는 모수에 대한 설명이 용이하고 다변량 경시적 자료에 대한 복잡한 상관관계를 모두 설명하며, 공분산 행렬의 양정치성과 이분산성을 만족하게 한다.

다변량 경시적 자료 분석에서 공분산 행렬은 항상 양정치성을 만족해야 하고, 그 차원이 고차원이기 때문에 모형화하기가 어렵다. 또한, 고차원의 공분산 행렬은 희박성(sparseness)을 가질 수 있는데, 이것은 차수가 높은 경우 행렬의 많은 성분들이 0에 가까운 값이 되는 것을 뜻한다. 앞서 기술한 방법들은 주로 정규화(regularization)를 통해 일반화 자기회귀모수의 개수를 감소시키지만 희박성 문제를 제대로 해결하지 못 한다. 따라서 최근 베이지안 변수 선택을 이용하여 일반화 자기회귀모수를 선택하는 방법이 제안되었다 (Lee 등, 2021). 일반화 자기회귀모수에 대해 스파이크와 슬랩 사전 분포(spike and slab prior)를 정의하여 이진 잠재변수(binary latent variable)를 통해 일반화 자기회귀모수의 중요한 변수를 선택한다. 이러한 방법을 통해 다변량 경시적 자료의 공분산 행렬의 희박성 문제와 복잡한 상관관계를 모두 설명할 수 있다. 이 방법의 우수성은 Kim과 Lee (2020)에서 모의실험을 통하여 제시되었다.

본 논문에서는 이러한 다변량 경시적 자료에 대해 베이지안 변수 선택을 이용하여 희박성을 가지는 공분산 행렬의 모형화와 양정치성 및 이분산성을 만족하는 모형을 이용하여 청소년 패널 데이터(Korean children and youth panel survey; KCYPS) 자료를 분석하고자 한다. 논문의 구성과 다음과 같다. 제 2절에서는 청소년 패널 데이터와 같은 다변량 경시적 자료 분석을 위해서 필요한 통계적 모형에 대한 문헌 고찰을 한다. 제 3절에서는 청소년 패널 데이터에 관한 자세한 설명과 기초적인 자료의 분석을 시행하며, 선행 연구를 기반으로 청소년 패널 데이터를 분석한다. 제 4절에서는 다변량 경시적 자료인 청소년 패널 데이터의 특성을 고려하여 베이지안 방법을 이용하여 분석한다. 마지막으로 5절에서는 앞서 분석한 내용을 요약한다.

2. 다변량 선형 모형

다변량 경시적 자료에서 공분산 행렬은 고차원이며, 양정치성을 만족해야 하므로 이를 모형화하는 것에 대해 살펴보고자 한다. 우선 다변량 경시적 자료 분석을 위해 기초가 되는 단변량 경시적 자료 분석에 수정된 콜레스키 분해 방법 (Pourahmadi, 1999)이 제안되었다. 따라서 이번 절에서는 단변량 경시적 자료의 공분산 행렬 모형화를 위한 수정된 콜레스키 분해와 다변량 경시적 자료의 공분산 행렬 모형화를 위한 수정된 콜레스키 블록분해 (Kim과 Zimmerman, 2012)를 알아본다. 나아가 다변량 경시적 자료의 상관계수 행렬 모형화에 대한 초구분해 (Lee 등, 2012)와 베이지안 방법을 이용하여 공분산 행렬의 중요한 모수를 결정하고 추정하는 방법 (Lee 등, 2021)을 살펴본다.

2.1. 수정된 콜레스키 분해(MCD)

우선 단변량 경시적 자료 분석에 제안된 수정된 콜레스키 분해 방법을 살펴본다. 경시적 자료는 여러 시점에 대해 반복하여 측정된 자료로 공분산 행렬이 고차원이며, 양정치성을 만족해야 한다. 이를 만족시키기 위해 수정된 콜레스키 분해를 이용한다. $Y_i = (Y_{i1}, \ldots, Y_{in_i})$는 i 번째 개체의 반응 변수 벡터로, Y_{it}는 t 번째 시점의 반응 변수이다. Y_{it}에 대해 이전 반응 변수들을 설명 변수로 하는 회귀식을 가정한다.

$$\begin{aligned} y_{i1} &= x_{i1}^T\beta + e_{i1}, \\ y_{it} &= x_{it}^T\beta + \sum_{j=1}^{t-1}\phi_{itj}\left(y_{ij} - x_{ij}^T\beta\right) + e_{it}, \\ e_i &= \left(e_{i1}, e_{i2}, \ldots, e_{in_i}\right)^T \sim N(0, D_i). \end{aligned} \tag{2.1}$$

여기서 ϕ_{itj}는 일반화 자기회귀모수(generalized autoregressive parameter; GARP)로 i 번째 개체의 현재 시점 t에서 j 번째 이전 시점의 반응 변수간 상관관계를 나타내며, 이것은 현재에 영향을 미치는 과거에 대한 효과를 나타내는 모수이다. 오차항의 공분산 행렬인 $D_i = \mathrm{diag}(\sigma_{it}^2, \ldots, \sigma_{in_i}^2)$의 t 번째 주대각 원소인 σ_{it}^2를 혁신분산(innovation variance; IV)이라 한다.

식 (2.1)을 다음과 같은 벡터와 행렬 형태로 쓸 수 있다.

$$T_i\left(Y_i - X_i^T\beta\right) = e_i. \tag{2.2}$$

여기서

$$T_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_{i21} & 1 & 0 & \cdots & 0 \\ -\phi_{i31} & -\phi_{321} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\phi_{in_i1} & -\phi_{in_i2} & -\phi_{in_i,3} & \cdots & 1 \end{pmatrix}, \quad X_i = \begin{pmatrix} x_{i1}^T \\ x_{i2}^T \\ \vdots \\ x_{in_i}^T \end{pmatrix}$$

이다. 식(2.2)에 분산을 취하면 공분산 행렬은 다음과 같이 분해 된다.

$$\begin{aligned} T_i\Sigma T_i^T &= D_i, \\ \Sigma_i = T_i^{-1}D_iT_i^{-T} &\iff \Sigma_i^{-1} = T_i^TD_i^{-1}T_i. \end{aligned}$$

이 결과로부터 반응 변수의 공분산 행렬을 GARP를 가지는 T_i와 혁신분산 (IV)을 가지는 D_i로 분해하여 모형화 할 수 있다. 행렬 T_i는 역행렬이 존재하고, 대각 행렬 D_i의 주대각 요소인 σ_{it}^2이 모두 양수이므로 공분산 행렬 Σ_i의 양정치성을 만족하게 된다.

MCD를 이용하여 공분산 행렬 Σ_i를 행렬 T_i의 요소인 GARP와 D_i의 주대각 요소인 σ_{it}^2로 분해하였지만 그 모수의 수가 많음을 알 수 있다. 모수의 수를 줄이면서 공분 행렬의 이분산성을 만족시키기 위해서 다음의 선형 및 로그선형 모형을 가정한다.

$$\begin{aligned} \phi_{i,tj} &= \omega_{i,tj}^T\alpha, \\ \log\left(\sigma_{it}^2\right) &= h_{it}^T\lambda. \end{aligned} \tag{2.3}$$

여기서 $w_{i,tj}$는 시간과 관련된 $a \times 1$ 차원의 공변량이며, h_{it}는 $b \times 1$ 차원의 공변량으로 α와 λ는 모르는 모수 벡터이다. 이와 같은 재모수화(reparameterization)를 통해 모수를 줄일 수 있다. GARP는 아무런 제약도 없으며 IV는 로그선형 모형을 통해 항상 양수가 되므로 자유로운 모형화가 가능하다.

2.2. 수정된 콜레스키 블록분해(MCBD)

수정된 콜레스키 분해(MCD)는 단변량 경시적 자료에 대한 분석 방법으로 같은 반응 변수의 다른 시점에 대한 상관관계만 고려한다. 따라서 다변량으로 확장할 경우, 더욱 복잡한 상관관계를 고려한 모형화가 필요하다. 같은 시점에 대한 상관관계와 다른 시점에 대한 상관관계를 모두 고려하기 위해 Kim과 Zimmerman (2012)은 수정된 콜레스키 블록분해(MCBD)를 제안하였다.

$y_i = (y_{i1}^T, \ldots, y_{in_i}^T)^T$는 i 번째 개체의 반응 변수 벡터이며 $y_{it}^T = (y_{it1}, \ldots, y_{itK})$는 t 번째 ($t = 1, 2, \ldots, n_i$) 시점에서 K개 속성의 반응 변수 벡터이다. 이 경우 다음의 모형을 가정한다.

$$\begin{aligned} y_{i1k} &= x_{i1}^T \beta_k + e_{i1k}, \\ y_{itk} &= x_{it}^T \beta_k + \sum_{j=1}^{t-1} \sum_{g=1}^{K} \phi_{itj,kg} \left(y_{ijg} - x_{ij}^T \beta_g \right) + e_{itk}, \\ e_i &= \left(e_{i1}, \ldots, e_{in_i} \right)^T \sim N(0, D_i). \end{aligned}$$

여기서 x_{it}는 $p \times 1$ 차원의 공변량 벡터이며, β_k는 k 번째 반응 변수에 관한 $p \times 1$ 차원의 모르는 모수 벡터이며, $e_{it} = (e_{it1}, \ldots, e_{itK})^T$이다. D_{it}를 혁신 공분산 행렬(innovation covariance matrix; ICM)이며 다음과 같다.

$$D_i = \operatorname{diag}\left(\operatorname{var}(e_{i1}), \ldots, \operatorname{var}(e_{it}) \right) = \operatorname{diag}\left(D_{i1}, \ldots, D_{it} \right).$$

단변량의 경우와 마찬가지로 위 식을 행렬과 벡터로 나타내면 다음과 같다.

$$T_i (Y_i - X_i^T \beta) = e_i. \tag{2.4}$$

여기서 T_i와 X_i는 다음과 같다.

$$T_i = \begin{pmatrix} I & 0 & \cdots & 0 \\ -\Phi_{i21} & I & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ -\Phi_{in_i 1} & -\Phi_{in_i 2} & \cdots & I \end{pmatrix}, \quad X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{in_i} \end{pmatrix}.$$

여기서 ϕ_{itj}는 다음과 같다.

$$\Phi_{itj} = \begin{pmatrix} \phi_{itj,11} & \cdots & \phi_{itj,1K} \\ \phi_{itj,21} & \cdots & \phi_{itj,2K} \\ \vdots & \ddots & \vdots \\ \phi_{itj,K1} & \cdots & \phi_{itj,KK} \end{pmatrix}.$$

Φ_{itj}는 일반화 자기회귀모수 행렬(generalized autoregressive parameter matrix; GARPM)라고 하며, 이것은 서로 다른 시점의 반응 변수 간 상관관계를 설명한다. 즉, 과거 시점의 반응 변수 값이 현재 시점의 반응 변수 값에 미치는 영향으로 다른 시점에서의 동일한 속성들 간의 상관관계와 다른 시점에서 다른 속성들 간의 상관관계를 나타낸다.

식 (2.3)과 동일하게 GARP는 다음의 선형 모형을 가정한다.

$$\phi_{itj,kg} = w_{itj}^T \alpha_{kg}. \tag{2.5}$$

여기서 w_{itj}는 식 (2.3)에 있는 w_{itj}와 동일하며, α_{kg}는 $a \times 1$의 모르는 모수 벡터이다.

식 (2.4)의 양변에 분산을 취하면 다음과 같은 식이 된다.

$$\Sigma_i = T_i^{-1} D_i T_i^{-T} \iff \Sigma_i^{-1} = T_i^T D_i^{-1} T_i.$$

따라서 MCBD는 공분산 행렬을 일반화 자기회귀모수 행렬과 혁신 공분산 행렬로 분해한다. GARPM의 요소들은 어떠한 제약 조건도 없으며, ICM이 양정치성을 만족하면 반응 변수의 공분산 행렬의 양정치성도 만족한다. 혁신 공분산 행렬의 양정치성을 만족시키기 위해 수정된 콜레스키 분해(MCD)를 한 번 더 적용하여 이 행렬을 다음과 같이 분해한다.

$$\overline{D_{it}} = \overline{T_{it}} D_{it} \overline{T_{it}}^T, \quad t = 1, 2, \ldots, n_i.$$

여기서 $\overline{T_{it}}$는 주대각 성분이 1이고 $\tau_{it,kl}$를 하삼각 성분으로 갖는 하삼각 행렬이며, $\overline{D_{it}}$=diag$(\delta_{it1}, \delta_{it2}, \ldots, \delta_{itK})$이다. 양정치성을 만족시키기 위해 $\log(\overline{D_{it}})$를 가정하고, D_{it}에 수정된 콜레스키 분해를 한 번 더 적용한 공분산 행렬은 다음과 같은 식이 된다.

$$\Sigma_i = T^{-1} \overline{T_i}^{-1} \left[\exp\left(\log \overline{D_i} \right) \right] \left(\overline{T_i} \right)^{-T} T_i^{-T}.$$

여기서 T_i, $\overline{T_i}$, $\log(\overline{D_{it}})$는 아무런 제약이 없다. 따라서 GARP에 대한 $\overline{T_i}$은 아무런 제약 조건이 없으므로 선형 모형을 적용할 수 있다. $\log(\overline{D_{it}})$의 대각 성분 δ_{itk} 또한 로그선형 모형을 적용하여 모형화할 수 있다.

$$\log \delta_{itk} = z_{it}^T \lambda_k,$$
$$\tau_{it,kl} = v_{it}^T \psi_{kl}.$$

여기서 z_{it}와 v_{it}는 시간에 의존하는 공변량 벡터이고, λ_k와 ψ_{kl}는 알려지지 않은 모수 벡터이다. 이러한 선형 모형을 이용한 모형화를 통해 추정해야 하는 모수의 개수를 줄일 수 있다.

MCBD에서는 D_{it}를 모형화할 때, 같은 시점에서 반응 변수들의 순서가 존재하지 않음에도 불구하고 반응 변수 간에 임의로 순서를 부여해야 한다는 한계가 있다. 이러한 한계를 극복하기 위해 Lee 등 (2020)은 초구분해를 제안하였다.

2.3. 초구분해(HD)

Lee 등 (2020)이 제안한 초구분해는 같은 시점에서 반응 변수 간에 불필요한 순서를 부여하는 대신에 D_{it}를 분산-공분산분해를 이용하여 분해한다.

$$D_{it} = S_{it} R_i S_{it}.$$

여기서 $S_{it} = \text{diag}\{\sigma_{it1}, \ldots, \sigma_{itK}\}$는 혁신 표준 편차(ISD)들을 주대각 원소로 가지는 대각 행렬이며, R_i는 ρ_{ilm} $(l \neq m = 1, \ldots, K)$로 이루어진 $K \times K$ 반응 변수들에 대한 상관계수 행렬이다. R_i도 S_{it}처럼 시간에 의존하도록 R_{it}로 모형화 할 수도 있으나, Lee 등 (2020)은 단순화를 위해 시간에 따라 변하지 않는다고 가정한다. 이를 통해 MCBD와 달리 반응 변수의 순서를 부여하지 않아도 공분산 행렬의 모형화가 가능하다.

D_{it}를 혁신 표준 편차(ISD)와 상관계수 행렬 R_i로 분해하였을 때, R_i는 상관계수 행렬로 대각 원소는 1, 대각 원소를 제외한 다른 원소들은 −1과 1사이의 값을 가져야 한다. 따라서 양정치성을 만족시키기 어려우므로 이를 해결하기기 위해 초구분해가 제안된다. 상관계수 행렬은 초구분해를 이용하여 다음과 같이 분해한다.

$$R_i = F_i F_i^T.$$

여기서 F_i는 다음과 같이 정의된 하삼각 행렬이다.

$$F_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ f_{i21} & f_{i22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{iK1} & f_{iK2} & f_{iK3} & \cdots & f_{iKK} \end{pmatrix}.$$

F_i의 각 요소는 다음의 식을 따른다.

$$f_{ilm} = \begin{cases} \cos(\omega_{ilm}), & \text{for } m = 1,\ l = 2, \ldots, K; \\ \cos(\omega_{ilm}) \prod_{r=1}^{m-1} \sin(\omega_{ilr}), & \text{for } 2 \le m < l \le K; \\ \prod_{r=1}^{m-1} \sin(\omega_{ilr}), & \text{for } l = m;\ m = 2, \ldots, K. \end{cases}$$

이때 $\omega_{ilm} \in (0, \pi)$이다.

혁신 표준 편차(ISD)를 로그선형 모형화하면 어떠한 제약 조건도 없으며, 추정해야 하는 모수의 개수를 줄일 수 있다. 그리고 R_i의 ω_{ilm} 또한 선형 모형화하면 다음과 같이 제약이 없는 모수 형태로 만들 수 있다. 그 모형은 다음과 같다.

$$\log(\sigma_{itk}) = h_{it}^T \lambda_k, \tag{2.6}$$

$$\log\left(\frac{\omega_{ilm}}{\pi - \omega_{ilm}}\right) = g_{ilm}^T \nu. \tag{2.7}$$

여기서 h_{it}는 시간 또는 개체 특징적 공변량 벡터, g_{ilm}은 단위 특징적 공변량 벡터이다. λ_k와 ν는 알려지지 않은 모수 벡터이다.

초구분해를 통해 R_i의 양정치성을 만족시키며, 로그선형 모형화를 통하여 모든 σ_{tik}의 양수를 만족시킴으로써 D_{it}의 양정치성을 만족시킨다. 따라서 결국에는 공분산 행렬의 양정치성을 만족하게 된다. 즉 GARP를 통해 서로 다른 시점의 반응 변수 간 상관관계를 설명하고, ICM에 MCD를 적용하는 대신 분산-공분산 분해와 초구분해를 적용시켜 동일한 시점의 반응 변수 간 상관관계를 설명한다. 따라서 다변량 경시적 자료의 세 가지 상관관계를 모두 설명할 수 있으며, 반응 변수에 임의로 순서를 부여하지 않아도 되기 때문에 자연스러운 모형화가 가능하다.

2.4. 베이지안 방법을 이용한 모수 추정과 변수 선택

앞서 살펴본 방법들은 모수에 대한 설명이 용이하고 다변량 경시적 자료에 대한 복잡한 상관관계를 모두 설명하며, 공분산 행렬의 양정치성과 이분산성을 만족하게 한다. 하지만 고차원의 공분산 행렬은 GARP가 급증하며 희박성의 문제를 가질 수 있다. 정규화를 통해 일반화 자기회귀모수의 개수를 감소시킬 수 있지만 희박성 문제를 해결하지 못 한다. 따라서 이를 해결하기 위해 최근 베이지안 변수 선택을 이용하여 일반화 자기회귀모수를 선택하는 방법이 제안되었다 (Lee 등, 2021).

Lee 등 (2021)은 반응 변수들 간 상관관계를 고려하기 위해 스파이크와 슬랩 사전 분포(spike and slab prior)의 이진 잠재 변수(binary latent variable)를 사용하며, 모형 추정을 위해 MCMC 알고리즘을 사용한다. 일반화 자기회귀모수에 대해 스파이크와 슬랩 사전 분포를 정의하여 이진 잠재 변수(binary latent variable)를 통해 일반화 자기회귀모수의 중요한 변수를 선택한다. 이를 통해 밴딩(banding)과 테이퍼링(tapering) 방법과는 다르게 GARP 수를 줄이고, 모수 선택의 불확실성을 해결할 수 있다.

베이지안 방법론에 따라, 모수들의 사전 분포가 다음과 같이 정의된다.

$$\begin{aligned}\beta &\sim N\left(0, \sigma_\beta^2 I\right),\\ \lambda_k &\sim N\left(0, \sigma_\lambda^2 I\right),\\ \nu &\sim N\left(0, \sigma_\nu^2 I\right).\end{aligned}$$

여기서 σ_β^2 ,σ_λ^2 그리고 σ_ν^2은 정해진 큰 값이다. β는 반응 변수에 대한 회귀 계수, λ_k와 ν는 식 (2.6)과 (2.7)에 있는 혁신 표준 편차와 초구분해의 모수이며, 계산의 편의를 위해 사전 분포들을 독립이라 가정한다. 따라서 β, λ_k, ν는 서로 독립이라고 가정할 수 있다.

SSVS 방법을 적용하여 이진 잠재 변수를 선택한다. 식(2.5)에 있는 상관관계에 대한 모수 α_{kgm}은 $\delta_{kgm} = 1$ 이면 $\alpha_{kgm} \neq 0$, $\delta_{kgm} = 0$이면 $\alpha_{kgm} = 0$ 으로 δ_{kgm}에 의해 결정된다.

$$\begin{aligned}\alpha_{kgm} &\sim \delta_{kgm} N\left(0, \sigma_\delta^2\right) + \left(1 - \delta_{kgm}\right)\eta_0,\\ \delta_{kgm} &\overset{indep.}{\sim} \text{Bernoulli}\left(\tilde{p}_{kgm}\right).\end{aligned}$$

여기서 σ_δ^2는 정해진 값이며, η_0는 0에 대한 확률 질량이다.

스파이크와 슬랩 사전 분포를 이용하면, 밴딩(banding)의 방법과 다르게 δ_{kgm}에 의해서 정규 분포의 형태를 띄는 α_{kgm}가 업데이트되며, 사전 분포는 GARP의 불확실성을 포함하므로 행렬의 희박성을 설명할 수 있다. 따라서 GARP에 대한 모수인 α_{kgm}를 선택하기 위해 δ_{kgm}을 추정해야 한다. 사후 분포를 통해 $\alpha_{kgm} \neq 0$ 이기 위해 $\delta_{kgm} = 1$일 확률과 이에 따라 결정되는 α_{kgm}의 기댓값을 추정한다. 이는 몬테칼로 방법(Monte Carlo method)을 사용하여 근사시키기 위해 몬테칼로 마르코프 연쇄(Monte Carlo Markov Chain; MCMC) 표집법을 시행한다. 이상의 알고리즘은 R package 'MLModelSelection'에서 구현되어 있으며 더 자세한 MCMC 방법은 Lee 등 (2021)을 참고하기 바란다.

3. 자료 분석

3.1. 청소년 패널 데이터(KCYP) 설명

본 논문에서 분석하는 청소년 패널(Korean children and youth panel; KCYP) 데이터는 다변량 형태의 경시적 자료로서, 2010년부터 2016년까지 2010년 기준 전국의 초등학교 4학년과 중학교 1학년 재학생을 표집하여 청소년의 성장 양상을 파악하고, 청소년 관련 정책을 수립 및 시행하기 위한 목적으로 수집된 것이다. 청소년 패널 데이터는 신체적, 정신적, 사회적으로 발달과 변화를 경험하는 청소년기에 청소년들이 올바르게 자립할 수 있도록 돕기 위해 조사되었으며, 청소년 조사 영역과 보호자 조사 영역으로 나눠져 있다. 청소년 조사 영역에는 개인 발달, 발달 환경으로 생활 시간, 진로, 사회/정서/역량발달 등의 정보가 포함되어 있으며, 보호자 조사 영역에는 청소년 발달 배경, 보호자 개인 영역으로 가정과 교육에 대한 정보가 포함되어 있다. 본 논문은 2011년부터 2015년까지의 자료를 분석대상으로 2011년 기준 중학교 1학년이 고등학교 2학년으로 성장할 때까지를 기준으로 분석하였다. 이 분석에서 첫 해인 2011년에 2,280명이 그리고 마지막 해인 2015년에는 1,927명이 이 조사를 마친 상태이며, 2015년의 기준으로 15%(353명)의 중도 탈락(dropout)이 있었다. 다른 많은 선행 연구에서 이 자료 분석을 무작위 결측(missing at random; MAR)으로 가정하였다. 무작위 결측은 이전에 관측된 변수의 값에 따라 중도 탈락 확률이 달라지는 결측형태이다. 이 논문에서 사용할 다변량 선형 모형을 위한 R 함수는 MAR가정 하에서 작용하기에 우리는 이 결측들을 MAR로 가정하여 분석하고자 한다.

반응 변수들로는 학업 성취도, 학교 적응력, 휴대전화 의존도를 고려하였다. 중학생의 휴대전화 의존도, 부모 양육 태도, 또래 애착이 학교 생활 적응에 미치는 영향 (Chang 등, 2011) 등과 같은 여러 선행 연구를 고려해 학교 적응도와 학업 성취도, 휴대전화 의존도 간의 상관성을 알아보기 위해 반응 변수로 채택하였다. 학교

Table 1: Summary for variables

Factor	Variable	Mean (Std.error)/Proprotion (Count)
Gender	Male (= 1)	0.51 (5351)
	Female (= 0)	0.49 (5156)
House	Income	8.290 (0.644)
Time	Sleep time	6.921 (1.278)
	Reading time	0.492 (0.743)
	Private education time	1.192 (1.437)
Parent factor	Neglect	7.495 (2.129)
	Abuse	6.881 (2.645)
Response variable	School adaptation	14.1 (4.160)
	Academic achievement	58.19 (4.160)
	Phone dependence	16.43 (4.868)

적응도는 학교 적응에 관한 설문 학습 활동, 학교 규칙, 교우 관계, 교사 관계 등을 모두 고려하여 측정되었으며 점수가 높을수록 학교 적응을 잘 하는 것을 의미한다. 학업 성취도는 학생 본인이 국어, 영어, 수학, 과학, 사회 성적을 주관적으로 평가한 문항으로 측정된다. 점수가 높을수록 학업 성취 결과에 대한 주관적인 평가가 높은 것을 의미한다. 휴대전화 의존도는 Chang 등 (2011)의 선행 연구를 참조하여 청소년 패널 조사 설문지의 휴대전화 의존도와 관련된 문항을 통하여 휴대전화 의존도를 측정하였다. 실제 문항의 답변은 1점을 '매우 그렇다', 4점을 '전혀 그렇지 않다'의 4점 척도로 측정되었으나, '점점 더 많은 시간을 휴대전화를 사용하며 보내게 된다', '휴대전화로 한참 동안 아무에게서도 연락이 오지 않으면 불안하다' 등과 같은 문항은 역코딩을 통해 점수가 높을수록 휴대전화 의존도가 높은 의미를 갖도록 하였다. 설명 변수는 관련 선행 연구들을 참고하여 유의미한 변수로 성별, 가정의 소득, 부모의 양육 방식, 사교육 시간, 독서 시간, 수면 시간을 고려하였다. 수면 시간은 등교일을 기준으로 주말과 공휴일을 제외한 월요일부터 금요일까지로 설정하였으며, 가정의 수익은 로그를 취한 값을 사용하여 분석을 진행하였다.

3.2. 선행 연구

청소년 발달에는 신체적, 정신적, 사회적으로 많은 요인에 의해 영향을 받는다. 그 중 성적은 청소년 성장에 큰 영향을 주는 것으로 나타난다. 성적이 학습 흥미와 우울감, 학교 적응도와 더 나아가 자아 개념 형성에까지 영향을 주는 것으로 나타난다 (Yoo, 1996). 학업 스트레스는 학교 생활의 부적응으로 이어질 수 있으며, 학교 적응도를 결정하는 요소가 될 수 있다. 또한 반대로 학교 적응도는 학업 성취도에 중요한 요소가 된다 (Kim, 2016). 한편 휴대전화 의존도의 문제가 청소년 발달의 새로운 문제로 부각되고 있으며, 휴대전화 의존도의 개념에는 청소년이 통제력을 상실하여 학교 생활에 적응하지 못하는 내용이 포함되어 있기 때문에 학교 적응도와 학업 성취도와 같은 학습 활동과 관련이 있는 것으로 나타난다 (Jang과 Cho, 2010). 이러한 선행 연구를 바탕으로 학교 적응도, 학업 성취도, 휴대전화 의존도 간의 종단적 변화에 어떤 요소가 영향을 미치는지 살펴보고자 한다. 이를 고려할 때 반응 변수 학교 적응도, 학업 성취도, 휴대전화 의존도를 단변량이 아닌 다변량의 형태로 구성하여 반응 변수들 간의 상관관계와 반응 변수에 대한 설명 변수들의 효과를 모두 모형화할 필요성이 존재한다. 또한, 청소년 패널 데이터는 한 학생의 성장 과정을 설문하는 자료로 경시적 구조를 고려해야하기 때문에 단변량 횡단 분석의 한계를 가진 선행 연구들을 고려한 다변량 경시적 자료 분석의 필요성이 요구된다. 이러한 다변량 경시적 자료 분석을 통하여 반응 변수들 간의 상관관계와 반응 변수와 설명 변수들 간의 상관관계를 모두 고려할 수 있을 것이다.

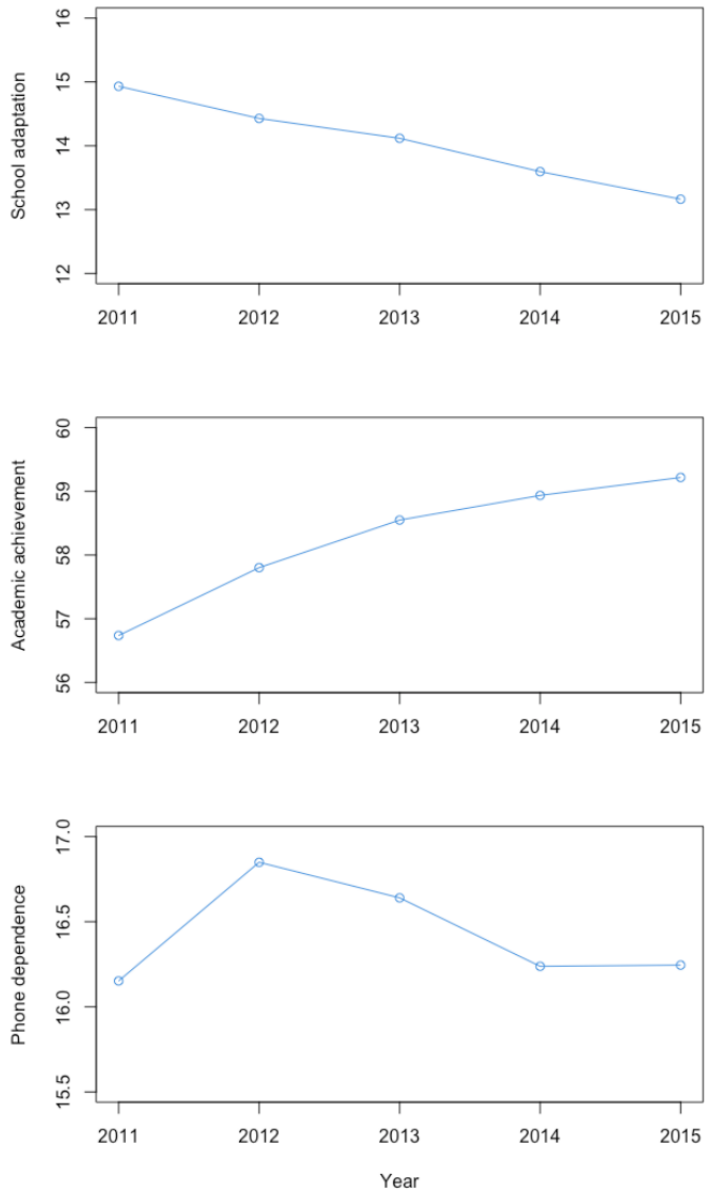

Figure 1: *The plots for response variables over year.*

3.3. 기초 분석

Lee 등 (2021)의 베이지안 방법을 적용하기 전에 본 패널자료의 설명 변수와 반응 변수들의 기초통계량을 요약하면 다음과 같다. 2011년부터 2015년까지 조사된 청소년 패널 데이터의 변수들에 대해 평균과 표준편차, 비율 등 기초통계량을 요약하여 Table 1과 같다.

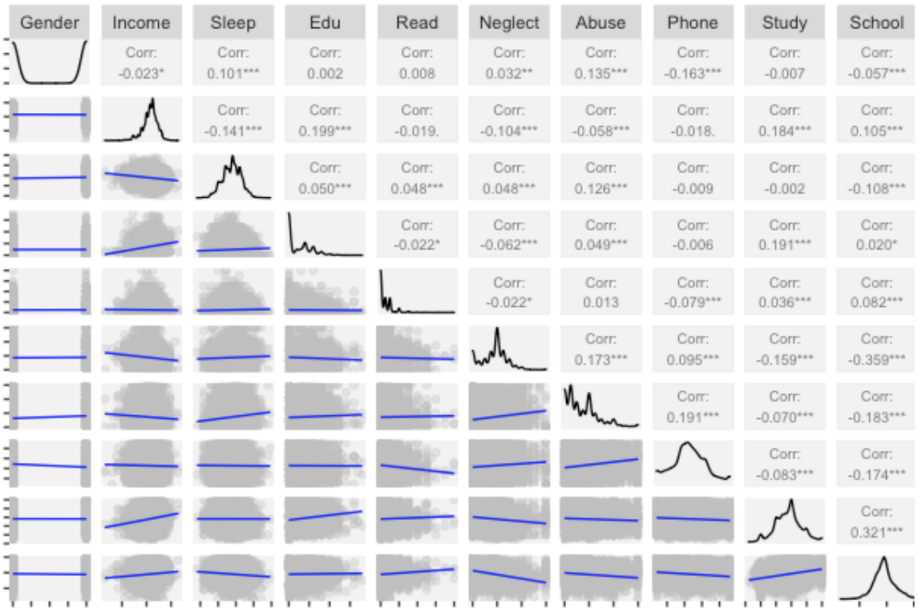

Figure 2: *The scatter plots for all variables.*

MAR dropout의 결측 비율이 15%로 5 년간 꾸준히 응답한 학생의 비율이 약 85%로 나타난다. 학생의 성별의 경우 남학생의 비율이 약 51%, 여학생의 비율이 약 49%로 나타났다. 수면 시간은 평균 약 6.9시간이고, 평균 독서 시간은 약 0.49시간으로 한 시간이 되지 않는 것으로 나타나지만 반응 변수들과 상관관계가 있는 것으로 나타난다. Table 1과 Figure 1을 통해 학교 적응도는 학년이 증가할수록 감소하며, 학업 성취도는 증가하는 것을 알 수 있다. 휴대전화 의존도의 평균은 중학교 2학년 (2012년)까지 증가하며 고등학교 1학년 (2013년)부터는 감소하는 추세를 보인다. 다음으로, Figure 2는 반응 변수 학교 적응도, 학업 성취도, 휴대전화 의존도와 설명 변수의 상관관계를 나타낸 것이다. 그리고 Table 2에서 각 반응변수의 연도별 표본평균과 더불어 표본 표준편차를 나타내고 있다. 세 반응 변수들 간 상관관계가 존재하며, 설명 변수와도 약한 상관관계가 존재한다는 것을 알 수 있다. 반응 변수들 중 학교 적응도와 학업 성취도는 약한 양의 상관관계, 학업 성취도와 휴대전화 의존도는 약한 음의 상관관계가 있는 것을 각 Figure 3와 Figure 4를 통해 알 수 있다. 또한 시차가 증가함에 따라 상관 계수가 줄어드는 것이 확인 되어 공분산 행렬을 고려한 모형이 필요하다. 따라서 이는 다변량 경시적 분석의 필요성이 있다는 것을 의미한다. 다변량 경시적 분석을 시작하기 전 분석에 대한 기초 조사를 실시하였다.

선행 연구를 기반으로 한 시점 (2015년 기준)을 대상으로 횡단적 분석을 시행해 반응 변수 별로 유의미한 변수를 확인하기 위해 선행 연구를 기반으로 회귀 분석을 실시하였다. 그 결과 반응 변수에 대한 설명 변수들 중 5% 유의수준에서 유의한 변수들을 확인하여 모형에 포함하였다. Table 3의 회귀 분석 결과 학교 적응도에는 모든 설명 변수가 5% 유의수준에서 매우 유의미한 것을 알 수 있다. 학업 성취도에서는 성별과 사교육 시간, 부모의 양육 방식 중 학대를 제외한 모든 설명 변수가 유의수준 5%에서 유의미한 것을 확인할 수 있다. 휴대전화 의존도는 성별, 수면 시간, 독서 시간, 부모의 양육 방식 중 학대가 유의미한 변수로 나타난다.

Table 4에서는 각각의 반응 변수별로 단변량 경시적 분석을 실시하였다. 분석에는 경시적 연속형 자료 분석에 일반적으로 사용되는 선형 모형(linear model)을 사용하였다. 일반 선형 모형은 반복 측정된 반응 변수들의 상관관계를 설명하면서 개체들의 측정시점이 동일하지 않아도 자료를 분석할 수 있다. 즉, 모든 개체들이

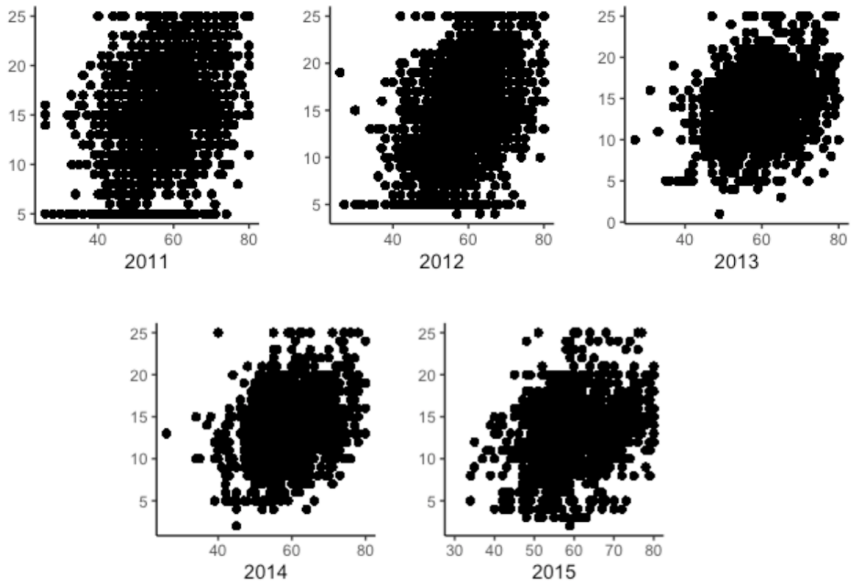

Figure 3: *Scatter plot between school adaptation and academic achievement.*

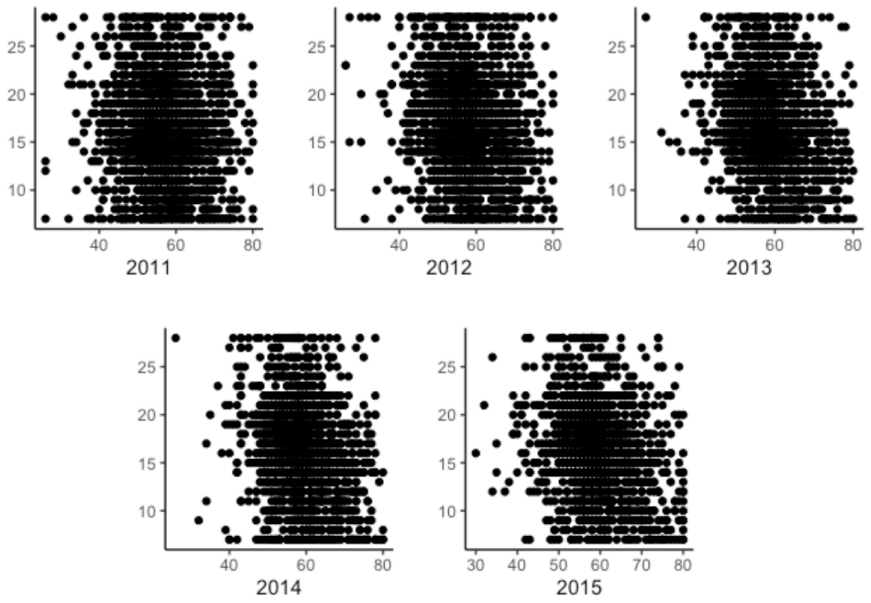

Figure 4: *Scatter plot between school adaptation and phone dependence.*

같은 시점과 같은 반복 수를 가질 필요가 없다. 이 분석에서 공분산 행렬은 일반적으로 R function에서 제공하는 AR(1)의 구조를 가정하였다. 경시적 자료 분석을 시행할 때에는 회귀 분석과 다르게 평균 모형과 분산 모형 모두 주의해야하며, 반복 측정의 상관관계의 설명이 필요하다. 일반 선형 모형을 적합한 결과는 다음과

Table 2: Mean and sd for response variables between 2011 and 2015

Year	School adaptation	Academic achievement	Phone dependence
2011	14.932	56.737	16.152
	(4.548)	(8.210)	(5.247)
2012	14.428	57.802	16.849
	(4.688)	(7.832)	(5.095)
2013	14.117	58.549	16.640
	(3.781)	(7.318)	(4.752)
2014	13.595	58.936	16.238
	(3.558)	(7.074)	(4.606)
2015	13.163	59.218	16.245
	(3.680)	(7.398)	(4.486)

Table 3: Result of regression analysis in 2015

Coefficients	Estimate	Std. Error	t-value	p-value
School adaptation				
(Intercept)	8.838	1.442	6.129	1.05e-09 ***
Gender	0.803	0.190	4.212	2.63e-05 ***
Income	1.188	0.133	8.902	<2e-16 ***
Sleep time	−0.266	0.095	−2.800	0.005 **
Private education time	0.564	0.062	9.105	<2e-16 ***
Reading time	0.375	0.130	2.876	0.004 **
Neglect	−0.262	0.039	−6.568	6.39e-11 ***
Abues	−0.179	0.033	−5.439	5.98e-08 ***
Academic achievement				
(Intercept)	61.340	2.609	23.503	<2e-16 ***
Gender	−0.201	0.345	−0.582	0.560
Income	0.711	0.241	2.952	0.003 **
Sleep time	−0.386	0.172	−2.247	0.024 *
Private education time	0.177	0.112	1.584	0.113
Reading time	1.445	0.236	6.105	1.22e-09 ***
Neglect	−1.046	0.072	−14.480	1.22e-09 ***
Abues	−0.093	0.059	−1.558	0.119
Phone dependence				
(Intercept)	17.079	1.761	9.694	<2e-16 ***
Gender	−2.366	0.231	−10.239	<2e-16 ***
Income	−0.013	0.165	−0.084	0.933
Sleep time	−0.348	0.115	−3.022	0.002 **
Private education time	0.037	0.075	0.502	0.615
Reading time	−0.807	0.162	−4.974	7.14e-07 ***
Neglect	0.039	0.048	0.824	0.409
Abues	0.439	0.040	10.971	<2e-16 ***

같다. 학교 적응도에는 사교육 시간을 제외한 모든 설명 변수가 5% 유의수준에서 매우 유의미한 것을 알 수 있다. 학업 성취도에서는 성별을 제외한 모든 설명 변수가 유의수준 5%에서 유의미한 것을 확인할 수 있으며,

Table 4: Results of univariate longitudinal analysis

Coefficients	Estimate	Std. Error	t-value	p-value
School adaptation				
(Intercept)	8.047	0.653	12.320	<2e-16 ***
Gender	0.040	0.082	0.489	0.624
Income	0.897	0.066	13.430	<2e-16 ***
Sleep time	0.062	0.032	1.904	0.056 *
Private education time	0.481	0.029	16.550	<2e-16 ***
Reading time	0.195	0.054	3.565	0.0003 ***
Abuse	−0.085	0.016	−5.370	8.05e-08 ***
Neglect	−0.253	0.019	−12.926	<2e-16 ***
Academic achievement				
(Intercept)	66.922	1.144	58.477	<2e-16 ***
Gender	−0.346	0.144	−2.389	0.016 *
Income	0.633	0.116	5.417	6.21e-08 ***
Sleep time	−0.441	0.057	−7.678	1.78e-14 ***
Private education time	0.009	0.051	0.179	0.857
Reading time	0.824	0.096	8.535	<2e-16 ***
Abuse	−0.317	0.028	−11.285	<2e-16 ***
Neglect	−1.200	0.034	−34.897	<2e-16 ***
Phone dependence				
(Intercept)	14.510	0.771	18.804	<2e-16 ***
Gender	−1.859	0.097	−19.050	<2e-16 ***
Income	−0.037	0.079	−0.468	0.639
Sleep time	−0.038	0.038	−0.988	0.323
Private education time	−0.066	0.034	−1.919	0.055 *
Reading time	−0.508	0.065	−7.793	7.23e-15 ***
Abuse	0.387	0.019	20.389	<2e-16 ***
Neglect	0.143	0.023	6.192 6.	21e-10 ***

휴대전화 의존도는 성별, 사교육 시간, 독서 시간, 부모의 양육 방식 중 학대와 방임 모두가 유의미한 변수로 나타난다. 그러나 청소년 패널 데이터는 다변량 경시적 자료로 단변량으로 자료를 분석하게 되면 복잡한 상관관계를 제대로 설명하지 못하게 된다. 따라서 반응 변수들 간의 상관관계와 반응 변수와 설명 변수들 간의 상관관계를 모두 고려한 분석이 요구된다.

3.4. 베이지안 방법을 이용한 자료 분석

앞 절에 살펴본 단변량 경시적 분석에 더불어 이번 절에서는 2011년부터 2015년까지 5년동안 약 2,300명의 자료를 다변량으로 고려하여 분석한다. 설명 변수로는 앞서 선택된 변수들을 토대로 단변량 경시적 분석과 동일하게 성별(남자 = 0, 여자 = 1), 가정의 수익, 사교육 시간, 독서 시간, 수면 시간, 부모의 양육 방식 중 방임과 학대의 정도가 있다. 여기서 가정의 수익은 로그를 취한 값을 사용하여 분석하였다. 앞 절에서 반응 변수들 간 산점도를 살펴 본 결과 반응 변수의 학교 적응도, 학업 성취도, 휴대전화 의존도는 서로 연관이 있으므로 다변량 경시적 모형을 적합한다. 2.3절에서 제시한 자기상관구조는 AR(1), AR(2), AR(3)을 가정하였다. ISD 구조는 설명 변수들의 조합으로 가정하여 고려한 모형들을 Table 5에 제시하였다. $\log(\sigma_{itk}) = \lambda_{k0}$를 기초 모형으로 하여 비교를 진행하였다.

Table 5: Description of models for GARP and ISD

IV	GARP	
	AR2	AR3
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Gender}_i$	Model1	Model11
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Edu}_i$	Model2	Model12
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Read}_i$	Model3	Model13
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Gender}_i + \lambda_{k2}\text{Edu}_i$	Model4	Model14
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Gender}_i + \lambda_{k2}\text{Read}_i$	Model5	Model15
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Read}_i + \lambda_{k2}\text{Edu}_i$	Model6	Model16
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Gender}_i + \lambda_{k2}\text{Edu}_i + \lambda_{k3}\text{Read}_i$	Model7	Model17
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Sleep}_i$	Model8	Model18
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Abuse}_i$	Model9	Model19
$\log\sigma_{itk} = \lambda_{k0} + \lambda_{k1}\text{Gender}_i + \lambda_{k2}\text{Edu}_i + \lambda_{k3}\text{Abuse}_i$	Model10	Model20

§ Edu_i is 'Private education time', Read_i is 'Reading time', and Sleep_i is 'Sleep time'.

스파이크와 슬랩 사전 분포를 사용하여 이진 잠재 변수를 통해 일반화 자기회귀모수의 중요한 변수를 선택할 수 있다. 이를 통해 세 반응 변수 학교 적응도, 학업 성취도, 휴대전화 의존도의 상관관계를 다변량 경시적 자료의 세 가지 상관관계로 모두 설명할 수 있다. 모수를 추정하기 위해 MLmodelselection MCMC 알고리즘 (Lee 등, 2021)을 사용한다. Table 5에는 Table 6에서 제안한 다양한 공변량 및 자기회귀모수 모형들에 대해 AIC (Akaike information criterion), BIC (Bayesian information criterion), DIC (Deviance information criterion), MPL (marginal predictive likelihood), MSPE (mean squared predictive error)를 비교하여 정리하였다. AIC, BIC, DIC 등과 같은 기준들은 다음과 같이 정의된다.

$$\begin{aligned}
\text{AIC}_M &= -2\log L\left(y|\hat{\theta_M}\right) + 2d_M,\\
\text{BIC}_M &= -2\log L\left(y|\hat{\theta_M}\right) + d_M\log(N),\\
\text{DIC}_M &= -2\log L\left(y|\hat{\theta_M}\right) + p_D.
\end{aligned}$$

$\hat{\theta}$은 각 모수의 사후 평균이고, p_D는 사후 평균 이탈도에서 모수의 사후 평균에서 평가된 이탈도를 뺀 값으로, 모수 개수의 관점에서 복잡성을 측정하는 데 사용된다.

DIC는 계층적 모형에 대한 AIC와 BIC 기준을 일반화한 형태로 MCMC 방법을 통해 얻어지는 사후 분포에서 계산된 값으로 베이지안 모형 선택 기준으로 자주 쓰인다. BIC는 선택된 모델의 복잡성에 더 큰 패널티가 부여된다. 정보 기준 외에도 한계 예측 가능성(marginal predictive likelihood; MPL)을 가진 후보 모델의 예측 분포 품질을 평가하기 위해 F-폴드 교차 검증(CV)을 고려하였다. MPL은 CPO (conditional predictive ordinate)를 통해 계산된다.

$$\text{MPL} = \sum_i\sum_t \text{CPO}_{it}.$$

CPO는 다음과 같이 표현된다.

$$\text{CPO}_{it} = f\left(Y_{it}|Y_{-it}\right) = \int f\left(Y_{it}|\theta\right)\pi\left(\theta|Y_{-it}, X_{-it}\right)\,d\theta.$$

Y_{it}와 X_{it}는 각각 Y_{it}와 X_{it}를 제외한 Y와 X의 관측값의 벡터이다.

$$\widehat{\text{CPO}_{it}} = \left[\frac{1}{S}\sum_{s=1}^{S}\frac{1}{f\left(Y_{it}|\theta^{(s)}, X_{it}\right)}\right]^{-1}.$$

Table 6: AIC, BIC, DIC, MPL, MSPE of the models defined in Table 5

Model	AIC	BIC	DIC	MPL	MSPE
1	196354.5	196646.9	125849.2	1.52E+52	36.91
2	198124.6	198434.1	126761.2	2.53E+54	38.81
3	196106.2	196433.0	129932.6	6.55E+52	37.67
4	195742.4	196086.3	**93543.6**	**9.04E+55**	**36.74**
5	195324.0	195685.1	97092.5	3.77E+57	37.85
6	199022.5	199400.8	97132.7	7.59E+56	36.82
7	197928.5	198324.0	95114.6	4.10E+57	37.46
8	200585.6	200998.3	96041.3	2.38E+79	39.81
9	201985.4	202380.9	133055.3	1.36E+85	36.77
10	218742.0	219068.8	97403.1	6.52E+93	37.83
11	**195054.4**	**195501.5**	97385.8	4.33E+61	37.08
12	196752.7	197113.8	95451.9	8.40E+53	36.94
13	195431.9	195810.3	96821.5	1.04E+53	37.19
14	196883.4	197278.9	95522.4	6.78E+54	37.36
15	195190.8	195603.5	97120.6	1.89E+57	37.23
16	204739.7	205169.6	97332.2	8.28E+51	36.77
17	195193.1	195537.1	96934.7	7.15E+52	36.83
18	208831.5	209295.7	97229.3	2.17E+25	40.28
19	213970.9	214418.0	97420.1	3.45e+126	40.13
20	205163.7	205542.0	98402.5	1.36E+70	40.23

여기서 S는 사후 표본의 개수이다. CPO 값이 클수록 모형의 예측이 더 우수함을 나타내므로 MPL 값이 큰 모형은 모형 적합도가 더 우수하다는 것을 의미한다.

또한 모델 비교를 위한 평균 제곱 예측 오차(MSPE)는 다음과 같이 나타낸다.

$$\text{MSPE} = \sum_i \sum_t \frac{\left(Y_{it} - \hat{Y}_{it}\right)^2}{\sum_{i=1}^{N} n_i}.$$

이러한 여러 기준들을 비교하여 Table 6에 나열된 다양한 모델에 대한 AIC, BIC, DIC, MPL 및 MSPE의 추정값를 제공한다. 다섯 가지 기준을 비교하였을 때, 모형 4를 가장 적합한 모형으로 선택할 수 있다. 실제 데이터에서는 모든 기준에서 우수함을 보이는 모형을 찾기 어려우므로 여러 기준을 비교하여 가장 적합한 모형을 선택할 수 있다. AIC와 BIC는 모형 11이 모형 4보다 우수하나, 다섯 가지 기준 중 세 가지 기준에서 우수함을 보이는 모형 4를 가장 적합한 모형으로 선택하여 분석을 진행하였다.

모형 4의 GARP와 ISD는 다음과 같이 정의된다.

$$\begin{aligned}
\phi_{itj,kg} &= \alpha_{kg0} I(|t-j|=1) + \alpha_{kg1} I(|t-j|=2), \\
\log(\sigma_{itk}) &= \lambda_{k0} + \lambda_{k1}\text{Gender}_i + \lambda_{k2}\text{Edu}_i, \\
\log\left(\frac{\omega_{ilm}}{\pi - \omega_{ilm}}\right) &= \nu_i + \nu_m.
\end{aligned}$$

모형 11은 모형 4를 제외하고 가장 적합한 모형으로 다음과 같이 정의된다.

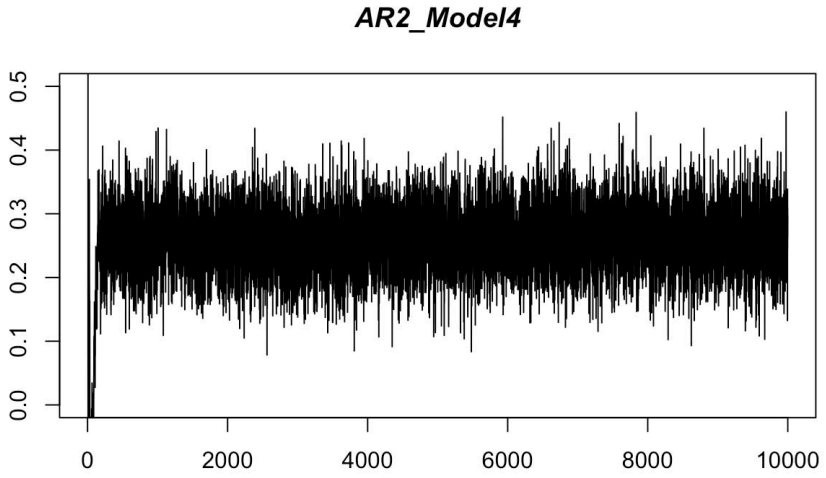

Figure 5: *Trace plot for the reading time coefficient in school adaptation (Model 4).*

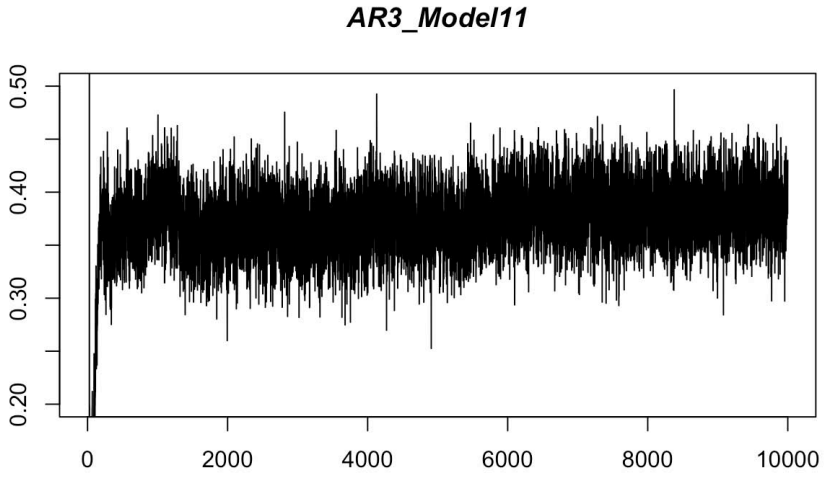

Figure 6: *Trace plot for the sleeping time coefficient in school adaptation (Model 11).*

$$\phi_{it\,j,kg} = \alpha_{kg0} I(|t-j|=1) + \alpha_{kg1} I(|t-j|=2) + \alpha_{kg2} I(|t-j|=3),$$
$$\log(\sigma_{itk}) = \lambda_{k0} + \lambda_{k1}\text{Gender}_i,$$
$$\log\left(\frac{\omega_{ilm}}{\pi - \omega_{ilm}}\right) = \nu_i + \nu_m.$$

Figure 5와 Figure 6는 각각 모형 4와 모형 11에 있는 학교 적응도(school adaption)의 독서 시간(reading time)과 수면 시간(sleeping time)의 시도표(trace plot)이다. 이 두 그림을 통해 두 모형 모두 추세가 보이지 않고, 평균이 일정한 상태로 마르코프 체인이 수렴하는 것을 알 수 있다. 그리고 Gelman-Rubin 통계량을 이용하여 마르코프 체인의 수렴성을 검정하였고, 모든 모수들의 Gelman-Rubin 통계량이 모두 1의 근처에 있음(1.00 1.007)을 알 수 있었고, 이는 수렴이 잘 하였음을 알 수 있었다.

Table 7: The posterior estimates and std, quantile of each β for Models 4 and 11

Variable	Model 4			Model 11		
	PM	PStd	$(2.5^{th}, 97.5^{th})$	PM	PStd	$(2.5^{th}, 97.5^{th})$
School adaptation						
Int.	8.809	0.126	(8.619, 8.968)*	8.695	0.167	(8.507, 8.969)*
Gender	0.137	0.046	(0.076, 0.198)*	0.138	0.014	(0.103, 0.149)*
Income	0.231	0.004	(0.224, 0.238)*	0.231	0.005	(0.228, 0.238)*
Sleep time	0.378	0.005	(0.372, 0.386)*	0.391	0.007	(0.380, 0.410)*
Private education time	0.397	0.008	(0.383, 0.408)*	0.406	0.003	(0.398, 0.408)*
Reading time	0.261	0.003	(0.257, 0.266)*	0.264	0.002	(0.263, 0.270)*
Neglect	−0.070	0.003	(−0.074, −0.063)*	−0.066	0.004	(−0.072, −0.060)*
Abuse	0.014	0.002	(0.010, 0.017)*	0.017	0.002	(0.012, 0.019)*
Academic achievement						
Int.	40.699	0.574	(39.950, 41.649)*	40.443	0.290	(40.207, 41.143)*
Gender	0.475	0.066	(0.366, 0.542)*	0.576	0.023	(0.516, 0.585)*
Income	0.757	0.018	(0.726, 0.783)*	0.760	0.009	(0.738, −0.767)*
Sleep time	1.457	0.032	(1.406, 1.500)*	1.4861	0.014	(1.452, 1.498)*
Private education time	0.166	0.013	(0.452, 0.185)*	0.130	0.003	(0.124, 0.134)*
Reading time	0.667	0.007	(0.656, 0.675)*	0.674	0.005	(0.670, 0.684)*
Neglect	−0.314	0.015	(−0.338, −0.294)*	−0.306	0.006	(−0.322, −0.301)*
Abuse	0.016	0.008	(0.007, 0.029)*	0.013	0.003	(0.003, 0.015)*
Phone dependence						
Int.	13.650	0.171	(13.446, 14.902)*	13.759	0.131	(13.682, 14.032)*
Gender	−2.113	0.018	(−2,138, −2.081)*	−2.152	0.010	(−2,169, −2.142)*
Income	0.068	0.005	(0.061, 0.077)*	0.063	0.004	(0.053, 0.067)*
Sleep time	−0.064	0.007	(−0.077, −0.054)*	−0.068	0.004	(−0.079, −0.065)*
Private education time	0.001	0.004	(−0.005, 0.007)	0.002	0.001	(0.001, 0.004)*
Reading time	−0.356	0.004	(−0.362, −0.349)*	−0.350	0.002	(−0.355, −0.348)*
Neglect	0.151	0.005	(0.143, 0.157)*	0.148	0.003	(0.141, 0.150)*
Abuse	0.296	0.002	(0.293, 0.300)*	0.297	0.002	(0.292, 0.298)*

*denotes that quantile interval does not contains 0.

Table 7은 모형 4와 모형 11의 모수 추정값을 비교한 것이다. 두 모형의 β 추정값은 비슷한 값을 가지는 것을 알 수 있다. 반응 변수에 대한 회귀 계수인 사후 추정값 β를 추정하면 다음 Table 7과 같다. 각 모수에 대해 2.5%와 97.5%의 구간 추정값를 확인해 보았을 때, 반응 변수 휴대전화 의존도에 대해 사교육 시간만이 구간에 0을 포함하고 있는 것으로 확인되었다. 즉, 학교 적응도와 학업 성취도에 대해 모든 설명 변수가 유의미하며, 휴대전화 의존도는 사교육 시간을 제외한 모든 변수가 유의미한 변수로 나타난다. 이는 반응 변수들이 서로 미치는 영향까지 고려하여 추정한 결과로, 각 반응 변수별로 단변량 경시적 분석을 진행한 결과와 차이가 있는 것을 볼 수 있다.

구체적인 분석은 가장 적합한 모형인 모형 4에 초점을 맞춰 해석을 진행한다. 상관관계에 대한 모수 α_{kgm}는 δ_{kgm}에 의해 업데이트 되며, δ가 0이면 α가 0이 된다. δ_{kgm}의 값이 1일 확률을 계산하였을 때, 그 확률이 0.5를 초과하는 값을 가지면 α_{kgm}이 0이 아닌 값으로 업데이트 된다. 그러나 확률이 0.5를 초과하지 않으면 α_{kgm}이 0으로 업데이트 된다. Table 8에서 이 확률을 사후 포함 확률(posterior inclusion probability; PIP)라고 표기한다. 이에 따라 다른 시점에 대한 반응 변수들 간의 상관관계를 설명할 수 있으며 동시에 희박성(sparseness) 문제를 해결할 수 있다. 모형 4에 대해 δ에 의해 결정된 값이 0이 아닌 α의 추정값은 Table 8와 같이 나타난다.

Table 8: Inclusion probability for α_{kgm}

	PIP			δ>0.5		
$I\{\|t-j\|=1\}$	0.950	0.950	0.122	TRUE	TRUE	
	0.950	0.950	0.104	TRUE	TRUE	
	0.132	0.564	0.950		TRUE	TRUE
$I\{\|t-j\|=2\}$	0.950	0.129	0.166	TRUE		
	0.111	0.950	0.148		TRUE	
	0.733	0.110	0.950	TRUE		TRUE

Table 9: The corresponding α_{kgm} estimates

	Estimation		
$I\{\|t-j\|=1\}$	0.500	0.033	
	0.196	0.615	
		−0.012	0.980
$I\{\|t-j\|=2\}$	0.072		
		0.195	
	−0.049		0.157

Table 10: Assuming AR(4), inclusion probability for α_{kgm}

	PIP			$\delta > 0.5$		
$I\{\|t-j\|=1\}$	0.750	0.750	0.086	TRUE	TRUE	
	0.750	0.750	0.088	TRUE	TRUE	
	0.109	0.555	0.750		TRUE	TRUE
$I\{\|t-j\|=2\}$	0.112	0.095	0.098			
	0.088	0.750	0.144		TRUE	
	0.440	0.084	0.750			TRUE
$I\{\|t-j\|=3\}$	0.750	0.096	0.086	TRUE		
	0.079	0.612	0.082		TRUE	
	0.089	0.081	0.750			TRUE
$I\{\|t-j\|=4\}$	0.451	0.092	0.920			
	0.795	0.090	0.083			
	0.080	0.080	0.091			

δ의 초모수(Hyperparameter)를 0.5와 0.1로 설정하여 민감도 분석을 진행하였을 때, 초모수에 따라 민감하게 반응하지 않는 것으로 나타나 δ의 초모수를 0.5로 설정하여 분석을 진행하였다. 비교를 위해 앞서 모형 비교를 진행한 모형 4의 공분산 행렬이 AR(2)나 AR(3)이 아닌 AR(4)의 구조를 가진다고 가정했을 때, α_{kgm}이 0이 아닌 값으로 업데이트될 확률을 나타내면 다음과 같다. Table 9에 따르면 $I\{|t-j|=4\}$일 때, α_{kgm}이 업데이트 되지 않기 때문에 공분산행렬은 AR(2)나 AR(3)으로 충분한 것을 확인해 볼 수 있다.

모형 4의 ISD는 다음과 같이 나타낸다.

$$\log(\sigma_{itk}) = \lambda_{k0} + \lambda_{k1}\text{Gender}_i + \lambda_{k2}\text{Edu}_i.$$

ISD의 상관계수 추정값 Table 11와 같다. λ는 y에 대한 분산으로 추정된 분산이 학교 적응도가 학업 성취도와 휴대전화 의존도에 비해 큰 것으로 나타난다. 학교 적응도와 학업 성취도, 휴대전화 의존도의 분산이 남학생일 때 여학생일 때보다 더 큰 것을 λ_{k2}를 통해 확인 할 수 있다 (남학생 = 1, 여학생 = 0). λ_{k3}를 통해서는

Table 11: Estimates of coefficients in $\log \sigma_{it}$ in the KCYPS data

Parameters	Estimates		
	School adaptation	Academic achievement	Phone dependency
λ_{k0}	0.019	−0.573	−0.642
λ_{k1}	1.286	0.295	1.441
λ_{k2}	−0.828	1.420	−0.114

사교육 시간이 증가할수록 학교 적응도와 휴대전화 의존도에 대한 분산은 감소하며, 사교육 시간이 증가할수록 학업 성취도에 대한 분산은 증가하는 것을 알 수 있다.

이에 따른 상관계수 행렬 R_i는 위의 식과 같이 추정되며, 이를 통해 학교 적응도와 학업 성취도는 양의 상관관계를 가진다는 것을 알 수 있다. 하지만 휴대전화 의존도는 학교 적응도, 학업 성취도와 약한 음의 상관관계가 있는 것으로 나타난다.

$$\hat{R}_i = \begin{pmatrix} 1.000 & 0.310 & -0.011 \\ 0.310 & 1.000 & -0.029 \\ -0.011 & -0.029 & 1.000 \end{pmatrix}.$$

혁신 공분산 행렬 D_{it}는 ISD를 주대각 원소로 가지는 대각 행렬 S_{it}와 상관계수 행렬 R_i로 분해되므로, $\hat{R}_i$로 추정이 가능하다. S_{it}는 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \hat{D}_{it} &= S_{it}\hat{R}_i S_{it}, \\ S_{it} &= \operatorname{diag}\{\sigma_{it1}, \sigma_{it2}, \sigma_{it3}\}, \\ \log(\sigma_{itk}) &= h_{it}^T \lambda_k. \end{aligned}$$

Table 10에서 구한 λ의 추정값을 통해 ISD를 로그선형모형화한 식을 다음과 같이 표현할 수 있다. 따라서 같은 시점에서 성별이 동일한 경우 사교육 시간이 증가할수록 학교 적응도와 휴대전화 의존도의 ISD는 감소하지만, 학업 성취도의 ISD는 증가하는 것을 알 수 있으며, 사교육 시간이 동일한 시간이라고 가정할 경우 모든 반응 변수에서 여학생들보다 남학생들의 ISD가 더 큰 것을 알 수 있다.

$$\begin{aligned} \log(\sigma_{it1}) &= 0.019 + 1.286\,\mathrm{Gender}_i - 0.828\,\mathrm{Edu}_i, \\ \log(\sigma_{it2}) &= 1.286 + 0.295\,\mathrm{Gender}_i + 1.441\,\mathrm{Edu}_i, \\ \log(\sigma_{it3}) &= -0.828 + 1.420\,\mathrm{Gender}_i - 0.114\,\mathrm{Edu}_i. \end{aligned}$$

4. 결론

본 논문에서는 청소년들의 발달에 관한 변수들을 경시적으로 분석하기 위하여 청소년 패널 데이터(KCYPS)에 다변량 경시적 모형을 적합한 후 그 결과를 해석하였다. 다변량 경시적 자료 분석에서 공분산 행렬은 고차원 행렬이며, 복잡한 상관관계를 가지고 있다. 다변량 경시적 자료에서 요구되는 세 가지 상관관계 즉, 다른 시점에서 같은 변수들 간의 상관관계, 다른 시점에서 다른 변수들 간의 상관관계, 같은 시점에서 다른 변수들 간의 상관관계를 모두 설명하기 위해 공분산 행렬 모형화에 대한 수정된 콜레스키 분해(MCD), 수정된 콜레스키 블록분해(MCBD), 초구분해(HD)의 특성을 고찰하였다. 그리고 모수 추정과 중요한 변수 선택 방법으로 베이지안 방법을 고려하여 청소년 패널 데이터를 분석하였다. 분석에는 무작위 결측(MAR) 방식을 사용하였으며, 청소년의 학교 적응도, 학업 성취도, 휴대전화 의존도에 대해 공분산 행렬을 추정하였다. 실제 경시적 자료의 경우 결측값이 없는 자료(complete data)는 거의 없다. 따라서 실제 데이터의 불완전성을 고려하여

이론의 적용이 가능함을 입증하였다. 고차원의 공분산 행렬은 GARP가 급증하며 희박성의 문제를 해결하기 위해 베이지안 방법을 통해 일반화 자기회귀모수를 선택한다. 일반화 자기회귀모수에 대해 스파이크와 슬랩 사전 분포를 정의하여 이진 잠재 변수를 이용해 GARP의 중요한 요소를 선택하도록 하였다. 베이지안 방법을 통해 깁스표본자와 메트로폴리스-해스팅스 알고리즘을 이용하여 모수를 추정하고 GARP를 선택할 수 있다. 청소년 패널 데이터에 가장 적합한 모형을 선택하기 위해 자기회귀모수와 혁신 표준 편차의 구조를 다양화하여 최적의 모형을 탐색하였으며 여러 비교 기준을 통해 GARP는 $\phi_{itj,kg} = \alpha_{kg0} I(|t-j|=1) + \alpha_{kg1} I(|t-j|=2)$, ISD는 $\log(\sigma_{itk}) = \lambda_{k0} + \lambda_{k1}\,\text{Gender}_i + \lambda_{k2}\,\text{Edu}_i$ 의 형태를 가진 모형 4를 최적의 모형으로 선정하였다. 모형 4는 반응 변수 휴대전화 의존도에 대해 사교육 시간만이 유의미하지 않은 변수로 나타났고, 이는 단변량 경시적 분석과 다른 결과로 나타난다. 이는 청소년 패널 데이터의 경시적 특성을 고려한 다변량 경시적 모형을 사용하였기 때문에 청소년들이 성장하며 반복 측정된 경시적 자료의 특성을 반영한 결과인 것으로 보인다. 추정된 상관 계수 행렬을 통해 학교 적응도와 학업 성취도는 양의 상관관계를 가지며, 휴대전화 의존도와 나머지 두 반응 변수와는 약한 음의 상관관계를 가지는 것을 알 수 있다. 같은 시점에서 모든 반응 변수에 대한 분산이 여학생보다 남학생이 더 크며, 성별이 동일한 경우 사교육 시간이 증가할수록 학교 적응도와 휴대전화 의존도의 분산은 감소하고 학업 성취도의 분산은 증가하는 것으로 나타난다. 다변량 경시적 분석에서는 단변량 경시적 분석과 달리 반응 변수들이 서로에게 미치는 영향까지 고려할 수 있기 때문에 보다 정확한 분석이 가능하다. 단변량으로 가정하여 분석한 결과와 다변량 분석의 결과를 비교해 본 결과, 유의미한 변수가 각 반응 변수에 대해 차이가 있었으며, 유의미한 변수에 대해서도 경향에 차이가 존재한다. 수면 시간에 경우 단변량에서는 음의 경향성을 띄지만 다변량의 경우 양의 경향성을 띈다. 이는 다변량 경시적 자료의 특성을 반영한 결과로 반응 변수들이 서로에게 미치는 영향까지 고려한 분석 결과로 단변량 분석에서 생기는 편향을 고려할 수 있다.

References

Kim C and Zimmerman DL (2012). Unconstrained models for the covariance structure of multivariate longitudinal data, *Journal of Multivariate Analysis*, **107**, 104–118.

Chang SJ, Song SW, and Cho MA (2011). The effects of mobile phone dependency, perceived parenting attitude, attachment to peer on school life adjustment of middle school students, *Korean Journal of Youth Studies*, **18**, 431–451.

Chang SJ, Song SW, and Cho MN (2018). The effects of mobile phone dependency, perceived parenting attitude, attachment to peer on school life adjustment of mobile school students, *Korean Journal of Youth Studies*, **18**, 432–451.

Huang JZ, Liu N, Pourahmadi M, and Liu L (2006). Covariance matrix selection and estimation via penalised normal likelihood, *Biometrika*, **93**, 85–98.

Jang SH and Cho KD (2010). Effects of depression scale, interaction anxiety and school adjustment on cellular phone addiction of teenagers, *The Journal of the Korea Contents Association*, **10**, 285–293.

Kim JA (2016). A longitudinal relationship between adolescents' school adaptation and academic achievement according to parenting attitudes, *Korea Journal of Counseling*, **17**, 303–326.

Kim SY and Hong SH (2014). Estimating adolescent's changes in mobile phone dependence: Testing for the effects of ecological factors on individual differences in the changes, *Studies on Korean Youth*, **25**, 101–123.

Kim Y and Lee K (2022). Comparison study for Bayesian multivariate linear model, *Journal of the Korean Data & Information Science Society*, **33**, 249–268.

Lee K and Cho H, Kwak MS, and Jang EJ (2020). Estimation of covariance matrix of multivariate longitudinal data using modified Cholesky and hypersphere decompositions, *Biometrics*, **76**, 75–86.

Lee KJ, Chen RB, Kwak MS, and Lee K (2021). Determination of correlations in multivariate longitudinal data with modified Cholesky and hypersphere decomposition using Bayesian variable selection approach, *Statistics in Medicine*, **40**, 978–997.

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

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

Yoo ES (1996). A study on the change in behavior of children with poor learning through group counseling, Graduate School of Education at Korea University, Master's Thesis, Seoul.

Verbeke G, Fieuws S, Molenberghs G, and Davidian M (2014). The analysis of multivariate longitudinal data: A review, *Statistical Methods in Medical Research*, **23**, 42–59.

Received July 5, 2022; Revised September 1, 2022; Accepted September 5, 2022

베이지안 다변량 선형 모형을 이용한 청소년 패널 데이터 분석

이인선[a], 이근백[1,a]

[a]성균관대학교 통계학과

요 약

다변량 경시적 자료 분석은 반복 측정된 자료에 존재하는 상관관계를 올바르게 추정하면서 자료를 분석해야 한다. 경시적 연구에서는 다변량 경시적 자료가 주로 생성되지만, 기존 통계적 모형은 대부분 단변량으로 분석되어 다변량 경시적 자료에 존재하는 복잡한 상관관계를 제대로 설명하지 못하게 된다. 따라서 본 논문에서는 복잡한 상관관계를 설명하기 위해 공분산 행렬을 모형화하는 다양한 방법에 대해 고찰한다. 그 중 수정된 콜레스키 분해, 수정된 콜레스키 블록분해와 초구분해를 살펴본다. 그리고 일반화 자기회귀모수 행렬이 가지는 희박성 문제를 해결하기 위해 베이지안 방법을 이용하여 청소년 패널 데이터를 분석한다. 청소년 패널 데이터는 다변량 경시적 자료이며, 반응 변수로는 학교 적응도, 학업 성취도, 휴대전화 의존도를 고려한다. 자기 상관 구조와 혁신 표준 편차 구조를 달리 가정하여 여러 모형을 비교한다. 가장 적합한 모형에 대해 학교 적응도와 학업 성취도에 대해 모든 설명 변수가 유의미하며, 휴대전화 의존도가 반응 변수일 때 사교육 시간을 제외한 모든 설명 변수가 유의미한 것으로 나타난다.

주요용어: 다변량 경시적 자료, 베이지안 변수 선택, 수정된 콜레스키 분해, 초구분해

이 논문은 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2022R1A2C1002752).
이 논문은 이인선의 석사논문의 일부를 발췌하였음.
[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: keunbaik@skku.edu