*The Korean Journal of Applied Statistics* (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.5.697
**26**(5), 697–712

# Comparison of GEE Estimation Methods for Repeated Binary Data with Time-Varying Covariates on Different Missing Mechanisms

Boram Park[a] · Inkyung Jung[b,1]

[a]Biometric Research Branch, National Cancer Center
[b]Department of Biostatistics, Yonsei University College of Medicine

(Received October 25, 2012; Revised July 31, 2013; Accepted September 6, 2013)

**Abstract**

When analyzing repeated binary data, the generalized estimating equations(GEE) approach produces consistent estimates for regression parameters even if an incorrect working correlation matrix is used. However, time-varying covariates experience larger changes in coefficients than time-invariant covariates across various working correlation structures for finite samples. In addition, the GEE approach may give biased estimates under missing at random(MAR). Weighted estimating equations and multiple imputation methods have been proposed to reduce biases in parameter estimates under MAR. This article studies if the two methods produce robust estimates across various working correlation structures for longitudinal binary data with time-varying covariates under different missing mechanisms. Through simulation, we observe that time-varying covariates have greater differences in parameter estimates across different working correlation structures than time-invariant covariates. The multiple imputation method produces more robust estimates under any working correlation structure and smaller biases compared to the other two methods.

Keywords: Generalized estimating equations, multiple imputation, weighted estimating equations, MCAR, MAR.

## 1. 서론

다시점 자료(longitudinal data)는 시간에 따라 같은 개체 내에서 반복 측정된 자료로 관측값들 사이에 종속성이 존재한다. 이러한 관측값들 사이의 상관관계를 고려하기 위해 일반화추정방정식(generalized estimating equations; GEE)이 많이 이용되고 있다. 일반화추정방정식은 가상관행렬(working correlation matrix)을 잘못 가정하더라도 모수의 일치추정량(consistent estimator)을 구할 수 있다 (Liang과 Zeger, 1986). 하지만, 일반화추정방정식은 결측 체계가 완전임의결측(MCAR)이 아닌 경우에 편의추정량을 제공하고 (Troxel 등, 1997), 시간-종속적 공변량(time-varying covariate)이 포함된 경우에는 가상관행렬에 따라 회귀계수 추정값이 다르게 나올 수 있다 (Liang과 Zeger, 1986). 결측 체계가 임의

[1]Corresponding author: Assistant Professor, Department of Biostatistics, Yonsei University College of Medicine, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-752, Korea. E-mail: ijung@yuhs.ac

결측(MAR)인 경우에 발생하는 문제를 해결하기 위해 가중 방법 (Robins 등, 1995)과 다중대체 (Rubin, 1987) 방법을 사용하는 것이 제안되었다. 본 논문에서는 모의실험을 통하여 자료의 상관구조, 결측 체계를 변화시키면서 세 가지 방법을 적용하였을 때 시간-독립적 공변량과 시간-종속적 공변량 추정값이 가상관행렬에 따라 어떤 양상으로 변화하는지 살펴보고자 한다. 또한, 가상관행렬에 따른 회귀계수 추정값 간에 차이를 통해 각 방법의 로버스트성(robustness)을 살펴보고 그 추정값의 편차, 분산, 평균제곱오차를 추정하여 각 방법별로 정확성을 비교하고자 한다. 2절에서는 일반화추정방정식, 결측 체계, 가중 방법, 그리고 다중대체 방법에 대해 소개하고, 3절에서는 간질 자료(epileptic data)에 세 가지 방법을 적용한 결과를 비교한다. 4절에서는 자료의 상관구조와 결측 체계를 다양하게 변화시켜 세 가지 방법을 적용한 모의실험에 대해 설명하고, 5절에서는 연구의 결과를 요약, 정리한다.

## 2. 이론적 배경

### 2.1. 일반화추정방정식

시간에 따라 반복 측정된 자료는 한 개체에서의 관측값들 사이에 종속성이 존재한다. 이러한 관측값들 사이의 상관관계를 고려하기 위해 Liang과 Zeger (1986)가 제안한 일반화추정방정식(GEE)을 이용하여 모형의 모수를 추정한다.

$i$번째 개체($i = 1, \ldots, K$)의 $t$번째 시간($t = 1, \ldots, n_i$)에서 반응변수 값을 $n_i \times 1$ 벡터인 $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$라고 하고, 설명변수 값을 $n_i \times p$ 행렬인 $X_i = (x_{i1}, \ldots, x_{in_i})^T$이라 하자. 지수족 분포인 $y_{it}$의 주변밀도함수를 식 (2.1)과 같이 정의한다.

$$f(y_{it}) = \exp\left[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\,\phi\right], \tag{2.1}$$

여기서 $a(\cdot), b(\cdot)$는 연결함수(link function), $\phi$는 척도모수(scale parameter), $\theta_{it} = h(\eta_{it})$, $\eta_{it} = x_{it}\beta$이고, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$은 알려지지 않은 모수인 $p \times 1$ 벡터이다. $y_{it}$의 1차, 2차 적률을 통해 평균과 분산은 각각 다음과 같다.

$$E(y_{it}) = a'(\theta_{it}), \qquad \mathrm{var}(y_{it}) = \frac{a''(\theta_{it})}{\phi}. \tag{2.2}$$

반복 측정된 관측값의 결합분포는 구체화하지 않고 $y_{it}$의 주변분포를 정의하고 반응변수 간의 상관관계를 나타내는 상관행렬을 가정한 후 모수를 추정한다. 이때, 가정하는 상관행렬을 가상관행렬(working correlation matrix)이라고 부른다.

만약 하나의 개체로부터 반복 측정된 관측값이 서로 독립이라면 추정방정식은 식 (2.3)과 같다 (Liang과 Zeger, 1986).

$$U_I(\beta) = \sum_{i=1}^{K} X_i^T \Delta_i S_i = 0, \tag{2.3}$$

여기서 $\Delta_i = \mathrm{diag}(d\theta_{it}/d\eta_{it})$인 $n_i \times n_i$ 행렬이고, $S_i = Y_i - a_i'(\theta)$인 $n_i \times 1$ 벡터이다. $D_i = d\{a_i'(\theta)\}/d\beta$라고 하면, $X_i^T\Delta_i = D_i^T V_i^{-1}$이 된다. 반복 측정된 관측값이 서로 독립이라고 가정하였으므로 $V_i$는 단지 $\beta$의 함수일 뿐 관측값 사이의 관계는 고려되지 않았다.

하지만 반복 측정된 관측값 사이에 상관관계가 존재하면 그 상관관계를 나타내는 가상관행렬을 $n_i \times n_i$ 대각행렬인 $R(\alpha)$라고 할 때, $y_{it}$의 분산-공분산 행렬 $V_i$는 다음과 같다.

$$V_i = \frac{A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}}{\phi}, \tag{2.4}$$

여기서 $A_i = \text{diag}\{a''(\theta_{it})\}$인 $n_i \times n_i$ 대각행렬, $\alpha$는 가상관행렬인 $R(\alpha)$를 설명하는 $s \times 1$ 벡터이고 만약 $R(\alpha)$가 $Y_i$의 참상관행렬이라면 $V_i$는 $\text{cov}(Y_i)$와 동일하다. 반복 측정된 관측값 사이에 존재하는 상관성을 고려한 분산-공분산 행렬을 이용해 일반화추정방정식을 표현하면 식 (2.5)와 같이 정의한다.

$$\sum_{i=1}^{K} D_i^T V_i^{-1} S_i = 0, \tag{2.5}$$

여기서 $D_i = d\{a_i'(\theta)\}/d\beta = A_i \Delta_i X_i$이고, $\Delta_i = \text{diag}(d\theta_{it}/d\eta_{it})$, $S_i = Y_i - a_i'(\theta)$이다. 이 때, 만약 $R(\alpha) = I$, 즉 관측값이 서로 독립이라면 식 (2.5)는 식 (2.3)과 동일하게 된다. 그리고 $U_i(\beta, \alpha) = D_i^T V_i^{-1} S_i$는 유사우도함수와 비슷하지만, 여기서는 관측값 사이에 상관성을 고려하므로 $V_i$는 $\beta$의 함수일 뿐만 아니라 $\alpha$의 함수이기도 하다. $\beta$와 $\phi$를 알고 있을 경우, 식 (2.4)의 $\alpha$를 $\alpha$의 $K^{1/2}$-일치추정량인 $\hat{\alpha}(Y, \beta, \phi)$로 대체함으로써 식 (2.5)를 $\beta$만의 함수로 다시 표현할 수 있다. $\beta$를 알고 있을 경우, $\phi$를 $\phi$의 $K^{1/2}$-일치추정량인 $\hat{\phi}(Y, \beta)$로 대체함으로써 식 (2.5)를 다음과 같은 방정식의 형태로 정의할 수 있다.

$$\sum_{i=1}^{K} U_i \left[\beta, \hat{\alpha}\left\{\beta, \hat{\phi}(\beta)\right\}\right] = 0. \tag{2.6}$$

식 (2.6)을 만족하는 해를 일반화추정방정식을 이용한 추정량 $\hat{\beta}_{GEE}$라고 하고, 이 추정량은 $y_{it}$가 정규분포를 따를 때 최대우도추정량과 일치하게 된다.

일반화추정방정식 방법은 가상관행렬을 잘못 가정하였을지라도 모형의 모수와 그 추정량의 분산을 일치적으로(consistently) 추정한다. 하지만, 모형에 시간-종속적 공변량이 포함되면 가상관행렬 선택에 따라 회귀계수 추정값이 다르게 추정되므로, 가상관행렬의 선택이 중요해진다 (Wall 등, 2005).

### 2.2. 결측 체계

다시점 자료는 같은 개체 내에서 시간에 따라 반복 측정된 자료로 결측(missing)이 많이 발생한다. 자료에서 결측이 발생하는 원리에 따라서 결측 체계(missing mechanisms)를 크게 완전임의결측(missing completely at random; MCAR), 임의결측(missing at random; MAR), 비임의결측(not missing at random; NMAR)으로 구분한다 (Little과 Rubin, 2002). 결측 체계에 따라서 분석 방법이 달라질 수 있으므로 결측 체계를 정확히 파악하는 것은 자료를 분석하는데 매우 중요한 의미가 있다.

$i$번째 개체의 $t$시점까지 반응변수를 $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{it})'$라고 할 때, $Y_i = (Y_i^O, Y_i^M)'$는 관측된 자료 $Y_i^O$와 관측되지 않은 자료 $Y_i^M$로 구성되어 있다. 또한, 반응 지시변수를 $R_i = (R_{i1}, R_{i2}, \ldots, R_{it})'$라고 할 때, $i$번째 개체가 $t$시점에서 관측되면 $R_{it} = 1$이고, 결측이면 $R_{it} = 0$으로 정의한다. $X_i$는 모두 관측된 설명변수라고 가정하면 각 결측 체계는 아래와 같은 식으로 표현할 수 있다.

$$\Pr\left(R_i | Y_i, X_i\right) = \Pr\left(R_i | X_i\right), \tag{2.7}$$

$$\Pr\left(R_i | Y_i, X_i\right) = \Pr\left(R_i | Y_i^O, X_i\right), \tag{2.8}$$

$$\Pr\left(R_i | Y_i, X_i\right) = \Pr\left(R_i | Y_i^O, Y_i^M, X_i\right). \tag{2.9}$$

완전임의결측은 $R_i$가 $Y_i^O$, $Y_i^M$ 모두와 독립적이므로 식 (2.7)과 같고, 임의결측은 $R_i$가 $Y_i^O$에는 종속적, $Y_i^M$에는 독립적이므로 식 (2.8)과 같고, 비임의결측은 $R_i$가 $Y_i^O$ 뿐만 아니라 $Y_i^M$에도 종속적이므로 식 (2.9)와 같이 표현할 수 있다.

### 2.3. 가중 방법

일반화추정방정식 방법은 자료의 결측 체계가 MCAR인 경우 모수에 대한 좋은 추정량을 제공하지만 MCAR이 아닌 결측 체계인 경우에는 편의추정량을 제공한다. 이러한 문제를 해결하기 위해 자료의 결측 체계가 MAR 일지라도 불편추정량을 제공해주는 가중 방법(weighted estimating equations)을 사용하는 것이 제안되었다 (Robins 등, 1995).

가중 방법은 $i$번째 개체가 $t$시점에서 관측될 확률의 역비율 값을 관측값에 가중을 주는 것이다. $i$번째 개체가 $t$시점에서 관측되었다면 $R_{it}=1$이고, 그 외에는 $R_{it}=0$으로 정의한다. 첫 번째 시점에서는 항상 관측이 되고, 한 번 결측이 발생하면 그 뒤 시점부터 끝 시점까지 모두 결측이 발생한 것으로 가정한다. 즉, 모든 개체에 대해서 $R_{i1}=1$이고, 만약 $R_{it}=0$이면 $R_{i(t+1)}=0$임을 의미한다.

시점 $t-1$까지 관측된 자료 중 시점 $t$에서 결측이 발생할 확률은 현재 및 미래 관측치인 $\{Y_{it},\ldots,Y_{iT}\}$와는 상관없고, 과거 관측치인 $D_{it}=\{X_i,Y_{i0},Y_{i1},\ldots,Y_{i(t-1)}\}$에 따라 달라진다. $i$번째 개체가 $t-1$시점에서 관측되었다는 조건하에 $t$시점에서도 관측될 확률은 $\lambda_{it}=P(R_{it}=1|R_{i(t-1)}=1,D_{it},Y_i)=P(R_{it}=1|R_{i(t-1)}=1,D_{it})$ 식을 만족한다고 가정한다. 그러나 MAR 체계에서는 $\lambda_{it}=P(R_{it}=1|R_{i(t-1)}=1,D_{i(T+1)})=P(R_{it}=1|R_{i(t-1)}=1,D_{it})$ 식을 만족한다고 가정한다 (Robins 등, 1994). 첫 번째 시점에서는 항상 관측된다고 가정하므로 모든 $i$에 대해서 $\lambda_{i1}=1$이다. 식 (2.10)을 통해 가중 방법에서의 $\hat{\beta}$을 구할 수 있다.

$$S(\beta)=\sum_{i=1}^{K}W_i\frac{\partial\mu_i}{\partial\beta'}\left(\phi A_i^{\frac{1}{2}}R_iA_i^{\frac{1}{2}}\right)^{-1}(y_i-\mu_i)=0, \tag{2.10}$$

여기서 $W_i=\mathrm{diag}\{R_{i1}w_{i1},\ldots,R_{iT}w_{iT}\}$이고, $w_{it}^{-1}=\hat{\lambda}_{i1}\times\cdots\times\hat{\lambda}_{it}$이다. $i$번째 개체의 $t$시점에서 가중값인 $w_{it}$는 시점 $t$에서 관측된 절대적인 확률의 역수이고 그 추정량은 조건부 확률의 누적 곱의 역수로써 위와 같이 구해진다. 이 가중값 $w_{it}$와 결측의 유무를 나타내는 $R_{it}$의 곱을 대각원소로 갖는 대각행렬 $W_i$가 일반화추정방정식에 추가된다. 시점 $t$에서 관측될 확률이 낮은 관측값은 큰 가중이 가해지고, 관측될 확률이 높은 관측값은 낮은 가중이 가해진다 (Kim, 2004). 이처럼 $W_i$을 통해 결측 처리를 함으로써 가중 방법은 MAR 가정하에서도 유효한 추정량을 제공한다.

### 2.4. 다중대체

결측값에 통계적 모형을 통하여 어떤 다른 값으로 채우는 것을 대체방법(imputation method)이라고 한다. 결측값에 하나의 값으로 채우는 방법을 단일 대체(single imputation)라고 하는데, 이 방법은 관측된 값과 대체된 값을 구분할 수 없어 정보의 양을 과다추정하고, 추정량의 분산을 과소추정 하는 문제가 발생한다. 이러한 문제점을 해결하기 위해 결측값에 여러 개의 값으로 대체하고 이 값들 간 차이의 분산이 추정량의 분산을 계산할 때 추가되어 분산이 과소추정 되지 않도록 하는 다중대체(multiple imputation) 방법이 Rubin (1987)에 의해 제안되었다.

다중대체를 $m$번 시행한 후 대체된 자료 각각에 대하여 $m$번 분석하여 얻어진 모수의 추정값들을 $\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_m$라 하고, 이 모수들의 추정된 분산을 $V_1,V_2,\ldots,V_m$이라 하면 $m$개의 통합된 모수의 추정값($\hat{\beta}^*$)과 그 모수의 분산 추정값($V^*$)을 식 (2.11)과 같이 정의한다.

$$\hat{\beta}^*=\frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i,\qquad V^*=W_m+\left(\frac{m+1}{m}\right)B_m, \tag{2.11}$$

여기서 통합된 모수의 분산 추정값 $V^*$은 식 (2.12)의 $W_m$과 $B_m$ 두 개의 분산 성분을 종합하여 계산된

다.

$$W_m = \frac{1}{m}\sum_{i=1}^{m} V_i, \qquad B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\beta}_i - \hat{\beta}^*\right)^2. \tag{2.12}$$

대체된 자료들로부터 추정된 $m$개 모수의 분산 추정값들 평균인 대체내분산(within- imputation variance; $W_m$)과 $m$개 모수의 추정값들 사이의 분산인 대체간분산(between-imputation variance; $B_m$)을 종합하여 통합된 모수의 분산 추정값을 구한다 (Rubin, 1987).

한 개의 결측값에 무한개의 값으로 대체한다면 추정량의 분산이 정확하게 추정될 것이다. 하지만 무한개의 값으로 대체하는 것은 불가능하므로 유한개의 값으로 대체하게 된다. 이 때 비록 대체 시행횟수가 작더라도 결측으로 인해 손실된 모수에 대한 정보량이 아주 크지 않다면 다중대체를 통해 모수의 분산이 거의 비슷하게 추정된다. 결측으로 인해 손실된 모수에 대한 정보량이란 결측이 없는 완전한 자료와 비교했을 때 결측으로 인해서 발생한 모수의 정밀도(precision)의 감소분을 의미한다 (Song과 An, 2009).

## 3. 실제 자료를 이용한 분석

새로 개발한 항발작제(anti-epileptic drug; AED)의 효능을 살펴보기 위해 수집한 총 89명의 간질환자 자료 (Faught 등, 1996)를 2절에서 소개한 세 가지 방법으로 분석해 보고자 한다. 모든 환자는 약을 복용하기 전 필요한 안정 기간으로 12주 동안은 약을 복용하지 않고 매주 병원을 방문하여 지난 일주일 동안의 발작 횟수를 측정하였고, 13주부터 28주까지 45명은 위약을 복용하고 44명은 신약을 복용하면서 매주 발작 횟수를 측정하였다. 전체 89명은 최소 2번에서 최대 27번씩 관측이 되었고, 총 관측치는 1,419로 한 사람당 평균적으로 16번씩 관측되었다. 아래와 같은 모형을 자료에 적용하고자 한다.

$$\text{logit}\left[P(Y_{it}=1)\right] = X_{it}'\beta, \quad i=1,\ldots,89,\ t=1,\ldots,28 \tag{3.1}$$
$$X_{it} = \left(1, x_i^{trt}, x_i^{sex}, x_i^{race}, x_i^{age}, x_i^{weight}, x_{it}^{drug}\right)',\ \beta = (\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_{drug})'.$$

$Y_{it}$는 지난 일주일 동안의 발작 여부로 발작이 일어났으면 1, 일어나지 않았으면 0이다. $X_{it}$는 시간-독립적 공변량(time-invariant covariate)인 치료그룹(위약, 신약), 성별, 인종, 나이, 몸무게, 그리고 시간-종속적 공변량(time-varying covariate)인 약 복용 여부가 포함된다. 여기서, 약 복용 여부는 $x_{it}^{drug} = I\{t \geq 13\}$인 지시변수이다. 시간-독립적 공변량인 치료그룹의 회귀계수는 위약군에 비해 신약군이 발작을 일으킬 위험을 나타내고, 시간-종속적 공변량인 약 복용여부의 회귀계수는 $t$시점에서 약 미복용 군에 비해 약 복용군이 발작을 일으킬 위험을 의미한다.

일주일 동안 발작이 일어났는지를 반응변수로 두고 일반화추정방정식 방법, 가중 방법, 그리고 다중대체 방법을 적용한 결과를 비교하였다. 실제 자료의 결측은 단조 패턴(monotone pattern)으로 다중대체는 식 (3.1) 로지스틱 회귀 모형을 기반으로 10번 시행하였다. 각 변수의 회귀계수 추정값($\hat{\beta}^w$), 로버스트 표준오차($\sqrt{V_s}$), 모형에 근거한 표준오차($\sqrt{V_m}$)를 이용하여 가상관행렬에 따라 달라지는 회귀계수 추정값 간의 차이(measure of difference)를 $\sum_{w=1}^{4}(\hat{\beta}^w - \bar{\hat{\beta}})^2/V_s(\hat{\beta}^w)$에 근거하여 Table 3.1에 나타내었다. 여기서, 회귀계수 추정값 간의 차이란 가상관행렬에 따라 도출된 회귀계수 추정값이 얼마나 다른지 정도를 나타내며, 그 값이 클수록 가상관행렬에 따라 회귀계수 추정값의 차이는 크다는 것을 의미한다. 가상관행렬은 독립적인(independent) 구조, 교환가능한(exchangeable) 구조, 자기상관(autoregressive; AR(1)) 구조, 2-종속적(2-dependent; Toep(2)) 구조 4가지를 고려하였다. 회귀계수 추정값 간의 차이가 시간-독립적 공변량 변수 5개가 비슷한 패턴을 보여 치료그룹(treatment) 변수의 결과만 표에 제시하였다.

**Table 3.1.** Coefficient estimates with robust and model-based standard errors ($\sqrt{V_s}$ and $\sqrt{V_m}$) for the epileptic data with four different working correlation matrices

| Method | Parameter | | Ind | Exch | AR(1) | Toep(2) | Measure of Difference |
|---|---|---|---|---|---|---|---|
| GEE | Treatment | $\hat{\beta}_{trt}$ | −0.443 | −0.366 | −0.440 | −0.427 | 0.051 |
| | | $\sqrt{V_s}$ | 0.279 | 0.275 | 0.277 | 0.277 | |
| | | $\sqrt{V_m}$ | 0.117 | 0.285 | 0.155 | 0.166 | |
| | Drug | $\hat{\beta}_{drug}$ | −0.577 | −0.414 | −0.504 | −0.437 | 1.041 |
| | | $\sqrt{V_s}$ | 0.128 | 0.117 | 0.124 | 0.130 | |
| | | $\sqrt{V_m}$ | 0.123 | 0.106 | 0.151 | 0.155 | |
| WGEE | Treatment | $\hat{\beta}_{trt}$ | −0.497 | −0.429 | −0.485 | −0.465 | 0.034 |
| | | $\sqrt{V_s}$ | 0.261 | 0.294 | 0.267 | 0.270 | |
| | | $\sqrt{V_m}$ | 0.026 | 0.056 | 0.032 | 0.034 | |
| | Drug | $\hat{\beta}_{drug}$ | −0.702 | −0.476 | −0.632 | −0.587 | 1.619 |
| | | $\sqrt{V_s}$ | 0.136 | 0.124 | 0.136 | 0.138 | |
| | | $\sqrt{V_m}$ | 0.028 | 0.026 | 0.034 | 0.035 | |
| MI | Treatment | $\hat{\beta}_{trt}$ | −0.438 | −0.446 | −0.438 | −0.437 | 0.001 |
| | | $\sqrt{V_s}$ | 0.195 | 0.192 | 0.195 | 0.195 | |
| | | $\sqrt{V_m}$ | 0.128 | 0.203 | 0.140 | 0.147 | |
| | Drug | $\hat{\beta}_{drug}$ | −0.563 | −0.563 | −0.545 | −0.524 | 0.042 |
| | | $\sqrt{V_s}$ | 0.159 | 0.159 | 0.159 | 0.159 | |
| | | $\sqrt{V_m}$ | 0.139 | 0.135 | 0.149 | 0.154 | |

세 가지 방법 모두에서 가상관행렬 선택에 따라 달라지는 회귀계수 추정값 간의 차이는 시간-독립적 공변량(treatment)에 비해 시간-종속적 공변량(drug)에서 상대적으로 더 크게 나타났다. 일반화추정방정식 방법, 가중 방법, 다중대체 방법의 시간-독립적 공변량(Treatment)에서 차이는 각각 0.051, 0.034, 0.001인데 반해, 시간-종속적 공변량(Drug)에서 차이는 각각 1.041, 1.619, 0.042로 더 크게 나타났다. 다중대체 방법이 가중 방법보다 일반화추정방정식 방법의 회귀계수 추정값과 더 유사하였고, 가상관행렬에 따른 회귀계수 추정값 간의 차이도 더 작게 나타났다. 가상관행렬에 따른 회귀계수 추정값의 차이는 시간-독립적 공변량은 일반화추정방정식 방법에서, 시간-종속적 공변량은 가중 방법에서 가장 크게 나타났다. 간질자료의 결측 체계를 정확하게 알고 있지 않으므로 더 다양한 결측 체계에 따라 각각의 방법을 적용했을 때 가상관행렬에 따른 회귀계수 추정값 간의 차이가 나타나는 패턴을 살펴보고자 4절에서 모의실험을 시행하였다.

## 4. 모의실험

### 4.1. 자료에 대한 개요

**4.1.1. 자료 생성** 간질환자 자료를 근거로 100개의 완전한 자료를 재구성하였다. 공변량인 성별, 인종, 약 복용 시작 시점은 균등분포(Uniform distribution)를 따르는 난수 발생을 통하여 생성하고, 나이, 몸무게는 정규분포(Normal distribution)를 따르는 난수 발생을 통하여 생성하였다. 전체 대상자는 위약군 100명, 신약군 100명으로 총 200명이 각각 10주씩 관측된다고 가정한다. 약을 복용하기 시작한 시점 $t_i^{drug}$은 4∼6주로 사람마다 약 복용 시작 시점을 다르게 설정하고, 시간-종속적 공변량인 약 복용 여부는 $x_{it}^{drug} = I\{t \geq t_i^{drug}\}$인 지시변수이다.

각 공변량을 생성한 후 상관성이 존재하는 이항 반응변수를 생성하기 위해서 다변량 이항분포(multi-

**Table 4.1.** Four different missing mechanisms determined by $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$

| Missing mechanism | $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ |
|---|---|
| MCAR | $\alpha = (\alpha_0, 0, 0, 0)$ |
| MAR-weak | $\alpha = (\alpha_0, -0.2, 0, 0)$ |
| MAR-strong | $\alpha = (\alpha_0, -0.5, 0, 0)$ |
| MAR-2-dep | $\alpha = (\alpha_0, -0.5, -0.2, 0)$ |

variate binary distribution)에 근거한 아래와 같은 방법을 이용하였다 (Preisser 등, 2002). 이 방법은 $n_i \times 1$ 평균 벡터 $\pi_i$, $n_i \times n_i$ 공분산 행렬 $V_i$, 그리고 $n_i \times n_i$ 상관행렬 $C_i$에 대한 가정이 필요하다. 먼저, 평균 벡터는 간질환자 실제 자료를 이용하여 일반화추정방정식 방법을 적용했을 때 도출된 로짓모형에 근거하여 다음과 같은 로짓모형 $\text{logit}[P(Y_{it} = 1)] = -1.508 - 0.543x_i^{trt} - 0.295x_i^{sex} + 0.199x_i^{race} + 0.006x_i^{age} + 0.036x_i^{weight} - 0.577x_{it}^{drug}$을 이용해서 구하였다. 공분산 행렬은 $V_i = A_i C_i A_i$으로 여기서 $A_i = \text{diag}\{v_{it}^{1/2}\}$, $v_{it} = \pi_{it}(1 - \pi_{it})$이다. 상관행렬 $C_i$은 자기상관(autoregressive; AR(1)) 행렬($\rho = 0.4, 0.6$)과 교환가능한(exchangeable) 행렬($\rho = 0.2, 0.4, 0.6$)로 총 5가지 형태를 가정하여 살펴보았다.

위와 같이 평균벡터, 공분산 행렬, 상관행렬을 가정하고 $Z_t = (Y_1, \ldots, Y_{t-1})^\top$, $\mu_t = E(Z_t)$, $G_t = \text{cov}(Z_t)$, $s_t = \text{cov}(Z_t, Y_t)$, $b_t = G_t^{-1} s_t$ $(t = 2, \ldots, T)$가 주어졌을 때, 조건부 평균 $\nu_t$을 식 (4.1)과 같이 정의한다.

$$\begin{aligned} \nu_t = \nu_t(z_t; \pi, V) &:= P(Y_t = 1 | Z_t = z_t) = \pi_t + b_t^\top (z_t - \mu_t) \\ &= \pi_t + \sum_{j=1}^{t-1} b_{tj}(y_j - \pi_j) \quad (t = 2, \ldots, T). \end{aligned} \tag{4.1}$$

발작 여부를 나타내는 이항 반응변수 $Y = (Y_1, \ldots, Y_{10})$에서 $Y_1$은 평균 $\pi_1$을 갖는 베르누이 분포(Bernoulli distribution)를 따르는 난수 발생을 통하여 생성하고, $Y_t\,(t = 2, \ldots, 10)$는 조건부 평균 $\nu_t$을 갖는 베르누이 분포(Bernoulli distribution)를 따르는 난수 발생을 통하여 생성하였다. 이처럼 첫 시점의 반응변수는 평균벡터를 이용하고 나머지 시점에서의 반응변수는 조건부 평균을 이용하여 조건부 선형 성질(conditional linear property)을 갖는 다변량 이항분포에 근거하여 상관성이 존재하는 이항 반응변수를 생성하였다.

**4.1.2. 결측생성** 본 논문에서는 공변량은 모두 관측되었다는 가정하에 반복 측정된 반응변수에서의 결측만 고려하여 모의실험을 시행하였다. 이항 반응변수에 결측을 생성하기 위하여 각 개체의 각 시점에서 관측될 확률($\lambda_{it}$)을 구한다. 첫 번째 시점에서는 항상 관측이 되고, 한 번 결측이 발생하면 그 뒤 시점부터 끝 시점까지 모두 결측이 발생한 것으로 가정한다. 즉, 모든 개체에 대해서 $R_{i1} = 1$이고, 만약 $R_{it} = 0$이면 $R_{i,t+k} = 0$, $k > 0$임을 의미한다. 아래와 같은 로짓모형을 이용하여 각 시점에서 관측될 확률($\lambda_{it}$)을 구하였다 (Preisser 등, 2002).

$$\text{logit}(\lambda_{it}) = \alpha_0 + \alpha_1 y_{i(t-1)}^\star + \alpha_2 y_{i(t-2)}^\star I(t > 2) + \alpha_3 y_{it}^\star, \quad t = 2, \ldots, 10, \tag{4.2}$$

여기서 $y_{it}^\star$는 $y_{it}^\star = 2y_{it} - 1$로 $i$번째 개체가 $t$시점에서 발작이 일어났으면 $y_{it}^\star = 1$, 발작이 일어나지 않았으면 $y_{it}^\star = -1$이다. 그리고 두 번째 시점 이후이면 $I(t > 2) = 1$, 나머지는 $I(t > 2) = 0$이며, 처음 시점에서는 반드시 관측된다는 가정을 통해 항상 $\lambda_{i1} = 1$이다.

$\alpha_1, \alpha_2, \alpha_3$의 값을 Table 4.1과 같이 결측 체계에 따라 각각 다르게 설정하였다. MCAR의 경우에는 현재 시점에서 관측될 확률은 이전 어느 시점의 관측값에도 영향을 받지 않는다. 그에 반해 결측 체계

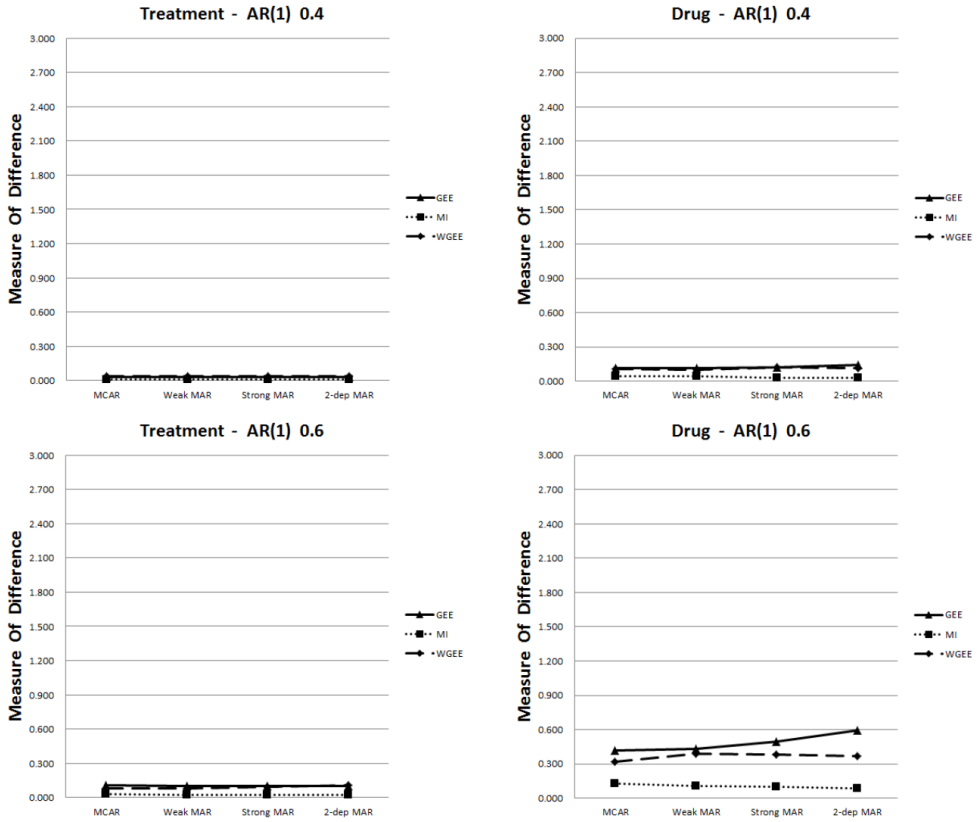


**Figure 4.1.** Measure of difference of parameter estimates for time-invariant and time-varying covariates across the different working correlation matrices (Autoregressive)

가 MAR일 때는 현재 시점에서 관측될 확률이 이전 시점의 관측값에 영향을 받는다. MAR-weak과 MAR-strong은 한 시점 전의 관측값에 따라 관측될 확률이 달라지며, MAR-2-dep는 두 시점 전과 한 시점 전 두 개의 관측값에 따라 관측될 확률이 달라진다. 식 (4.2) 모형에서 이전 시점까지 관측되었다는 조건하에 현재 시점에서 결측이 될 평균 조건부 확률(average conditional probability)과 관련된 $\alpha_0$을 결측률에 따라 다르게 지정한다. 즉, 결측률이 낮을 때는 $\alpha_0$에 큰 값을 결측률이 높을 때는 작은 값을 지정하는데, 결측률이 5%, 10%, 20%, 40%일 때 각각의 $\alpha_0$값을 3.0, 2.2, 1.4, 0.4로 지정한다 (Preisser 등, 2002). 본 논문에서는 모든 결측 체계에서 결측률을 5%라고 가정하여 $\alpha_0 = 3.0$으로 지정하고 모의실험을 시행하였다.

### 4.2. 모의실험 결과

반응변수의 다양한 상관구조와 여러 가지 결측체계에서의 일반화추정방정식 방법(GEE), 가중 방법(WGEE), 다중대체 방법(MI)의 로버스트성(robustness)을 살펴보고자 한다. 각각의 결측 체계별로 세 가지 방법을 적용하여 가상관행렬에 따른 회귀계수 추정값 간의 차이를 비교하였다. 여기서, 회귀계수 추정값 간의 차이(measure of difference)는 100번의 모의실험에서 $\sum_{w=1}^{4}(\hat{\beta}^w - \bar{\hat{\beta}})^2/V_s(\hat{\beta}^w)$에 근거하여 도출된 100개의 추정값을 평균내어 제시하였고, 회귀계수 추정값 간의 차이가 시간-독립적 공변

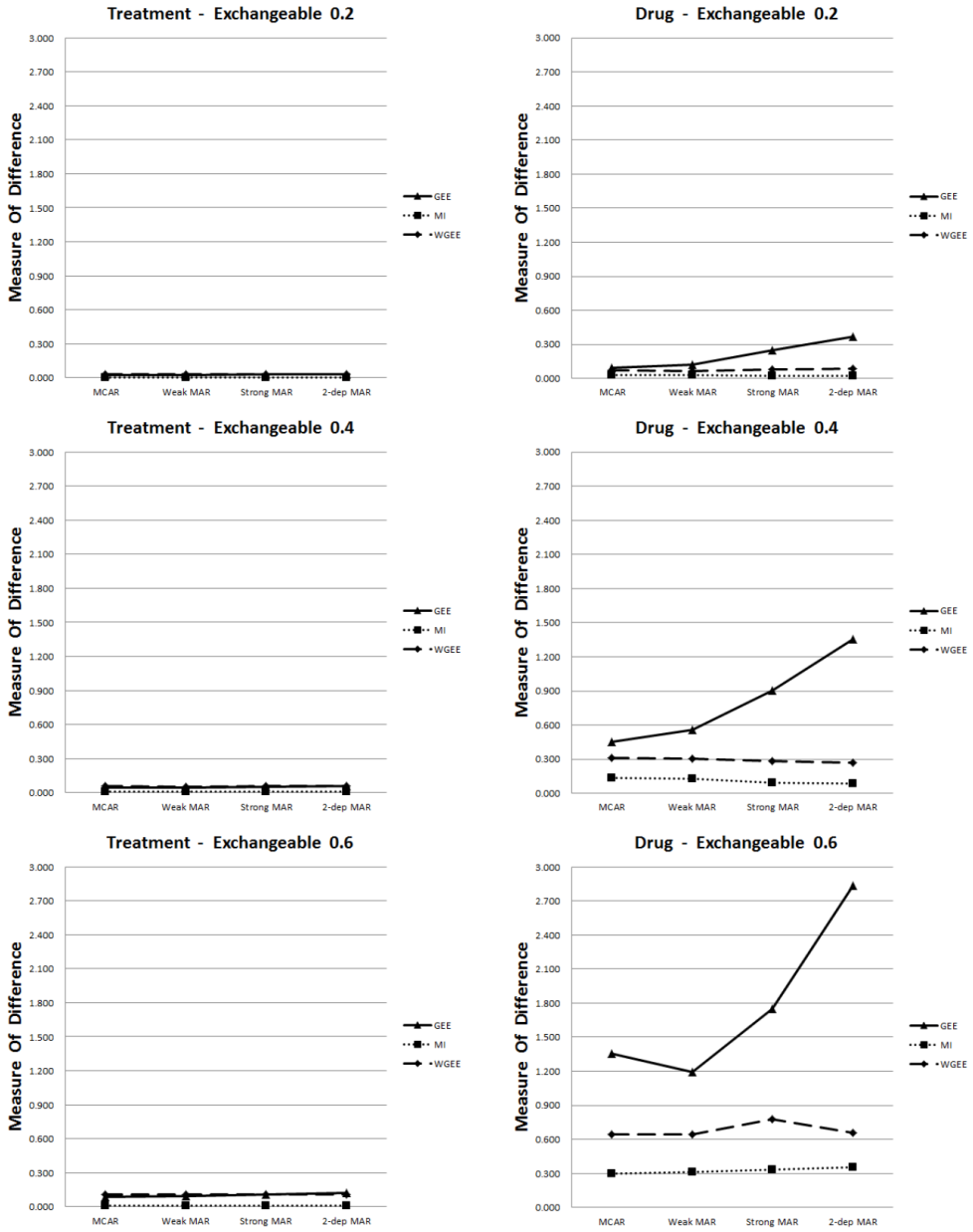


**Figure 4.2.** Measure of difference of parameter estimates for time-invariant and time-varying covariates across the different working correlation matrices (Exchangeable)

량 변수 5개가 비슷한 패턴을 보여 치료그룹(treatment) 변수의 결과만 그림에 나타내었다.

Figure 4.1, Figure 4.2를 보면 결측 체계, 적용 방법에 상관없이 시간-종속적 공변량(drug)은 시간-독립적 공변량(treatment)에 비해 가상관행렬에 따른 회귀계수 추정값의 차이가 항상 더 크게 나타났다.

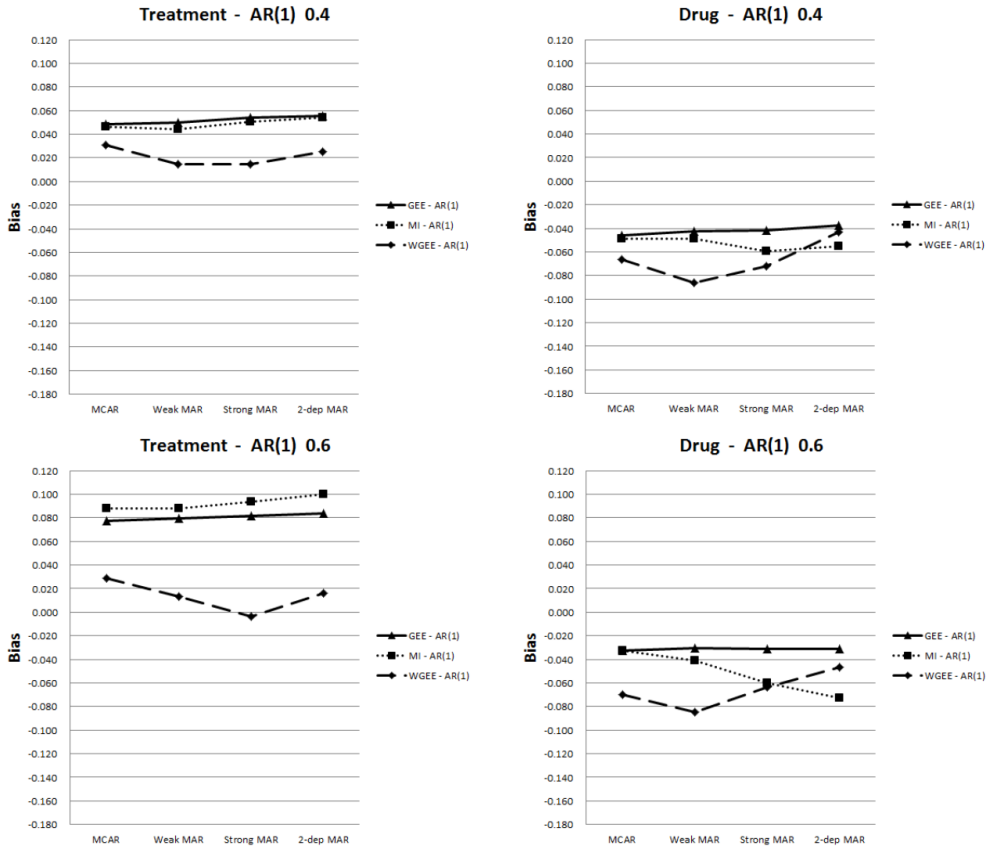


**Figure 4.3.** Bias of parameter estimates for time-invariant and time-varying covariates (Autoregressive)

자료의 상관행렬이 자기상관(AR(1)) 구조일 때보다 교환가능한(exchangeable) 구조일 때, 그리고 동일한 상관행렬 내에서는 $\rho$값이 증가할수록 시간-종속적 공변량과 시간-독립적 공변량 간에 차이가 더 뚜렷하게 나타나는 것을 확인할 수 있다. 시간-독립적 공변량은 회귀계수 추정값의 차이가 자료의 상관행렬, 결측 체계, 적용하는 방법에 크게 의존하지 않지만, 시간-종속적 공변량은 결측 체계에 따라 각각의 방법별로 가상관행렬에 따른 회귀계수 추정값의 차이가 다른 경향을 보였다. 결측 체계가 MCAR에서 2-dep MAR로 갈수록 가상관행렬에 따른 회귀계수 추정값의 차이가 가중방법과 다중대체 방법에서는 크게 변화가 없었지만, 일반화추정방정식을 적용했을 때는 점점 증가하였다. 즉, 일반화추정방정식 방법은 결측 발생이 이전 관측값과 상관성이 강할수록 가상관행렬에 따른 회귀계수 추정값의 차이가 증가하는 경향이 있었다. 또한, 동일한 결측 체계에서 가중 방법이 일반화추정방정식 방법보다 시간-종속적 공변량의 회귀계수 추정값 차이를 더 감소시켰다. 다중대체 방법은 전반적으로 가상관행렬에 따른 회귀계수 추정값의 차이가 작아 가상관행렬의 형태에 로버스트함을 확인할 수 있었다.

각각의 결측 체계별로 세 가지 방법을 적용했을 때 추정의 정확성을 살펴보고자 참값과 각 방법을 통해 구한 추정값을 이용해서 편차(bias), 분산(variance), 평균제곱오차(mean squared error; MSE)를 구해 각각의 경우를 비교하였다. 이 때, 각각의 결측 체계별로 세 가지 각 방법 내에서 가상관행렬에 따른 편차 및 평균제곱오차가 거의 비슷한 결과가 도출되어서 사전에 가정한 상관행렬인 자기상관(AR(1)) 구

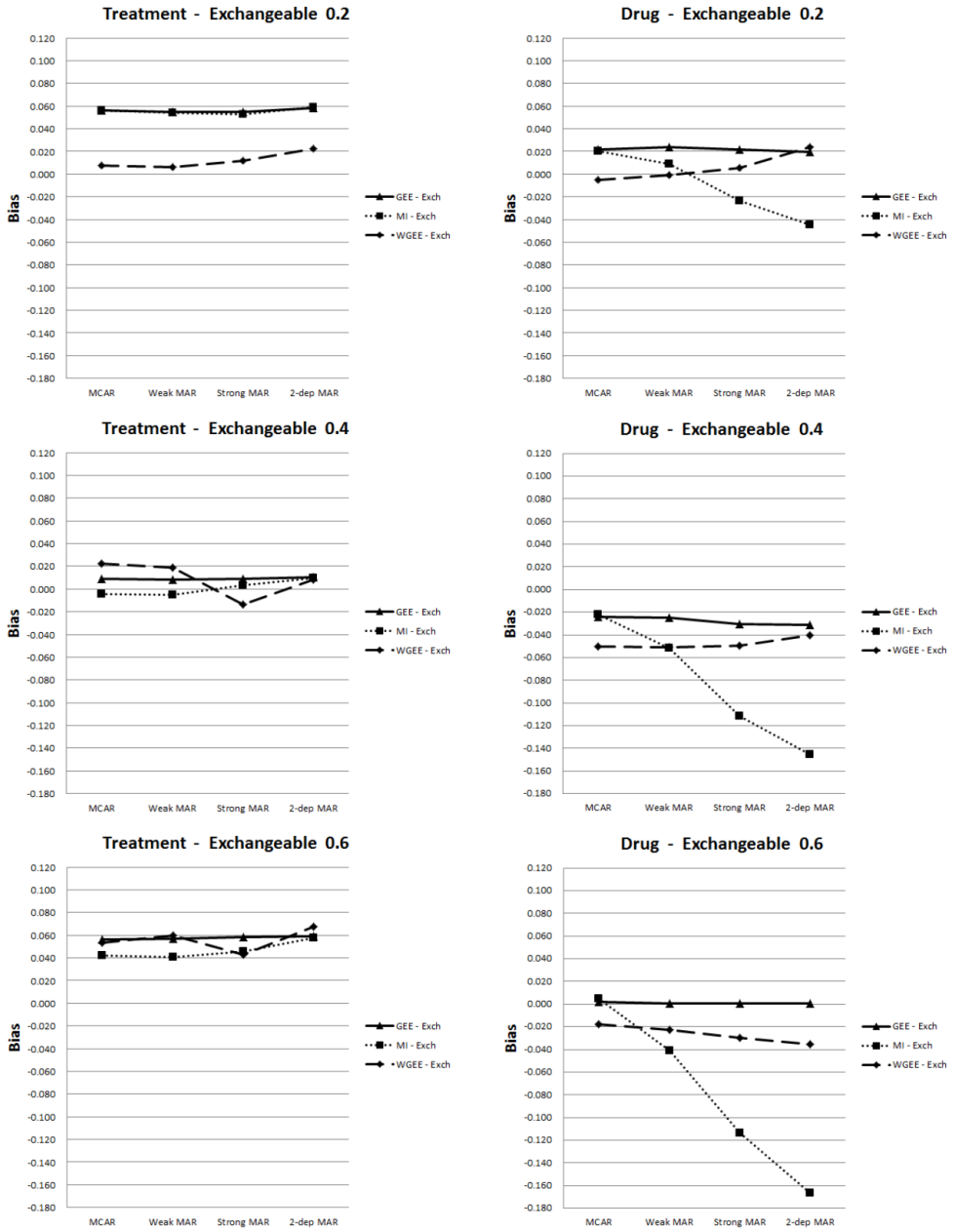


**Figure 4.4.** Bias of parameter estimates for time-invariant and time-varying covariates (Exchangeable)

조와 교환가능한(exchangeable) 구조만 그림에 제시하였다.

Figure 4.3, Figure 4.4를 보면 결측 체계와 적용한 방법에 따라 편차는 다른 경향을 보였다. 시간-독립적 공변량인 경우에 가중 방법은 일반화추정방정식 방법이나 다중대체 방법보다 편차가 더 작게 나타

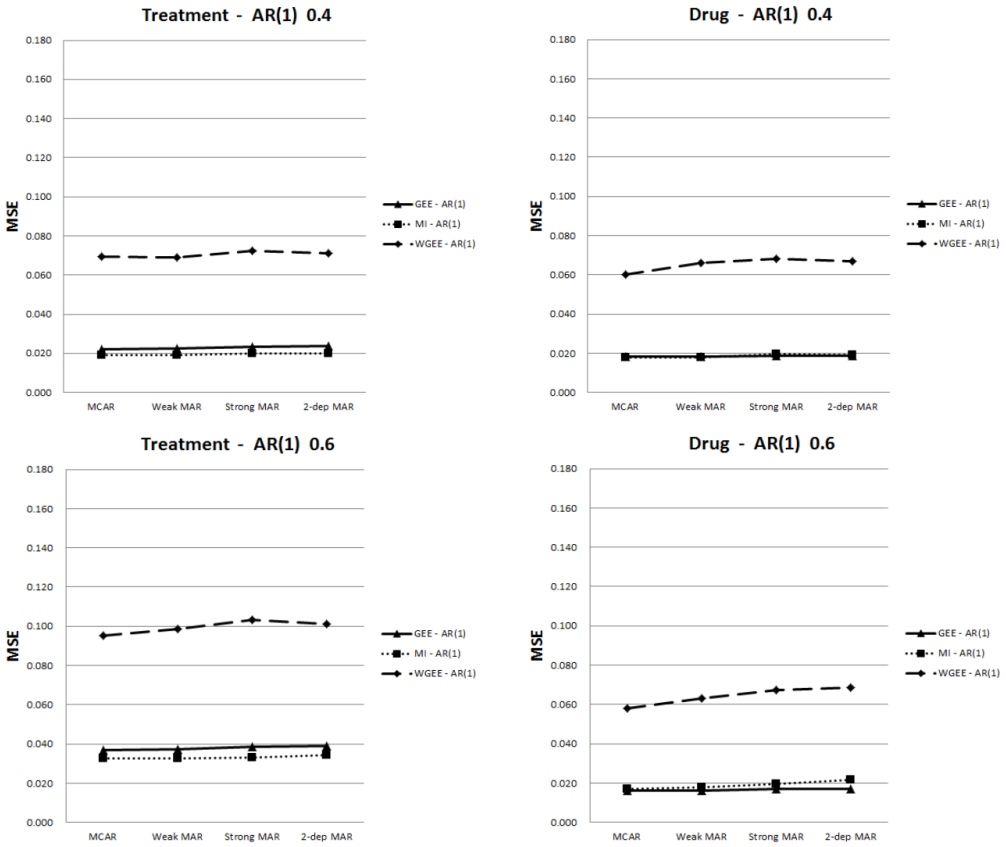


**Figure 4.5.** Mean squared error (MSE) of parameter estimates for time-invariant and time-varying covariates (Autoregressive)

났고, 일반화추정방정식 방법과 다중대체 방법은 비슷한 경향을 보였다. 시간-종속적 공변량인 경우에 MCAR이나 Weak MAR 가정하에서는 가중 방법이 편차가 가장 크게 나타났지만, Strong MAR 또는 2-dep MAR 가정으로 갈수록 다중대체 방법의 편차가 점점 증가하는 경향을 보였다. 즉, 결측 발생이 이전 관측값과 상관성이 강할수록 다중대체 방법은 추정값의 정확성이 떨어지는 것을 확인할 수 있다. 위와 같은 경향은 자료의 상관행렬이 자기상관 구조일 때보다 교환가능한 구조일 때, 그리고 동일한 상관행렬 내에서는 $\rho$값이 증가할수록 더 뚜렷한 형태를 보이고 있다. 또한, 시간-종속적 공변량이 시간-독립적 공변량에 비해 편차의 변화폭이 더 큰 경향을 보였다.

Figure 4.5, Figure 4.6을 보면 결측 체계나 자료의 상관구조에 상관없이 전반적으로 일반화추정방정식 방법과 다중대체 방법보다 가중 방법을 통해 구한 추정값의 평균제곱오차가 더 크게 나타났다. 또한 다중대체 방법이 일반화추정방정식 방법에 비해 평균제곱오차가 시간-독립적 공변량에서는 더 작게 추정되었지만, 시간-종속적 공변량에서는 더 크게 추정되었다. 이는 자료의 상관행렬이 자기상관(AR(1)) 구조일 때보다 교환가능한(exchangeable) 구조일 때, 그리고 동일한 상관행렬 내에서는 $\rho$값이 증가할수록 더 뚜렷하게 나타났다. 분산은 전반적으로 평균제곱오차와 비슷한 패턴을 보였다. 다른 방법에 비해 가중 방법을 적용했을 때 분산이 크게 추정되었고, 이는 가중 방법이 평균제곱오차가 크게 추정되는데

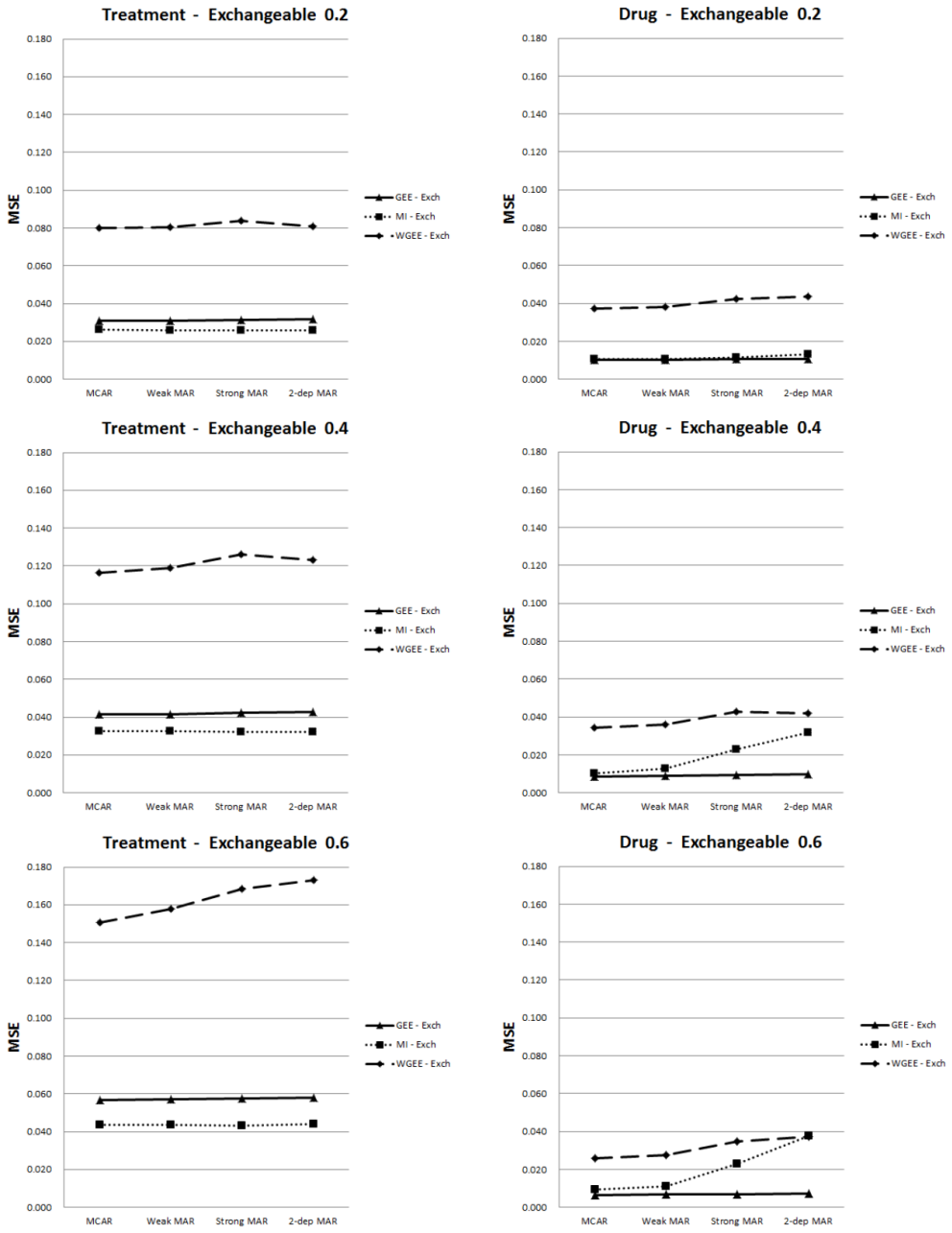


**Figure 4.6.** Mean squared error (MSE) of parameter estimates for time-invariant and time-varying covariates (Exchangeable)

영향을 미쳤다. 전반적으로 가상관행렬이 정확하게 가정되었을 때 잘못 가정된 상관행렬에 비해 편차가 더 낮게 추정되었고, 로버스트 표준오차와 모형에 근거한 표준오차 간에 차이도 더 작게 나타났다.

## 5. 결론

다시점 자료 연구에서 일반화추정방정식 방법은 가상관행렬을 잘못 가정하더라도 모수에 대한 일치추정량을 구할 수 있어서 많이 쓰이고 있다. 하지만 결측 체계가 완전임의결측이 아닌 경우에 편의추정량을 제공하고, 시간-종속적 공변량이 포함된 경우에는 가상관행렬에 따라 회귀계수 추정값이 다르게 도출될 수 있다. 본 논문에서는 일반화추정방정식 방법, 가중 방법, 다중대체 방법을 이용하여 GEE 분석에서 시간-독립적 공변량과 시간-종속적 공변량의 추정값이 가상관행렬에 따라 어떤 양상으로 변화하는지 연구하였다. 반응변수의 다양한 상관구조와 여러 가지 결측체계에서 각 방법의 로버스트성(robustness)을 살펴보고, 참값과 각 방법을 통해 구한 추정값을 이용하여 정확성(accuracy)을 비교하였다. 정확성의 척도로는 편차(bias), 분산(variance), 평균제곱오차(MSE)를 살펴보았고 편차와 분산의 정보를 모두 이용한 평균제곱오차에 근거하여 정확성을 비교하였다.

시간-독립적 공변량은 반응변수의 상관 구조, 결측 체계, 적용하는 방법에 크게 의존하지 않으며 가상관행렬의 형태에 로버스트한 반면, 시간-종속적 공변량은 가상관행렬에 따라 회귀계수 추정값이 다르게 도출되었다. 시간-종속적 공변량에 일반화추정방정식 방법을 적용하면 결측 체계가 이전 시점의 관측값과 상관성이 강할수록 가상관행렬에 따른 회귀계수 추정값의 차이가 크게 나타났다. 하지만 다중대체 방법을 적용하면 가상관행렬에 따른 회귀계수 추정값의 차이가 작아 전반적으로 가상관행렬의 형태에 로버스트하였다.

시간-독립적 공변량은 가중 방법을 적용하였을 때 편차는 가장 작게 분산은 가장 크게 추정되었고, 시간-종속적 공변량인 경우에는 다중대체 방법을 적용하였을 때 편차는 가장 크게 분산은 가장 작게 추정되었다. 그래서 편차와 분산의 정보를 모두 이용한 평균제곱오차에 근거하여 세 가지 방법의 정확성을 비교하였다. 그 결과, 전반적으로 일반화추정방정식 방법과 다중대체 방법에 비해 가중 방법을 적용하였을 때 평균제곱오차가 가장 크게 추정되었다. 시간-독립적 공변량은 다중대체 방법에서 평균제곱오차가 가장 작게 추정되었고, 시간-종속적 공변량은 일반화추정방정식 방법에서 평균제곱오차가 가장 작게 추정되었다.

본 논문에서 반응변수의 상관구조, 결측 체계를 변화시키면서 세 가지 방법을 모두 적용해보았다. 일반화추정방정식 방법은 평균제곱오차가 작아 정확성은 입증하였지만 가상관행렬에 따른 회귀계수 추정값의 차이가 크므로 가상관행렬의 선택이 중요해진다. 가중 방법은 가상관행렬의 형태에 크게 의존하지는 않지만 평균제곱오차가 크게 추정되어 정확성이 떨어지는 경향이 잇다. 다중대체 방법은 가상관행렬에 따른 회귀계수 추정값의 차이도 가장 작고, 평균제곱오차도 작게 추정되어 다른 방법에 비해 더 좋은 추정량을 제공해주는 것을 확인할 수 있었다. 본 논문에서는 반응변수에만 결측이 있는 경우에 대해 살펴보았지만, 실제 자료에서는 공변량에도 결측이 많이 발생한다. 이러한 자료에서는 가중 방법보다 더 정확한 추정을 하는 다중대체 방법을 선호한다 (Beunckens 등, 2008). 또한 시간-독립적 공변량에 비해 시간-종속적 공변량은 가정된 가상관행렬이 모수의 추정에 영향을 미친다. 모형에 시간-종속적 공변량이 포함되어 있을 때, 비대각(non-diagonal) 가상관행렬을 사용하면 GEE와 주변모형의 가정에 어긋나므로 편의추정량을 도출할 수 있다 (Pepe와 Anderson, 1994). 반면에, 독립적 구조인 가상관행렬을 사용하면 시간-종속적 공변량의 회귀계수 추정에 효율(efficiency)이 떨어진다 (Fitzmaurice, 1995). 가상관행렬에 따라 회귀계수 추정값이 달라지는 시간-종속적 공변량은 가상관행렬을 선택하는 데 있어서 구체적인 방법이 필요하리라 여겨진다.

## References

Beunckens, C., Sotto, C. and Molenberghs, G. (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data, *Compu-*

*tational Statistics & Data Analysis*, **52**, 1533–1548.

Faught, E., Wilder, B. J., Ramsay, R. E., Reife, R. A., Kramer, L. D., Pledger, G. W. and Karim, R. M. (1996). Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages, *Neurology*, **46**, 1684–1690.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimation equations with multiple multivariate binary data, *Biometrics*, **51**, 309–317.

Kim, T. H. (2004). *Handling data in GEE with missing response*, Sungkyunkwan University.

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

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, John Wiley & Sons.

Pepe, M. S. and Anderson, G. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data, *Communication in Statistics B*, **23**, 939–951.

Preisser, J. S., Lohman, K. K. and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random, *Statistics in Medicine*, **21**, 3035–3054.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **189**, 846–866.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, **90**, 106–121.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.

Song, J. W. and An, H. (2009). *Handling and Analysis of Missing Data*, Statistical Training Institute, Seoul.

Troxel, A. B., Lipsitz, S. R. and Brennan, T. A. (1997). Weighted estimating equations with nonignorably missing response data, *Biometrics*, **53**, 857–869.

Wall, M. M., Dai, Y. and Eberly, L. E. (2005). GEE estimation of a misspecified time-varying covariate: An example with the effect of alcoholism treatment on medical utilization, *Statistics in Medicine*, **24**, 925–939.

# 시간-종속적 공변량이 포함된 이분형 반복측정자료의 GEE를 이용한 분석에서 결측 체계에 따른 회귀계수 추정방법 비교

박보람[a] · 정인경[b,1]

[a]국립암센터 바이오메트릭연구과, [b]연세대학교 의학통계학과

(2012년 10월 25일 접수, 2013년 7월 31일 수정, 2013년 9월 6일 채택)

**요 약**

다시점 자료 연구에서 일반화추정방정식은 가상관행렬을 잘못 가정하더라도 모수의 일치추정량을 도출하므로 많이 이용된다. 하지만, 결측 체계가 완전임의결측이 아닌 경우에는 편의추정량을 제공하고, 시간-종속적 공변량이 포함된 경우에는 가상관행렬에 따라 회귀계수 추정값이 다르게 도출될 수 있는 문제점이 있다. 결측 체계가 임의결측인 경우에 발생하는 문제를 해결하기 위해 가중 방법과 다중대체 방법을 사용하는 것이 제안되었다. 본 논문에서는 시간-종속적 공변량이 포함된 이분형 반복측정자료를 GEE를 이용하여 분석할 때 다양한 결측 체계에서 일반화추정방정식 방법, 가중 방법, 다중대체 방법의 회귀계수 추정에 대한 로버스트성과 정확성을 모의실험을 통하여 비교해 보았다. 세 가지 방법 모두에서 시간-종속적 공변량의 회귀계수가 시간-독립적 공변량의 회귀계수에 비해 가상관행렬에 따라 추정값의 차이가 크게 나타났다. 다른 두 방법에 비해 다중대체 방법이 가상관행렬의 형태에 대해 더 로버스트하고 편의도 작은 추정치를 도출하였다.

주요용어: 일반화추정방정식, 다중대체, 가중 방법, 완전임의결측, 임의결측.

[1]교신저자: (120-752) 서울시 서대문구 연세로 50, 연세대학교 의학통계학과, 조교수. E-mail: ijung@yuhs.ac