응용통계연구 제 16권 2호, 2003년, pp. 407-426

대체방법별 GEE추정량 비교

김동욱[1] 노영화[2]

요 약

본 연구에서는 범주형 반복측정자료의 일반화 추정방정식(GEE)모형에서 결측이 발생할 경우 결측값 대체(imputation)방법들에 대한 성능을 비교하고자 한다. 설명변수 X가 부분적으로 결측을 갖는 경우 GEE추정량을 계산할 수 없다. 본 논문에서는 시점에 따라 값이 변하는 설명변수에 결측이 있는 경우 GEE모형에서 결측값을 추정하는 7가지의 대체방법을 다루며, 실제자료와 모의실험을 통하여 대체방법별 GEE추정량의 성질을 연구한다. 대체방법별 GEE추정량의 성능을 비교하기 위해 우리는 반응변수가 범주형인 반복측정모형에서 완전자료의 GEE추정량과 완전자료에서 결측을 생성하여 결측값에 각 대체방법을 적용하여 대체한 후 구한 GEE추정량을 비교한다. 대체방법으로는 (1) 단순삭제 (2) 표본 평균대체 (3) 행 평균대체 (4) 횡 시점 회귀대체 (5) 이월대체 (6) 베이지안 붓스트랩 (7) 근사적 베이지안 붓스트랩에 대해서 살펴본다. 결측과정(missing mechanism)은 무시할 수 있는 무응답(ignorable nonresponse)을 가정하며, 결측 발생에 대해서는 원자료의 시점 무응답 패턴(wave nonresponse pattern)을 고려하여 발생시키거나 또는 시점 무응답 패턴을 고려하지 않고 단순임의추출로 결측을 발생시키는 방법을 각각 고려한다.

주요용어: 결측값, 다시점자료, 대체방법, 무시할 수 있는 무응답, 무응답 패턴, 일반화 추정방정식.

1. 서론

시간에 따라 동일한 실험단위의 반복측정이나 집락내의 실험단위들에 대한 측정은 관찰값들 사이에 종속성이 존재하므로 관찰값들 사이의 상관관계를 고려한 모형이 필요하게 된다. Wedderburn(1974)은 유사우도함수(Quasi-likelihood function)를 제안하였으며, Liang과 Zeger(1986)는 이산형과 연속형인 다시점 자료(longitudinal data)의 분석에 이러한 유사우도함수의 장점을 이용한 일반화 추정방정식(Generalized Estimating Equations; GEE)모형을 제안하였다. 또한, 반복 측정된 다항의 범주형 자료의 분석에 GEE모형을 사용할 수 있다(Lipsitz 외, 1994; 김동욱 외, 2002). GEE모형은 모수의 추정을 위해 반복 측정된 관찰값의 결합분포는 구체화시키지 않고 단지 반응변수의 주변분포를 정의하고 반응변수들간의 상관관계를 나타내는 상관행렬의 구조만을 가정한다.

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 부교수
E-mail : dkim@skku.ac.kr
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과
E-mail : hjyh7775@intizen.com

GEE추정량은 시간의 종속성에 대한 약한 가정하에서 회귀모수 추정량과 그 분산의 추정량이 일치추정량이 되며, 근사적으로 정규분포를 따른다(Liang과 Zeger, 1986). 그러나 결측이 있는 경우 GEE모형에서 모수를 추정할 수 없다. 무응답을 처리하기 위해 일반적으로 많이 사용되는 방법은 대체(imputation)방법이다. 대체방법은 단일대체(single imputation)와 다중대체(multiple imputation)로 나눠지며 단일대체는 무응답에 대한 모형하에서 각 결측에 대해 단일대체값으로 대체하는 것이다(Little과 Rubin, 1987; Lepkowski, 1989). 또한 Rubin과 Schenker(1986)와 Rubin(1987)은 다중대체 방법들을 연구하였다. 단일대체는 추정값의 분산이 과소 추정되는 단점이 있으나 다중대체는 알려져 있지 않은 결측값으로부터 기인된 변동을 반영하므로 정확한 분산 추정값을 제공한다.

반복측정 자료에서 결측이 자주 발생하여 결측값 처리를 위해 대체방법이 연구되었으며, 패널조사(panel survey)에서도 시점 무응답 패턴의 결측값 처리가 연구되었다(Lepkowski, 1989; Park, 2002). GEE모형에 결측이 발생한 경우 대체방법에 관한 연구로는 종속변수에 결측이 발생한 경우 결측값 대체 방법이 연구되었으며(Robins 외, 1995; Fitzmaurice 외, 1995; Paik, 1997), Xie와 Paik(1997a, b)은 반응변수가 이진자료일 때 설명변수에 결측이 발생한 경우 결측값 대체 방법을 연구하였다.

이진자료인 반응변수와 단조(monotone)패턴을 갖는 공변량에서 결측 문제가 많이 연구되었으나, 본 논문에서는 반응변수가 세 개 이상의 범주를 갖는 순서자료인 범주형 반복측정 자료이며 공변량에서 시점 무응답 패턴을 고려한 경우와 고려하지 않은 경우의 결측 모형에서 대체방법을 다룬다. 즉 실제자료와 모의실험(simulation)을 통하여 완전한 자료에서 설명변수에 결측을 생성하여 여러 가지 대체방법을 통해 결측값을 대체한 후 일반화 추정방정식모형에 적용하여 구한 GEE 추정값들의 성질을 연구한다. 특히 다시점 자료에서 행 평균 대체방법의 사용을 제안하며, 결측과정(missing mechanism)은 무시할 수 있는 무응답(ignorable nonresponse)을 가정한다. 2절에서는 Liang과 Zeger가 제안한 GEE모형과 결측이 있는 GEE모형에 대해 비교한다. 3절에서는 반복측정모형에서 결측이 있는 경우 대체방법에 대하여 알아본다. 4절에서는 실제자료를 통하여 완전한 자료에서 시점 무응답 패턴(wave nonresponse pattern)을 고려한 경우와 고려하지 않은 경우 각각에 대해 결측을 발생시킨 후 3절에서 언급한 7가지 대체방법을 사용하여 결측값을 대체시켜 대체값과 실제값의 비교 및 GEE 추정량들의 성질을 비교한다. 그리고 5절에서는 모의실험을 통하여 공변량의 시점간 상관계수가 높은 경우와 낮은 경우 각각에 대해 대체방법의 성능을 비교한다.

2. GEE모형

2.1. 결측이 없는 GEE모형

반복 측정된 실험에서 측정된 자료 $y_{ij}(i = 1, \ldots, K, j = 1, \ldots, n_i)$를 i번째 실험단위의 j번째 관찰값이라 하자. 여기서 주어진 실험단위의 관찰값들 사이에는 종속성이 있지만 다른 실험단위의 관찰값과는 서로 독립이다. 그리고 $\mathbf{x}_{ij}$는 i번째 실험단위의 j번째 설명벡터라 하며 y_{ij}의 주변분포는 다음의 지수족 분포를 가정한다.

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

여기서, $\theta_{ij} = h(\eta_{ij}), \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta}$ 이다. 또한 y_{ij}의 평균과 분산은 $E(y_{ij}) = a'(\theta_{ij}), Var(y_{ij}) = a''(\theta_{ij})/\phi$ 이다.

GEE모형은 반복 측정된 관찰값의 결합분포는 구체화하지 않고 y_{ij}의 주변분포만을 고려하며 관찰값들간의 상관관계를 나타내는 가상관(working correlation)행렬을 필요로 한다.

$\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$라 할 때 $\mathbf{Y}_i$ 의 분산-공분산 행렬을 구체화하면 $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}/\phi$ 이다. 여기서, $\mathbf{R}(\boldsymbol{\alpha})$는 가상관행렬이며, $\mathbf{R}(\boldsymbol{\alpha})$가 $\mathbf{Y}_i$의 참상관행렬이면 $\mathbf{V}_i$는 $Cov(\mathbf{Y}_i)$와 동일하다. $\mathbf{A}_i = diag\{a''(\theta_{ij})\}$는 대각행렬이며, ϕ는 산포모수이다. 유사우도함수에 대한 추정방정식을 확장함으로써 일반화 추정방정식은 $\sum_{i=1}^{K} \mathbf{D}_i{}^T\mathbf{V}_i{}^{-1}(\mathbf{Y}_i - \mathbf{a}_i{}'(\theta)) = \mathbf{0}$ 으로 정의되며, $\mathbf{D}_i = d\{\mathbf{a}_i{}'(\theta)\}/d\boldsymbol{\beta}$ 이다.

GEE 모형을 이용하여 추정한 $\hat{\boldsymbol{\beta}}_G$에 대해 $K^{1/2}(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta})$는 근사적으로 다변량 정규분포를 따르며, 평균은 $\mathbf{0}$이고 분산-공분산 행렬 $\mathbf{V}_G$는 다음과 같다.

$$\mathbf{V}_G = \lim_{K\to\infty} K(\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i)^{-1}\{\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}Cov(\mathbf{Y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\}(\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i)^{-1} \tag{2.2}$$

$\mathbf{V}_i = Cov(\mathbf{Y}_i)$이면, 식(2.2)는 다음과 같이된다.

$$\mathbf{V}_G = \lim_{K\to\infty} K(\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i)^{-1} \tag{2.3}$$

식 (2.2)는 로버스트(robust) 분산 추정량 행렬이라 하고 식 (2.3)은 모형에 근거한(model-based) 분산 추정량 행렬이라고 한다(Liang과 Zeger, 1986). 로버스트 분산 추정량은 평균모형이 정확하면 일치추정량이 되며, 모형에 근거한 분산 추정량은 평균 모형과 공분산 모형 둘 다 정확해야 일치추정량이 된다.

2.2. 결측이 있는 GEE모형

GEE모형에서 공변량 $\mathbf{x}_{ij}$에 결측이 있는 경우를 고려하자. 반복측정 자료에서 결측은 흔히 발생하며, 특히 실험단위가 사람이거나 관측이 여러 번에 걸쳐서 일어날 때 결측은 자주 발생된다. 반응변수벡터 $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$에서 각각의 y_{ij}는 세 개 이상의 범주를 갖는 순서자료이며 공변량은 $C_i^T = (Z_i, X_i)$이라 하자. 여기서, Z는 완전하게 관측된 공변량이고 X는 부분적으로 관측된 공변량이다. 만약 X_i가 관측되었으면 $r_i = 1$이고 그렇지 않으면 $r_i = 0$라 하면 r_i는 관측 지시(observation indicator)변수이다. X가 관측될 확률이 X에는 의존하지 않으나 Y와 Z에만 의존한다고 가정하자. 그러면 이 관측 확률은 $Pr(r = 1|Y, Z)$로

표시할 수 있다. 결측과정(missing mechanism)은 X가 관측될 확률로 설명되며, X가 관측될 확률이 X에는 의존하지 않으므로 랜덤결측(missing at random : MAR)이라 한다. 만약 X가 관측될 확률이 X뿐만 아니라 Y와 Z 에도 의존하지 않으면 완전 랜덤결측(missing completely at random : MCAR)이라 한다(Rubin, 1976; Little과 Rubin, 1987; Laird, 1988; Kenward 외, 1994).

Y와 Z가 범주형 자료인 경우 Y와 Z의 각 층 안에서 결측이 랜덤하게 발생하면 완전 랜덤결측보다 약한 가정인 랜덤결측을 사용할 수 있으나 결측발생 과정이 더 세분된다. 또한 결측값은 관측값인 Y와 Z가 주어졌을 때 조건부 분포로부터 결측값을 대체한다. 한편 랜덤결측과 완전 랜덤결측을 무시할 수 있는 무응답(ignorable nonresponse)이라 하며, 무시할 수 없는 무응답(non-ignorable nonresponse)은 결측과정이 결측 자료와 관련이 있는 경우를 의미한다(Rubin, 1976). 본 논문에서는 무시할 수 있는 무응답을 가정한다. GEE모형은 Y의 주변평균과 분산으로 특정되어져 있으며, Y는 연결함수 $g(\cdot)$를 통하여 공변량C와 관련된다. 즉 $\mu = E[Y|C] = g(C\beta)$ 이다. 그런데 X가 부분적으로 결측값을 갖는 경우 일반화 추정방정식에서 $\mathbf{D}^T\mathbf{V}^{-1}$및 $\mathbf{D}^T\mathbf{V}^{-1}\mu$는 X의 함수이므로 계산할 수 없게 된다. 우리는 결측이 있는 반복측정 자료를 GEE모형에 적용하기 위하여 대체방법을 사용하여 결측값을 대체값으로 대체한다.

3. 대체방법

본 논문에서는 다음의 대체방법들을 고려한다.

[1] 단순삭제 (Naive Deletion : ND)

단순삭제 방법(Little과 Rubin, 1987)은 어떤 변수에서 실험단위에 결측이 있을 때 결측값이 있는 실험단위를 제거한 후에 오직 응답이 모두 있는 실험단위들을 가지고 분석을 하는 가장 간단한 방법이다.

[2] 표본 평균대체 (Sample Average Imputation : SA)

표본 평균대체(Little과 Rubin, 1987; Xie와 Paik, 1997a)는 전체 표본을 보조변수의 값에 따라 여러 개 층으로 나눈 후에 각 층 내에서 얻어진 응답자의 평균을 그 층 내의 모든 결측값에 대체하는 방법이다. 즉 Y와 Z의 값에 따라 층을 나눈 후 각 층 내에서 관찰된 X의 평균값으로 결측값을 대체한다.

[3] 행 평균대체 (Row Average Imputation : RA)

행 평균대체는 i번째 실험단위의 어떤 한 시점에서 결측이 있는 경우 i번째 실험단위의 결측을 제외한 나머지 시점들의 응답값의 평균으로 대체하는 방법이다.

[4] 횡 시점 회귀대체 (Cross-wave Regression Imputation: C-W)

Lepkowski(1989)에 의하면 횡 시점 회귀대체는 조사시점의 순서에 따라 표본을 정렬한 후 무응답이 있는 j번째 시점에서 x_j를 종속변수라 두고 응답이 있는 가장 가까운 과거 $j-1$번째 시점의 응답값 x_{j-1}을 독립변수로 하여 결측이 있는 실험단위를 제외한 후 회귀모형을 추정하여 j번째 시점의 i번째 대체값 $\hat{x}_{i,j} = \hat{\beta}_0 + \hat{\beta}_1 x_{i,j-1} + \hat{e}_{i,j}$을 결측값 $x_{i,j}$에 대체한다. 단, $\hat{e}_{i,j} = 0$을 간주한다. 연속해서 j번째 시점과 $j-1$번째 시점에 결측이 있는 경

우 먼저 $j-1$번째 시점의 응답값을 종속변수라 두고 $j-2$번째 시점의 응답값을 독립변수로 하여 회귀모형을 추정한 후, 추정값 $\hat{x}_{i,j-1}$을 결측값 $x_{i,j-1}$에 대체한다. 같은 방법을 사용하여 j번째 시점의 i번째 결측값 $x_{i,j}$에 $\hat{x}_{i,j-1}$을 사용하여 구한 추정값 $\hat{x}_{i,j}$을 대체한다.

[5] 이월대체 (Carry-over Imputation : C-O)

Lepkowski(1989)에 의하면 이월대체는 횡 시점 회귀대체방법에서 구한 모형 $\hat{x}_{i,j}=\hat{\beta}_0+\hat{\beta}_1 x_{i,j-1}+\hat{e}_{i,j}$에서 오차항 없이 $\hat{\beta}_0=0$과 $\hat{\beta}_1=1$을 대입하여 결측값 $x_{i,j}$에 응답값 $x_{i,j-1}$을 대체하는 방법으로서 횡 시점 회귀대체의 간편한 방법이다. 만약 연속해서 j번째 시점과 $j-1$번째 시점에 무응답이 있는 경우 응답이 있는 가장 가까운 과거 $j-2$번째 시점의 응답값 $x_{i,j-2}$를 결측값 $x_{i,j-1}$과 $x_{i,j}$에 대체한다.

다음 두 방법은 베이지안 관점에서 쉽게 이해될 수 있으며, 기본 개념은 자료와 관련된 모수 θ를 그것의 사후분포에서 추출한 후, θ의 값이 주어졌을 때 결측값들이 독립적으로 동일한 분포에서 뽑혀진다는 것이다.

[6] 베이지안 붓스트랩 (Bayesian Bootstrap : BB)

베이지안 붓스트랩 방법(Rubin과 Schenker, 1986; Rubin, 1987; Xie와 Paik, 1997b)은 동일한 확률을 가지는 붓스트랩 방법과는 대조적으로 디리슈레(Dirichlet) 분포를 사용하여 일반화된 확률을 가지고 응답된 X값들로부터 복원 추출한 임의의 표본에서 결측 개수만큼 임의로 뽑는 방법이다.

모든 실험단위, $i=1,\ldots,K$를 고려하여 결측값을 가지는 공변량 X에서 결측값이 m개가 있으며 전체 N개의 자료에서 o개의 응답값을 $X_1,\ldots,X_o$라 하자. 즉 $N=o+m$이다. 균일분포(uniform distribution) $U(0,1)$에서 $o-1$개의 임의의 값을 발생시켜 크기 순서대로 $u(1)<\cdots<u(o-1)$과 같이 배열한 후, 차이 g_q를 구하면 $g_q=u(q)-u(q-1), q=1,\ldots,o$이다. 여기서, $u(0)=0$이고 $u(o)=1$이다. 다음으로 이전에 구한 $g=(g_1,\ldots,g_o)$의 확률을 사용하여 $X_1,\ldots,X_o$로 부터 m개만큼 임의로 추출하여 결측값에 대체한다. 즉 새로운 m개의 균일난수 u'를 뽑아서 만약 $u_{q-1}<u'\leq u_q$ 이면 X_q값을 대체값으로 한다.

[7] 근사적 베이지안 붓스트랩 (Approximate Bayesian Bootstrap : ABB)

근사적 베이지안 붓스트랩 방법(Rubin과 Schenker, 1986; Rubin, 1987; Xie와 Paik, 1997b)은 동일한 확률 $1/o$ (o는 관측값의 갯수)을 가지고 관측된 X값들로부터 복원 추출한 임의의 표본에서 결측값의 수만큼 임의로 뽑는 방법이다. 이 방법은 베이지안 붓스트랩 방법과 유사하며, 이 두 방법의 차이는 관찰된 X값들로부터 붓스트랩 표본을 추출할 때 동일 확률 또는 다른 확률을 사용하는가이다. 베이지안 붓스트랩과 근사적 베이지안 붓스트랩은 Y와 Z가 범주형 자료일 때 쉽게 적용할 수 있다.

위의 대체방법을 다음의 두 가지로 구분할 수 있다.

i) 실험단위 전체의 X자료에서 결측된 자료를 제외한 X의 응답값을 이용하여 대체 : [2], [4], [6], [7]

ii) i번째 실험단위에 결측값이 있는 경우 i번째 실험단위의 나머지 응답값을 이용하여 대체 : [3], [5]

4. 대체방법을 적용한 GEE모형

반복 측정된 순서자료에 대해 결측이 있는 경우 GEE모형을 적용할 때 2절에서도 설명했듯이 X가 부분적으로 결측을 갖는 경우 일반화 추정방정식에서 $\mathbf{D}^T\mathbf{V}^{-1}$및 $\mathbf{D}^T\mathbf{V}^{-1}\mu$는 X의 함수이므로 모수를 추정할 수 없다. 우리는 결측 발생시 무응답 패턴을 고려한 경우와 고려하지 않은 경우의 대체방법의 성능 비교를 위해 실제자료를 사용하며, 결측값을 각각의 대체방법에 따라 대체한 후 모수 추정값들을 비교하여 어떠한 대체방법이 더 효율적인지를 알아보고자 한다.

4.1. 결측 모형

본 논문에서 사용된 자료는 Lipsitz, Kim과 Zhao(1994)가 사용한 자료로서 2개의 도시에서 어머니의 평균 흡연량에 따른 어린이의 천식 정도 (1=천식 없음, 2=추위에 의한 천식, 3=추위와 상관없는 천식)를 1년마다 한번씩 4년간 반복 측정한 자료이다. 이 자료에서 어린이의 천식정도를 반응변수 Y라 두고, 도시(0=Kingston, 1=Portage)를 공변량 Z, 그리고 어머니의 평균 흡연량을 공변량 X라 하였다. 반응변수 Y는 세 개 범주를 갖는 순서자료이며, X는 0에서 60까지의 값을 가지며 0이 많이 발생하는 희박자료(sparse data)이다. 평균 흡연량에 결측이 있는 297명에 대한 자료에서 매년 측정한 자료가 모두 응답이 있는 경우, 즉 297명 중 결측이 없는 완전한 자료인 91명의 자료를 사용하였고 4번씩 반복 측정을 하였으므로 총 관측수는 364개이다.

91명의 결측이 없는 완전한 자료를 평균 흡연량(X)에 대해 결측을 임의로 만들어서 각각의 대체방법들을 비교한다. 따라서 반응변수 Y와 공변량 Z는 결측이 없는 완전한 자료이다. 그리고 시점 무응답 패턴과 무응답 비율이 대체방법 성능에 미치는 영향을 알아보기 위해 X에 결측을 임의로 만들 때 결측이 있는 297명에 대한 전체 자료의

i) 시점 무응답 패턴을 고려한 방법

ii) 시점 무응답 패턴을 고려하지 않은 방법

각각을 사용하여 비교하였다. 다음은 4개 시점에 대한 무응답 패턴을 고려한 방법으로 표 4.1은 위의 자료에서 전체자료의 무응답 패턴의 비율을 알아본 것이다.

표 4.1에서 나타난 무응답 패턴 중 비율이 높게 나온 패턴들만 선택하면 4개 시점이 모두 응답된 패턴 xxxx와 소멸패턴(attrition pattern) 3가지 (xxxo, xxoo, xooo), 그리고 비소멸패턴(non-attrition pattern) 중에서 2가지(oxxx, ooxx)를 선택할 수 있었고, 나머지 패턴들은 큰 영향을 미치지 않으므로 고려하지 않았다. 비율이 높은 6가지 패턴을 전체로 하여 다시 비율을 구한 결과 각각의 무응답 패턴의 비율은 xxxx는 37.0%, xxxo는 15.8%, xxoo는 12.2%, xooo는 8.6%, oxxx는 15.8%, 그리고 ooxx는 10.6%로 나타났다. 이 비율을 91명의 완전한 자료에 적용시켜 결측을 만들었으며 결측률은 25.8% (94/364)가 되었다. 또한 무응답 패턴을 고려하지 않은 방법으로는 단순임의추출법(Simple Random Sampling)을 사용하였으며, 결측률은 무응답 패턴을 고려한 것과 비교하기 위해서 25.8%로 하였다. 그리고 대체방법으로서는 3절에서 설명한 7가지 방법을 사용하여 비교하였다.

표 4.1 전체 자료에서 평균 흡연량의 4개 시점에 대한 무응답 패턴과 그 비율 (응답 : x , 무응답 : o)

시점 무응답 패턴		도수	비율(%)
응답	xxxx	91	30.6
소멸패턴	xxxo	39	13.1
	xxoo	30	10.1
	xooo	21	7.1
비 소멸패턴	xxox	5	1.7
	xoxx	2	0.7
	xoox	2	0.7
	xoxo	2	0.7
	oxxx	39	13.1
	ooxx	26	8.7
	oxox	1	0.3
	oxxo	8	2.7
	ooox	14	4.7
	ooxo	8	2.7
	oxoo	9	3
합계		297	100(%)

4.2. 실제값과 대체값의 비교

3절의 대체방법을 비교하기 위해 각 대체방법별로 결측을 만들기 이전의 실제값과 대체값 사이의 차이제곱의 평균, 절대차이의 평균, 그리고 결측값을 발생하기 이전의 실제값의 표본표준편차와 대체값의 표본표준편차의 차이의 절대값을 각각 500번 반복하여 그 평균을 비교하였다.

표 4.2의 결과를 보면 무응답 패턴을 고려한 경우와 고려하지 않은 경우 모두 평균적으로 RA방법으로 구한 대체값들이 실제값에 가장 근사하며 C-O방법과 C-W방법도 근사하게 나타났다. 특히 SA와 BB, 그리고 ABB는 실제값과 대체값 사이에는 큰 차이가 나타났다. 따라서, 다시점 자료인 경우는 자료가 서로 연관되어 있으므로 결측이 있을 경우 RA나 C-O가 변동이 작으면서 실제값에 가까운 대체값을 제공하며, 특히 공변량 값들이 많이 변하지 않을 때 또는 시점간 상관관계가 높을 때 RA나 C-O가 좋은 대체방법이 된다. 그리고 C-W도 실제값에 가까운 대체값을 제공한다.

표 4.2 차이제곱의 평균, 절대차이의 평균, 표본표준편차의 차이의 절대값
(500번 반복 시행한 평균)

	차이제곱의 평균		절대차이 평균		$\lvert s_i$(실제값) $- s_i$(대체값)$\rvert$	
대체방법	고려	무시	고려	무시	고려	무시
SA	103.811	101.065	8.153	8.078	7.841	7.951
RA	24.776	22.973	2.102	2.044	0.800	0.776
C-W	31.989	54.973	3.177	3.957	1.924	1.544
C-O	31.584	59.970	2.269	3.433	0.829	0.801
BB	206.737	205.248	9.185	9.104	1.553	1.425
ABB	206.864	203.020	9.204	9.085	1.523	1.358

4.3. 대체 후 GEE추정량 비교

다시점 자료에서 대체값이 실제값에 얼마나 가까운지를 각각의 대체방법에 대해서 비교해 보았다. 이제는 4.1절의 자료를 사용하여 공변량X에 결측이 있는 경우 결측값을 대체값으로 대체한 후 GEE모형에 적용하여 GEE 추정량의 성질을 연구한다.

GEE추정량을 구하기 위해서 반응변수 Y의 주변분포와 상관행렬을 구체화해야 한다. 반응변수가 순서자료인 경우 비례오즈모형(proportional odds model)이 사용되며, 반복 측정된 순서자료인 경우 다음과 같은 누적 로짓 모형(cumulative logit model)을 사용한다.

$$\log(\frac{F_{ikt}}{1-F_{ikt}}) = \delta_k + \mathbf{C}_{it}\boldsymbol{\beta}, \qquad k = 1,2, \qquad t = 1,2,3,4. \tag{4.1}$$

여기서, F_{ikt}는 시점 t에서 i실험단위에 대해 반응이 k이하일 확률이다. 시간의 추세에 관한 효과를 고려하기 위해 $T = \{-3,-1,1,3\}$을 모형에 두었다. δ_k는 절편모수(intercept parameter)이고, $\boldsymbol{\beta} = [\beta_C, \beta_S, \beta_T]^T$이다. β_C는 공변량 Z인 도시(city)에 대한 효과, β_S는 공변량 X인 평균 흡연량(smoke)에 대한 효과, 그리고 β_T는 공변량 T인 시간(time)에 대한 효과를 나타낸다.

가상관행렬의 형태는 독립(independent)인 구조와 교환 가능한(exchangeable) 구조, 무리 지어진(banded) 구조, 그리고 비구조적인(unstructured) 가상관행렬을 각각 사용하였다. 그리고 공변량 중 결측값을 대체한 변수가 평균 흡연량이므로 우리가 관심 있는 모수 추정값은 $\hat{\beta}_S$이다.

4.1절의 동일한 자료를 사용하여 실험은 완전한 자료에서 i) 무응답 패턴 고려 또는 ii) 무응답 패턴 무시의 두 가지 방법으로 결측값을 생성하여 각 대체방법에 따라 결측값을 대체한 후 GEE모형에 적용하여 모수 추정값 $\hat{\beta}_S$를 구하였다. 각 대체방법별로 위의 과정을 500번 반복 시행하여 구한 500개 $\hat{\beta}_S$들의 평균, 표준편차, 그리고 평균제곱오차(mean-squared error)를 구하여 실제 완전자료의 모수 추정값 $\hat{\beta}_S$에 얼마나 근사한지를 비교하였다. 각 상관행렬별로 위의 방법을 적용한 결과, 가상관행렬별로 결과가 유사하므로 관심있는 가상관행렬인 교환 가능한 구조에 대해서만 표 4.3에 나타냈다.

표 4.3 각 대체방법에 대해 500번 반복 시행하여 구한 GEE추정값들의 평균, 표준편차, 평균제곱오차 (가상관행렬 : 교환 가능)

대체방법	실제자료	대체자료					
		추정값의 평균		추정값의 표준편차		추정값의 평균제곱오차	
		고려	무시	고려	무시	고려	무시
ND	$\hat{\beta}_s = 0.0025$	0.0027	0.0049	0.019	0.026	0.00036104	0.00068176
SA		0.0054	0.0014	0.014	0.013	0.00020441	0.00017021
RA		-0.0003	0.0008	0.006	0.006	0.00004384	0.00003889
C-W		0.0022	0.0028	0.008	0.008	0.00006409	0.00006409
C-O		0.0006	0.0031	0.007	0.007	0.00005261	0.00004936
BB		0.0021	0.0005	0.010	0.010	0.00010016	0.00010400
ABB		0.0020	0.0006	0.011	0.010	0.00012125	0.00010361

패턴고려 수렴횟수 : ND=430, SA=500, RA=500, C-W=500, C-O=500, BB=500, ABB=500
패턴무시 수렴횟수 : ND=408, SA=500, RA=500, C-W=500, C-O=500, BB=500, ABB=500

표 4.3에서 가상관행렬의 형태가 교환 가능한 구조일 때 완전한 자료의 모수 추정값 $\hat{\beta}_S$은 0.0025로 나타났다. 먼저, 무응답 패턴을 고려한 경우를 살펴보면 ND, C-W, BB, ABB 방법으로 구한 추정값의 평균이 실제 모수 추정값 $\hat{\beta}_S$에 근사하게 나타났다. 그러나 추정량의 평균제곱오차를 살펴볼 때 RA와 C-O 및 C-W방법이 작게 나타났다. 또한 무응답 패턴을 고려하지 않은 경우에도 C-W방법과 C-O방법으로 구한 추정값의 평균이 실제 모수 추정값 $\hat{\beta}_S$에 가장 근사하며, 추정량의 평균제곱오차를 살펴볼 때 RA와 C-O 및 C-W방법이 작게 나타났다. 그러므로 가상관행렬이 교환 가능한 구조인 실험에서 무응답 패턴 고려 유무에 상관없이 RA와 C-O 및 C-W방법이 더 효율적이다. 그리고 다른 가상관행렬(독립, 무리 지어진 구조, 비구조)에 대한 실험에서도 무응답 패턴 고려 유무에 상관없이 동일한 결론을 얻었다.

다음은 각각의 가상관행렬별로 무응답 패턴을 고려한 경우와 고려하지 않은 경우에 대해 C-O방법에 대한 다른 대체방법의 상대효율(relative efficiency)을 비교하였다. 표 4.4를 보면 상대효율 측면에서 C-O방법이 RA방법을 제외한 다른 대체방법들 보다 효율적이며, 가상관행렬이 교환가능한 구조인 경우에는 무응답 패턴 고려 유무에 상관없이 RA방법이 C-O방법보다 더 효율적이다. 그리고 C-W방법도 C-O방법과 효율성에 큰 차이가 나지 않았다. 따라서, 가상관행렬 구조에 상관없이 RA와 C-O 및 C-W방법이 효율적으로 나타났다.

표 4.4 가상관행렬별로 이월대체(C-O)방법에 대한 다른 대체방법의 상대효율

가상관행렬	$\frac{MSE(ND)}{MSE(C-O)}$ 고려	$\frac{MSE(ND)}{MSE(C-O)}$ 무시	$\frac{MSE(SA)}{MSE(C-O)}$ 고려	$\frac{MSE(SA)}{MSE(C-O)}$ 무시	$\frac{MSE(RA)}{MSE(C-O)}$ 고려	$\frac{MSE(RA)}{MSE(C-O)}$ 무시
독립	29.990	48.966	6.516	5.035	1.000	0.993
교환가능	6.863	13.812	3.885	3.448	0.833	0.788
무리	5.176	16.656	3.260	3.459	0.902	1.126
비구조	9.979	27.607	2.624	3.235	1.103	0.923

가상관행렬	$\frac{MSE(C-W)}{MSE(C-O)}$ 고려	$\frac{MSE(C-W)}{MSE(C-O)}$ 무시	$\frac{MSE(BB)}{MSE(C-O)}$ 고려	$\frac{MSE(BB)}{MSE(C-O)}$ 무시	$\frac{MSE(ABB)}{MSE(C-O)}$ 고려	$\frac{MSE(ABB)}{MSE(C-O)}$ 무시
독립	0.990	1.000	5.015	4.035	5.028	3.983
교환가능	1.218	1.298	1.904	2.107	2.305	2.099
무리	1.129	1.311	1.674	2.197	1.684	2.648
비구조	1.094	1.272	2.031	2.412	2.062	2.763

5. 모의실험

반복측정 모형에서 대체방법의 성능을 일반적인 상황에서 비교하기 위해 많은 모의실험을 행했다. 4절에서 설명한 7가지 대체방법의 성능을 결측 생성 이전의 완전한 반복측정 순서자료에서의 결과와 비교하였다. 많은 연구에서 반응변수는 이진자료를, 결측 발생은 단조(monotone)패턴을 가정하였다. 그러나 우리는 종속변수의 범주의 수가 세 개인 반복측정 순서자료이며, 시점에 따라 값이 변하는 공변량에서 무응답 패턴을 고려하여 결측을 발생시키거나 무응답 패턴을 고려하지 않고 랜덤하게 결측을 발생시키는 경우를 각각 연구한다. 그리고 결측이 발생하는 공변량 X의 시점간 상관이 높은 경우와 시점간 상관이 낮은 경우 각각에 대해 대체방법의 성능을 비교한다.

5.1. 자료생성과 모형설정

모의실험에서 실험단위 N=100 이고 반복 관찰횟수 $T_i = 4,\ i = 1,\ldots,N$, 그리고 반응범주 수 K=3 으로 모형을 설정했다. 공변량 $Z_{ij},\ j = 1,\ldots,4$는 성공확률 $p = 0.5$인 베르누이(Bernoulli) 시행에서 생성하였고 각 실험단위 내에서 고정되었다. 공변량 X_{ij}는 표준정규난수 4개를 발생시켜 4×1벡터 $\boldsymbol{N}$을 만든 후 상관관계가 강한 경우와 약한 경우 각각의 X를 생성하기 위하여 다음의 상관행렬 $\boldsymbol{R}$을 각각 사용하였다.

$$\mathbf{R(s)} = \begin{pmatrix} 1 & 0.8 & 0.6 & 0.5 \\ & 1 & 0.8 & 0.6 \\ & & 1 & 0.8 \\ & & & 1 \end{pmatrix} \qquad \mathbf{R(w)} = \begin{pmatrix} 1 & 0.4 & 0.3 & 0.2 \\ & 1 & 0.4 & 0.3 \\ & & 1 & 0.4 \\ & & & 1 \end{pmatrix} \tag{5.1}$$

분산 공분산행렬이 $\boldsymbol{R}$인 다변량 정규난수를 생성하기 위해 $\boldsymbol{R}$의 정규화된 고유벡터들이 열로 구성된 행렬 $\boldsymbol{Q}$와 이 고유벡터에 대응하는 고유값들이 대각요소인 대각행렬 $\boldsymbol{D}$를 구하여 공변량 $\mathbf{X} = \mathbf{QD}^{1/2}\mathbf{N}$를 생성하였다. 그 후 $\boldsymbol{X}$에 $\boldsymbol{\mu}' = [3,3,3,3]$을 더하여 $\mathbf{X} \sim N(\mathbf{3},\mathbf{R})$을 생성하였다.

반응범주의 개수가 세 개인 순서자료 Y를 생성하기 위해 우리는 반복비례적합(Iterative Proportional Fitting) 알고리즘을 이용한 Gange(1995)방법을 사용하였으며, 순서난수 Y를 생성하기 위해 Y의 결합분포를 먼저 생성하여야 한다. Y의 결합분포를 생성하기 위해 모집단 회귀모수 $\boldsymbol{\beta}$, 설계행렬(design matrix) X_Z 그리고 상관행렬 R을 정의해야 한다. 비례오즈모형에서 절편모수는 $\delta_1 = -0.5$, $\delta_2 = 0.5$ 로 두며 회귀모수 $\boldsymbol{\beta} = [0.5, 0.3]^T$라 두고 설계행렬 X_Z는 공변량 Z의 값에 따라 다음과 같이 두 그룹으로 나누었다.

$$\mathbf{X}_0 = \begin{pmatrix} 0 & 3 \\ 0 & 3 \\ 0 & 3 \\ 0 & 3 \end{pmatrix} \qquad \mathbf{X}_1 = \begin{pmatrix} 1 & 3 \\ 1 & 3 \\ 1 & 3 \\ 1 & 3 \end{pmatrix} \tag{5.2}$$

그리고 상관행렬 R은 8×8행렬로서 교환 가능구조를 사용하였으며, 대각블록은 2×2인 단위행렬(identity matrix)이며 비대각블록은 4절에서 사용된 자료에서 구해진 결과를 다음과 같이 사용하였다(김동욱 외, 2002).

$$\mathbf{r}_{ij}(\text{비대각}) = \begin{pmatrix} 0.38 & 0.13 \\ 0.13 & 0.1 \end{pmatrix} \tag{5.3}$$

여기서, 구체화된 $\boldsymbol{\beta}$, X_Z, R을 이용하여 공변량 $Z = 0$ 혹은 1에 따라 Y의 결합분포를 각각 생성하였으며 순서자료 Y를 생성할 때는 공변량 Z의 값에 따라 각각 생성된 결합분포에서 0, 1, 2의 순서난수를 발생시켰다.

이와 같이 Z, X, Y 에 대한 완전자료를 발생한 후 결측 발생은 4절에서 사용된결측률 25.8%를 적용하였으며, 생성된 반복측정 순서자료 100명에 대해 무응답 패턴을 고려한 경우에는 표 4.1의 무응답 패턴과 비율에 따라 결측을 만들었고 무응답 패턴을 고려하지 않은 경우에는 단순임의추출법을 사용하였다. 따라서 결측 수는 103개이며 이와 같은 모의실험 과정을 500번 반복 시행하였다.

생성된 자료의 반응변수가 순서자료이므로 비례오즈모형이 사용되며, 반복 측정된 순서자료 모형은 식 (4.1)과 같은 누적 로짓 모형을 사용한다. 여기서, $\boldsymbol{\beta} = [\beta_Z, \beta_X]^T$이다. β_Z는 공변량 Z에 대한 효과, β_X는 공변량 X에 대한 효과를 나타낸다. 그리고 공변량 중 결측값을 대체한 변수가 X이므로 우리가 관심 있는 모수 추정값은 $\hat{\beta}_X$이다.

5.2. 모의실험 결과

5.2.1. 실제값과 대체값의 비교

5.1절에서 생성된 반복측정 자료에서 공변량 X의 실제값과 결측 발생 후 대체값 사이의 차이제곱의 평균 및 절대차이의 평균, 그리고 실제값의 표본표준편차와 대체값의 표본표준편차의 차이의 절대값을 각각 500번 씩 반복하여 그 평균을 비교하였다.

표 5.1 차이제곱의 평균, 절대차이의 평균, 표본표준편차의 차이의 절대값
(500번 반복 시행한 평균)

대체방법	차이제곱의 평균		절대차이 평균		$\lvert s_i$(실제값) $- s_i$(대체값)$\rvert$	
	고려	무시	고려	무시	고려	무시
	$corr(X_i, X_j)$: 강한경우					
SA	1.011	1.009	0.801	0.799	0.834	0.843
RA	0.651	0.477	0.636	0.540	0.087	0.092
C-W	0.499	0.620	0.555	0.596	0.260	0.220
C-O	0.583	0.722	0.596	0.636	0.072	0.057
BB	1.989	1.998	1.124	1.127	0.093	0.090
ABB	2.002	2.008	1.129	1.128	0.087	0.085
	$corr(X_i, X_j)$: 약한경우					
SA	1.023	1.016	0.806	0.803	0.864	0.855
RA	1.151	1.007	0.851	0.797	0.159	0.189
C-W	0.915	0.943	0.762	0.773	0.653	0.594
C-O	1.307	1.372	0.910	0.930	0.084	0.083
BB	1.988	2.005	1.126	1.130	0.094	0.092
ABB	1.989	1.984	1.126	1.122	0.080	0.084

표 5.1의 모의실험 결과를 보면 시점간의 상관관계가 강한 경우와 약한 경우 둘 다 무응답 패턴을 고려하는 것에 상관없이 RA와 C-W 및 C-O방법으로 구한 대체값들이 실제값에 근사하였으며 C-O방법으로 구한 대체값의 표본표준편차가 실제값의 표본표준편차에 가장 근사하였다. 그리고 시점간의 상관관계가 약한 경우 SA방법도 실제값에 근사하였다.

그림 5.1과 그림 5.2는 대체값이 실제값에 어느 정도 근사한 지를 산점도(scatter plot)를 통해 시점간의 상관관계가 강한 경우와 약한 경우에 따라 각각 알아보았다. 무응답 패턴을 고려한 경우와 무시한 경우간에 차이가 나지 않으므로 무응답 패턴을 고려하지 않은 경우에 대해서 나타냈다. 산점도에서 수평축은 각 방법으로 구한 대체값이며 수직축은 실제값을 나타낸다. 시점간의 상관관계가 강한 경우는 RA와 C-W 및 C-O방법이 직선에 가깝게 나타나며 시점간의 상관관계가 약한 경우는 강한 경우보다 많이 퍼져 있다.

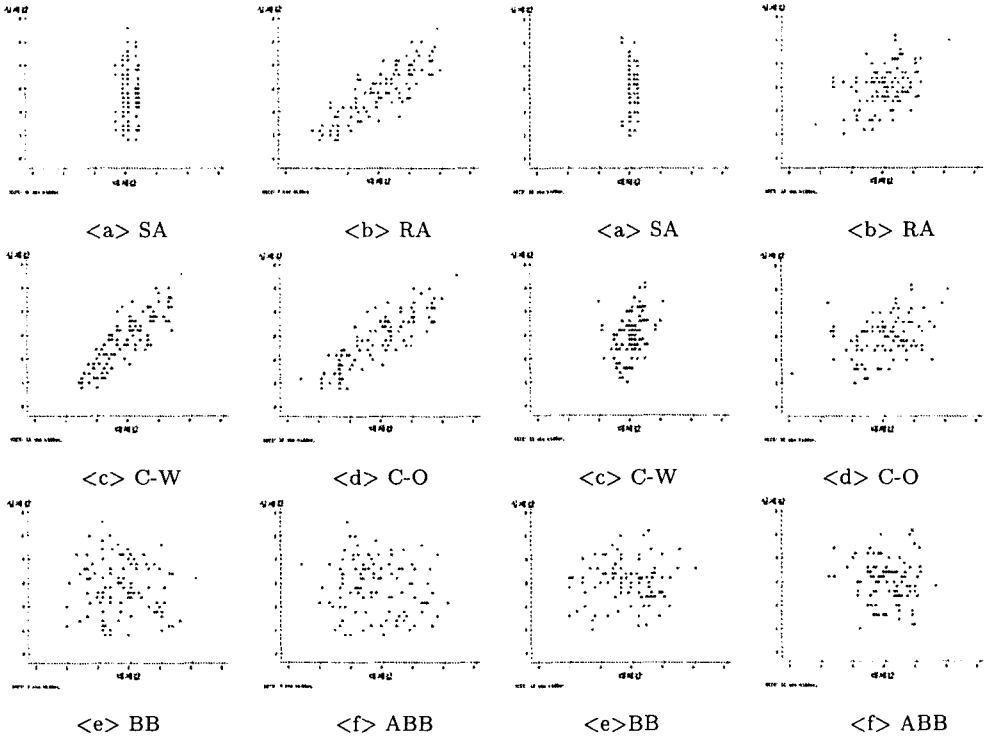


그림 5.1 대체방법별 실제값과 대체값의 산점도
$corr(X_i, X_j)$:강한경우,무응답 패턴무시

그림 5.2 대체방법별 실제값과 대체값의 산점도
$corr(X_i, X_j)$:약한경우,무응답 패턴무시

본 모의실험에서 실제값과 대체값을 비교해 본 결과, 무응답 패턴고려 유무에 상관없이 시점간에 상관관계가 강한 경우는 RA와 C-W 및 C-O방법이 실제값에 가까운 대체값을 제공하며 시점간의 상관관계가 약한 경우에도 이 방법들이 다른 방법들보다 실제값에 더 가까운 대체값을 제공한다.

5.2.2. GEE추정량 $\hat{\beta}_X$의 성질

공변량의 시점간 상관관계가 강한 경우와 약한 경우에 따라 각 대체방법별 GEE추정량의 성질을 알아보았다. 무응답 패턴을 고려한 경우와 고려하지 않은 경우 각각에 대해 공변량 X에 결측을 발생시켜 각 대체방법으로 결측값을 대체한 후 종속변수가 순서자료인 GEE모형에 적용하여 GEE추정량을 구하였다. 위의 과정을 500번 반복 시행하여 구한 모수 추정값 $\hat{\beta}_X$들의 평균과 표준편차를 구하여, 이들 결과가 결측을 발생하기 전의 완전자료(full data; FUL)의 결과에 얼마나 근사한지를 알아보았다. 각각의 상관행렬별로 위의 방법을 적용한 결과 가상관행렬별로 유사한 결과가 나와서 관심있는 교환 가능한 구조에 대

표 5.2 GEE추정값 $\hat{\beta}_X$들의 평균, 표준편차 (500번 반복시행, 가상관행렬: 교환가능)

대체방법	추정값의 평균		추정값의 표준편차	
	고려	무시	고려	무시
	$corr(X_i, X_j)$: 강한경우			
FUL	-0.0050	-0.0011	0.117	0.121
ND	-0.0048	0.0049	0.207	0.239
SA	-0.0028	-0.0006	0.181	0.176
RA	-0.0116	-0.0011	0.129	0.133
C-W	-0.0082	-0.0003	0.131	0.133
C-O	-0.0097	-0.00004	0.125	0.128
BB	0.0029	-0.0017	0.099	0.103
ABB	-0.0005	0.0007	0.105	0.103
	$corr(X_i, X_j)$: 약한경우			
FUL	-0.0013	0.0071	0.127	0.102
ND	0.0016	0.0108	0.264	0.198
SA	0.0006	0.0197	0.191	0.153
RA	-0.0038	0.0110	0.143	0.119
C-W	-0.0024	0.0135	0.142	0.119
C-O	-0.0035	0.0094	0.137	0.114
BB	-0.0034	0.0118	0.115	0.099
ABB	-0.0003	0.0079	0.108	0.096

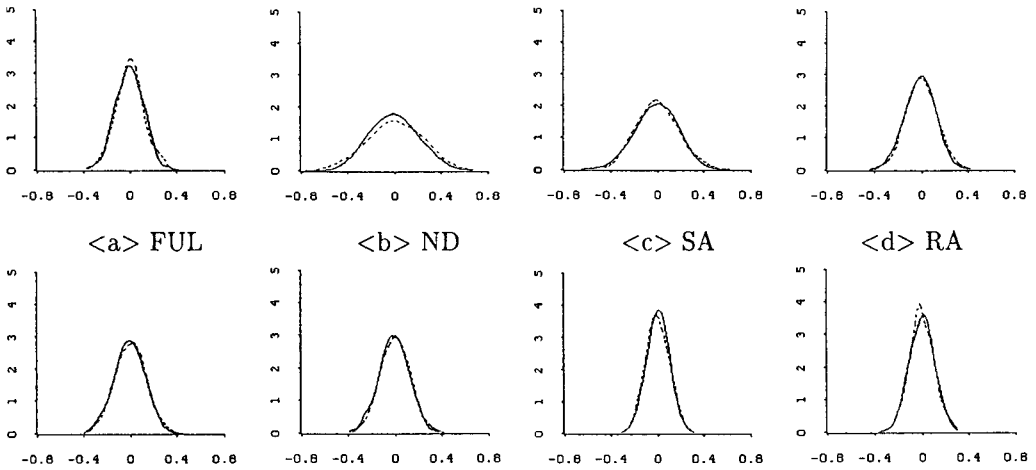
FUL: 생성된 반복측정 순서자료

해서만 나타냈다.

표 5.2의 결과와 다른 가상관행렬인 경우의 결과를 종합하면 무응답 패턴을 고려하였을 때 시점간의 상관관계가 강하거나 약한 경우 둘 다 C-W와 ABB방법으로 구한 추정값의 평균이 FUL방법으로 구한 추정값의 평균에 근사하였고 추정값의 표준편차도 작았다. 그리고 무응답 패턴을 고려하지 않은 경우 시점간의 상관관계가 강한 경우에는 RA방법이 가장 근사하였으며, 시점간의 상관관계가 약한 경우에는 ABB방법이 가장 근사하였다. 그러나 ND와 SA방법은 추정값들의 표준편차가 FUL방법의 결과와 큰 차이가 났다.

다음은 무응답 패턴고려 유무에 따라 표 5.2에서 구한 GEE추정값 $\hat{\beta}_X$들의 표본분포를 시점간의 상관관계가 강한 경우와 약한 경우에 따라 커널밀도함수(kernel density function)를 이용하여 가상관행렬이 교환 가능한 구조인 경우를 그림 5.3과 그림 5.4에 나타냈다.

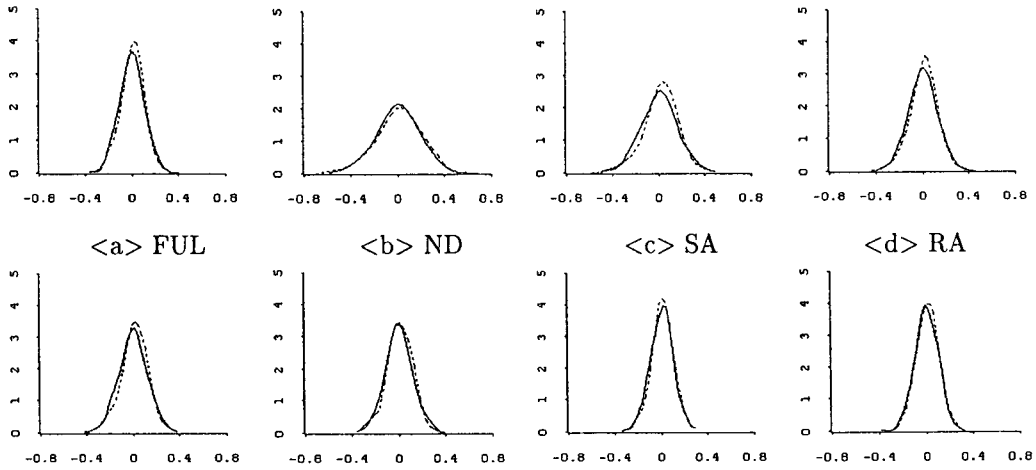
그림 5.3의 시점간의 상관관계가 강한 경우 무응답 패턴고려 유무에 상관없이 RA와 C-W 및 C-O방법으로 대체한 후 구한 GEE추정값 $\hat{\beta}_X$들의 분포가 FUL방법으로 구한 $\hat{\beta}_X$들의 분포에 가장 근사하며, 그림 5.4의 시점간의 상관관계가 약한 경우에는 BB와 ABB방법으로 구한 분포가 FUL방법의 결과에 가장 근사하였다.



<a> FUL <b> ND <c> SA <d> RA

<e> C-W <f> C-O <g> BB <h> ABB

그림 5.3 대체방법별 GEE 추정값 $\hat{\beta}_X$들의 분포 (실선: 패턴고려 , 점선 : 패턴무시)
($corr(X_i, X_j)$: 강한경우 , 가상관행렬: 교환가능, 500번 반복시행)



<a> FUL <b> ND <c> SA <d> RA

<e> C-W <f> C-O <g> BB <h> ABB

그림 5.4 대체방법별 GEE 추정값 $\hat{\beta}_X$들의 분포 (실선: 패턴고려 , 점선 : 패턴무시)
($corr(X_i, X_j)$: 약한경우 , 가상관행렬: 교환가능, 500번 반복시행)

5.2.3. GEE추정량 $\hat{\beta}_X$의 분산 추정값

GEE추정량 $\hat{\beta}_X$뿐 아니라 $\hat{\beta}_X$의 분산 추정값의 정확성은 모수 추론시 매우 중요하다. 따라서 우리는 시점간에 상관관계가 강한 경우와 약한 경우에 따라 무응답 패턴을 고려한 경우와 고려하지 않은 경우 각각의 $\hat{\beta}_X$의 로버스트 분산 추정값들과 모형에 기초한 분산 추정값들을 500번 반복 시행하여 평균을 각 대체 방법별로 비교하였다. 각각의 상관행렬별로 분산 추정값들의 평균을 비교한 결과 가상관행렬별로 유사한 결과가 나와서 교환 가능한 구조에 대해서만 나타냈다.

표 5.3 GEE추정값 $\hat{\beta}_X$의 분산 추정값들의 평균 (500번 반복시행, 가상관행렬: 교환가능)

	로버스트				모형			
	강한경우		약한경우		강한경우		약한경우	
대체방법	고려	무시	고려	무시	고려	무시	고려	무시
FUL	0.0142	0.0141	0.0099	0.0099	0.0142	0.0142	0.0100	0.0100
ND	0.0395	0.0487	0.0275	0.0338	0.0396	0.0500	0.0280	0.0346
SA	0.0172	0.0164	0.0130	0.0128	0.0166	0.0160	0.0129	0.0127
RA	0.0170	0.0166	0.0129	0.0128	0.0171	0.0169	0.0130	0.0129
C-W	0.0176	0.0164	0.0131	0.0129	0.0175	0.0167	0.0132	0.0130
C-O	0.0162	0.0152	0.0117	0.0113	0.0162	0.0155	0.0118	0.0114
BB	0.0111	0.0106	0.0090	0.0088	0.0109	0.0106	0.0091	0.0089
ABB	0.0108	0.0105	0.0089	0.0090	0.0108	0.0105	0.0091	0.0090

표 5.3을 보면 무응답 패턴고려 유무에 상관없이 ND방법으로 구한 분산 추정값의 평균이 FUL방법으로 구한 분산 추정값의 평균과 가장 차이가 많이 났다. 무응답 패턴고려 유무에 상관없이 시점간의 상관관계가 강한 경우는 RA, C-O, 그리고 C-W방법으로 구한 분산 추정값들의 평균이 FUL방법보다 조금 크게 추정되었으나 FUL방법의 분산 추정값의 평균에 가장 근사하였다. 시점간의 상관관계가 약한 경우는 BB와 ABB방법으로 구한 분산 추정값들이 평균적으로 FUL방법보다 작게 추정되었으나 이들 평균이 FUL방법의 분산 추정값의 평균에 가장 근사하였다.

6. 결론

GEE모형에서 시점에 따라 값이 변하는 공변량에 결측이 있는 경우 반복측정 순서자료에서 대체방법에 따른 추정량의 성질을 Lipsitz 외(1994)가 사용한 실제자료와 몬테칼로 시뮬레이션(Monte Carlo simulation)을 통하여 각각 알아보았다.

먼저, 실제자료를 이용한 결과를 종합하면 RA와 C-O 및 C-W방법으로 구한 대체값이 다른 대체방법으로 구한 대체값보다 실제값에 더 근사하게 나타났다. GEE모형을 적합했을 때 각 대체자료로서 구한 GEE추정값 $\hat{\beta}_S$들이 완전한 자료에서 구한 실제 GEE추정값 $\hat{\beta}_S$에

얼마나 근사하는지를 비교해 본 결과, 가상관행렬 구조에 상관없이 RA와 C-O 및 C-W방법의 효율성이 다른 방법에 비해 매우 높게 나타났다. 또한 지면관계상 본 논문에서 결과를 나타내지 않았지만 결측률이 20%와 30%인 경우 각각의 대체방법의 성능을 비교하였다. 각 결과는 결측률의 변화에 민감하지 않았으며, 본 결과와 유사한 결과를 얻었다.

다음으로 모의실험 결과를 종합하면 실제값과 대체값을 비교하였을 때 RA와 C-W 및 C-O방법으로 구한 대체값이 실제값에 가장 근사하게 나타났으며, 특히 시점간의 상관관계가 강한 경우가 실제값에 더 근사하게 나타났다. 그리고 각 대체방법별로 구한 GEE추정량 $\hat{\beta}_X$가 FUL방법으로 구한 GEE추정량 $\hat{\beta}_X$에 얼마나 근사한지를 비교해 본 결과, 추정값의 평균과 표준편차를 고려할 때 무응답 패턴을 고려하지 않은 일반적인 경우에서 시점간 상관이 강하면 RA방법이, 그리고 시점간 상관이 약하면 ABB방법이 FUL방법에 가장 근사하였다. 또한 GEE추정량 $\hat{\beta}_X$의 표본분포와 분산 추정값을 비교할 때 FUL방법으로 구한 분포와 분산 추정값에 가장 근사한 방법은 무응답 패턴고려 유무에 상관없이 시점간의 상관관계가 강한 경우에는 RA와 C-W 및 C-O방법이며, 시점간의 상관관계가 약한 경우에는 BB와 ABB방법이다.

따라서, 본 실제자료와 모의실험 결과를 종합하면 사용된 실제자료처럼 0이 많이 발생하는 희박자료인 경우 7가지의 대체방법 중에서 효율적인 대체방법으로는 RA와 C-O 및 C-W방법이라고 할 수 있으며, 또한 이 자료에서 공변량의 시점간 상관계수는 높게 나타났다. 모의실험에서는 공변량의 시점간 상관관계가 강한 경우는 RA와 C-W 및 C-O방법을 사용하는 것이 매우 좋게 나타났다. 이는 반복측정 자료인 경우 시점에 영향을 받으므로 시점간 상관관계가 높은 경우 결측값을 대체할 때 각 실험단위에서 관찰값 평균을 사용하여 대체하는 RA방법과 단지 결측값 이전 시점의 관찰값으로 대체하는 C-O방법, 그리고 두 시점간의 회귀모형을 사용하여 대체하는 C-W방법을 사용하는 것이 매우 좋게 나타났다. 따라서 반복측정 순서자료에서 공변량 값들이 많이 변하지 않는 자료 구조를 가지거나 결측이 있는 공변량들간에 상관관계가 높으면 RA나 C-O 및 C-W방법이 추천된다. 또한 시점간의 상관관계가 약한 경우는 ABB방법이 좋게 나왔다. 그리고 본 논문에서는 전반적으로 무응답 패턴을 고려한 경우와 고려하지 않은 경우 추정값에 차이가 나지 않았다.

모의실험에 사용된 프로그램을 소개하면, 순서자료 난수 생성은 Gange(1995)가 S-plus로 작성한 프로그램을 참조하였으며, GEE추정량에 대한 계산부분은 Lipsitz, Kim과 Zhao(1994)가 SAS매크로를 이용하여 작성한 프로그램을 수정하여 사용하였다.

참고문헌

[1] 김동욱, 김재직 (2002). 범주형 반복측정자료를 위한 일반화 추정방정식의 소표본 특성, <응용통계연구>, 제15권 2호, 297-310.

[2] Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs, *Journal of the Royal Statistical Society*, B, Vol. 57, 691-704.

[3] Gange, S. J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm, *The American Statistician,* Vol. 49, No 2, 134-138.

[4] Kenward, M. G., Lesaffre, E. and Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random, *Biometrics,* Vol. 50, 945-953.

[5] Laird, N. (1988). Missing data in longitudinal studies, *Statistics in Medicine,* Vol. 7, 305-315.

[6] Lepkowski, J. M. (1989). Treatment of wave nonresponse in panel surveys, in *Panel Survey* (D. Kasprzyk, ed.), John Wiley & Sons, 348-374.

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

[8] Lipsitz, S. R., Kim, K. and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations, *Statistics in Medicine,* Vol. 13, 1149-1163.

[9] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data,* John Wiley & Sons.

[10] Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random, *Journal of the American Statistical Association,* Vol. 92, 1320-1329.

[11] Park, J. (2002). A combined method compensating for wave nonresponse, *Journal of the Korean Statistical Society,* Vol. 31, 1-14.

[12] 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,* Vol. 90, 106-121.

[13] Rubin, D. B. (1976). Inference and missing data, *Biometrika,* Vol. 63, 581-590.

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

[15] Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *Journal of the American Statistical Association,* Vol. 81, 366-374.

[16] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika,* Vol. 61, 439-447.

[17] Xie, F. and Paik, M. C. (1997a). Generalized estimating equation model for binary

outcomes with missing covariates, *Biometrics,* Vol. 53, 1458-1466.

[18] Xie, F. and Paik, M. C. (1997b). Multiple imputation methods for the missing covariates in generalized estimating equation, *Biometrics,* Vol. 53, 1538-1546.

[2003년 3월 접수, 2003년 6월 채택]

Comparison of GEE Estimators Using Imputation Methods

Donguk Kim [1] Younghwa Noh [2]

ABSTRACT

We consider the missing covariates problem in generalized estimating equations(GEE) model. If the covariate is partially missing, GEE can not be calculated. In this paper, we study the performance of 7 imputation methods to handle missing covariates in GEE models, and the properties of GEE estimators are investigated after missing covariates are imputed for ordinal data of repeated measurements. The 7 imputation methods include i) Naive Deletion ii) Sample Average Imputation iii) Row Average Imputation iv) Cross-wave Regression Imputation v) Carry-over Imputation vi) Bayesian Bootstrap vii) Approximate Bayesian Bootstrap. A Monte-Carlo simulation is used to compare the performance of these methods. For the missing mechanism generating the missing data, we assume ignorable nonresponse. Furthermore, we generate missing covariates with or without considering wave nonresponse patterns.

Keywords: Missing value; longitudinal data; imputation method; ignorable nonresponse nonresponse pattern; generalized estimating equations.

1) Associate Professor, Department of Statistics, Sungkyunkwan University.
E-mail: dkim@skku.ac.kr
2) Department of Statistics, Sungkyunkwan University.
E-mail: hjyh7775@intizen.com