Confounding of Time Trend with Dropout Process in Longitudinal Data Analysis

Ji-Hyun Kim[1], Hye-Hyun Choi[2]

Abstract

In longitudinal studies, outcomes are repeatedly measured over time for each subject. It is common to have missing values or dropouts for longitudinal data. In this study time trend in longitudinal data with dropouts is of concern. The confounding of time trend with dropout process is investigated through simulation studies. Some simulation results are reported for binary responses as well as continuous responses with patterns of dropouts varying. It has been found that time trend is not confounded with random dropout process for binary responses when it is estimated using GEE.

Keywords: binary response, GEE, random dropouts, informative dropouts.

1. 서 론

여러 개의 관측 대상에 대해 관측을 하되 각 관측대상의 특성을 여러 시점에 걸쳐서 관측한 자료를 다시점 자료(longitudinal data)라고 하는데, 한 관측대상으로부터 반복측정된 자료들은 독립적이지 않다. 다시점 자료를 통한 연구(longitudinal study)에서는 횡단면 연구(cross-sectional study)에서와 달리 관측대상의 효과(cohort effect)와 시간에 따른 효과(age effect)를 분리해서 추정할 수 있다는 장점이 있다 (Diggle et al., 1994). 따라서, 시간에 따른 효과에 관심이 있거나 관측대상에 따라 반응변수의 값의 변동이 클 때 고려할 수 있는 연구방법이다. 다시점 자료에 대한 국내 사례연구로 박태성 등 (1998)이 있다.

다시점 자료를 연속형(continuous longitudinal data)과 이산형(discrete longitudinal data)으로 구분해 볼 수 있다. 연속형 다시점 자료에는 흔히 다변량정규분포를 이용한 일반선형모형(genearal linear model)을 적용하는데, Laird & Ware (1982) 이후 많은 연구가 있었고, 최근 Verbeke & Molenberghs (2000)에 그 응용사례와 함께 분석 방법이 잘 소개되어 있다.

한편, 이산형 다시점 자료의 분석에는 확장된 일반화 선형모형(generalized linear model)이 적용되는데 (Diggle et al. (1994), Fitzmaurice et al. (1993)), 크게 가능도(likelihood)에 기초한 방법과 그렇지 않은 방법으로 구분할 수 있다. 가능도에 기초한 방법은 다루기가 힘들고, 많은 장애모수(nuisance parameter)의 추정이 필요하다. 그래서, 정확한 가능도를 명시할 필요가 없는

---

1) Professor, Department of Statistics, Soongsil University, Dongjak-Ku Sangdo-Dong, Seoul 156-743, KOREA.
E-mail: jhkim@stat.soongsil.ac.kr

2) Consumer Credit Risk Management Team, Koram Bank HQ, #39 Da-dong Jung-ku, Seoul, KOREA 100-180.

GEE(Generalized Estimating Equations) 접근방법을 흔히 사용한다 (Liang & Zeger, 1986). GEE 방법은 일반화선형모형(generalized linear models)에서의 준가능도방법(quasi-likelihood approach)을 확장해 다시점 자료에 적용시킨 방법으로서, 하나의 관측대상에서 반복 관측된 값으로 이루어진 확률벡터에 대해 분포에 대한 가정 없이 평균과 분산-공분산 구조에 대한 가정만으로, 평균에 대한 회귀모형의 회귀계수를 추정할 수 있다. 이 때 분산-공분산 구조를 잘못 가정하더라도 회귀계수와 그 추정량의 분산에 대한 일치추정량을 얻을 수 있다는 특성이 있다.

본 연구에서는 연속형과 이산형 자료를 모두 고려하되, 이산형 다시점 자료의 경우 이항형(binary) 자료만 고려하며 GEE 방법에 대해서만 다루기로 한다.

이러한 다시점 자료는 자료의 특성상 불균형(unbalanced), 즉, 관측대상에 따라 관측값의 수가 일정하지 않은 경우가 많으며 결측값(missing values)도 흔히 발생한다. 다시점 자료에서 결측값은, 간헐적으로 발생하는(intermittent) 경우와 특정 시점 이후의 값이 모두 관측되지 않는 중도탈락(dropout)의 경우로 나눌 수 있는데, 본 연구에서는 흔히 발생하는 중도탈락의 경우만 고려하였다. 중도탈락(또는 결측)의 발생 여부가 관측값의 크기에 의존하는가, 의존한다면 어느 시점의 관측값에 의존하는가 등에 따라 그 유형을 완전랜덤(completely random), 랜덤(random), 정보적(informative) 등 세 가지로 흔히 구분한다. (Rubin (1976), Diggle & Kenward (1994)). 일반적으로 중도탈락의 유형이 완전랜덤일 때는 추정에 편향이 생기지 않는다. 결측값이 있는 연속형 다시점 자료에 대한 연구로서 Diggle & Kenward (1994)를 들 수 있으며, 결측값이 있는 이산형 자료에 대한 연구로 Robins et al. (1995)과 Paik (1997) 등을 들 수 있는데, 이들 연구는 결측 또는 중도탈락의 형태가 완전랜덤이 아닐 때 대처하는 방법에 대해 논의하였다.

본 연구의 목적은, 연속형과 이항형 다시점 자료에서 시간에 따른 효과가 주요 관심사일 때, 이러한 시간에 따른 효과가 중도탈락의 효과와 중첩되는(confounded) 현상에 대해 알아보는 것이다. 중도탈락이 발생할 때, 시간에 따른 효과가 없음에도 불구하고 중도탈락의 유형에 따라 시간에 따른 효과가 유의하게 나타나는 현상을 모의실험을 통해서 확인해 보았다. 그 결과로, 기존의 연구결과에 따른 예상과 달리, 이항형 다시점 자료에서 랜덤 중도탈락이 일어날 때, GEE 방법으로 추정하면 시간 효과가 중도탈락 효과와 중첩되지 않는다는 사실을 발견할 수 있었다.

먼저, 제2절에서는 다시점 자료의 형태가 연속형일 때와 이항형(binary)일 때의 두 가지 경우로 나누어서 모의실험을 위한 자료생성 모형을 소개하였다. 제3절에서는 모의실험 결과를 보고하고 중도탈락의 영향에 대해 이미 알려진 사실과 비교하여 논의하였다.

2. 모의실험을 위한 자료생성

연속형 다시점 자료의 경우, 모수 추정방법이 가능도에 기초한 방법이므로, 중도탈락 유형이 랜덤이면 문제점이 발생하지 않는다 (Diggle & Kenward, 1994). 그러나, 중도탈락 유형이 정보적이면 그렇지 못하다. 한편, 이항형 다시점 자료의 경우 GEE 방법을 적용하면, 가능도에 기초한 방법이 아니므로 정보적 중도탈락, 랜덤 중도탈락 두 가지 형태 모두에서 문제점이 나타난다고 알려져 있다 (Fitzmaurice et al., 1995). 본 연구에서는 이를 모의실험을 통해 실증적으로 확인해 보았는데, 이항형 다시점 자료의 경우 새로운 사실을 발견할 수 있었다. 먼저 모의실험을 위한 자료생성 방법을 연속형과 이항형으로 나누어 각각 설명한다.

2.1 연속형 다시점 자료 생성 방법

연속형 반응변수 Y에 대해 식 (2.1)과 같은 모형을 가정한다. 여기서, m은 개체의 수, T는 시점의 수이다. 설명변수 x는 그룹 변수인데, 개체가 처리 그룹(treatment group)에 속하면 1을, 그렇지 않으면 0을 갖는다.

$$y_{it} = \beta_0 + \beta_1 x_i + \varepsilon_{it}, \ i = 1, \ldots, m, \ t = 1, \ldots, T \tag{2.1}$$

$$x_i = \begin{cases} 1, & i \text{ 번째 개체가 처리그룹에 속할 때} \\ 0, & i \text{ 번째 개체가 제어그룹에 속할 때} \end{cases}$$

$$\varepsilon_{it} = \rho \varepsilon_{i,t-1} + \delta_{it}, \ \ \varepsilon_{i1} \sim N(0, \sigma^2)$$

$$\delta_{it} \sim N(0, \sigma^2 (1 - \rho^2))$$

시간에 따른 효과가 없는 위 모형에서 계수를 $\beta_0 = 1$, $\beta_1 = 1$ 로 두었으며, $\sigma = 0.5$, $\sigma = 2.0$인 경우에 대해서 각각 살펴보았다. 독립적인 자료는 $\rho = 0$ 이며, 의존적인 자료는 $\rho = 0.5$, $\rho = 0.9$ 두 가지 경우에 대해 그 결과를 살펴보았다. 생성된 자료의 개체 수와 시점의 수는 $m = 100$ (각 그룹에 50개씩), $T = 5$ 으로 설정하였고, 따라서 총 관측값의 수 $N = 500$ 이다.

생성된 자료에 대해 다음과 같은 방법으로 랜덤 중도탈락과 정보적 중도탈락을 적용시켰다. 먼저, 중도탈락시킬 반응변수의 기준값 c를 설정한다. 그룹별 중도탈락 비율의 균형을 맞추기 위해 그룹별로 c를 다르게 주었다. 그룹이 1일 경우에는 $2 + \sigma$ 로 주었고, 그룹이 0인 경우에는 $1 + \sigma$로 설정하였다. 반응변수 값이 설정된 기준값 이하가 되면 중도탈락확률 p에 의해 탈락여부가 결정되도록 하였는데, p는 0.3, 0.5, 0.7로 변화시켜가며 그 결과를 살펴보았다. 예를 들어 $\sigma = 0.5$, $p = 0.3$ 으로 정하였을 때, 처리 그룹에서 생성된 반응변수의 값이 2.5보다 작으면 중도탈락될 확률이 0.3이 된다. 랜덤 중도탈락인 경우는 어떤 시점에서의 중도탈락이 바로 전 시점의 반응변수 값에 의해 결정되게 하였고, 정보적 중도탈락인 경우는 현재 시점의 반응변수 값에 의해 결정되게 하였다. 예를 들면, 한 개체가 있다고 할 때, 4번째 시점에서 반응변수 값이 c보다 작고 중도탈락으로 결정되었으면, 랜덤 중도탈락인 경우에는 5번째 시점부터의 관측값들이 중도탈락되고, 정보적 중도탈락인 경우에는 4번째 시점부터의 관측값들이 모두 중도탈락된다. 랜덤 중도탈락되는 비율과 정보적 중도탈락되는 비율의 균형을 맞추기 위해 랜덤 중도탈락은 첫 번째 시점부터 ($t = 1, \cdots, T-1$) 중도탈락 여부를 결정하게 하고 정보적 중도탈락은 두 번째 시점부터 ($t = 2, \cdots, T$) 적용하였다.

- 랜덤 중도탈락에서의 중도탈락 방법

  만약 $y_{it} < c$ 이면, 중도탈락확률 p에 의해 y_{ik}를 중도탈락시킴, $k = t+1, \ldots, T$.
- 정보적 중도탈락에서의 중도탈락 방법

  만약 $y_{it} < c$ 이면, 중도탈락확률 p에 의해 y_{ik}를 중도탈락시킴, $k = t, \ldots, T$.

2.2 이항형 다시점 자료 생성 방법

이항형 반응변수 Y에 대해 다음 식 (2.2)와 같은 모형을 가정한다.

$$\log[P(Y_{it} = 1)/P(Y_{it} = 0)] = \beta_0 + \beta_1 x_i, \ i = 1, \ldots, m, \ t = 1, \ldots, T \tag{2.2}$$

$$x_i = \begin{cases} 1, & i \text{ 번째 개체가 처리그룹에 속할 때} \\ 0, & i \text{ 번째 개체가 제어그룹에 속할 때} \end{cases}$$

모형에서의 계수를 $\beta_0 = -1$, $\beta_1 = 2$ 으로 설정하였다. 개체 수 $m = 100$, 시점의 수 $T = 5$로 설정하여, 생성된 관측값의 총수 N은 500이다. 한 개체에서 측정된 값들, $Y_{it}, t = 1, \ldots, T$ 가 독립적인 경우와 종속적인 경우에 대해 각각 모의실험을 실시하였다. 먼저 독립적인 자료는, $P(Y_{it} = 1) = 1/\left(1 + e^{-(\beta_0 + \beta_1 x_i)}\right)$ 의 관계식을 이용하여 $Y_{it}, t = 1, \ldots, T$ 의 값을 독립적으로 결정하였다. 의존적인 자료는 $P(Y_{i1} = 1) = 1/\left(1 + e^{-(\beta_0 + \beta_1 x_i)}\right)$ 에 의해 각 개체의 첫 관측값을 생성한 다음, 다음 시점의 관측값 $Y_{it}, j = 2, \ldots, T$ 는 확률 0.5로 직전의 값 $Y_{i,t-1}$과 같아지든지, $P(Y_{it} = 1) = 1/\left(1 + e^{-(\beta_0 + \beta_1 x_i)}\right)$ 에 의해 독립적으로 정해지도록 하였다.

그리고, 생성된 자료에 대해 다음과 같은 방법으로 랜덤 중도탈락과 정보적 중도탈락을 적용시킨다. 먼저, 반응변수의 값이 0이면 중도탈락확률 p에 의해 중도탈락 여부를 결정하였으며, 각 시점에서 중도탈락이 일어날 확률 p는 0.3, 0.5, 0.7 로 각각 변화시켜가며 그 결과를 살펴보았다.

· 랜덤 중도탈락에서의 중도탈락 방법

만약 $y_{it} = 0$ 이면, 중도탈락확률 p에 의해 y_{ik}를 중도탈락시킴, $k = t + 1, \ldots, T$.

· 정보적 중도탈락에서의 중도탈락 방법

만약 $y_{it} = 0$ 이면, 중도탈락확률 p에 의해 y_{ik}를 중도탈락시킴, $k = t, \ldots, T$.

3. 모의실험 결과

앞 절에서 설명한 대로 생성한 연속형 자료와 이항형 자료를, 랜덤 중도탈락인 경우와 정보적 중도탈락인 경우로 나누어 중도탈락 효과가 시간 효과와 중첩이 되는지 알아보았다. 즉, 존재하지 않는 시간 효과가 중도탈락에 의해 유의하게 나타나게 되는 현상을 모의실험을 통해 확인해 보고자 한다. 연속형 다시점 자료의 분석을 위해 SAS 8.01 (SAS Institute Inc., Cary, NC) PROC MIXED를 이용하였고, 이항형 다시점 자료의 분석을 위해 GEE방법을 지원하는 PROC GENMOD를 이용하였다.

3.1 연속형 다시점 자료에 대한 모의실험 결과

설정된 모형 (2.1)에 의해 생성된 자료에는 시간에 따른 효과가 없었는데, 중도탈락의 영향으로 인해 시간 효과가 유의하게 나올 수 있다. SAS PROC MIXED를 이용하여, 모형 (3.1)을 적용시켜 보았다.

$$y_{it} = \beta_0 + \beta_1 x_i + \beta_2 t, \; i = 1, \ldots, m, \; t = 1, \ldots, T \tag{3.1}$$

즉, 시간에 따른 효과가 있을 수 있다고 가정하고, 모형에 시간 효과를 추가하였다. 시간 효과를 나타내는 모수 β_2가 0과 같다는 귀무가설을 검정하였는데, 총 100회의 모의실험 중에서 유의하게 0과 다르게 나오는 회수를 표로 정리해 보았다.

$\sigma = 0.5$이고 독립적인 자료에 대해 랜덤 중도탈락인 경우와 정보적 중도탈락인 경우의 결과가 [표 3-1]이다. 각 개체 내의 공분산 구조를 독립일(Independent) 때와 AR(1)으로 주었을 때의 결

과인데, 랜덤 중도탈락인 경우에는 각 개체 내에서의 공분산 구조를 어떻게 주는가에 상관없이 검정 결과에 문제가 없다 (100회의 모의실험 중에서 검정 결과가 유의하게 나오는 회수가 $np \pm 2\sqrt{npq} = 5 \pm 4.3$ 사이에 속하면, 즉 1에서 9 사이이면 별 문제가 없는 것으로 볼 수 있다). 그러나, 정보적 중도탈락인 경우에는 중도탈락되는 관측값이 많아질수록 시간 효과가 유의하다고 결론을 잘못 내리게 되는 회수가 많아짐을 알 수 있다. $\sigma = 2.0$ 일 때에도 비슷한 결과를 얻었다 ([표 3-2]).

[표 3-3]과 [표 3-4]는 $\sigma = 2.0$ 이고 각각 $\rho = 0.5, 0.9$ 인 의존적인 자료에 대한 결과이다. 두 경우 모두 각 개체 내에서의 공분산 구조를 독립으로 잘못 주면 랜덤 중도탈락과 정보적 중도탈락 두 가지 모두에서 문제점이 발생하지만, AR(1)으로 제대로 줄 경우에는 랜덤 중도탈락에서는 문제가 없다. 그러나, 정보적 중도탈락인 경우 지정하는 공분산 구조에 따라 차이가 있긴 하지만, 시간 효과가 유의하게 존재한다고 결론을 잘못 내리게 됨을 알 수 있다. $\sigma = 0.5$ 일 때에도 비슷한 결과를 얻었으나 표를 생략하였다.

[표 3-1] 연속형 독립적 자료 ($\sigma = 0.5$)

중도탈락 유형	중도탈락 확률	총 100회의 실험 중 귀무가설이 기각되는 회수 ($H_0 : \beta_2 = 0$, 유의수준 : 0.05)	
		Independent	AR(1)
랜덤 중도탈락	0.3	2	2
	0.5	4	6
	0.7	6	2
정보적 중도탈락	0.3	9	10
	0.5	29	22
	0.7	53	45

[표 3-2] 연속형 독립적 자료 ($\sigma = 2.0$)

중도탈락 유형	중도탈락 확률	총 100회의 실험 중 귀무가설이 기각되는 회수 ($H_0 : \beta_2 = 0$, 유의수준 : 0.05)	
		Independent	AR(1)
랜덤 중도탈락	0.3	5	6
	0.5	4	3
	0.7	7	3
정보적 중도탈락	0.3	4	12
	0.5	21	30
	0.7	46	58

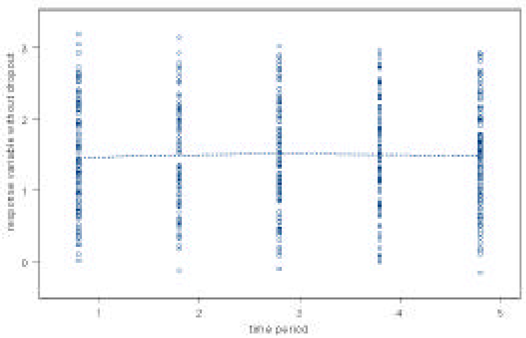
[표 3-3] 연속형 의존적 자료 ($\rho = 0.5$, $\sigma = 2.0$)

중도탈락 유형	중도탈락 확률	총 100회의 실험 중 귀무가설이 기각되는 회수 ($H_0 : \beta_2 = 0$, 유의수준 : 0.05)	
		Independent	AR(1)
랜덤 중도탈락	0.3	13	3
	0.5	25	8
	0.7	32	5
정보적 중도탈락	0.3	19	6
	0.5	67	22
	0.7	86	42

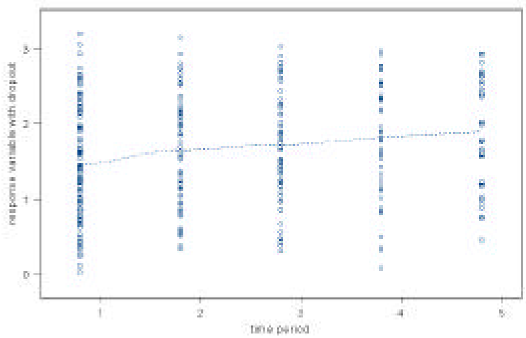
[표 3-4] 연속형 의존적 자료 ($\rho = 0.9$, $\sigma = 2.0$)

중도탈락 유형	중도탈락 확률	총 100회의 실험 중 귀무가설이 기각되는 회수 ($H_0 : \beta_2 = 0$, 유의수준 : 0.05)	
		Independent	AR(1)
랜덤 중도탈락	0.3	51	3
	0.5	97	5
	0.7	100	3
정보적 중도탈락	0.3	70	9
	0.5	95	15
	0.7	100	48

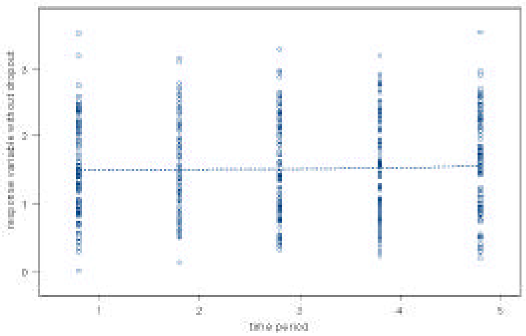
랜덤 중도탈락과 정보적 중도탈락일 때 검정결과가 서로 다른 사실을 탐색적 자료분석만으로는 알아내기 어렵다. 랜덤 중도탈락과 정보적 중도탈락 각 경우에 대해서 중도탈락이 일어나기 전과 후를 그림으로 비교해 보았다. 중도탈락 확률이 0.3이고, $\sigma = 0.5$, $\rho = 0.9$인 의존적인 자료의 경우, 랜덤 중도탈락 전후의 결과가 [그림 3-1]과 [그림 3-2]이고, 정보적 중도탈락 전후의 결과가 [그림 3-3]과 [그림 3-4]이다. 각 그림에서 점선으로 표시된 곡선은 LOESS(locally weighted regression scatter-plot smoothing) 방법으로 평활시킨(smoothing) 것이다. [그림 3-1]과 [그림 3-2]의 두 그림을 비교해보면, 랜덤 중도탈락이 일어나면서 시간 효과가 마치 있는 것처럼 보이는 것을 알 수 있다. 이런 현상은 정보적 중도탈락의 경우인 [그림 3-3]과 [그림 3-4]에서도 마찬가지이다. 그러나, 모수 추정에 있어서 랜덤 중도탈락의 경우에는 문제가 없지만, 정보적 중도탈락인 경우에는 그렇지 않다. [그림 3-2]와 [그림 3-4]의 각 자료에 대해 β_2의 신뢰구간이나 검정 결과에서 확인할 수 있듯이, 외견상으로는 보이지만 실제로는 존재하지 않는 시간 효과가 랜덤 중도탈락인 경우에는 추론 과정에서 구별되지만, 정보적 중도탈락인 경우는 구별되지 못한다 (Diggle et al. (1994) 11.3절 참조).



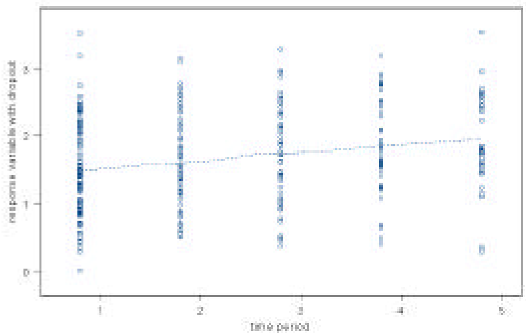
[그림 3-1] 랜덤 중도탈락 전



[그림 3-2] 랜덤 중도탈락 후



[그림 3-3] 정보적 중도탈락 전



[그림 3-4] 정보적 중도탈락 후

그룹효과에 대한 편향(bias)에 대해서도 살펴보았다. 자료를 생성할 때 설정된 그룹효과를 나타내는 모수 β_1의 값은 1이었다. 독립적인 자료와 의존적인 자료에 대해 중도탈락 유형별로 β_1의 신뢰구간을 구해 보았는데 모든 모의실험에서 신뢰구간이 참값 1을 포함하였다. 정보적 중도탈락

일 때, 시간 효과가 잘못 탐지되었듯이 그룹효과에 대해서도 잘못된 추론이 일어나리라고 예상했었으나 중도탈락 유형에 따른 차이는 발견되지 않았다.

마지막으로 시간 효과에 대한 편향(bias)을 총 100회의 실험에서 얻어진 β_2의 추정값들로 신뢰구간을 구하여 살펴보았다. 자료를 생성할 때 설정된 모수 β_2의 값은 0이었다. [표 3-5]와 [표 3-6]은 독립적인 자료에서 $\sigma = 0.5$일 때와 $\sigma = 2.0$일 때 시간 효과를 나타내는 모수 β_2의 95% 신뢰구간을 구한 결과이다. 검정 결과와 마찬가지로, 추정 과정에서 지정하는 공분산 구조에 상관없이, 랜덤 중도탈락에서는 편향이 발생하지 않지만, 정보적 중도탈락에서는 편향이 생기는 것을 알 수 있다. [표 3-7]은 $\sigma = 2.0$ 이고 $\rho = 0.5$인 의존적인 자료에서의 결과이다. 공분산 구조를 AR(1)으로 제대로 주면 랜덤 중도탈락에서는 편향이 생기지 않지만, 정보적 중도탈락에서는 그렇지 않다는 것을 알 수 있다. ρ가 0.9인 [표 3-8]에서도 마찬가지 결과를 보여준다. σ가 0.5일 때에도 편향의 절대적 크기가 작을 뿐 마찬가지 결과를 얻었는데, 지면관계상 보고는 생략하였다.

[표 3-5] 연속형 독립적 자료($\sigma = 0.5$): 시간 효과(β_2)에 대한 신뢰구간

중도탈락 유형	중도탈락 확률	공분산구조					
		Independent			AR(1)		
		평균	표준편차	95% 신뢰구간	평균	표준편차	95% 신뢰구간
랜덤 중도탈락	0.3	-0.003	0.019	[-0.007, 0.001]	-0.002	0.021	[-0.006, 0.002]
	0.5	0.001	0.028	[-0.004, 0.007]	0.004	0.029	[-0.002, 0.010]
	0.7	-0.006	0.045	[-0.015, 0.003]	-0.004	0.038	[-0.011, 0.003]
정보적 중도탈락	0.3	0.016	0.023	[0.011, 0.020]	0.013	0.022	[0.009, 0.018]
	0.5	0.038	0.038	[0.030, 0.045]	0.036	0.033	[0.029, 0.042]
	0.7	0.089	0.047	[0.079, 0.098]	0.086	0.045	[0.077, 0.095]

[표 3-6] 연속형 독립적 자료($\sigma = 2.0$): β_2에 대한 신뢰구간

중도탈락 유형	중도탈락 확률	공분산구조					
		Independent			AR(1)		
		평균	표준편차	95% 신뢰구간	평균	표준편차	95% 신뢰구간
랜덤 중도탈락	0.3	-0.002	0.087	[-0.020, 0.015]	-0.010	0.091	[-0.028, 0.007]
	0.5	0.004	0.120	[-0.020, 0.027]	0.006	0.106	[-0.015, 0.026]
	0.7	-0.020	0.200	[-0.059, 0.020]	-0.019	0.177	[-0.053, 0.016]
정보적 중도탈락	0.3	0.040	0.083	[0.023, 0.056]	0.072	0.085	[0.056, 0.089]
	0.5	0.141	0.121	[0.117, 0.165]	0.158	0.121	[0.135, 0.182]
	0.7	0.351	0.198	[0.312, 0.390]	0.379	0.207	[0.338, 0.419]

[표 3-7] 연속형 의존적 자료(ρ = 0.5, σ = 2.0): β_2에 대한 신뢰구간

중도탈락 유형	중도탈락 확률	공분산 구조					
		Independent			AR(1)		
		평균	표준편차	95% 신뢰구간	평균	표준편차	95% 신뢰구간
랜덤 중도탈락	0.3	0.043	0.095	[0.025, 0.062]	-0.003	0.100	[-0.023, 0.016]
	0.5	0.129	0.133	[0.103, 0.155]	0.017	0.126	[-0.008, 0.042]
	0.7	0.240	0.167	[0.207, 0.273]	-0.006	0.174	[-0.040, 0.028]
정보적 중도탈락	0.3	0.094	0.090	[0.076, 0.111]	0.034	0.060	[0.022, 0.046]
	0.5	0.269	0.132	[0.243, 0.295]	0.079	0.069	[0.065, 0.092]
	0.7	0.546	0.195	[0.508, 0.584]	0.157	0.095	[0.139, 0.176]

[표 3-8] 연속형 의존적 자료(ρ = 0.9, σ = 2.0): β_2에 대한 신뢰구간

중도탈락 유형	중도탈락 확률	공분산 구조					
		Independent			AR(1)		
		평균	표준편차	95% 신뢰구간	평균	표준편차	95% 신뢰구간
랜덤 중도탈락	0.3	0.177	0.080	[0.161, 0.193]	-0.004	0.054	[-0.015, 0.006]
	0.5	0.413	0.105	[0.393, 0.434]	-0.010	0.071	[-0.024, 0.004]
	0.7	0.699	0.137	[0.672, 0.726]	-0.011	0.088	[-0.028, 0.006]
정보적 중도탈락	0.3	0.207	0.081	[0.191, 0.223]	0.035	0.063	[0.023, 0.047]
	0.5	0.440	0.121	[0.416, 0.464]	0.077	0.066	[0.064, 0.090]
	0.7	0.797	0.130	[0.772, 0.823]	0.161	0.096	[0.143, 0.180]

3.2 이항형 다시점 자료에 대한 모의실험 결과

이항형 자료에 대해서도 마찬가지로, 시간에 따른 효과가 없음에도 불구하고 중도탈락의 영향으로 인해 유의하게 나오는 현상을 총 100회의 모의실험을 통해 살펴보았다. 모형 (2.2)에 의해 생성된 자료에 대해 SAS PROC GENMOD를 이용하여, 다음 모형을 적합해 보았다.

$$\log \frac{P(Y_{it}=1)}{P(Y_{it}=0)} = \beta_0 + \beta_1 x_i + \beta_2 t, \quad i = 1, \ldots, m, \ t = 1, \ldots, T$$

즉, 시간에 따른 효과가 있을 것이라고 추측하고, 모형에 시간에 따른 효과를 포함시켜 적합해 보았다.

독립적인 자료와 의존적인 자료에 대하여 각각 랜덤 중도탈락과 정보적 중도탈락이 발생했을 때의 결과가 [표 3-9], [표 3-10]과 같다. [표 3-9]의 결과를 보면, 독립적인 자료의 경우 공분산 구조를 어떻게 지정해주는가에 상관없이 랜덤 중도탈락인 경우에는 문제가 없어 보이지만 정보적 중도탈락인 경우에는 문제가 있다. 의존적인 자료의 경우 [표 3-10]의 결과를 보면, 우선 정보적

중도탈락의 경우에는 지정하는 공분산 구조의 종류에 상관없이 시간 효과에 대한 결론을 잘못 내리게 됨을 알 수 있다. 반면에 랜덤 중도탈락의 경우, 공분산 구조를 독립으로 잘못 지정해주면 문제가 나타나지만, 의존적인 자료임에도 불구하고 AR(1)으로 '제대로' 지정해주면 크게 문제가 없다. (2.2절에서 생성된 자료의 공분산 구조가 정확히 AR(1)은 아니다. PROC GENMOD에서 정형화된 공분산 구조(또는 가상관 구조, working correlation structure)를 지정해 주어야 하는데, 2.2절에서 생성된 자료에 대해 Corr(Y_{it}, $Y_{i,t-k}$)가 k가 커짐에 따라 감소하고, AR(1)일 때 Corr(Y_{it}, $Y_{i,t-k}$)= α^k 의 형태로서 역시 k가 커짐에 따라 감소한다. 그러므로, PROC GENMOD에서 지정할 수 있는 정형화된 공분산 구조 중에서는 AR(1)이 가장 '가까운' 구조이다. 공분산 구조에 아무런 제약이 없는 UNSTRUCTURED 구조는 추정해야 할 모수가 너무 많아 수렴성에 문제가 있어 적용하지 않았다.)

Fitzmaurice et al.(1995)은 이항형 자료의 경우 정보적 중도탈락일 때는 물론 랜덤 중도탈락일 때도 시간 효과에 편향이 존재한다고 하였으나, 적어도 본 연구의 제한된 모형과 모의실험 조건에서는, 랜덤 중도탈락일 때 공분산 구조만 제대로 지정해주면 편향은 무시할 정도임을 발견할 수 있었다.

[표 3-9] 이항형 독립적 자료

중도탈락 유형	중도탈락 확률	총 100회의 실험 중 귀무가설이 기각되는 회수 ($H_0 : \beta_2 = 0$, 유의수준 : 0.05)	
		Independent	AR(1)
랜덤 중도탈락	0.3	4	7
	0.5	4	5
	0.7	7	4
정보적 중도탈락	0.3	15	20
	0.5	50	50
	0.7	86	84

[표 3-10] 이항형 의존적 자료

중도탈락 유형	중도탈락 확률	총 100회의 실험 중 귀무가설이 기각되는 회수 ($H_0 : \beta_2 = 0$, 유의수준 : 0.05)	
		Independent	AR(1)
랜덤 중도탈락	0.3	14	9
	0.5	36	4
	0.7	46	12
정보적 중도탈락	0.3	44	36
	0.5	89	72
	0.7	100	98

시간 효과의 편향을 β_2에 대한 신뢰구간을 통해 살펴보았다. 각 구간이 처음 설정된 값 0을 포함하고 있는가를 살펴보도록 하자. 총 100회의 모의실험 결과에서 나온 β_2의 추정값들로부터 신뢰구간을 구하였다. 독립적인 자료의 경우 [표 3-11]의 결과를 보면, 랜덤 중도탈락이 일어날 때 편향이 생기지 않는다 (단, 중도탈락 확률이 0.5이고 독립으로 공분산 구조를 주었을 때 신뢰구간이 참값 0을 포함하지 않지만 그 편향의 크기는 미미하다). 그러나, 정보적 중도탈락일 경우에는 지정하는 공분산 구조에 상관없이 전반적으로 편향이 생기는 것을 볼 수 있다. 의존적 자료의 경우 [표 3-12]의 결과를 보면, 공분산 구조를 AR(1)으로 지정해 주었을 때 랜덤 중도탈락인 경우에는 편향이 발생하지 않는 반면에 공분산 구조를 독립으로 잘못 지정해 주면 랜덤 중도탈락인 경우에도 편향이 발생한다. 그러나, 정보적 중도탈락인 경우에는 공분산 구조를 어떻게 지정해 주는

가에 상관없이 전체적으로 편향이 발생하며 중도탈락 확률이 커질수록 편향이 커짐을 알 수 있다.

[표 3-11] 이항형 독립적 자료: 시간 효과(β_2)에 대한 신뢰구간

중도탈락 유형	중도탈락 확률	공분산 구조					
		Independent			AR(1)		
		평균	표준편차	95% 신뢰구간	평균	표준편차	95% 신뢰구간
랜덤 중도탈락	0.3	-0.004	0.084	[-0.020, 0.012]	-0.016	0.089	[-0.034, 0.001]
	0.5	0.019	0.084	[0.002, 0.035]	-0.011	0.098	[-0.030, 0.008]
	0.7	-0.015	0.115	[-0.037, 0.008]	-0.015	0.100	[-0.035, 0.005]
정보적 중도탈락	0.3	0.077	0.088	[0.060, 0.094]	0.089	0.095	[0.071, 0.108]
	0.5	0.192	0.120	[0.169, 0.216]	0.204	0.090	[0.187, 0.222]
	0.7	0.421	0.145	[0.392, 0.449]	0.412	0.159	[0.381, 0.443]

[표 3-12] 이항형 의존적 자료: 시간 효과(β_2)에 대한 신뢰구간

중도탈락 유형	중도탈락 확률	공분산 구조					
		Independent			AR(1)		
		평균	표준편차	95% 신뢰구간	평균	표준편차	95% 신뢰구간
랜덤 중도탈락	0.3	0.076	0.010	[0.056, 0.095]	-0.003	0.102	[-0.023, 0.016]
	0.5	0.152	0.114	[0.130, 0.174]	0.010	0.102	[-0.010, 0.030]
	0.7	0.220	0.123	[0.196, 0.244]	0.018	0.141	[-0.010, 0.046]
정보적 중도탈락	0.3	0.173	0.091	[0.156, 0.191]	0.132	0.101	[0.113, 0.152]
	0.5	0.363	0.129	[0.338, 0.388]	0.286	0.120	[0.262, 0.309]
	0.7	0.704	0.208	[0.663, 0.744]	0.615	0.161	[0.584, 0.647]

4. 결 론

지금까지 다시점 자료에서 시간에 따른 효과가 주관심사일 때, 중도탈락 유형별로 그 추정에 있어 나타나는 문제점에 대해서 알아보았다. 실제로 시간 효과가 없을 때 중도탈락의 효과에 의해 시간 효과가 마치 있는 것처럼 나타나는 현상을 중도탈락 유형별로 살펴보았다.

연속형 다시점 자료에서는 기존의 연구 결과와 마찬가지로 랜덤 중도탈락이 발생할 때 별 문제점이 없고, 정보적 중도탈락인 경우에는 그렇지 않다는 것을 실증적으로 확인하게 되었다. 특히, 의존적인 자료의 경우, 지정해 주어야 하는 공분산 구조가 시간에 따른 효과의 추정에 있어서 중요한 영향을 미친다는 사실을 알 수 있었다.

랜덤 중도탈락이나 정보적 중도탈락이 발생한 이항형 다시점 자료에서, 가능도에 기초하지 않은 GEE 방법이 편향된 추정량을 초래한다는 기존의 연구가 있었다 (Fitzmaurice et al. 1995). 본 연

구에서는 이러한 문제점을 모의실험을 통해 실증적으로 확인하고자 하였다. 이항형 다시점 자료에서 랜덤 중도탈락과 정보적 중도탈락 모두에서 문제점이 발생할 것으로 예상하였으나, 공분산 구조만 제대로 지정하면 연속형 다시점 자료에서와 마찬가지로 랜덤 중도탈락일 때는 문제점이 발생하지 않았고, 정보적 중도탈락일 경우에는 문제가 발생하는 것을 알 수 있었다. 물론 선형 시간효과 β_2가 0인 경우만 고려하였고, 특수한 랜덤 중도탈락과 정보적 중도탈락의 경우에 대해서만 모의실험을 행하였으므로 일반화하는 데에 어려움이 있으나, GEE 방법이 완전랜덤 중도탈락은 물론 랜덤 중도탈락이 발생할 때에도 편향되지 않은 추정을 한다는 새로운 사실을 확인할 수 있었다. 아울러, 연속형 자료에서와 마찬가지로 공분산 구조가 매우 중요한 영향을 미치는 것을 알 수 있었다.

시간에 따른 효과가 없는 자료임에도 불구하고 중도탈락으로 인해 시간 효과가 나타날 수 있듯이, 시간 효과가 있는 자료임에도 불구하고 중도탈락으로 인해 그 효과가 희석되는 경우도 있을 수 있으나, 그러한 경우를 찾는 것이 너무 작위적이라 연구에 포함시키지 않았다.

참 고 문 헌

[1] 박태성, 이승연, 성건형, 강종명, 강경원 (1998). 반복측정자료 분석에 대한 고찰: 신장이식 환자의 신기능 부전 연구를 중심으로. <응용통계연구>, 제11권 제2호, 205-219.

[2] Diggle, P.J., Liang, K., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford Science Publications.

[3] Diggle, P.J. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with Discussion). *Applied Statistics*, 43, 49-93.

[4] Fitzmaurice, G.M., Laird, N.M., and Rotnitsky, A.G. (1993). Regression Models for Discrete Longitudinal Responses (with Discussion). *Statistical Science* 8, 284-309.

[5] 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, Series B*, 57, 691-704.

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

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

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

[9] 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.

[10] Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

[11] Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer.