한국통계학회논문집 제2권 2호
1995년 12월 pp. 155-165

반복측정된 포아송 자료의 GEE 분석에서 산포모수의 역할에 관한 연구1)

박태성2), 신민웅3)

요 약

반복측정자료의 분석을 위해 제안된 Liang and Zeger(1986)의 회귀모형은 일반화추정식(generalized estimating equations, GEE)을 이용하여 모형의 모수를 추정한다. 이 모형은 반복측정된 반응변수와 설명변수들과의 관계를 추정하는 것이 주된 목적이기 때문에 회귀모수는 중요한 모수로 간주되나 산포모수는 중요하지 않은 장애모수(nuisance parameters)로 간주된다. 일반적으로 GEE 분석에서 회귀모수의 추정량은 산포모수에 상관없이 일치적(consistent)으로 얻어진다고 알려져 있다. 그러나 본 논문에서는 포아송분포를 따르는 반복측정자료에 대한 사례연구와 모의실험을 통해서 일반적으로 믿어져왔던 것과는 달리 GEE 방법이 산포모수에 민감하게 영향을 받고 있음을 보였다. 특히 산포모수의 값이 일정하지 않은 경우에는 GEE 방법에서 밝혀진 회귀모수 추정량의 일치성에도 문제가 발생할 수 있음을 보였다.

1. 서 론

반복측정자료는 한 개체로부터 다른 시간이나 다른 실험 조건으로부터 반복적으로 관측치가 얻어진 자료를 말한다. 이러한 반복측정자료의 분석은 생물학을 비롯한 기초과학 분야에서, 또 공장에서의 여러 공정 과정에서, 또 새로운 치료 방법을 개발하기 위한 의학 분야 등에서 광범위하게 사용된다. 특히 자료를 얻을 수 있는 개체(subject)의 수가 적어서 많은 자료를 얻을 수 없을 때 한 개체로부터 반복적인 관측을 통해 많은 자료를 얻어 효과적인 분석을 하기 위해 사용된다.

처음 반복측정자료를 분석하기 위한 통계적 모형은 어린이들의 성장 과정을 통계적 모형으로 추정하기 위해 Potthoff 와 Roy(1964)에 의해 제안된 성장곡선모형이 주종을 이루었으나 최근에는 그 응용 분야가 보다 다양해지면서 보다 일반화된 형태의 통계 모형이 사용되고 있다. 반복측정자료는 반응변수의 형태에 따라 다양한 통계모형이 사용된다. 반응변수가 연속형인 경우는 다변량 정규분포에 기초한 최대우도추정법이 많이 사용되며 이들 추정량을 구하기 위한 알고리즘들이 최근에 개발되어 SAS PROC MIXED나 BMDP의 5V등과 같은 프로그램에서 사용되고 있다 (Laird 와 Ware, 1982; Jennrich 과 Schluchter, 1986; Laird, Lange, Stram, 1987).

1) 이 연구는 1994년도 한국과학재단지원 연구비지원에 의한 결과임(과제번호: 94-0701-01-01-3).
2) (449-071) 경기도 용인군 모현면 한국외국어대학교 자연과학대학 통계학과
3) (449-071) 경기도 용인군 모현면 한국외국어대학교 자연과학대학 통계학과

반응변수가 연속형이 아니고 이산형으로 두개의 값을 갖거나 범주형 자료의 값을 갖는 경우에 표본의 크기가 클 때는 가중최소제곱법(weighted least squares method)에 근거한 통계 모형이 사용된다. 이 모형은 일변량 범주형 자료를 분석하기 위해 제안된 모형(Grizzle, Starmer, Koch, 1964)을 범주형 반복측정자료를 분석할 수 있도록 확장한 모형이다(Koch, et. al, 1972; Koch, et. al, 1977). 최근에 이 모형을 기초로 범주형 반복측정자료를 분석하기 위한 여러 기법들이 제안되었다(Landis, Miller, Davis, Koch, 1988; Park and Davis, 1993). 이러한 가중최소제곱법에 근거한 통계 모형은 SAS의 PROC CATMOD를 이용하여 쉽게 추정과 검정을 할 수 있는 장점이 있다. 그러나 이 모형은 반응변수와 독립변수가 모두 범주형일 때에만 사용할 수 있다.

최근 들어 연속형의 반응변수와 이산형 혹은 범주형의 반응변수를 모두 다룰 수 있는 새로운 통계 모형이 개발되어 이에 대한 관심이 깊어지고 있다. 이 모형은 일변량 자료의 분석에 널리 사용되는 일반화선형모형(generalized linear model)을 반복측정자료의 분석을 위해 확장한 모형이다(Liang and Zeger, 1986; Zeger and Liang, 1986; Wei and Stram, 1988; Moulton and Zeger, 1989). 이 모형의 특징은 한 개체에서 여러 다른 실험 조건 혹은 다른 시간에 관측된 반응변수들 간의 결합확률분포에 대한 아무런 가정 없이 단지 각각의 주변확률분포에 대한 가정만 가지고 모형에 대한 추정이 가능한 점이다. Liang과 Zeger(1986)는 이 모형을 추정하기 위하여 우도방정식과 같은 역할을 하는 일반화추정방정식(generalized estimating equations, GEE)을 제시하고 이 방정식을 이용한 추정방법을 제시하였다. 이러한 모형은 정규분포 외에 포아송분포, 이항분포, 감마분포를 따르는 여러 반응변수에 모두 응용이 될 수 있는 장점을 가지고 있다.

Liang과 Zeger의 모형은 반복측정된 반응변수에 관심있는 설명변수들이 미치는 영향을 추정하는 것이 주된 목적이기 때문에 회귀모형의 모수는 중요한 모수로 간주되나 그 외에 산포모수나 반응변수들 간의 상관관계를 나타내는 공상관행렬(correlation matrix)은 장애모수(nuisance parameters)로 간주된다. 공상관행렬은 여러 형태의 구조적(structured) 행렬이나 비구조적(unstructured)행렬이 사용될 수 있다. 구조적 행렬은 상관행렬이 특정의 구조를 갖고 있음을 가정하는 것이고 비구조적 행렬은 아무런 구조를 갖고 있지 않음을 가정한 것이다. 실제로 여러 연구들을 통하여 회귀모형의 추정이 공상관행렬의 형태에 로버스트(robust)하다는 사실이 알려져 있다. 그러나 산포모수가 회귀모형에 미치는 영향에 관한 연구는 전무한 상태이다. 그러나 일반적으로 공상관행렬과 마찬가지로 회귀모형의 추정이 산포모수에 대해서도 로버스트 하다고 믿어지고 있다.

본 연구에서는 특별히 반복측정된 반응변수가 포아송분포를 따르는 경우에 대해서 산포모수가 회귀모형의 추정에 미치는 영향을 고찰해 보았다. 포아송분포를 따르는 반복측정자료에 대한 사례연구와 모의실험을 통해서 산포모수의 값이 시간에 따라 일정하지 않은 경우에는 GEE 방법이 산포모수에 민감하게 영향을 받고 있음을 보였다. 특히 산포모수의 값이 일정하지 않은 경우에는 GEE 방법에서 밝혀진 회귀모수 추정량의 일치성에도 문제가 발생할 수 있음을 지적하였다.

2절에서는 회귀모형과 GEE 추정방법을 정리하였고 3절에서는 포아송분포를 따르는 반복측정자료 분석을 통해서 산포모수의 영향력을 예시하였고 4절에서는 모의실험을 통해서 산포모수가 회귀모수의 추정에 미치는 효과를 측정하였다. 마지막 절에서는 결론을 제시하였다.

2. 회귀모형과 GEE 추정방법

먼저 모형을 설명하기 위해 y_{ij}는 i번째 ($i=1,...,n$) 개체로부터 관측시간 혹은 관측조건 $j(j=1,...,t)$에 얻어진 반응변수를 나타낸다고 하자. 또 같은 시간 혹은 조건에서 관측된 설명변수의 $p\times1$ 벡터를 $x_{ik}=(x_{ij1},...x_{ijp})'$로 표시하고 이 경우에 반복은 t개의 서로 다른 시간에 관측되어졌거나 t번의 반복된 실험을 통해 실시된 것으로 간주하자. 한 개체로부터 얻어진 반응변수들 간에는 상관관계가 존재하나 다른 개체들로부터 얻은 반응변수들은 서로 독립이다.

y_{ij}의 주변확률분포는

$$f(y_{ij};\theta_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij}-b(\theta_{ij})}{a(\phi)}+c(y_{ij},\phi)\right\}$$

와 같은 지수족분포를 따른다고 가정하자. 이 때

$$E(y_{ij}) = \mu_{ij} = b'(\theta_{ij}) \quad \text{이고}$$

$$Var(y_{ij}) = b''(\theta_{ij})a(\phi)$$

이다. 여기서 ϕ는 미지 혹은 기지의 산포모수가 된다. 반응변수가 포아송분포를 따르는 경우는 $a(\phi)=\phi=1$이나 실제로 관측된 반응변수가 도수(frequency)를 나타내는 변수라도 포아송분포보다 더 과분포되어(over-dispersed) 평균값보다 더 큰 분산을 가지게 되는 경우는 $\phi>1$이 성립하고 ϕ는 관측되는 반응변수가 과분포된 정도를 나타내는 산포모수로 사용된다.

Liang과 Zeger(1986)가 제안한 주변모형은 $\mu_{ij}=E(y_{ij})$ 에 대해

$$h(\mu_{ij}) = x_{ij}\beta$$

이다. 여기서 $\beta=(\beta_1,...\beta_p)'$ 는 미지의 $p\times1$ 모수벡터이고 h는 기지의 연결함수(link function)이다. 만약 y_{ij}가 포아송분포를 따르는 경우에는 정준연결함수(canonical link function)인 로그함수 $h(x)=\log(x)$ 가 주로 사용된다. 연결함수에 대한 자세한 설명은 McCullagh and Nelder(1989)에 서술되어 있다.

i번째 개체로부터 관측된 반응벡터를 $y_i=(y_{i1},\cdots,y_{it})'$ 라고 표시하면 y_i의 공상관행렬은 $t\times t$ 행렬로 주어진다. 이 공상관행렬을 미지의 모수 벡터 α의 함수라고 가정하고 $R(\alpha)$라고 표현하자. 흔히 사용되는 구조적 $R(\alpha)$의 종류는 단위행렬(identity matrix)과 일차자기상관(first-order autocorrelation, AR-1) 행렬과 교환가능(exchangeable) 행렬 등이 있다. 단위행렬은 같은 개체로부터 다른 시간에 얻어진 반응변수들이 서로 독립임을 가정한다. AR-1행렬은 1개의 모수 α의 함수로 임의의 두 관측시간 j, j' ($=1$, ... t)에 대해 $corr(y_{ij},y_{ij})=\alpha^{|j-j'|}$ 임을 가정하며 교환가능

행렬도 역시 1개의 모수 α의 함수로 j, j'에 대해 $corr(y_{ij}, y_{ij}) = \alpha$ 가 성립함을 가정한다. 비구조적상관(unstructured correlation)행렬은 공상관행렬에 특정의 구조를 가정하지 않고 $t(t-1)/2$개의 모든 원소를 모두 모수로 간주한다. Liang and Zeger(1986) 방법의 장점은 잘못된 형태의 공상관행렬 $R(\alpha)$을 사용하더라도 β에 대한 추정량이 일치적으로(consistently) 얻어질 수 있다는 것이다.

Liang and Zeger(1986)은 다음의 두 단계를 반복하여 β 및 α와 ϕ를 추정하는 방법을 제안하였다.

<1단계> (β 추정)

<2단계>에서 추정된 ϕ 와 α로부터 모형 $h(\mu_{ij}) = x_{ij}\beta$ 를 만족하는 β를 다음의 일반화방정식(GEE)의 해로 구한다.

$$\sum_{i=1}^{n} D_i' \; V_i^{-1} S_i = 0 \,,$$

여기서 $V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} / a(\phi)$ 이고 $A_i = diag\{b''(\theta_{ij})\}$ 이고 $D_i = \partial b'(\theta_i)/\partial\beta$이다.

<2단계> (ϕ 와 α의 추정)

<1단계>에서 추정된 β를 이용해 얻어진 Pearson의 잔차 $r_{ij} = \dfrac{y_{ij} - \widehat{\mu_{ij}}}{\sqrt{g(\widehat{\mu_{ij}})}}$ 로부터 ϕ와 α를 로버스트 방법으로 추정한다. 여기서 $g(\widehat{\mu_{ij}})$는 V_i의 j번째 대각선 원소를 나타낸다.

여러 종류의 $R(\alpha)$에 대하여 α를 추정하는 구체적인 방법은 Liang and Zeger(1986)에 기술되어 있다. 위의 방정식을 만족하는 최종해를 $\hat{\beta}$ 라고 표현하면 Liang and Zeger(1986)의 정리2에 의해 $\hat{\beta}$는 근사적으로 다음의 대표본 분포를 따르게 된다.

$$\sqrt{n}\,(\hat{\beta} - \beta) \rightarrow N(0, V_\beta).$$

여기서

$$V_\beta = \lim_{n\to\infty} (\sum_{i=1}^{n} D_i' \; V_i^{-1} D_i)^{-1} (\sum_{i=1}^{n} D_i' \; V_i^{-1} cov(y_i) V_i^{-1} D_i) (\sum_{i=1}^{n} D_i' \; V_i^{-1} D_i)^{-1}$$

이다. 이 때 ϕ는 서로 상쇄되어 V_β의 추정에 아무런 영향을 미치지 않게 된다. 즉 산포모수 ϕ는 $\hat{\beta}$의 일치성(consistency)과 $\hat{\beta}$의 분산과는 무관하게 된다.

산포모수 ϕ는 일반적으로 과포화된 정도를 나타내는 모수로 해석이 될 수 있다. 반복측정자료

는 같은 개체로부터 다른 관측시간에 (혹은 다른 실험조건하에서) 자료를 얻기 때문에 반응변수 y_{ij}의 변동이 j에 따라 다르게 나타날 수 있다. 만약 y_{ij}가 포아송분포를 따르고 μ_{ij}가 일정하다면 이 변동은 ϕ에 반영이 되어야 한다. 그러나 Liang and Zeger의 방법은 'ϕ가 모든 j에 대해 일정하다'는 강한 가정을 전제로 하고 있다. 3절에서는 실제 예제를 통해서 이 가정이 만족되지 않는 경우를 제시하고 4절에서는 소표본 모의실험 연구를 통해서 ϕ의 효과를 수리적으로 측정해 보았다.

3. 포아송반복측정자료의 분석

<표 1>에 있는 자료는 73명의 유아들을 대상으로 1년 (4분기) 동안 병원을 방문하는 횟수를 조사한 것이다 (Karim, 1989). 이 횟수에 영향을 미칠 것으로 생각되는 설명변수들은 성별, 엄마의 흡연 여부, 연구 시작시 아동의 나이이다. 그러나 Karim의 분석 결과에 의하면 이 모든 설명변수들은 방문횟수에 유의한 영향을 미치지 않은 것으로 나타났다. 여기서는 방문횟수가 분기별로 증가하는 지를 보기 위해 분기를 나타내는 설명변수 $x_{ij}=j$ 만을 갖는 간단한 모형을 생각해보자. 즉 log 연결함수를 사용하여 $\log\mu_{ij}=\beta_0+\beta_1 j,\ j=1,\cdots,4,\ i=1,\cdots,73$ 와 같은 모형을 추정해보자.

표 1
4분기 동안 병원 방문 횟수

방문회수	분기			
	1	2	3	4
0	25	39	38	52
1	22	16	15	13
2	14	7	12	6
3	5	8	1	1
4	2	3	4	0
5	3	0	2	1
6	0	0	0	0
7	2	0	1	0

<표 2>는 여러 상관행렬에 대하여 GEE방법으로 모형을 추정한 결과들을 정리한 표이다. <표 2>에서 알 수 있듯이 β_0와 β_1에 대한 추정값은 상관행렬의 형태에 상관없이 일정한 값을 가짐을 알 수 있다. 또한 방문횟수는 선형적으로 줄어드는 경향을 보이고 있고 줄어드는 속도는 대략적으로 $\exp(-0.3)=0.7408$ 가 된다. 다음은 산포모수의 추정값을 구하여 <표 3>에 정리하였다. 첫번째 열은 상관행렬의 형태를 나타내고 두번째 열은 산포모수 ϕ가 j에 상관없이 일정하다는 가정 하에서 구한 추정값이고 세번째 열에 있는 4개의 값은 ϕ가 j에 따라 변한다고 가정한 후에 Park (1993)이 제안한 방법으로 추정한 값이다. ϕ의 추정값은 상관행렬의 형태에 관련없이 일정

하게 추정이 되었으나 j=3 일 때가 다른 경우에 비해 2배 정도 큰 값을 갖는 것을 알 수 있다. 즉 j=3 일 때 관측된 자료가 다른 경우에 비해 과분포된 정도가 심함을 나타내고 있다. 그러면 이렇게 서로 다른 ϕ가 모수 β의 추정에 미치는 영향은 어떠한가? 다음 절에서는 ϕ의 효과를 모의실험 연구를 통해서 이 문제에 대해 조사해보았다.

표 2
추정된 결과

모수	상관행렬	추정값	표준오차	Z-값
절편 β_0	단위행렬	0.625	0.152	4.10
	교환가능	0.626	0.152	4.11
	AR-1	0.648	0.155	4.18
	비구조적	0.678	0.143	4.75
기울기 β_1	단위행렬	-0.294	0.067	-4.40
	교환가능	-0.294	0.067	-4.40
	AR-1	-0.307	0.067	-4.59
	비구조적	-0.314	0.064	-4.92

표 3 ϕ의 추정값

상관행렬	ϕ의 추정값	
	ϕ가 일정할 때	ϕ가 j에 따라 변할 때
단위행렬	1.864	(1.842, 1.444, 3.145, 1.491)
교환가능	1.863	(1.877, 1.473, 3.247, 1.534)
AR-1	1.886	(1.836, 1.462, 3.264, 1.556)
비구조적	1.861	(1.806, 1.452, 3.267, 1.573)

4. 모의실험

이 절에서는 산포모수 ϕ가 모수 β의 추정에 미치는 영향을 조사하기 위해 모의실험 연구를 실시하였다. 2개의 서로 다른 그룹에서 각 개체로부터 3번(t=3) 반복되어 자료가 얻어졌다고 가정한 후 설정한 모형은 $E(y_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{1ij} * x_{2ij}$이다. 여기서 x_{1ij}는 첫번째 그룹이면 0이고 두번째 그룹이면 1의 값을 갖고 $x_{2ij} = j$, $i = 1, \cdots, n$, j=1,2,3,이다. 자료를 생성하기 위해서 $(\beta_0, \beta_1, \beta_2, \beta_3)' = (0.1, 0.2, 0.2, 0.1)'$ 이고 각 그룹의 개체수는 25와 50으로 가정하고 다음과 같은 세종류의 상관행렬을 사용했다.

교환가능 AR-1 비구조적

$$\begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0.5 & 0.1 \\ 0.5 & 1 & 0.3 \\ 0.1 & 0.3 & 1 \end{bmatrix} .$$

산포모수의 효과를 보기 위해 ϕ값이 j에 따라 변할 수 있다는 가정하에서 $(\phi_1,\phi_2,\phi_3)=(1,1,1)$, $(2^{1/3},2^{2/3},2)$, $(2,2,2)$ 의 값을 갖도록 설정하였다. 상관된 포아송 확률변수값을 생성시키기 위해 Park, Park, and Shin(1995)이 제안한 알고리즘을 사용하였다. 상대편차(relative bias), 평균제곱오차(MSE)와 95% 신뢰구간의 포함확률(coverage probability)를 구해서 모의실험의 결과를 <표 4>(n=25)와 <표 5>(n=50)에 정리하였다.

<표 4>와 <표 5>에서 볼 수 있는 바와 같이 모의실험 결과는 상관행렬의 형태에 상관없이 일정한 결과를 보여주고 있다. 일반적으로 $(\phi_1,\phi_2,\phi_3)=(1,1,1)$인 경우에는 안정적인 결과를 보여주나 (ϕ_1,ϕ_2,ϕ_3)가 $(2^{1/3},2^{2/3},2)$ 와 $(2,2,2)$ 로 변함에 따라 상대편차의 절대값과 MSE의 값은 늘어나고 포함확률은 거꾸로 줄어드는 경향을 보이고 있다. 이 경향은 $\hat{\beta}$와 $\widehat{var}(\hat{\beta})$에서 공통적으로 관찰되고 있다. 여기서 유의할 점은 $(\phi_1,\phi_2,\phi_3)=(2^{1/3},2^{2/3},2)$인 경우가 $(\phi_1,\phi_2,\phi_3)=(2,2,2)$인 경우보다 $\hat{\beta}_0$와 $\hat{\beta}_2$의 상대편차가 상당히 크고 포함확률은 상당히 작게 됨을 알 수 있다. 즉 산포모수의 값이 시간에 따라 일정하지 않은 경우에는 GEE 방법에서 밝혀진 추정량의 일치성에도 문제가 발생할 수 있음을 암시해 준다. 따라서 일반적으로 믿어져왔던 것과는 달리 GEE 방법이 산포모수에 민감하게 영향을 받고 있음을 보여 주고 있다.

5. 결론

본 연구에서는 GEE 방법에서 반복측정된 반응변수가 포아송분포를 따르는 경우에 산포모수가 회귀모형의 추정에 미치는 영향을 고찰해 보았다. 본 연구를 통해서 일반적으로 믿어져왔던 것과는 달리 GEE 방법이 산포모수에 아주 민감하게 영향을 받고 있다는 사실을 보였다. 따라서 산포모수가 일정하지 않고 관측시간에 따라 조금씩 변하게될 때 GEE추정량의 일치성이 만족되지 않을 수 있기 때문에 GEE 방법을 적용하기에 앞서서 산포모수가 일정한 지를 조사해 보는 과정이 필요하게 된다. 이 산포모수가 일정한지를 검정하기 위한 검정법을 개발하는 것도 바람직한 연구과제라고 생각된다. 참고로 반복측정된 반응변수가 정규분포를 따르는 경우에 산포모수 ϕ는 모든 y_{ij}의 공통된 분산이 된다. 즉 GEE 방법은 모든 가 공통의 분산을 갖고 있다는 가정을 전제로 하고 있는 것이다. 그러나 실제 관측된 자료가 이런 강한 조건을 만족시키기 어렵기 때문에 이 조건을 완화시키는 보다 일반화된 GEE 모형의 개발이 필요하다. Park(1993)이 제안한 방법도 GEE 방법을 일반화시킨 모형으로 볼 수 있을 것이다.

표4 모의실험 결과 (n=25)

상관행렬	Φ	모수	$\hat{\beta}$			$\widehat{Var}(\hat{\beta})$	
			상대편차	MSE	포함확률	상대편차	MSE
AR-1	(1, 1, 1)	β_0	-0.24720	0.06121	93.8	-0.01957	0.00172
		β_1	0.07955	0.09762	95.19	-0.01258	0.00161
		β_2	0.01548	0.00956	94.08	-0.02046	0.00028
		β_3	-0.01580	0.01402	95.99	-0.01253	0.00025
	$(2^{\frac{1}{3}}, 2^{\frac{2}{3}}, 2)$	β_0	-3.12018	0.17134	80.4	-0.097311	0.0026
		β_1	0.19882	0.12397	94.0	-0.10876	0.0032
		β_2	1.27818	0.07551	25.8	-0.21648	0.00095
		β_3	0.08655	0.016317	93.8	-0.22368	0.00143
	(2, 2, 2)	β_0	2.08105	0.11287	82.8	-0.26844	0.01027
		β_1	0.79457	0.12973	93.5	-0.31275	0.02186
		β_2	0.44009	0.01765	84.4	-0.30307	0.00195
		β_3	-0.35688	0.01591	92.5	-0.33575	0.00373
교환가능	(1, 1, 1)	β_0	-0.31645	0.05042	95.19	-0.01051	0.00134
		β_1	0.09085	0.08450	95.69	-0.00591	0.00117
		β_2	0.03628	0.00623	95.19	-0.00940	0.00020
		β_3	-0.03895	0.00993	96.49	-0.00239	0.00017
	$(2^{\frac{1}{3}}, 2^{\frac{2}{3}}, 2)$	β_0	-0.304910	0.15608	77.30	-0.06616	0.00205
		β_1	0.14052	0.10122	94.50	-0.07506	0.00216
		β_2	1.26750	0.07166	12.00	-0.18728	0.00057
		β_3	0.13499	0.01142	95.10	-0.18565	0.00076
	(2, 2, 2)	β_0	2.11819	0.10006	81.6	-0.25117	0.00784
		β_1	0.78074	0.11243	92.7	-0.29886	0.01670
		β_2	0.44316	0.01412	83.5	-0.27870	0.00118
		β_3	-0.35250	0.01121	94.2	-0.31552	0.00227
비구조적	(1, 1, 1)	β_0	-0.29832	0.07048	94.96	-0.00440	0.00189
		β_1	0.09609	0.11895	96.17	-0.00330	0.00163
		β_2	0.03979	0.01112	94.76	-0.00519	0.00030
		β_3	-0.06144	0.01890	95.27	-0.00333	0.0026
	$(2^{\frac{1}{3}}, 2^{\frac{2}{3}}, 2)$	β_0	-3.08001	0.18218	83.6	-0.08251	0.00279
		β_1	0.16277	0.13733	94.5	-0.09232	0.00311
		β_2	1.26170	0.07708	35.8	-0.17954	0.00086
		β_3	0.13883	0.02053	93.8	-0..18842	0.00126
	(2, 2, 2)	β_0	1.98337	0.11280	86.2	-0.25830	0.01096
		β_1	0.82309	0.14642	93.1	-0.30341	0.02332
		β_2	0.46583	0.01989	88.1	-0.28958	0.00216
		β_3	-0.38527	0.01939	94.1	-0.32512	0.00422

표5 모의실험 결과 (n=50)

상관행렬	ϕ	모수	$\hat{\beta}$			$\widehat{Var}(\hat{\beta})$	
			상대편차	MSE	포함확률	상대편차	MSE
AR-1	(1, 1, 1)	β_0	-0.16329	0.02934	95.4	-0.00829	0.00040
		β_1	0.05715	0.04668	96.1	-0.00518	0.00036
		β_2	0.01424	0.00449	94.9	-0.00961	0.00006
		β_3	-0.01402	0.00664	97.0	-0.00621	0.00005
	$(2^{\frac{1}{3}}, 2^{\frac{2}{3}}, 2)$	β_0	-3.20980	0.14023	60.7	-0.084351	0.00077
		β_1	0.26342	0.06326	93.3	-0.09660	0.00108
		β_2	1.29946	0.07261	3.7	-0.20402	0.00037
		β_3	0.05218	0.00801	94.1	-0.21422	0.00061
	(2, 2, 2)	β_0	2.01236	0.07377	78.5	-0.25052	0.00420
		β_1	0.85701	0.08017	88.7	-0.29988	0.00976
		β_2	0.45614	0.01294	75.5	-0.28528	0.00081
		β_3	-0.37609	0.00846	93.5	-0.32296	0.00167
교환가능	(1, 1, 1)	β_0	-0.19560	0.02571	94.8	0.00144	0.00037
		β_1	0.07003	0.04425	94.4	0.00094	0.00031
		β_2	0.02136	0.00323	95.1	-0.00156	0.00005
		β_3	-0.01948	0.00525	94.1	0.00143	0.00004
	$(2^{\frac{1}{3}}, 2^{\frac{2}{3}}, 2)$	β_0	-3.12109	0.12637	55.4	-0.04867	0.00049
		β_1	0.19477	0.05057	94.8	-0.05925	0.00055
		β_2	1.28400	0.06946	0.5	-0.17207	0.00020
		β_3	0.10480	0.00578	94.0	-0.17530	0.00029
	(2, 2, 2)	β_0	1.88837	0.06695	76.5	-0,23423	0.00311
		β_1	0.89645	0.07910	86.4	-0.28666	0.00739
		β_2	0.47762	0.01276	65.3	-0.26379	0.00049
		β_3	-0.39626	0.00699	90.8	-0.30457	0.00102
비구조적	(1, 1, 1)	β_0	-0.15349	0.03450	95.4	-0.00184	0.00047
		β_1	0.02390	0.05896	95.3	0.00220	0.00042
		β_2	0.01292	0.00553	94.9	0.00394	0.00007
		β_3	0.00727	0.00943	95.7	0.00544	0.00006
	$(2^{\frac{1}{3}}, 2^{\frac{2}{3}}, 2)$	β_0	-3.13871	0.13992	67.8	-0.06927	0.00079
		β_1	0.20948	0.06994	94.1	-0.08324	0.00103
		β_2	1.28436	0.07228	7.9	-0.16920	0.00032
		β_3	0.10131	0.01009	94.8	-0.18226	0.00054
	(2, 2, 2)	β_0	2.03580	0.08033	79.6	-0.25265	0.00477
		β_1	0.82286	0.08623	89.7	-0.29970	0.01092
		β_2	0.45990	0.01445	76.7	-0.28122	0.00095
		β_3	-0.36359	0.01026	93.4	-0.31907	0.00197

최근 들어 GEE를 응용한 여러 방법들이 주로 이진형의 반응변수를 갖는 반복측정자료의 분석을 중심으로 많이 개발되었으나 대부분의 방법들이 Liang and Zeger의 방법과 마찬가지로 'ϕ가 모든 j에 대해 일정하다'는 강한 가정을 전제로 하고 있다(Prentice, 1988; Zhao and Prentice, 1990; Lipsitz, Laird, and Harrington, 1991). 본 논문에서 밝힌 바와 같이 이진형의 반복측정자료에 대해서도 산포모수가 모든 관측시간에 따라 변하게 될 때 같은 형태의 문제가 생길 것으로 짐작된다. 따라서 산포모수가 일정한지에 대한 사전 검토없이 무턱대고 GEE추정량을 구하고 이를 기초로 결론을 내리는 것을 심각한 오류를 범할 위험이 있다는 것을 명심해야할 것이다.

참고문헌

[1] Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969). Analysis of categorical data by linear models, *Biometrics*, Vol. 25, 489-504.

[2] Jennrich, R.I. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics*, Vol. 42, 805-820.

[3] Karim, M.R. (1989). *GEE1 PC SAS*. Technical Report #674. Department of Biostatistics, The Johns Hopkins University, Baltimore, MD.

[4] Koch, G.G., Imrey, P.B., and Reinfurt, R.G. (1972). Linear model analysis of categorical data with incomplete response vectors, *Biometrics*, Vol. 33, 133-158.

[5] Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data, *Biometrics*, Vol. 33, 133-158.

[6] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, Vol. 38, 963-974.

[7] Laird, N.M., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association*, Vol. 82, 97-105.

[8] Landis, J.R., Miller, M.E., Davis, C.S., and Koch, G.G. (1988). Some general methods for the analysis of categorical data in longitudinal studies, *Statistics in Medicine*, Vol. 7, 109-315.

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

[10] Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, Vol. 78, 153-160.

[11] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. 2nd Edition, Chapman and Hall, London and New York.

[12] Moulton, L.H. and Zeger, S. L. (1989). Analyzing repeated measures on generalized linear models via the bootstrap, *Biometrics*, Vol. 45, 381-394.

[13] Park, C.G, Park, T., and Shin, D.W. (1995). An approach for generating correlated binary variates having fixed marginal distributions. (under revision for American Statistician)

[14] Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine*, Vol. 12, 1723-1732.

[15] Park, T. and Davis, C.S. (1993). A test for the missing data mechanism for the incomplete repeated categorical data, *Biometrics*, Vol. 49, 631-638.

[16] Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika*, Vol. 51, 313-326.

[17] Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics*, Vol. 44, 1033-1048.

[18] Wei, L.J. and Stram, D. O.(1988). Analyzing repeated measurements with possibly missing observations by modelling marginal distributions, *Statistics in Medicine*, Vol. 7, 139-148.

[19] Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model, *Biometrika*, Vol. 77, 642-648.

[20] Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, Vol. 42, 121-130.