The Korean Journal of Applied Statistics
2024, Vol. 37, No. 4, 467–484

DOI: https://doi.org/10.5351/KJAS.2024.37.4.467
Print ISSN1225-066X / Online ISSN 2383-5818

Analysis of medical panel binary data using marginalized models

Chaeyoung Oh[a], Keunbaik Lee[1,a]

[a]Department of Statistics, Sungkyunkwan University

Abstract

Longitudinal data are measured repeatedly over time from the same subject, so there is a correlation from the repeated outcomes. Therefore, when analyzing this correlation, both serial correlation and between-subject variation must be considered in longitudinal data analysis. In this paper, we will focus on the marginalized models to estimate the population average effect of covariates among models for analyzing longitudinal binary data. Marginalized models for longitudinal binary data include marginalized random effects models, marginalized transition models, and marginalized transition random effect models, and in this paper, these models are first reviewed, and simulations are conducted using complete data and missing data to compare the performance of the models. When there were missing values in the data, there is a difference in performance depending on the model in which the data was generated. We analyze Korea Health Panel data using marginalized models. The Korean Medical Panel data considers subjective unhealthy responses as response variables as binary variables, compares models with several explanatory variables, and presents the most suitable model.

Keywords: KHP data, longitudinal binary data, MTLVM, self-rated health

1. 서론

경시적 자료는 생물학, 의학, 약학, 사회학, 경제학 등의 다양한 연구에서 많이 만들어지는 자료이다. 이 자료에는 시간의 흐름에 따라 각 개체의 속성을 반복 측정함으로써 반복 측정된 자료들 간의 상관관계가 존재하며, 각 개체들 간의 변동성이 또한 존재한다. 이 둘을 '개체 내 상관관계(serial correlation)'와 '개체 간 변동(between-subject variability)'이라고 하며, 이것을 올바르게 설명하며 공변량의 반응변수의 효과를 추정해야 한다. 그렇지 않으면 공변량의 효과의 추정값에 편향(bias)가 발생할 수 있다 (Lee, 2022). 경시적 자료분석에서는 앞서 제시한 개체 내 상관관계와 개체 간 변동을 설명하는 구조를 가진 통계적 모형이 제안되어 왔다. 본 논문에서는 한국의료패널 자료(Korea Health Panel Survey; KHPS) (KHP, 2023) 중 주관적 건강수준에 미치는 요인들을 분석하기 위하여 경시적 이진 자료를 분석을 위한 통계모형을 다루고자 한다.

경시적 이진 자료분석에서 일반적으로 많이 사용되는 모형은 조건부 모형(conditional models)과 주변 모형(marginal model)이 있다. 조건부 모형은 주로 개체 특성적 효과(subject-specific effect)를 추정할 때 사용되며 최대가능도(maximum likelihood) 방법을 이용하여 모수를 추정한다. 주변 모형은 모집단 평균 효과(population-averaged effect)를 추정할 때 사용되며 주로 일반화추정방정식(generalized estimating equation)(Liang과 Zeger, 1986)을 이용하여 모수를 추정한다.

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 Chaeyoung Oh'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 August 2024 / journal homepage: http://kjas.or.kr

조건부 모형에는 임의효과(random effect) 변수를 이용하여 반복 측정된 자료의 상관관계를 설명하는 일반화선형혼합모형(generalized linear mixed models; GLMM)이 있다 (Breslow와 Clayton, 1993). 일반화선형혼합모형은 임의효과가 주어진 상태에서 반응변수(response variable)의 조건부평균과 공변량(covariate)의 관계를 설명하며 주로 최대가능도 방법을 이용해 모수를 추정한다. 따라서 최대가능도 추정량(maximum likelihood estimator)의 특성인 일치성(consistency)과 점근적 정규성(asymptotic normality)를 따른다.

주변 모형은 최대 가능도 방법 대신에 일반화추정방정식(generalized estimating function; GEE)을 이용하여 모수를 추정한다 (Liang과 Zeger, 1986). 이때 반복 측정된 응답변수들의 상관관계는 가상관행렬(working correlation matrix)을 이용하여 설명한다. GEE를 사용해서 구한 추정량은 최대가능도 추정량과 비슷하게 일치성과 점근적 정규성을 가진다. 부가적으로 가상관행렬이 잘못 가정되어도 평균모수의 추정량은 강건성을 가지는 장점이 있다. 하지만 임의결측치(missing at random; MAR)가 존재하는 경시적 자료에서는 전형적인 GEE를 이용하여 모수를 추정하게 되면 그 추정값에 편향이 존재할 수 있다. 또한 GEE는 최대가능도 원리를 기반으로 하지 않았기 때문에 흔히 쓰이는 벌점화 모형선택 기준(penalized model selection criterion)인 Akaike information criterion (AIC)나 Bayesian information criterion (BIC)를 사용할 수 없다는 한계가 있다 (Lee, 2022).

본 논문에서는 모집단 평균 효과를 추정하면서 가능도 방법을 이용한 모형인 주변화 모형(marginalized model) (Heagerty, 1999, 2002)에 대해 집중하고자 한다. 주변화 모형은 주변 모형과 조건부 모형의 장점을 모두 이용할 수 있는 모형이다. 앞서 설명한 것같이 주변화 모형은 가능도 방법을 이용하기에 위해서 두 개의 부모형(sub models)인 주변평균 모형(marginal mean model)과 의존모형(dependence model)으로 구성되어 있다. 주변평균 모형은 공변량 모집단의 평균적인 효과를 설명할 수 있고, 의존 모형은 반복 측정된 결과들의 상관관계를 설명하면서 가능도 함수를 구성할 수 있게 한다. 주변화 모형은 의존모형의 형태에 따라서 주변화 임의효과 모형(marginalized random effects model) (Heagerty, 1999), 주변화 전이모형(marginalized transition model) (Heagerty, 2002) 그리고 주변화 전이 임의효과모형(marginalized transition and latent variable model) (Schildcrout와 Heagerty, 2007)이 있다. 세 모형 모두 주변평균 모형은 동일하나 의존모형에서 주변화 임의효과 모형은 임의효과를, 주변화 전이모형은 마코프(Markov) 구조를, 그리고 주변화 전이 임의효과 모형은 마코프 구조와 임의효과를 둘다 넣어서 반복 측정된 결과들의 상관관계를 설명하는 모형이다.

주변화 모형은 앞서 제시한 것같이 주변평균 모형과 의존모형이 분리됨으로써 공변량의 응답변수의 대한 모집단 평균 효과를 직접적으로 설명할 수 있다. 이것은 또한 모집단의 평균적인 효과에 대한 해석을 할 때 의존 모형에 영향을 받지 않아 잘못된 의존모형의 가정으로 인한 편향의 영향을 적게 받는다. 그리고 앞서 제시한 것같이 주변화 모형은 가능도 방법을 이용하기에 최대 추정량의 일치성과 점근적 정규성을 가질 수 있으며 최대가능도 원리에 기초한 모형선택의 방법을 이용할 수 있다. 마지막으로 무시할 수 있는 결측값(ignorable missingness)을 가정한 경우, GEE 추정량과 다르게 주변화 모형의 경우 가능도 원리에 기초한 모형이므로 모수 추정에 제약이 없다 (Daniels와 Hogan, 2008).

본 논문에서는 주변화 모형인 주변화 임의효과 모형, 주변화 전이 모형 그리고 주변화 전이 임의효과 모형을 이용하여 한국의료패널 자료 (KHP, 2023)를 분석하고자 한다. 한국의료패널은 한국보건사회연구원과 국민건강보험공단이 공동으로 수행하는 조사로 보건의료비용과 의료비 지출 수준의 변화를 파악하여 보건의료정책 및 건강보험정책 수립을 목표로 한다. 1기 한국의료패널은 2005년 등록센서스를 모집단으로 확률비례 2단 층화집락추출 통해 선별한 전국 8,000여 가구와 그 가구원을 대상으로 한다. 분석에는 2011년부터 2014년까지 4년간의 조사자료 (Version 1.7.3)를 이용했다. 한국의료패널의 주관적 건강수준에 관한 선행연구를 참고하여 주관적 불건강 응답률을 이진 변수로 설정해 반응변수로 고려하였다. 설명변수로는 미충족 의료 경험여부, 미충족 의료 경험 이유, 소득, 직업, 성별, 나이, 교육수준, 만성질환여부, 장애여부, 흡연을 설정했다. 본 논문은 세 종류의 주변화 모형을 이용해 한국의료패널 자료를 경시적 이진 자료 하에서 심층적으로 분석하고자 한다.

논문의 순서는 다음과 같다. 우선 2절에서는 경시적 이진 자료를 분석하기 위해 제안된 모형들을 소개한다. 3절에서는 세 종류의 주변화 모형을 비교하는 모의실험을 진행한다. 4절에서는 한국의료패널 데이터를 설명하고, 선행연구와 기초분석 결과를 제시한다. 이후 한국의료패널 자료에 주변화 모형을 적용하고 각 모델들을 비교하여 선택된 최종 모형에 대한 결과를 설명한다. 5절은 4절의 자료 분석 결과를 토대로 결론을 도출하여 본 논문을 요약한다.

2. 경시적 이진자료 분석을 위한 모형

경시적 이진자료를 분석하기 위한 다양한 모형이 제안되어 왔다. 공변량의 모집단 평균 효과를 추정하고자 할 때 사용되는 주변 모형과 주변화 모형을 이 절에서 소개하고자 한다. 구체적으로 GEE를 사용한 주변모형, 주변화 임의효과 모형, 주변화 전이 모형 그리고 주변화 전이 임의효과 모형을 소개한다.

2.1. GEE를 이용한 주변모형

주변모형은 개체 간(between-subject) 공변량 효과와 개체 내(within-subject) 변동 효과의 모형화를 분리한다. 주변평균을 통해 개체 간 공변량 효과를 모형화하며, 공분산 구조(covariance structure)를 통해 개체 내 변동 효과를 모형화한다.

i 번째 개체의 j 번째 이진 응답변수를 y_{ij}라고 하고, 그것에 상응하는 $p \times 1$ 공변량을 x_{ij}라고 하자 ($i = 1, \ldots, m$; $j = 1, \ldots, n_i$). y_{ij}는 개체가 다른 경우 서로 독립이라고 가정한다. Liang과 Zeger (1986)는 주변모형에서 GEE를 아래와 같이 제안하였다.

$$g(\mu_{ij}) = x_{ij}^T \beta, \tag{2.1}$$

여기서 $\mu_{ij} = E(y_{ij}|x_{ij})$이며, $g(\cdot)$는 연결함수로 logit 또는 probit 함수 등을 가정할 수 있다. 반복 측정된 응답변수들의 상관관계는 장애모수 (nuisance parameter)로 취급하며 아래와 같은 상관계수 모형(correlation model)을 가정하였다.

$$\begin{aligned} &\operatorname{cov}(y_i \mid x_i) = V_i(\phi, \alpha) = C_i^{\frac{1}{2}} R_i C_i^{\frac{1}{2}}, \\ &\operatorname{var}\left(y_{ij} \mid x_i\right) = v_{ij} = \phi v(\mu_{ij}), \quad \operatorname{corr}\left(y_{ij}, y_{ik} \mid x_i\right) = \rho_{ijk}, \end{aligned}$$

여기서 R_i는 $n_i \times n_i$의 상관계수행렬이며 $C_i = \operatorname{diag}\{v_{i1}, \ldots v_{in_i}\}$이다. 모수 α는 상관계수를 위한 모수이며, ϕ는 분산의 척도모수(scale parameter)이다.

GEE방법은 불편향 추정함수(unbiased estimating equation)를 구성함으로써 모수추정을 할 수 있다. GEE는 일반적으로 평균모수 β의 추정에 관심이 있고, 나머지 분산 척도모수 ϕ와 상관계수 모수 α는 장애모수로 간주한다. α와 ϕ가 주어진 상태에서, 평균모수 β 추정방정식은 다음과 같다.

$$\Sigma_{i=1}^m U_i(\beta) = \Sigma_{i=1}^m D_i^T V_i^{-1} \{y_i - \mu_i(\beta)\} = 0, \tag{2.2}$$

여기서 $y_i = (y_{i1}, \ldots, y_{in_i})^T$, $\mu_i(\beta) = (\mu_{i1}(\beta), \ldots, \mu_{in_i}(\beta))^T$, $D_i = \partial \mu_i / \partial \beta = (D_i(j,k))$이고 $D_i(j,k) = \partial \mu_{ij} / \partial \beta_k$, $j = 1, \ldots, n_i$, $k = 0, \ldots, p-1$, $V_i = V_i(\beta, \alpha, \phi) = C_i^{\frac{1}{2}} R_i(\alpha) C_i^{\frac{1}{2}}$이다. 식 (2.2)는 평균모수 β의 GEE이며, 이 방정식을 만족하는 해는 피셔-스코어링 알고리즘을 이용하여 구할 수 있다.

$$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} + \left(\sum_{i=1}^m D_i^T V_i^{-1} D_i \right)^{-1} \sum_{i=1}^m D_i^T V_i^{-1} (y_i - \mu_i).$$

척도모수 ϕ는 적률방법(method-of-moments)를 이용하여 추정할 수 있고, 상관계수 모수 α는 잔차를 이용한 적률방법을 이용하여 추정할 수 있다.

GEE는 가상관계수행렬 R_i를 틀리게 가정하여도 GEE 추정량은 강건하다는 특성을 가진다. 이는 $\hat{\beta}$이 β에 대한 일치성은 평균모형이 정확하게 기술된 경우에 상관계수행렬이 틀리게 가정되어도 유지된다는 것을 의미한다. 또한 $\hat{\beta}$은 점근적 정규성을 따른다. 하지만 GEE는 결측값이 완전 임의결측(missing completely at random; MCAR)인 경우에만 사용가능하며 경시적 자료에서 많이 발생하는 임의결측의 경우에는 사용이 불가능하다는 단점이 있다.

2.2. 주변화 모형

Heagerty (1999, 2002)는 경시적 이진자료를 분석하기 위한 주변화 임의효과 모형과 주변화 전이모형을 제안하였다. 주변화 모형은 주변모형의 장점인 공변량의 모집단 평균효과에 대한 해석을 가능하게 할 뿐만 아니라 최대가능도 추정법을 적용할 수 있다. 즉, 최대 가능도 추정량의 특성인 일치성, 점근적 정규성, 가능도비 검정 사용 가능이라는 장점이 있다. 또한 잘못된 임의효과 분포 가정으로 인한 공변량 효과의 편향에 강건하다는 특징이 있다. 본 논문에서는 경시적 이진자료를 분석하기 위한 3가지 주변화 모형에 대해 살펴볼 것이다. 각각의 주변화 모형은 공통의 평균모형을 가지며, 의존모형에서 다른 구조를 가진다.

2.2.1. 주변화 임의효과 모형

Heagerty (1999)는 경시적 이진자료를 분석하기 위해 주변화 임의효과모형(marginalized random effects model; MREM)을 제안했다. MREM은 의존 모형에 임의효과를 삽입하여 반복 측정값들의 상관관계를 설명한다. 경시적 이진자료를 위한 주변화 임의효과모형은 다음과 같다.

$$\text{주변평균모형} : \text{logit}P\left(y_{ij}=1 \mid x_{ij}\right) = x_{ij}^T\beta, \tag{2.3}$$

$$\text{의존모형} : \text{logit}P\left(y_{ij}=1 \mid x_{ij}, b_{ij}\right) = \Delta_{ij} + b_{ij}, \tag{2.4}$$

$$b_i = (b_{i1}, \ldots, b_{in_i})^T \sim N(0, \Sigma_i),$$

여기서 β는 $p \times 1$ 평균모형의 모수벡터이며, Δ_{it}는 절편이며, $\Sigma_i = \sigma^2 R_i$이다. 여기서 $i = 1, \ldots, m$, $j = 1, \ldots, n_i$이며, R_i는 상관계수행렬이며 (j,k) 번째 요소는 $e^{-|j-k|\alpha}$이다.

주변모형 (2.3)은 공변량의 모집단에 대한 평균효과를 설명하고, 의존모형 (2.4)은 반복된 결과값들의 상관관계를 설명한다. Δ_{it}는 주변 평균 $\mu_{ij}^M = P(y_{ij}=1|x_{ij})$과 조건부 평균 $\mu_{ij}^c(b_i) = P(y_{ij}=1|x_{ij}, b_{ij})$의 함수이므로 Δ_{it}는 (β, σ^2)의 함수이며, 다음의 관계가 성립한다.

$$\begin{aligned}\mu_{ij}^M &= E_{b_i}\left(\mu_{ij}^c(b_i)\right), \\ &= \int P\left(y_{ij}=1 \mid b_i\right) f\left(b_{ij}\right) db_{ij},\end{aligned}$$

여기서 $f(b_{ij})$는 평균이 0이며 분산이 σ^2인 정규분포이며 위의 관계식을 통해 뉴턴-라프슨(Newton-Raphson) 알고리즘을 이용하여 Δ_{it}가 계산된다.

MREM의 주변평균 모형과 의존모형을 통해 공변량의 한계효과(marginal effect) 해석 시 의존모형에 직접적인 영향을 받지 않음을 알 수 있다. 따라서 의존모형을 이용해 최대가능도 추정량을 이용해 모수 추정을 할 수 있다. 더 자세한 추정방법은 Heagerty (1999)을 참고하기 바란다.

2.2.2. 주변화 전이모형

Heagerty (2002)는 주변화 전이모형(marginalized transition model; MTM)에서 의존모형은 마코프 구조를 이용해 반복 측정된 반응변수들의 상관관계를 설명한다. 경시적 이진자료를 분석하기 위한 p차 주변화

전이모형은 아래와 같다.

$$\text{주변평균모형} : \text{logit}P\left(y_{ij}=1 \mid x_{ij}\right)=x_{ij}^{T}\beta, \tag{2.5}$$

$$\text{의존모형} : \text{logit}P\left(y_{ij}=1 \mid x_{ij}, y_{ij}\right)=\Delta_{ij}+\Sigma_{k=1}^{p}\gamma_{ijk}y_{ij-k}, \tag{2.6}$$

여기서 β는 평균모형의 모수벡터, $\gamma_{ijk}=z_{ij}^{T}\alpha_k$이고, α_k는 그 이전의 반응변수 y_{ij-k}의 효과를 나타내는 모수벡터이다. z_{ij}는 x_{ij}의 부분집합이며, Δ_{ij}는 절편이면서 β, α의 함수이다. MREM과 마찬가지로 모형 (2.6)의 Δ_{it}는 (β,α)의 함수이며, 다음의 관계가 성립한다.

$$P\left(y_{ij}=1 \mid x_{ij}\right)=\sum_{k=0}^{1}P\left(y_{ij}=1 \mid y_{ij-1}=k\right)P\left(y_{ij-1}=k \mid x_{ij-1}\right).$$

주변평균모형 (2.5)은 독립변수 모집단에 대한 평균효과를 설명하고, 의존모형 (2.6)은 반복측정된 결과값들의 상관관계를 설명하기 위해서 앞 시점의 결과값을 공변량으로 사용한다. MTM 또한 MREM과 동일하게 주변평균 모형과 의존모형이 분리되어 있어서 공변량 한계효과를 해석할 때 의존모형에 영향을 받지 않는다. 그리고 의존모형을 통해 가능도 함수를 구성할 수 있으므로 최대가능도 방법을 이용할 수 있다. 더 자세한 추정방법은 Heagerty (2002)을 참고하기 바란다.

2.2.3. 주변화 전이 임의효과모형

주변화 전이 임의효과모형(marginalized transition and latent variable model; MTLVM)은 1차 주변화 전이모형의 의존모형과 주변화 임의효과모형의 의존모형을 결합해 반복측정된 개체들의 장기적인 상관관계는 임의효과로 단기적인 상관관계는 전이구조(transition structure)로 설명하는 모형이다 (Schildcrout와 Heagerty, 2007). 이 모형은 특히 개체 당 반복 측정수가 매달 이루어진 것과 같이 많은 경우에 유용하다. 경시적 이진자료를 분석하기 위한 주변화 전이 임의효과모형은 다음과 같다.

$$\text{주변평균모형: } \text{logit}P(y_{ij}=1 \mid x_{ij})=x_{ij}^{T}\beta^{m}, \tag{2.7}$$

$$\text{의존모형: } \text{logit}P(y_{ij}-1 \mid y_{ij-1}, b_i)=\Delta_{ij}+\gamma y_{ij-1}+b_i, \tag{2.8}$$

$$b_i \sim N\left(0,\sigma^{2}(X_i)\right),$$

여기서 $\sigma^{2}(X_i)$는 설명변수에 의존하는 이분산성(heteroscadasticity)을 나타내는 임의효과의 분산이다.

주변평균모형 (2.7)는 x_{ij}가 주어졌을 때 y_{ij}의 평균을 의미하며 의존모형 (2.8)은 바로 앞의 결과 (y_{ij-1}) 뿐만 아니라 장기적인 상관관계를 임의효과 b_i를 통하여 설명한다. Δ_{ij}는 주변평균모형과 의존모형을 연결해주는 값으로 전이와 임의효과를 포함하는 조건부 평균 로지스틱 회귀 모형의 형태라고 볼 수 있다. MTLVM의 주변평균모형과 의존모형은 다음과 같은 관계를 가진다.

$$\begin{aligned}\mu_{ij}^{m}&=E_{Z_i,Y_{ij-1},}\left(\mu_{ij}^{c}\right)=E_{Z_i}\left[E_{Y_{ij-1}},\left\{\text{logit}^{-1}\left(\Delta_{ij}+\gamma Y_{i,j-1}+\sigma Z_i\right)\right\}\right]\\&=\int\underbrace{\left\{\text{logit}^{-1}\left(\Delta_{ij}+\sigma z_i\right)\left(1-\mu_{ij-1}^{pc,z}\right)+\text{logit}^{-1}\left(\Delta_{ij}+\gamma+\sigma z_i\right)\mu_{ij-1}^{pc,z}\right\}}_{\mu_{ij}^{pc,z}}\phi\left(z_i\right)dz_i,\end{aligned} \tag{2.9}$$

여기서 $\mu_{ij}^{m}=P(y_{ij}=1|x_{ij})$이며 $\mu_{ij}^{c}=P(y_{ij}=1|y_{ij-1},b_i)$이다. $\phi(\cdot)$는 표준정규분포를 의미하며 $\mu_{ij}^{pc,z}$는 한 시점 전 반응변수의 분포에 대한 조건부 평균 μ_{ij}^{c}의 기댓값을 취해 얻은 부분 조건부 평균이다.

식 (2.9)은 MREM과 MTM에서 Δ_i를 구할 때와 비슷하지만 (Y_{ij-1},b_i)의 결합분포에서 주변화가 발생했기 때문에 μ_{ij}^{c}는 μ_{ij-1}^{c}에 의존하며 반복된 개체 내 측정에 대한 순차적인 업데이트가 필요하다. 자세한 계산과정은 Schildcrout와 Heagerty (2007)를 참고하면 된다.

Table 1: Simulation results for mean parameter estimates of MTLVM, MTM, MREM from the MTLVM data set

True model		MTLVM			MTM			MREM		
Fitted model		MTLVM	MTM	MREM	MTLVM	MTM	MREM	MTLVM	MTM	MREM
N = 100	$\overline{\text{SE}}$	0.410	0.419	0.396	0.391	0.394	0.381	0.390	0.4	0.390
	$\overline{\lvert\text{PRB}\rvert}$	6.222	5.930	5.930	15.836	14.345	16.387	7.380	8.393	7.127
	$\overline{CP}$	94.250	94.250	93.500	93.750	94.250	93.250	94.500	95.000	94.500
N = 300	$\overline{\text{SE}}$	0.410	0.419	0.397	0.410	0.419	0.398	0.389	0.398	0.389
	$\overline{\lvert\text{PRB}\rvert}$	8.524	8.464	8.666	8.549	8.464	8.666	8.833	8.469	8.372
	$\overline{\text{CP}}$	93.997	94.912	93.995	93.997	94.912	93.077	94.495	95.330	94.912
N = 500	$\overline{\text{SE}}$	0.407	0.416	0.395	0.394	0.397	0.385	0.387	0.396	0.387
	$\overline{\lvert\text{PRB}\rvert}$	8.479	9.013	8.433	7.348	7.392	7.012	6.473	6.892	6.727
	$\overline{\text{CP}}$	94.200	95.050	93.350	94.450	95.100	93.950	94.450	95.350	94.950

500 complete data sets with sample size of $N = 100$, 300 and 500 with $T = 5$. The average standard error ($\overline{\text{SE}}$), average of the absolute value of percent relative bias ($\overline{|\text{PRB}|}$) and average of the absolute value of coverage percentage ($\overline{\text{CP}}$) of 500 estimates.

MTLVM도 다른 주변화 모형과 같이 주변평균 모형과 의존모형이 분리되어 있어서 공변량 한계효과를 해석할 때 의존모형에 영향을 받지 않는다. 그리고 의존모형을 이용해 가능도 함수를 구성할 수 있으므로 최대 가능도 추정량을 피셔-스코링을 이용하여 계산한다. 더 자세한 추정방법은 Schildcrout와 Heagerty (2007)을 참고하기 바란다. 경시적 자료의 특성 상 같은 개체 내에서 짧은 시간 간격으로 반복측정된 경우에 서로 더 높은 상관관계가 있을 가능성이 높다. 따라서 반복의 수가 많은 자료의 경우 MTLVM을 통한 분석이 유의미한 결과를 낼 것으로 기대된다.

3. 모의실험

이 절에서는 모의실험을 통해 2절에서 설명한 세 가지 종류의 주변화 모형의 성능을 비교하고자 한다. 모의실험마다 500개의 자료 집합(data set)을 생성하였다. 자료 집합이 주변화 임의효과모형에서 생성된 경우, 주변화 전이모형에서 생성된 경우, 주변화 전이 임의효과모형에서 생성된 경우를 고려하여 자료를 생성하였다. 또한 자료 집합에 결측값이 없는 경우와 임의 결측값이 있는 경우를 고려하였고 자료의 크기(sample size)는 $N = 100$, 300, 500인 경우를 고려하여 자료를 생성하였다. 각 모형에서 생성된 자료를 가지고 세 가지 모형의 회귀계수를 추정하였고, 이를 통해 각 모형의 성능을 비교하고자 한다.

각 모형에서 난수를 발생시키기 위하여 다음과 같은 주변 평균모형을 고려하였다.

$$\text{logit}P(y_{ij} = 1 \mid x_{ij}) = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 x_i + \beta_3 \text{time}_{ij} \times x_i,$$

여기서 time_{ij}는 1에서 10 사이의 값을 랜덤하게 가지고, x_i은 binomial$(N, 1/2)$에서 랜덤추출 하여 0 또는 1의 값을 가진다. 각 그룹은 같은 크기의 표본을 가진다. 그리고 $i = 1, \ldots, N$, $j = 1, \ldots, T$이다. N는 100, 300, 500인 경우와 T는 5, 10인 경우로 나누어 모의실험을 시행하였다. 회귀계수의 참값은 다음과 같다.

$$(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 0.25, 0.25, 0.1).$$

3.1. 결측값이 없는 자료의 경우

우선 자료 집합에 결측값이 없는 경우를 고려하자. 주변화 전이 임의효과모형, 주변화 전이모형, 주변화 임의효과모형 각각에서 표본의 크기가 $N = 100$, 300, 500 이고 $T = 5$, 10인 자료를 생성하여 3개의 모형에

Table 2: Simulation results for mean parameter estimates of MTLVM, MTM, MREM from the MTLVM data set

True model		MTLVM			MTM			MREM		
Fitted model		MTLVM	MTM	MREM	MTLVM	MTM	MREM	MTLVM	MTM	MREM
N = 100	$\overline{\text{SE}}$	0.261	0.266	0.249	0.245	0.246	0.230	0.238	0.240	0.238
	$\overline{\text{PRB}}$	8.650	11.372	6.250	9.990	10.491	8.822	4.688	7.593	5.337
	$\overline{\text{CP}}$	97.250	97.000	95.250	96.000	96.000	94.250	96.500	97.500	96.500
N = 300	$\overline{\text{SE}}$	0.264	0.269	0.251	0.248	0.248	0.233	0.240	0.242	0.241
	$\overline{\text{PRB}}$	4.508	4.412	3.192	3.550	3.873	2.742	2.266	2.232	2.224
	$\overline{\text{CP}}$	95.749	96.166	94.249	94.916	94.749	93.499	95.527	96.499	95.749
N = 500	$\overline{\text{SE}}$	0.268	0.264	0.251	0.248	0.248	0.233	0.241	0.242	0.240
	$\overline{\text{PRB}}$	1.520	1.620	1.890	4.902	5.114	4.646	1.385	1.433	1.377
	$\overline{\text{CP}}$	96.550	96.050	94.100	95.550	95.450	94.305	95.7	96.500	95.900

500 complete data sets with sample size of $N = 100$ and $N = 500$ with $T = 10$. The average standard error ($\overline{\text{SE}}$), average of the absolute value of percent relative bias ($\overline{|\text{PRB}|}$) and average of the absolute value of coverage percentage ($\overline{\text{CP}}$) of 500 estimates.

적용해 모수를 추정하였다. Table 1 은 표본의 크기가 $N = 100, 300, 500$ 이고 $T = 5$ 일 때 500개의 자료 집합을 이용하여 모수를 추정한 결과를 요약한 것이다. Table 2은 표본의 크기가 $N = 100, 300, 500$ 이고 $T = 10$ 일 때 500개의 자료 집합을 이용하여 모수를 추정한 결과를 요약한 것이다. 전체적인 값을 파악하기 위해 추정값들의 표준오차의 평균 ($\overline{SE}$), 실제 β와의 상대 편향(percent relative bias; PRB) 절댓값의 평균 ($\overline{|PRB|}$), 그리고 모수들의 추정값이 실제 β값의 95% 신뢰구간에 속할 확률인 포함확률(coverage probability; CP)의 평균 ($\overline{CP}$)을 포함했다. 표본의 크기가 상대적으로 작은 경우인 $N = 100$에서 다른 모형에서 발생시킨 난수를 MTM에 적합시킬 경우 상대 편향이 다른 모형에 비해서 크다는 것을 알 수 있다. 하지만 표본의 크기가 커짐에 따라 작아짐을 알 수 있다. 표준오차와 포함확률의 경우 각 모형별로 비슷한 값을 가짐을 알 수 있다. 그리고 표본의 크기가 커짐에 따라서 표준오차는 줄어들고, 포함확률은 95% 근처의 값을 가짐을 알 수 있다.

Figure 1는 이 중에서 평균 상대 편향의 크기를 각 경우에 따라서 비교한 것이다. 모든 경우에서 대체적으로 표본의 크기가 커질수록 $\overline{|PRB|}$값이 작아짐을 알 수 있다. $T = 10$이고 MTM에서 생성된 자료의 경우에서만 $N = 500$일 때 $\overline{|PRB|}$가 조금 증가한다. 어떤 모형에서 생성된 자료인지에 상관없이 $N = 100, 300, 500$과 $T = 5, 10$인 모든 경우에서 각 모형의 $\overline{|PRB|}$ 또한 비슷한 경향을 띤다. 하지만 난수를 MTM에서 생성한 경우에는 전체적으로 $\overline{|PRB|}$이 다른 모형에서 발생한 난수들의 결과와 비교해서 표본의 크기가 작은 경우 다소 크다는 것을 알 수 있다. 하지만 표본의 크기가 커짐에 따라서 모든 모형의 $\overline{|PRB|}$가 비슷해 진다. 이것은 작은 표본의 경우 난수발생 과정에서의 임의잡음(random noise)로 인한 것으로 사료된다. 하지만 모형간 비교에서는 MTM과 MTLVM은 비슷한 성능을 가지며 MREM에서 조금 더 큰 $\overline{|PRB|}$를 가짐을 알 수 있다.

3.2. 결측값이 있는 자료의 경우

결측값이 존재하는 경우 각 모형에서 생성된 자료를 세 가지 모형에 적용했을 때의 모수 추정에 대해 비교하기 위해 임의결측(missing at random; MAR)을 가지는 자료를 이용하였다. 본 모의실험에서는 결측값 발생 확률을 다음과 같은 식을 사용하여 설정하였다.

$$\text{logit } P(\text{dropout} = t \mid \text{dropout} \geq t) = a_0 + a_1 Y_{it-1},$$

여기서 결측값 발생 확률은 전체 개체의 5% ($a_0 = 2.3, a_1 = -0.2$), 10% ($a_0 = 2.4, a_1 = -0.21$), 20% ($a_0 = 2.5, a_1 = -0.23$)로 하는 세 가지 경우로 나누어 설정하였다. 앞에서의 결측값이 없는 데이터의 경우와 같이 같은 모형에서 500개의 자료 집합을 생성하고 결측값 발생 확률에 따라 각 모형에서 결측값을 생성한 다음,

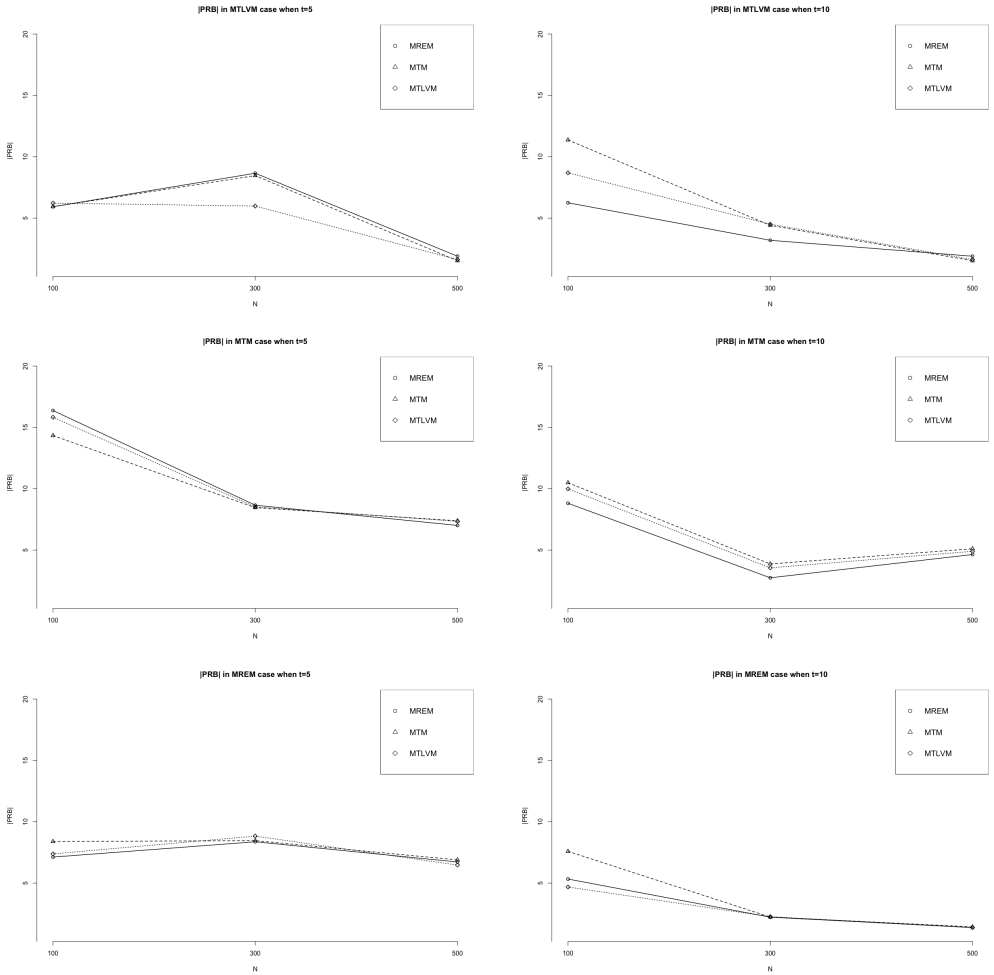

Figure 1: *Comparing the mean absolute percent relative bias (mean(|PRB|)) across three scenarios, with sample sizes* $N = 100, 300,$ *and* 500*, and time points* $T = 5$ *and* 10*, 500 complete data sets were generated from three different models: MTLVM, MTM, and MREM.*

결측값이 존재하는 데이터에 세 가지 모형을 적용하여 평균 모수를 추정하였다. 그 외에 표본크기와 반복수, 모수의 참값은 결측값이 없는 자료의 경우와 동일하게 설정하였다.

Table 3는 각각 3개의 모형을 적합한 결과를 제시하고 있다. 결측값 발생 확률이 증가함에 따라서 $\overline{|\text{PRB}|}$의 값이 커짐을 알 수 있다. 편향에서는 모든 모형에서 비슷함을 알 수 있다. 그리고 $\overline{\text{SE}}$와 $\overline{\text{CP}}$는 모든 모형에서 비슷함을 알 수 있다.

Figure 2는 $N = 500$이고 $T = 10$인 경우 MTLVM, MTM, MREM에서 각각 결측율이 0%, 5%, 10%, 20%일 때 자료를 생성해 각 모형에 적용한 결과의 $\overline{|\text{PRB}|}$를 비교한 것이다.. Figure 2을 보면 모든 경우에서 대체적으로 결측율이 높아질수록 $\overline{|\text{PRB}|}$값이 커짐을 알 수 있다. 어떤 모형에서 생성된 자료인지에 상관없이 결측율이 20%로 가장 높을 때 모든 경우에서 각 모형의 $\overline{|\text{PRB}|}$의 값이 가장 크다. 다만 MTLVM에서 생성된 자료의

Table 3: Simulation results for mean parameter estimates of MTLVM, MTM, MREM from the MTLVM data set with MAR missing

True model		MTLVM			MTM			MREM		
Fitted model		MTLVM	MTM	MREM	MTLVM	MTM	MREM	MTLVM	MTM	MREM
5%	$\overline{\text{SE}}$	0.279	0.283	0.266	0.262	0.263	0.249	0.257	0.259	0.257
	$\overline{\lvert\text{PRB}\rvert}$	5.600	4.085	6.306	6.25	6.095	6.926	3.501	3.616	3.448
	$\overline{\text{CP}}$	95.3	95.9	93.35	95.85	95.9	94.1	94.35	93.5	94.3
10%	$\overline{\text{SE}}$	0.296	0.300	0.283	0.279	0.279	0.265	0.275	0.277	0.275
	$\overline{\lvert\text{PRB}\rvert}$	4.899	5.471	4.720	6.461	6.214	6.653	3.88	3.985	3.837
	$\overline{\text{CP}}$	94.8	95.35	92.95	95.95	96.1	94.35	94.8	94.9	94.95
20%	$\overline{\text{SE}}$	0.337	0.342	0.324	0.319	0.320	0.306	0.316	0.318	0.316
	$\overline{\lvert\text{PRB}\rvert}$	5.695	6.758	5.721	7.755	7.183	8.036	4.158	5.436	4.284
	$\overline{\text{CP}}$	94.9	95.55	93.45	95.05	95.35	94.1	95.00	94.55	95.20

500 complete data sets from true models with sample size of $N = 500$ and $T = 10$ with MAR missing rate 5%, 10%, 20%. The average standard error ($\overline{\text{SE}}$), average of the absolute value of percent relative bias ($\overline{|\text{PRB}|}$) and average of the absolute value of coverage percentage ($\overline{\text{CP}}$) of 500 estimates.

경우에서만 결측율이 5%에서 10%로 증가할 때 MTLVM, MREM에 적용했을 때 $\overline{|\text{PRB}|}$의 값이 근소하게 감소함을 알 수 있다. 또한 결측값이 없는 경우 (0%)와 비교해보았을 때 결측율이 높아질수록 $\overline{|\text{PRB}|}$값이 증가하는 추세를 보인다. 결측값이 없는 경우에 비해 결측값이 있을 때 크게는 $\overline{|\text{PRB}|}$가 5정도 차이남을 확인했다.

4. 한국의료패널자료 분석

4.1. 자료 설명

한국의료패널자료는 2008년 1차 조사를 시작으로 현재까지 한국보건사회연구원과 국민건강보험공단이 공동으로 '한국의료패널 컨소시엄'을 구축하여 매년 수행하는 조사로 의료이용 및 의료비 지출에 영향을 미치는 요인들을 포괄적이고 심층적으로 분석할 수 있는 패널 데이터를 구축하는데 주요 목적을 두고 있다. 또한 국민의료비 산출 및 변화 양상 추적, 의료비 지출 양상과 패턴에 대한 지속적인 모니터링, 데이터 분석을 통한 의료비의 흐름 분석, 그리고 보건의료 정책 수립 및 평가를 위한 동태적 보건복지관련 지표 생산을 중장기목적으로 한다.

제1기 한국의료패널은 2008년 1차 조사를 시작으로 2019년 총 14차 조사까지 진행되었으며, 14차 조사를 끝으로 2020년부터는 개편된 제2기 한국의료패널 조사가 시작되었다. 본 논문에서 이용한 1기 한국의료패널의 원표본은 2005년 등록센서스를 추출 틀로 한 전국 약 8,000가구와 그 가구에 속해 있는 가구원이다. 조사의 현실성을 반영하고자 전국 16개 시도에 거주하고 있는 일반가구만 대상으로 했다. 패널조사의 특성상 조사 진행 중 지속적으로 표본이 탈락했고 이로 인한 대표성 문제를 완화하고자 2010년 등록센서스를 추출 틀로 해 2012년에 전국 2,500가구와 그 가구에 속해 있는 가구원을 추가표본으로 추출하였다. 2008년과 2010년에는 상반기와 하반기 각 1회씩 조사가 진행되고 2011년부터는 연 1회씩 조사가 진행되었다. 이에 따라 본 논문에서는 자료의 정확성을 위해 2011년 이후의 자료를 활용하였다.

자료는 가구 대상 조사내용과 가구원 개인을 대상으로 하는 개인 조사내용으로 나누어져 있으며 본 논문에서는 개인 조사내용을 사용했다. 또한 중간에 관측이 누락된 경우 누락 이전까지의 데이터만 분석에 포함하고 그 이후 결측값은 임의결측(missing at random; MAR)을 가정하였다. 최종 연구대상 인원은 2011년 한국의료패널조사에 응답한 3,534명이다. 자료 분석에는 10개의 선행연구를 고려해 주관적 불건강 응답률을 이진 반응변수로 고려하였으며 선행연구논문을 참고해 선택한 여러 요인들을 고려하여 주관적 불건강에

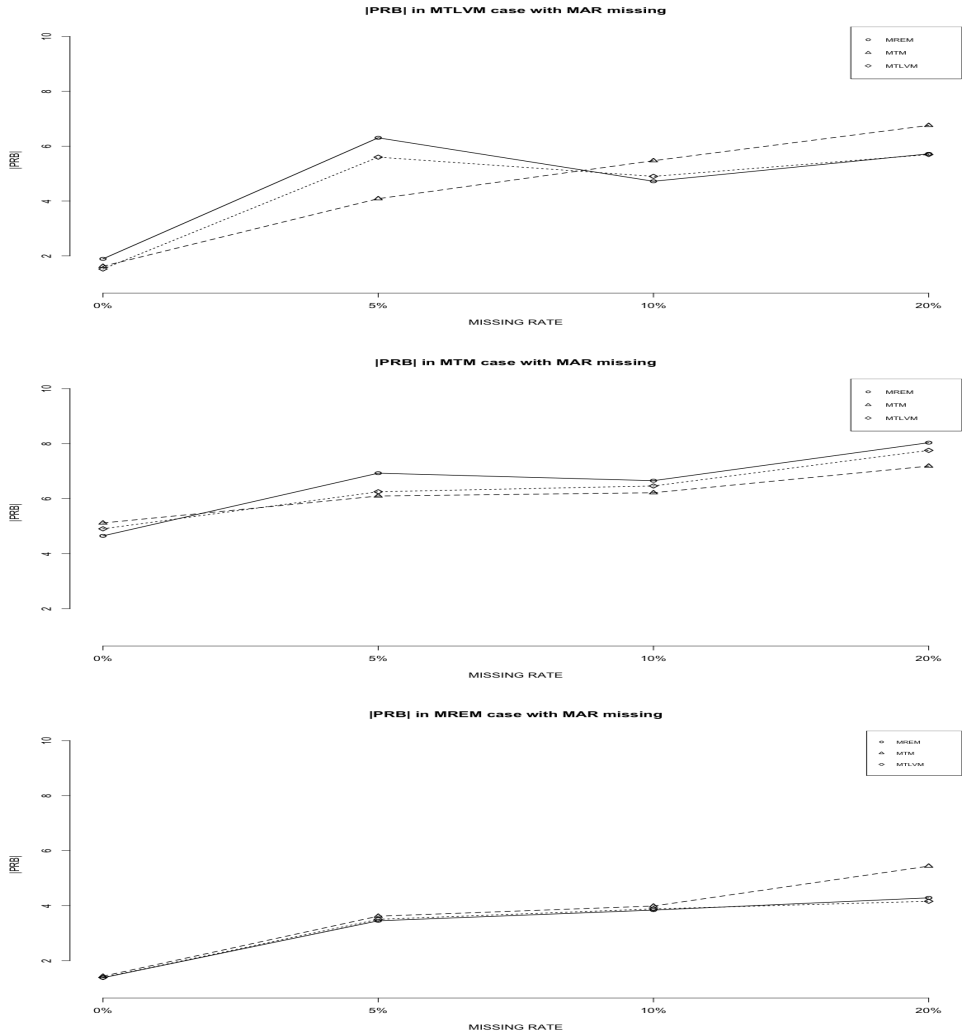

Figure 2: *Comparing the mean absolute percent relative bias (mean(|PRB|)) across three scenarios with missing rates of 0%, 5%, 10%, and 20%, and with sample size $N = 500$ and $T = 10$, 500 data sets were respectively generated from MTLVM, MTM, and MREM under MAR missingness.*

영향을 미치는 요인을 파악하고자 했다.

4.2. 한국의료패널자료 분석

4.2.1. 선행연구

주관적 불건강 응답률과 관련된 여러 선행연구가 있다. 그 중에서 Yoon (2016)은 주관적 건강수준을 이용해 사회경제적 계층간 건강수준의 차이를 성별에 나눠 분석했다. 주관적 불건강 수준에 관한 연구는 주로 노인, 암 환자, 장애인 등을 대상으로 이루어졌다 (Song 등, 2003; Oh 등, 2006; Kim, 2021; Moon 등, 2017;

Table 4: Summary of variables (2011)

Variables	Category	Ratio (Counts) / Mean (SD)
Response variable		
Subjective unhealthy	Poor	0.15 (528)
	Good	0.85 (3006)
Explanatory variables		
Uunmet medical needs experience	Yes	0.169 (597)
	No	0.831 (2937)
Unmet medical needs reason	Supply factor	0.068 (41)
	Demand factor	0.397 (597)
	User choice	0.266 (159)
Income	Low	0.131 (466)
	Middle	0.626 (2214)
	High	0.241 (854)
Job	Full-time	0.269 (952)
	Temporary	0.231 (817)
	Day laborer	0.150 (533)
	Employer	0.341 (1208)
	Side job	0.006 (24)
Sex	Male	0.551 (1949)
	Female	0.448 (1585)
Education	Below elementary school	0.262 (926)
	Below high school	0.474 (1677)
	Above university	0.263 (931)
Chronic diseases	Yes	0.885 (3128)
	No	0.114 (406)
Disorder	Yes	0.064 (229)
	No	0.935 (3305)
Smoke	Smoking	0.234 (827)
	Quit	0.231 (819)
	Never	0.534 (1888)
Age	Continuous	53.668 (12.877)

Jung, 2014). 또는 사회경제적 수준이 건강에 미치는 영향이 크다는 사실을 이용해 경제수준과 교육수준을 복합적으로 고려한 주관적 불건강에 관한 연구가 진행되어 왔다 (Kim, 2005). 금연이나 운동과 같은 개인적 요인이 주관적 건강에 얼마나 영향을 주는지에 관한 연구도 있었다 (Kang, 2007; Kim, 2005). 하지만 각각의 연구들은 특정한 대상을 상대로 진행되었기 때문에 하나의 연구결과로 통합하기가 어렵고, 주로 횡단 자료를 이용하거나 각 개체 별 변동을 고려하기 어려운 고정 효과 모형을 사용하였다.

Park 등 (2018)의 연구에 따르면 미충족 의료와 소득이 상호작용을 이루어 주관적 불건강 수준에 영향을 미친다. 해당 선행연구를 살펴보면 미충족 의료 경험 여부와 경험을 했을 때 그 이유 중 이용자 선택을 제외한 교육수준, 고용형태, 소득, 장애여부, 만성질환여부는 유의한 것으로 나타났다. 그럼에도 불구하고 미충족 의료 경험 이유는 주관적 불건강에 간접적으로 영향을 미치기 때문에 분석에 포함시켰다. 추가적으로 흡연 여부 또한 개인의 건강 행태에 포함되어 주관적 불건강에 영향을 미친다는 선행 연구인 Park 등 (2015)을 참고해 분석에 추가하여 진행했다.

앞서 살펴본 선행연구는 GEE를 이용해 주관적 불건강에 영향을 미치는 요인을 분석하였다. 하지만 일반

Table 5: Chi-square statistics for response variable and categorical explanatory variables

Response variable	Explanatory variable (categorical)	χ^2 Statistic (p-value)
Subjective unhealthy	Unmet medical needs experience	192.676 (≤ 0.001)
	Unmet medical needs reason	243.601 (≤ 0.001)
	Income	287.978 (≤ 0.001)
	Job	256.069 (≤ 0.001)
	Sex	69.08 (≤ 0.001)
	Education	367.989 (≤ 0.001)
	Chronic diseases	68.693 (≤ 0.001)
	Disorder	149.908 (≤ 0.001)
	Smoke	9.656 (0.008)

화 추정 방정식은 데이터에 임의 결측값이 존재하는 경우에는 적용이 불가능하며 AIC나 BIC와 같은 모형선택 기준을 사용할 수 없다는 한계가 있다. 한국의료패널데이터에는 시간의 흐름에 따라 발생하는 결측이 필연적으로 존재하기 때문에 모집단 평균효과를 추정할 때 더 효과적인 주변화 모형을 이용해 분석을 진행하고자 한다.

4.2.2. 자료요약 및 기초분석

주변화 모형에 자료를 적합하기 전에, 본 패널자료의 반응변수와 설명변수들의 기초통계량을 요약하여 제시한다. 우선 제 1기인 2011년에 대한 변수들의 기초통계량 전반을 Table 4에 정리하였다. 연구에서 반응변수로 고려한 주관적 불건강 응답률(self-rated health rate)에 대해서는 조사항목 중 주관적 건강상태 문항을 이용하였다. '귀하께서는 현재 본인의 건강상태가 어떠하다고 생각하십니까?' 라는 질문에 매우 좋음, 좋음, 보통, 나쁨, 매우 나쁨 5개의 선지 중 매우 좋음, 좋음, 보통을 선택한 경우에는 '좋음(good)'으로, 나쁨, 매우 나쁨을 선택한 경우에는 '나쁨(poor)'으로 구별한 이분변수로 변환해 주관적 불건강 응답률 변수를 활용했다. 자료의 설명변수는 1개의 연속형 설명변수와 9개의 범주형 변수를 사용하였으며 설명변수는 앞서 제시한 선행연구에서 유의미하거나 주관적 불건강에 유의미한 영향을 미치는 변수를 토대로 미충족 의료 경험여부(unmet medical needs experience = 1: 경험, 0: 미경험), 미충족 의료 경험 이유(unmet medical needs reason = 0: 공급자 자원, 1: 이용자 자원, 2: 이용자 선택), 소득(income = 0: 저소득, 1: 중간 소득, 2: 고소득), 직업(job = 1: 상용직, 2: 임시직, 3: 일용직, 4: 고용주, 5: 부업소득), 성별(sex = 0: 남성, 1: 여성), 교육수준(education = 0: 초등 이하, 1: 고등 이하, 2: 대학 이상), 만성질환여부(chronic diseases = 1: 만성질환 있음, 0: 만성질환 없음), 장애여부(disorder = 1: 장애 있음, 0: 장애 없음), 흡연(smoke = 1: 흡연, 2: 흡연했으나 현재는 금연, 3: 피운 적 없음), 나이(age)를 설명변수로 고려하였다. 2011년을 예시로 들면 주관적 불건강이 좋다고 응답한 비율은 85%로 주관적 불건강이 나쁘다고 응답한 비율보다 높았다. 미충족 의료를 경험한 비율은 약 17%로 미충족 의료 미경험보다 훨씬 더 적은 비율을 차지하였다. 미충족 의료를 경험한 인원 중 그 이유로는 공급자 자원이 6%, 이용자 자원이 39%, 이용자 선택이 26%를 차지한다.

반응변수인 주관적 불건강 응답률(training participation rate)과 범주형의 설명변수 간의 관계를 카이-제곱 검정을 실시하였다 (Table 5). 그 결과 모든 범주형 설명변수들이 유의미함을 알 수 있었다. 즉, 반응변수와 범주형 설명변수 간의 주변 관련성(marginal association)이 있음을 알 수 있었다.

4개의 관측시점에 대한 반응변수 중 주관적 불건강에 나쁨이라고 응답한 비율추세를 Figure 3에 나타내었다. 각 연도마다 그 비율은 비슷한 값을 가지지만 2014년에 주관적 불건강이 나쁘다고 응답한 사람의 비율이 조금 상승함을 알 수 있다.

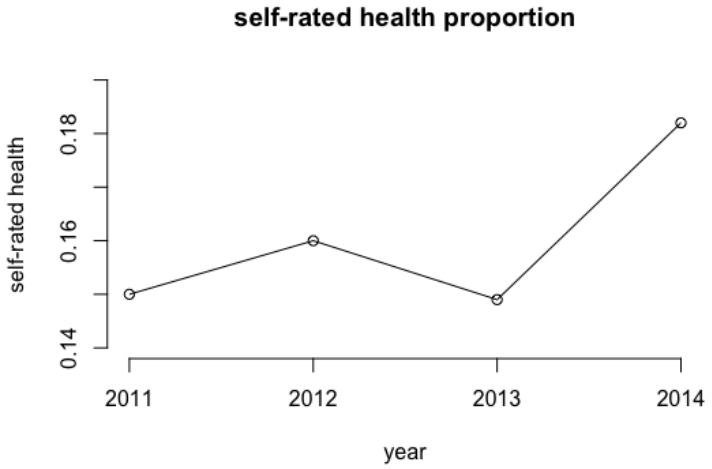

Figure 3: *Profile of response variable (self-rated health).*

Table 6: Models

Model number	Model	Explanatory variables	AIC	BIC
1		All variables without JOB	8399.986	8504.879
2	MREM	All variables without SMOKING	8399.701	8504.594
3		All variables	8393.836	8511.069
4		All variables without JOB	8515.342	8620.235
5	MTM	All variables without SMOKING	8516.854	8621.747
6		All variables	8511.189	8628.423
7		All variables without JOB	8393.559	**8504.623**
8	MTLVM	All variables without SMOKING	8393.959	8505.022
9		All variables	**8388.261**	8511.665

4.2.3. 모형적합

이제 주변화 모형에 한국의료패널 자료를 적합하여 주관적 불건강 응답률에 영향을 미치는 요인을 탐색한다. 이 모형을 적합하기 위하여 CRAN (The Comprehensive R Archive Network)에서 제시한 R library 'binaryMM'을 이용하였다. 적합할 모형과 그 모형의 성능을 비교하기 위해 AIC (Akaike information criterion)와 BIC (Bayesian information criterion)를 Table 6에 제시하였다. 모형의 비교할 때, AIC와 BIC의 값은 작은 모형일수록 그 자료에 더 적합한 모형임을 의미한다.

모형1은 직업 변수를 제외한 모든 변수 (미충족의료 경험여부, 미충족의료 경험 이유, 만성질환여부, 성별, 나이, 직업, 장애여부, 소득, 흡연여부)를 포함한 주변화 임의효과모형, 모형 2는 흡연여부 변수를 제외한 모든 변수를 포함한 MREM, 모형 3은 모든 설명변수를 포함한 MREM, 모형 4는 직업 변수를 제외한 모든 변수를 포함한 주변화 전이모형, 모형 5는 흡연여부 변수를 제외한 모든 변수를 포함한 MTM, 모형 6은 모든 설명변수를 포함한 MTM, 모형 7은 직업 변수를 제외한 모든 변수를 포함한 주변화 전이 임의효과모형, 모형 8은 흡연여부 변수를 제외한 모든 변수를 포함한 MTLVM, 모형 9는 모든 설명변수를 포함한 MTLVM이다.

실제 데이터에서는 모든 기준에서 우수한 모형을 선택하는 것이 어려우므로 여러 기준을 비교하여 가장

Table 7: Likelihood ratio tests

Model	LRT statistic	df	p-value
1 vs 3	28.324	4	< 0.001
4 vs 6	31.351	4	< 0.001
7 vs 9	28.142	4	< 0.001

Table 8: Data analysis results for for Models 3, 6 and 9. Estimates and p-values are displayed

	Model 3		Model 6		Model 9	
Parameter	Estimate	p-value	Estimate	p-value	Estimate	p-value
(Intercept)	3.031	< 0.001	3.012	< 0.001	3.005	< 0.001
EXP1	−0.228	0.073	−0.268	0.036	−0.219	0.036
REA0	−0.588	0.020	−0.490	0.058	−0.594	0.255
REA1	−0.410	0.003	−0.367	0.008	−0.411	0.003
CHR1	−0.580	< 0.001	−0.639	< 0.001	−0.583	< 0.001
JOB2	−0.302	0.001	−0.320	< 0.001	−0.310	0.001
JOB3	−0.535	< 0.001	−0.566	< 0.001	−0.539	< 0.001
JOB4	−0.431	< 0.001	−0.460	< 0.001	−0.435	< 0.001
JOB5	0.027	0.948	−0.056	0.896	0.024	0.955
SEX1	−0.640	< 0.001	−0.610	< 0.001	−0.634	< 0.001
AGE	−0.012	< 0.001	−0.011	0.001	−0.012	0.001
EDU1	0.271	0.002	0.283	< 0.001	0.275	0.001
EDU2	0.527	< 0.001	0.525	< 0.001	0.533	< 0.001
DIS1	−0.821	< 0.001	−0.825	< 0.001	−0.821	< 0.001
INC1	0.245	0.002	0.256	0.001	0.253	0.001
INC2	0.479	< 0.001	0.493	< 0.001	0.488	< 0.001
SMO2	0.065	0.486	0.067	0.463	0.059	0.528
SMO3	0.350	0.001	0.327	0.002	0.343	0.002

우수한 모형을 선택해야 한다. 모형에 대한 기준들을 정리한 표6에 따르면 MREM으로 생성한 모형 중에서는 모형 2가 가장 작은 BIC를 가지고 모형 3이 가장 작은 AIC를 가진다. MTM에서 생성한 모형 중 모형 4가 가장 작은 BIC, 모형 6이 가장 작은 AIC를 가진다. MTLVM에서 생성한 모형 중에서는 모형 9가 가장 작은 AIC를 가지고 모형 7이 최적의 BIC를 갖는다. 각 모델에서 최적의 AIC, BIC를 갖는 모형들은 지분되어(nested) 있으므로 가능도비검정을 실시했다. Table 7는 검정 결과를 나타내고 있다. 모든 경우에서 p-value가 < 0.01로 0.05 이하이므로 더 복잡한 모형인 모형3, 모형6, 모형9를 최종 모형으로 선택했다. 각 모형의 추정 결과를 Table 8에 제시하였다. 모형9가 세 가지 모형 중 가장 작은 AIC, BIC값을 가지므로 한국의료패널자료를 분석하기 위한 최종모형으로 선택했다.

4.3. 최종모형 선택 및 분석

MREM, MTM, MTLVM에서 각각 이 자료를 잘 적합시킨 모형인 모형 3, 6, 9의 적합된 추정량을 Table 8에 제시하였다. 모든 모형들의 평균모수의 추정값이 비슷함을 알 수 있다. 이것은 주변화 모형이 주변평균모형과 의존모형이 분리되어 있어서 의존모형의 가정에 강건함을 보이는 것으로 기인한 결과이다. 특히 모형 3과 9의 추정값이 모형 6에 비해서 매우 비슷함을 알 수 있다. 이것은 두 모형의 AIC와 BIC의 값이 비슷함을 통해서도 알 수 있다. 우리는 이 중에서 모형 9를 집중해서 분석을 진행한다. Table 8의 마지막 열은 모형 9의 모수추정값 및 Chi-검정통계량의 p-값을 보여준다. 그리고 추정된 회귀계수를 사용하여 모형을 적합하면

다음과 같다.

$$\begin{aligned}\text{logit(SRH)} = {} & 3.005 - 0.219 \times EXP1 - 0.594 \times REA0 - 0.411 \times REA1 + 0.583 \times CHR1 \\ & - 0.310 \times JOB2 - 0.539 \times JOB3 - 0.435 \times JOB4 + 0.024 \times JOB5 \\ & - 0.634 \times SEX1 - 0.012 \times AGE + 0.275 \times EDU1 + 0.533 \times EDU2 \\ & - 0.821 \times DIS1 + 0.253 \times INC1 + 0.488 \times INC2 + 0.059 \times SMO2 + 0.343 \times SMO3.\end{aligned}$$

최종모형의 분석 결과를 살펴보면 직업 중 부업소득(JOB5), 흡연여부 중 흡연 후 금연(SMO2)를 제외한 모든 변수가 주관적 불건강에 영향을 미치는 요인임을 확인할 수 있다. 결과를 해석해보면 미충족의료 경험이 있을 때 (EXP = 1) 경험이 없는 경우보다 주관적으로 건강하지 못하다고 판단할 로그오즈(log odds)가 0.219만큼 감소한 것을 알 수 있고, 미충족의료를 경험한 이유가 공급자 자원(REA = 0)일 때 이용자 선택(REA = 2)인 경우에 비해 주관적 불건강 로그오즈가 0.59만큼 감소하였으며 이용자 자원(REA = 1)일 때엔 0.41만큼 감소한다. 또한 만성질환이 있을 때(CHR = 1) 없는 경우에 비해 주관적 불건강 로즈오즈가 0.58만큼 감소하고 남성에 비해 여성(SEX = 1)의 불건강 로즈오즈가 0.634만큼 감소한다. 직업의 경우 상용직(JOB = 1)에 비해 임시직(JOB = 2)의 주관적 불건강 로그오즈가 0.31만큼, 일용직(JOB = 3)일 때에는 0.539만큼 감소하였으며 고용주(JOB = 4)는 0.024만큼 증가한다. 교육수준의 경우 고등이하일 때(EDU = 1) 초등이하인 경우보다 주관적 불건강 로그오즈가 0.275만큼 높으며 대학이상인 경우(EDU = 2)엔 0.533만큼 높다. 장애가 있는 사람(DIS = 1)의 주관적 불건강 로그오즈는 없는 사람에 비해 0.82만큼 낮다. 중간소득인 사람(INC = 1)은 저소득자에 비해 주관적 불건강 로그오즈가 0.25만큼 높으며 고소득자는 (INC = 2)는 0.488만큼 높은 것을 알 수 있다. 마지막으로 비흡연자(SMO = 3)는 흡연자에 비해 주관적 불건강 로그오즈가 0.37만큼 낮다고 볼 수 있다.

5. 결론

본 논문에서는 경시적 이진 자료분석을 위한 주변화 임의효과모형, 주변화 전이모형, 주변화 전이 임의효과모형을 고찰하였다. 주변화모형은 공변량의 모집단 평균효과에 대한 해석이 가능하고, 최대가능도 추정법을 적용할 수 있다는 장점이 있다. 주변화 모형은 모집단의 평균 효과를 추정하고자 할 때 주로 사용하는 모형으로, 평균모형과 의존모형 2가지 구조를 가진다. 주변화 임의효과모형은 의존 모형에 임의효과를 삽입하여 반복 측정값들의 상관관계를 설명하는 모형이며, 주변화 전이모형의 의존모형은 마코프 구조를 이용해 반복 측정된 반응변수들의 상관관계를 설명한다. 주변화 전이 임의효과모형은 1차 주변화 전이모형의 의존모형과 주변화 임의효과모형의 의존모형을 결합해 반복측정된 개체들의 장기적인 상관관계는 임의효과로 단기적인 상관관계는 전이구조로 설명하는 모형이다. 세 가지 주변화모형은 공통적인 평균모형을 가지며 모집단 평균 효과 추정 외에도 최대가능도 추정법을 적용할 수 있다는 장점이 있다.

세 가지 모형의 성능을 비교하기 위해 진행한 모의실험은 자료에 결측값이 없는 경우와 결측값이 있는 경우로 나누어 진행하였다. 결측값이 없는 경우에서는 자료의 수가 어느정도 크다면 회귀 계수의 추정은 비슷한 결과를 보임을 알 수 있었다. 반면에 결측값이 있는 경우에서는 자료가 생성된 모형에 따른 성능 차이가 있음을 보였다. 즉, MTM에서 결측값이 있는 자료를 생성해 MREM, MTLVM에서 평균 모수를 추정했을 때보다 MTM에서 평균 모수를 추정했을 때의 성능이 더 뛰어나다. 결측값이 없는 경우와 비교해보았을 때, 결측값이 존재할 때 성능이 저하되는 것을 확인했다. 결측값이 없는 경우에 비해 결측값이 있을 때 상대편향이 증가함을 알 수 있었다.

한국의료패널자료 분석에서 주관적 불건강 응답률에 미치는 요인들을 파악하기 위하여 앞서 제시한 3가지 주변화 모형(MREM, MTM, MTLVM)을 적용하였다. 설명변수의 선택은 사전 문헌고찰과 설명변수와

반응변수의 카이제곱 검정을 통한 주변관계를 체크한 이후에 주변화 모형에 투입하였다. 그 결과 주변평균 모형의 모수들의 추정값이 비슷함을 알 수 있었다. 추정값을 토대로 흡연여부 중 흡연 후 금연 범주를 제외한 모든 설명변수(미충족의료 경험, 미충족의료 이유, 소득, 성별, 나이, 교육수준, 만성질환 여부, 장애여부, 흡연여부)가 유의미함을 알 수 있었다.

본 논문에서 제시한 모의실험과 자료분석은 모두 임의결측(missing at random)을 가정한 상태에서 분석한 결과이다. 하지만 경시적 자료에서 비임의결측(missing not at random)이 자주 발생한다. 제시된 3가지 주변화 모형은 무시가능한 결측(ignorable missingness) 하에서 바로 적용할 수 있는 모형이다. 하지만 비임의결측 하에서는 모수 추정에 편향이 발생할 수 있다. 따라서 비임의결측 하에서 주변화 모형을 적용하려면 결측 자료 메카니즘(missing data mechanism)을 고려한 모형인 선택 모형(selection model) (Ten Have 등, 1998) 또는 패턴 혼합 모형(pattern mixture model) (Little과 Rubin, 2002)을 고려하여 자료를 분석하여야 한다.

References

Breslow NE and Clayton DG (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9–25.

Daniels MJ and Hogan JW (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, CRC Press, Boca Raton, FL.

Heagerty PJ (1999). Marginally specified logistic-normal models for longitudinal binary data, *Biometrics*, **55**, 688–698.

Heagerty PJ (2002). Marginalized transition models and likelihood inference for longitudinal categorical data, *Biometrics*, **58**, 342–351.

Jung D-J (2014). The Effects of self-rated health on depression to disabled elderly: The moderating effects of psychosocial resources, *Health and Social Welfare Review*, **34**, 247–275.

Kang E-J (2007). Clustering of lifestyle behaviors of Korean adults using smoking, drinking, and physical activity, *Health and Social Welfare Review*, **27**, 44–66.

KHP (2023). Korea health panel survey, Available from: https://www.khp.re.kr:444/

Kim GS (2021). Self-rated health, depression and anxiety in family caregivers of terminal cancer patients: The mediating effects of bonding social capital and bridging social capital, *Health and Social Welfare Review*, **41**, 212–233.

Kim H-R (2005). The relationship of socioeconomic position and health behaviors with morbidity in Seoul, Korea, *Health and Social Welfare Review*, **25**, 3–35.

Lee K (2022). *Longitudinal Data Analysis: Using R*, Free Academy, Paju.

Liang K-Y and Zeger SL (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

Little RJA and Rubin DB (2002). *Statistical Analysis with Missing Data (2nd ed)*, Wiley, New York.

Moon S-J, Sohn M-S, and Choi M-K (2017). The effects of changes in economic activity of the physically disabled on the self-rated health: Focusing on the mediating effect of self-esteem, *Disability & Employment*, **27**, 217–239.

Oh Y-H, Bae H-O, and Kim Y-S (2006). A study on physical and mental function affecting self-perceived health of older persons in Korea, *Journal of the Korea Gerontological Society*, **26**, 461–476.

Park E-J, Jun J, and Kim N-S (2015). The association of multiple risky health behaviors with self-reported poor health, stress, and depressive symptom, *Health and Social Welfare Review*, **35**, 136–157.

Park Y-K, Kim C-Y, and Hwang S-S (2018). Interaction effects of income and unmet healthcare needs to subjective health status: Using the Korea health panel, 2009–2014, *Health and Social Science*, **47**, 57–83.

Schildcrout JS and Heagerty PJ (2007). Marginalized models for moderate to long series of longitudinal binary response data, *Biometrics*, **63**, 322–331.

Song M-S, Song H-J, and Mok J-Y (2003). Community based cross-sectional study on the related factors with perceived health status among the elderly, *Journal of the Korea Gerontological Society*, **23**, 127–142.

Yoon B-J (2016). Differential effects on self-rated health by socioeconomic class, *Journal of Health Informatics Statistics*, **41**, 35–42.

Received January 10, 2024; Revised March 27, 2024; Accepted April 01, 2024

주변화 모형을 이용한 의료 패널 이진 데이터 분석

오채영[a], 이근백[1,a]

[a]성균관대학교 통계학과

요 약

경시적 자료는 같은 개체를 반복 측정함으로써 시간의 흐름에 따른 반복 측정된 자료들 간의 상관관계가 존재한다. 따라서 경시적 자료분석에서는 이 상관관계를 분석할 때 개체 내 상관관계와 개체 간 변동성 모두를 고려해야 한다. 본 논문에서는 경시적 이진 자료를 분석하기 위한 모형 중 공변량의 모집단 평균 효과의 추정을 위해 주변화 모형에 집중하고자 한다. 경시적 이진 자료분석을 위한 주변화 모형으로는 주변화 임의효과, 주변화 전이, 주변화 전이 임의효과 모형이 있으며, 본 논문에서 이들 모형을 먼저 고찰하고, 그리고 모형들의 성능을 비교하기 위해 결측치가 없는 자료와 결측치가 있는 자료로 나눠서 모의실험을 진행한다. 모의실험에서 자료에 결측치가 있는 경우에 자료가 생성된 모형에 따른 성능 차이가 있음을 확인하였다. 마지막으로 주변화 모형을 이용하여 한국의료패널자료를 분석한다. 한국의료패널자료는 반응변수로 주관적 불건강 응답을 이진변수로 고려하였고, 여러 설명변수를 가진 모형을 비교하고 가장 적합한 모형을 제시한다.

주요용어: 경시적 이진 자료분석, 한국의료패널자료, 주관적 건강수준, 주변화 전이 임의효과 모형

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

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