The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.4.605
25(4), 605–612

A Comparison Study of Multivariate Binary and Continuous Outcomes

Daewoo Pak[1] · HyungJun Cho[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University

(Received February 10, 2012; Revised March 19, 2012; Accepted April 30, 2012)

Abstract

Multivariate data are often generated with multiple outcomes in various fields. Multiple outcomes could be mixed as continuous and discrete. Because of their complexity, the data are often dealt with by separately applying regression analysis to each outcome even though they are associated the each other. This univariate approach results in the low efficiency of estimates for parameters. We study the efficiency gains of the multivariate approaches relative to the univariate approach with the mixed data that include continuous and binary outcomes. All approaches yield consistent estimates for parameters with complete data. By jointly estimating parameters using multivariate methods, it is generally possible to obtain more accurate estimates for parameters than by a univariate approach. The association between continuous and binary outcomes creates a gap in efficiency between multivariate and univariate approaches. We provide a guidance to analyze the mixed data.

Keywords: Multivariate methods, mixed outcomes, factorization estimation, GEE.

1. 서론

의학, 생물학, 사회학을 아우르는 다양한 분야에서 수집되는 다변량 자료는 그 형태가 매우 다양하다. 개체의 정보를 여러 변수로 하여 취합한 자료가 있는 반면, 여러 반응변수들과 설명변수들 간의 인과관계를 적절히 파악하는 데 주안을 두는 자료도 있다. 본 논문에서는 후자와 같이 여러 개의 반응변수들과 그에 따른 설명변수를 가지는 자료를 다루되 반응변수가 여러 가지 분포로 혼합되어 있을 경우의 분석 방법에 대해 논의하고자 한다. 반응변수를 가지는 다변량 자료를 분석할 경우, 반응변수들의 분포를 다변량 분포로 가정하는데, 반응변수들의 주변확률분포가 다양하게 혼합되어 있게 되면 반응변수의 다변량 분포의 결정이 쉽지 않아진다. 이는 다변량 분포를 어떻게 구성하느냐에 따라 모수의 해석 방법과 분석 결과가 크게 달라지기 때문이다. 반응변수가 혼합되어 있는 경우는 예전부터 다루어졌다. Olkin과 Tate (1961)는 반응변수가 이산형과 연속형으로 혼합되어 있는 경우를 처음으로 다루었는데 반응변수의 다변량 분포를 이산형 분포와 이산형 반응변수의 조건 하의 연속형 분포의 곱으로 나타내는 우도함수를 제안했다. Fitzmaurice와 Laird (1995)는 이를 군집을 고려한 회귀 모형으로 확장하였고 연이어,

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0007936).
[2]Corresponding author: Associate Professor, Department of Statistics, Korea University, Anam-dong 5-ga, Seongbuk-gu, Seoul 136-701, Korea. E-mail: hj4cho@korea.ac.kr

Table 2.1. Heart Disease Data

Binary responses	
Responses (Y_d)	Disease, FBS
Covariates (X_d)	Age, Sex
Continuous responses	
Responses (Y_c)	CHOL, MHR
Covariates (X_c)	Age, Sex, Vessel

Fitzmaurice와 Laird (1997)에서 혼합된 반응변수가 범주 별로 2개이상일 경우의 접근 방법을 상세히 다루었다. Fitzmaurice와 Laird (1997)에서는 다수의 이항형, 연속형 종속변수를 가지는 다변량 분석만을 제시하였지만 이산형 변수가 여러 이항형 변수로 표현가능하기 때문에 제시한 모형을 이산형과 연속형 종속변수를 가지는 다변량 접근 방법으로 제안하였다.

한편 Cox (1972)는 Olkin과 Tate (1961)가 제시한 모형과 반대인 경우인 연속형 분포와 연속형 반응변수의 조건 하의 이산형 분포의 곱으로 가정하는 방법을 제시하기도 하였고 Prentice와 Zhao (1991)와 Zhao 등 (1992)은 Liang과 Zeger (1986)가 제시한 준가능도방법(quasi-likelihood method)인 일반화추정방정식(Generalized Estimating Equation; GEE) 방법을 이산형과 연속형이 혼합된 반응변수의 경우로 확장하기도 하였다.

위에서 언급된 다양한 방법 중, Fitzmaurice와 Laird (1995)와 Zhao 등 (1992)은 동일한 반응변수의 다변량 분포를 가정하고 있다. 이에 Pinto와 Normand (2009)은 이 두 가지 방법이 반응변수들의 연관성을 무시한 단변량 모형과 비교하여 어느 정도의 효율을 가지는지를 밝히기도 하였다. 하지만 Pinto와 Normand (2009)은 반응변수가 이항형, 연속형이 각각 하나씩 존재할 때의 비교에 그치면서 다양한 관점에서 접근하지 못한 제한 점이 있다. 혼합된 반응변수가 범주 별로 2개 이상으로 늘어나게 된다면, 혼합된 반응변수 간의 연관성, 설명변수의 반응변수 기인 형태 측면에서 다변량 방법의 효율에 대한 다양한 검증이 요구된다. 이에 본 논문에서는 Fitzmaurice와 Laird (1995)와 Zhao 등 (1992)에서 제시한 모형을 혼합된 반응변수가 범주 별로 2개씩 가지는 경우로 확장하여 모의실험을 통해 여러 방면을 자세히 비교 분석하였다. 나아가 자료의 형태에 따라 접근 방법의 가이드라인을 제시함으로써 분석 시 혼합된 반응변수를 다루는데 있어 효율적인 모수추정이 가능하도록 하였다.

2. 이항형과 연속형이 혼합된 반응변수의 분석방법

Table 2.1은 이항형과 연속형이 혼합된 반응변수와 그에 따른 설명변수를 잘 보여주는 자료로 UCI Machine Learning Repository 사이트에서 제공하는 Stalog 데이터베이스 중 심장병 자료이다. 반응변수 Disease는 심장병 발병 여부, FBS는 공복시혈당이 120mg/dl보다 이상인지의 여부이며 CHOL은 콜레스테롤 수치, MHR은 최대심박수를 나타낸다. Table 2.1처럼 반응변수는 각각의 설명변수를 가지고 있는데 어떤 설명변수는 두 가지 반응변수에 모두 기인하기도 한 반면, 한가지에만 반응변수에만 포함되는 설명변수도 있다. 이 때 Y_d를 이항형 반응변수, Y_c를 연속형 반응변수라고 하자. 이에 따른 설명변수는 이항형 반응변수의 경우 X_d로 표시하고 연속형 반응변수는 X_c로 표시한다.

2.1. 단변량 방법

단변량 방법(univariate approach)은 반응변수 Y_d와 Y_c간의 연관성을 무시하고 개별적으로 분석하는 방법이다. 본 논문에서는 다변량 방법의 효율을 보여줄 때 비교척도로 사용된다. 이항형 반응변수와 연

속형 반응변수를 개별적으로 추정하되 이항형 반응변수의 경우 로짓모형을, 연속형의 경우 선형회귀모형을 다음과 같이 가정한다.

$$\begin{aligned}\text{logit}\{E(y_{d_j i}|\mathbf{x}_{d_j i})\} &= \mathbf{x}'_{d_j i}\boldsymbol{\beta}_d,\\ y_{c_j i}|\mathbf{x}_{c_j i} &= \mathbf{x}'_{c_j i}\boldsymbol{\beta}_c + \varepsilon_i,\end{aligned}$$

여기서 $\boldsymbol{\beta}_d$와 $\boldsymbol{\beta}_c$는 추정해야하는 모수이며 $y_{d_j i}$와 $y_{c_j i}$는 각각 i번째 개체의 j번째 이항형, 연속형 반응변수로 이항형 반응변수일 때는 $j = 1, \ldots, T$, 연속형 반응변수일 때는 $j = 1, \ldots, S$이다. 마지막으로 $\varepsilon_i \sim N(0, \sigma_c^2)$를 따른다.

2.2. 인수분해식 접근 방법

인수분해식 접근 방법(factorization approach)은 이항형과 연속형 반응변수를 가지는 다변량 자료를 우도함수(likelihood function)를 통해 분석하는 방법으로 우도함수 설정 방법에 따라 다양한 형태로 가지게 된다. 본 논문에서는 이항형 반응변수를 조건부로 하는 우도함수를 고려하였다. 이 방법은 Olkin과 Tate (1961)에 의해서 처음 제안되었으며 Fitzmaurice와 Laird (1997)에 의해 확장된 방법으로 반응변수의 결합분포를 이항형 다변량 분포와 이항형 반응변수가 주어졌을 때의 연속형 다변량 분포의 곱으로 표현한다. 각 개체 i가 독립을 가정하면 우도함수는 다음과 같다.

$$L(Y_d, Y_c) = \prod_{i=1,\ldots,N} f_{Y_d,Y_c}(\mathbf{y}_{di}, \mathbf{y}_{ci}) = f_{Y_d}(\mathbf{y}_{di}) f_{Y_c|Y_d}(\mathbf{y}_{ci}|\mathbf{y}_{di}),$$

여기서 $\mathbf{y}_{di} = (y_{d_1 i}, \ldots, y_{d_T i})$, $\mathbf{y}_{ci} = (y_{c_1 i} \ldots, y_{c_S i})$이며 개체 i에서 각각 $T \times 1$, $S \times 1$ 벡터를 가진다.

주변밀도 함수 $f_{Y_d}(\mathbf{y}_{di})$은 이항형 다변량 분포로 다양하게 정의될 수 있다 (Cox, 1972; Bishop 등, 1975). 그 중에서 Fitzmaurice와 Laird (1997)은 추정된 모수의 해석이 비교적 쉬운 로그 선형형태를 사용한다.

$$f(\mathbf{y}_{di}, \boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i) = \exp\left\{\boldsymbol{\Psi}'_i \mathbf{y}_{di} + \boldsymbol{\Omega}'_i \boldsymbol{\tau}_i - A(\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i)\right\}$$

여기서 $\boldsymbol{\tau}_i$은 두개 이상의 이항형 반응변수의 교차 곱 벡터로 이항형 반응변수 $y_{d_j i}$에서 $j = 1, 2, \ldots, T$이면 $(y_{d_1 i} y_{d_2 i}, \ldots, y_{d_{(T-1)} i} y_{d_T i}, \ldots, y_{d_1 i} y_{d_2 i} \cdots y_{d_{(T-1)} i} y_{d_T i})$으로 표현될 수 있다. 벡터 $\boldsymbol{\Omega}_i$는 벡터 $\boldsymbol{\tau}_i$에 대한 모수로 이항형 반응변수 간의 조건부 로그오즈비로 해석가능하다 (Fitzmaurice와 Laird, 1997). 마지막으로 $A(\boldsymbol{\Psi}_i, \boldsymbol{\Omega}_i)$는 정규화상수이다.

반면에 주변밀도 함수 $f_{Y_c|Y_d}(\mathbf{y}_{ci}|\mathbf{y}_{di})$은 다변량 정규분포로 가정한다.

$$f(\mathbf{y}_{ci}|\mathbf{y}_{di}) = (2\pi)^{-\frac{T}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{-\left(\frac{1}{2}\right)\left(\mathbf{y}_{ci} - \boldsymbol{\mu}_{(c|d)i}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_{ci} - \boldsymbol{\mu}_{(c|d)i}\right)\right\}.$$

위 식에서 $\boldsymbol{\Sigma} = \text{cov}(\mathbf{Y}_{ci}|\mathbf{Y}_{di})$, $\boldsymbol{\Gamma}$는 $S \times T$ 행렬로 이루어진 $\mathbf{y}_{di}$와 $\mathbf{y}_{ci}$의 선형회귀계수이다. $\boldsymbol{\mu}_{di}$, $\boldsymbol{\mu}_{ci}$, $\boldsymbol{\mu}_{(c|d)i}$는 각각의 설명변수와 여러 가지 연결함수로 정의될 수 있지만 본 논문에서는 $\boldsymbol{\mu}_{ci} = \mathbf{X}_{ci}\boldsymbol{\beta}_c$, $\text{logit}(\boldsymbol{\mu}_{di}) = \mathbf{X}_{di}\boldsymbol{\beta}_d$로 정의하고 $\boldsymbol{\mu}_{(c|d)i}$는 다음과 같이 정의한다.

$$\boldsymbol{\mu}_{(c|d)i} = \boldsymbol{\mu}_{ci} + \boldsymbol{\Gamma}(\mathbf{y}_{di} - \boldsymbol{\mu}_{di}), \tag{2.1}$$

여기서 $\boldsymbol{\Gamma}$는 연속형 반응변수와 이항형 반응변수 간의 연관성과 관련 있으며 $S \times T$ 행렬을 가진다.

2.3. 일반화추정방정식 접근 방법

Prentice와 Zhao (1991)와 Zhao 등 (1992)는 혼합된 반응변수의 다변량 분포가 지수족에 속한다는 가정 아래 준가능도 방법인 일반화 추정방정식(Generalized Estimating Equation; GEE)을 이용하여 모수의 일치추정량을 구할 수 있음을 보였다. 혼합된 반응변수의 다변량 분포는 다음과 같이 표현될 수 있다.

$$f(\mathbf{y}_i^*;\boldsymbol{\theta}_i) = \boldsymbol{\Delta}_i \exp\left\{\mathbf{y}_i^{*'}\boldsymbol{\theta}_i + c_i\left(\mathbf{y}_i^*, \boldsymbol{\lambda}\right)\right\},$$

여기서 $\mathbf{y}_i^*$는 i개체에서의 이항형과 연속형 반응변수를 모두 포함하는 벡터이며 $\boldsymbol{\theta}_i$은 자연모수, Δ_i는 적분상수, c_i는 모수 λ를 가지는 형태 함수(shape function)이다. 일반화 추정방정식을 이용하여 모수를 추정할 경우에 반응변수가 평균과 어떻게 연결되어 구체화하여야 하는데 본 논문에서는 인수분해식 접근 방법에서 사용한 것과 동일하게 이항형 반응변수에는 로짓모형을, 연속형 반응변수에는 선형모형을 사용하기로 한다. 이 때, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_d, \boldsymbol{\beta}_c)$의 추정식은

$$\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1}\left(\mathbf{y}_i^* - \boldsymbol{\mu}_i^*\right) = 0$$

이며, $\boldsymbol{\mu}_i^*$은 $\mathbf{y}^*$의 평균, $\mathbf{D}_i$는 $\partial\boldsymbol{\mu}_i^*/\partial\boldsymbol{\beta}^*$, $\mathbf{V}_i$는 분산행렬이다. 혼합된 반응변수를 다루는 일반화 추정방정식에서는 인수분해식 접근 방법과 다르게 식 (2.1)을 가정하지 않지만 평균에 의존하는 적절한 $\mathbf{V}_i$을 정의해야만 한다. 이는 일반화추정방정식을 이용하여 모수를 추정할 때 필요한 가정 중 하나이지만 다행스럽게도 $\mathbf{V}_i$가 잘못 정의되더라도 $\boldsymbol{\beta}^*$의 추정치는 항상 일치성을 갖는다.

3. 모의 실험

모의 실험 자료는 반응변수 간의 연관성, 설명변수가 반응변수에 기인하는 형태에 따라 식 (3.1)∼(3.3)으로부터 다양한 모의실험 자료를 생성하였다.

$$\text{logit}\left[P(y_{d_1 i} = 1 | x_{d_1 i})\right] = \beta_{d_1 0} + \beta_{d_1 1} x_{d_1 i}, \tag{3.1}$$

$$\text{logit}\left[P(y_{d_2 i} = 1 | x_{d_2 i})\right] = \beta_{d_2 0} + \beta_{d_2 1} x_{d_2 i}, \tag{3.2}$$

$$\begin{pmatrix} y_{c_1 i}|y_{d_1 i} \\ y_{c_2 i}|y_{d_2 i} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \beta_{c_1 0} + \beta_{c_1 1} x_{c_1 i} + \boldsymbol{\Gamma}_1(\mathbf{y}_{di} - \boldsymbol{\mu}_{di}) \\ \beta_{c_2 0} + \beta_{c_2 1} x_{c_2 i} + \boldsymbol{\Gamma}_2(\mathbf{y}_{di} - \boldsymbol{\mu}_{di}) \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \tag{3.3}$$

여기서 $\boldsymbol{\Gamma}_k$는 $\boldsymbol{\Gamma}$의 k번째 행 벡터이다. $\boldsymbol{\Gamma} = \left(\begin{smallmatrix} \boldsymbol{\Gamma}_1 \\ \boldsymbol{\Gamma}_2 \end{smallmatrix}\right)$는 연속형 반응변수와 이항형 반응변수 간의 연관성, ρ는 연속형 반응변수의 조건부 연관성으로 고려하였고 이항형 반응변수간의 연관성의 경우 오즈비(Odds Ratio; OR)를 이용하였다. 그러므로 생성된 자료는 반응변수 간의 연관성을 다음과 같은 세 가지 범주로 나누어 해석가능하고 본 논문의 모의실험에서는 각각을 두 가지 수준으로 나누어 자료를 생성하였다.

1. 반응변수의 다른 범주 간의 연관성 (없는 경우: $\boldsymbol{\Gamma} = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, 있는 경우: $\boldsymbol{\Gamma} = \left(\begin{smallmatrix} 4 & 4 \\ 4 & 4 \end{smallmatrix}\right)$).
2. 연속형 반응변수간의 연관성 (없는 경우: $\rho = 0$, 있는 경우: $\rho = 0.6$).
3. 이항형 반응변수간의 연관성 (없는 경우: OR $= 1$, 있는 경우: OR $= 2$).

한편, 반응변수에 기인하는 설명변수의 형태는 두 가지를 나누어 고려하였다.

1. 반응변수가 모두 동일한 설명변수를 가지는 경우: $x_{d_1 i} = x_{d_2 i} = x_{c_1 i} = x_{c_2 i} \sim N(0, 3)$.

Table 3.1. The MSE ratio of an univariate approach to multivariate approaches when the set of predictors for the continuous and binary responses are the same

Association between responses			Method	Efficiency of the multivariate approach							
Categories	Binary	Continuous		$\beta_{d_1 0}$	$\beta_{d_1 1}$	$\beta_{d_2 0}$	$\beta_{d_2 1}$	$\beta_{c_1 0}$	$\beta_{c_1 1}$	$\beta_{c_2 0}$	$\beta_{c_2 1}$
none	none	none	M1†	1.00*	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			M2‡	1.01	1.00	1.03	1.04	1.00	1.00	1.00	1.00
none	none	some	M1	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00
			M2	1.02	0.99	1.07	1.04	1.00	1.00	1.00	1.00
none	some	none	M1	1.09	0.97	1.01	1.00	1.00	0.99	1.08	1.01
			M2	1.00	1.00	1.01	1.02	1.00	0.99	1.03	1.02
none	some	some	M1	1.00	0.98	1.00	1.06	1.00	0.99	1.00	0.99
			M2	1.00	1.01	1.08	1.01	0.99	0.98	1.00	1.00
some	none	none	M1	0.99	0.99	1.00	0.99	1.00	1.00	1.00	1.00
			M2	1.01	1.02	1.02	1.03	1.00	1.00	1.00	0.99
some	none	some	M1	1.02	1.01	1.01	0.99	1.00	1.00	1.00	1.00
			M2	1.03	1.01	1.02	1.00	1.00	1.00	1.00	1.00
some	some	none	M1	1.00	0.99	1.00	0.01	1.00	1.01	1.01	1.02
			M2	1.02	0.99	1.00	1.03	0.99	1.01	1.01	1.02
some	some	some	M1	1.00	1.03	1.00	1.00	0.92	1.00	1.00	1.00
			M2	1.02	1.02	1.00	1.00	0.99	0.99	1.01	1.01

† Factorization approach, ‡ GEE approach.

* Lower value indicates that the multivariate approach is more efficient.

2. 반응변수가 각기 다른 설명변수를 가지는 경우: $x_{d_1 i} \sim N(0,1)$, $x_{d_2 i} \sim N(0,2)$, $x_{c_1 i} \sim N(0,3)$, $x_{c_2 i} \sim N(0,4)$.

또한 $\sigma^2 = 3$으로 하였고 추정하고자 하는 모수인 $(\beta_{d_1 0}, \beta_{d_1 1}, \beta_{d_2 0}, \beta_{d_2 1}, \beta_{c_1 0}, \beta_{c_1 1}, \beta_{c_2 0}, \beta_{c_2 1})$은 각각 $(-1, 0.5, 1, -1, -2, 1, 1, -0.5)$으로 설정하였다. 모의실험에서 생성된 자료는 각 2가지 수준을 가지고 있는 3가지 형태의 연관성, 2가지 형태로 이루어진 반응변수에 대한 설명변수의 기인형태를 모두 고려한 $2^3 \times 2 = 16$가지이다. 식 (3.3)에서의 ρ는 β와 $\mathbf{\Gamma}$의 값에 따라 관측된 연속형 반응변수 간의 상관계수와 차이가 날 수 있다. 그러므로 ρ를 높게 설정하여도 다른 조건으로 인한 변동으로 달라질 수 있기 때문에 강함의 정도가 아닌 '있음'과 '없음'으로 표기하기로 한다. 생성된 모의 실험자료에서 다른 범주의 반응변수 간의 연관성을 나타내는 $\mathbf{\Gamma}$가 $\left(\begin{smallmatrix} 4 & 4 \\ 4 & 4 \end{smallmatrix}\right)$인 경우 대략적으로 이항형과 연속형 반응변수의 상관계수가 대략 $(0.4 \sim 0.6)$정도가 되었다. 또한 $\rho = 0$이라도 $\mathbf{\Gamma} = \left(\begin{smallmatrix} 4 & 4 \\ 4 & 4 \end{smallmatrix}\right)$인 경우 관측된 연속형 반응변수 사이의 상관계수는 $(0 \sim 0.2)$정도가 나타났다.

Table 3.1과 Table 3.2는 반응변수에 기인하는 설명변수의 형태와 반응변수간의 연관성을 모두 고려하여 얻은 모의실험 결과로, 다변량 방법(인수분해식 접근 방법, 일반화추정방정식 접근 방법)으로 추정된 모수가 단변량방법에 의해 추정된 모수보다 얼마나 모수와 가까이 추정되었는지 보여준다. Table 안의 값은 500개체 자료를 100번 생성하여 구한 다변량 모형의 MSE와 단변량 모형의 MSE의 비를 나타낸 것이다. 즉, 작은 값을 가지는 모형일수록 모수를 잘 추정하는 모형이라고 할 수 있다.

$$\text{MSE의 비} = \frac{\text{다변량 접근 방법 추정치의 MSE}}{\text{단변량 접근 방법 추정치의 MSE}}$$

또한 Table 3.1과 Table 3.2에서는 나타나지 않았지만 추정된 모든 회귀계수들이 방법론, 반응변수의 연관성에 상관없이 신뢰구간에 95% 가까이 포함되어 있어 일치성을 가짐을 확인할 수 있었다.

Table 3.2. The MSE ratio of an univariate approach to multivariate approaches when the set of predictors for the continuous and binary responses are not the same.

Association between responses			Method	Efficiency of the multivariate approach							
Categories	Binary	Continuous		$\beta_{d_1 0}$	$\beta_{d_1 1}$	$\beta_{d_2 0}$	$\beta_{d_2 1}$	$\beta_{c_1 0}$	$\beta_{c_1 1}$	$\beta_{c_2 0}$	$\beta_{c_2 1}$
none	none	none	M1†	1.00*	1.00	0.99	0.99	1.00	1.00	1.00	1.00
			M2‡	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.00
none	none	some	M1	1.00	1.01	1.00	0.99	1.00	1.00	1.00	1.00
			M2	1.00	1.01	1.01	1.01	1.00	1.00	1.00	1.00
none	some	none	M1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			M2	1.00	1.00	1.01	1.02	1.00	1.00	1.00	1.00
none	some	some	M1	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00
			M2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
some	none	none	M1	0.73	0.94	0.99	1.00	0.66	1.00	1.00	1.00
			M2	1.00	0.98	0.98	0.98	0.99	1.00	0.99	0.99
some	none	some	M1	0.77	0.95	0.99	1.00	0.70	1.01	0.97	1.00
			M2	1.00	0.99	0.98	0.98	1.00	1.00	1.00	0.99
some	some	none	M1	0.75	0.91	1.00	0.98	0.68	1.00	0.98	1.00
			M2	1.00	0.98	0.97	0.99	1.00	1.01	0.99	1.00
some	some	some	M1	0.74	0.94	1.00	0.99	0.69	1.01	1.00	1.00
			M2	1.00	1.00	0.99	1.00	0.98	0.99	1.00	0.99

† Factorization approach, ‡ GEE approach

* Lower value indicates that the multivariate approach is more efficient

Table 3.1은 모든 반응변수가 범주와 상관없이 동일한 설명변수를 가지는 경우에서의 MSE비를 보여준다. 모든 다변량 방법에서 값들이 1을 크게 벗어나지 못하고 있음을 살펴볼 수 있다. 이는 다변량 모형에서 추정치 MSE가 단변량 모형 시의 MSE와 비교했을 때와 큰 차이를 보여주는 못함을 뜻한다. 이러한 결과는 반응변수 사이의 여러 연관성 형태에 상관없이 거의 모든 값에서 동일하게 나타나고 있다. 그러므로 모든 반응변수가 동일한 설명변수를 가지는 경우에는 반응변수 간의 연관성과 관계없이 다변량 모형과 단변량 모형의 모수 추정치 사이의 효율은 비슷하다고 할 수 있다.

반면에, Table 3.2는 반응변수가 서로 다른 설명변수를 가지고 있는 경우를 다루었다. 이항형간, 연속형간, 두 범주간 연관성이 모두 없는 경우 두 다변량 접근 방법은 단변량 접근 방법에 비해 큰 효율을 얻지 못하였다. 또한 두 범주간 연관성이 없는 경우에는 이항형간, 연속형간 연관성의 유무에 상관없이 두 다변량 접근 방법은 단변량 접근 방법에 비해 좋은 효율을 얻지 못하였다. 반면에 범주간의 연관성이 있는 경우에는 이항형, 연속형 반응변수 간의 연관성과 관계없이 두 단변량 분석방법으로 얻어진 MSE가 다변량 접근 방법의 MSE보다 현저히 작은 회귀계수들을 여럿 관찰할 수 있었다. 효율의 정도는 인수분해식 접근 방법이 일반화추정방정식 접근 방법보다 우수했으며 연속형 반응변수의 회귀계수보다는 이항형 반응변수에 대한 계수들이 그 영향을 크게 받았다. 결국, 다변량 접근 방법의 장점은 반응변수들에 기인하는 설명변수의 형태가 다를 때 나타나며 같은 범주 내 연관성보다는 다른 범주 간의 연관성에 관련이 깊다. 또한 같은 범주의 반응변수끼리 동일한 설명변수가 가지고 다른 범주 사이에는 다른 설명변수를 갖는 경우도 생각해볼 수 있는데 Table를 첨부하지 않았지만 Table 3.2와 비슷한 결과를 얻었다.

4. 예제

이항형과 연속형 변수로 혼합되어 있는 Table 2.1의 다변량 자료를 본 논문에서 다루었던 3가지 방법에

Table 4.1. Correlation matrix of binary and continuous responses

	Disease	FBS
CHOL	0.12	−0.42
MHR	0.03	0.02

Table 4.2. Parameter estimates and P-values for Heart Disease Data.

Model	Parameter	Univariate approach		Factorization approach		GEE approach	
		Estimate	P-value	Estimate	P-value	Estimate	P-value
Disease	Intercept	−4.864	<0.001	−4.786	<0.001	−4.664	<0.001
	Age	0.064	<0.001	0.063	<0.001	0.061	<0.001
	Sex	1.622	<0.001	1.599	<0.001	1.583	<0.001
FBS	Intercept	−4.331	<0.001	−4.086	0.003	−3.514	<0.001
	Age	0.042	0.037	0.039	0.067	0.029	0.049
	Sex	0.353	0.361	0.366	0.224	0.294	0.226
log(CHOL)	Intercept	5.335	<0.001	5.325	<0.001	5.314	<0.001
	Age	0.004	0.009	0.004	0.011	0.004	0.005
	Sex	−0.070	0.006	−0.066	0.023	−0.069	0.006
	Vessel	0.018	0.173	0.005	0.342	0.015	0.191
log(MHR)	Intercept	5.384	<0.001	5.415	<0.001	5.419	<0.001
	Age	−0.006	<0.001	−0.007	<0.001	−0.007	<0.001
	Sex	−0.038	0.057	−0.043	0.012	−0.044	0.009
	Vessel	−0.021	0.048	0.001	0.476	0.002	0.431

적용해보고자 한다. 이 다변량 자료는 심장병(heart disease)에 관련된 자료로서 결측치가 없는 270명을 개체 정보를 포함하고 있다. 심장병은 순환기 질환 중 심장에 관련된 질환으로 고지혈증, 심근경색, 협심증 등을 포함한다. 분석 목적은 심장병의 발병여부와 설명변수인 성별과 나이와의 관계를 관련있는 다른 반응변수들과의 연관관계와 함께 고려함으로써 더 나은 추정치를 얻고자 하는데 있다. 자료의 이항형 반응변수 간의 오즈비는 0.91, 연속형 반응변수 간의 상관계수는 −0.02이며 다른 범주간의 상관계수는 다음과 같다.

같은 범주의 반응변수간 상관관계는 없다고 판단되지만 다른 범주의 반응변수 간 연관성에서는 이항형 반응변수인 FBS와 연속형 반응변수 CHOL간의 상관계수가 −0.42로 높다. 반응변수가 다른 설명변수를 가지고 있고 다른 범주간의 연관성이 있다고 판단되므로 다변량 분석이 단변량 분석보다 모수추정에 있어 좋은 효율을 보일 것이 예상된다. Table 4.2는 다변량 방법인 인수분해식 접근 방법, 일반화 추정방정식 접근 방법, 단변량 접근 방법으로 분석하여 얻은 회귀계수 추정치와 그에 대한 표준오차를 보여준다. 인수분해식 접근 방법의 경우 여성의 심장병 발병 여부(disease = 1)의 오즈비가 남성보다 4.95배 정도 높은 반면, 단변량 분석방법은 5.06배 정도 높게 나왔으며 그 차이는 둘다 유의하다. 반응변수 FBS의 경우에는 3가지 방법 모두에서 비슷한 회귀계수 추정치를 얻을 수 있었다. 콜레스테롤 수치를 나타낸 연속형 반응변수(CHOL)에서는 세 방법 모두 나이가 많을수록, 남성일수록 많음을 보여주었으며 설명변수 VH는 모든 방법에서 유의하지 않았다. 반면에, FBS의 경우에는 나이와 성별은 세 방법에서 모두 비슷한 계수를 보여주었지만 설명변수 VH의 경우에서 상이한 결과를 보여주었다. 단변량 방법으로 분석했을 때 VH의 회귀계수는 −0.021로 유의하여 major vessel의 수(Vessel)가 많을수록 MHR가 줄어듦을 보여주었지만 두 다변량 분석방법에서는 이 설명변수가 유의하지 않게 나왔다. 전체적으로 다변량 방법인 인수분해식 접근 방법과 일반화추정방정식 접근 방법은 서로 비슷한 추정치를 보여주었고 단변량 접근 방법은 몇몇 추정치에서 다변량 접근 방법과 상이한 값들을 보여주었다. 이는 다

변량 접근 방법이 단변량 접근 방법과 다르게 다른 범주의 반응변수간 연관성을 고려하여 추정하였기 때문이다.

5. 결론

이항형과 연속형 반응변수가 범주별로 2개일 경우 각각의 설명변수에 대한 회귀계수를 어떠한 접근 방법으로 추정해야 효율적인 추정치를 얻을 수 있는지 살펴보았다. 모의실험은 반응변수들 간의 연관성을 같은 범주내의 연관성과 다른 범주끼리의 연관성으로 나누었고 반응변수에 기인하는 설명변수의 형태는 반응변수의 설명변수가 동일한 경우와 다른 경우로 나누어 설정하였는데 이는 이항형과 연속형 반응변수 각각 하나씩을 다룬 Pinto와 Normand (2009)보다 확장된 관점이다. 모의실험 결과, 다른 범주의 반응변수 사이의 연관성은 다변량 접근 방법과 단변량 접근 방법의 차이를 가장 많이 내는 요인임을 보여주었고 그 효율은 인수분해식 접근 방법이 일반화추정방정식 접근 방법보다 더 우세함을 보였다. 하지만 모든 반응변수가 같은 설명변수를 가지게 될 경우 이러한 차이는 나타나지 않았다. 즉, 모든 반응변수의 설명변수가 같은 경우 반응변수간의 어떠한 연관성도 다변량 접근 방법과 단변량 접근 방법의 차이를 보여주지 못하였다. 그러므로 반응변수가 혼합되어 있는 자료를 분석할 때, 모든 반응변수가 같은 설명변수를 가지는 경우에는 다변량 접근 방법과 단변량 접근 방법 사이의 효율차이가 크지 않기 때문에 단변량 접근 방법을 사용하는 것이 시간과 편의성을 고려할 때 좋다. 반면에, 반응변수가 다른 설명변수를 가지는 경우에는 다변량 접근 방법을 사용하는 것이 효율적인 추정치를 얻을 수 있다. 특히, 반응변수의 다른 범주간 연관성이 높을 경우에는 추정치의 효율면에서 큰 차이를 보이기 때문에 단변량 접근 방법보다는 다변량 접근 방법을 사용하는 것이 바람직하다.

본 논문에서는 다변량 접근 방법 중에서 주로 사용되면서도 비교적 다루기 쉬운 반응변수의 다변량 분포를 사용하여 단변량 접근 방법과 비교하였다. 하지만 상황에 따라서는 일반화선형모형 등의 방법을 통해 여러 반응변수 모수들 사이의 연관성을 규명하여 조금 더 나은 추정치를 얻을 수 있다. 이는 추후 과제로 남겨두도록 한다.

References

Bishop, Y. M. D., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*, Theory and Practice, The MIT Press, Cambridge, MA, London.

Cox, D. R. (1972). The analysis of multivariate binary data, *Applied Statistics*, **21**, 113–120.

Fitzmaurice, G. M. and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering, *Journal of the American Statistical Association*, **90**, 845–852.

Fitzmaurice, G. M. and Laird, N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values, *Biometrics*, **53**, 110–122.

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

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables, *Annals of Mathematical Statistics*, **32**, 448–465.

Pinto, A. T. and Normand, S. L. T. (2009). Correlated bivariate continuous and binary outcomes: Issues and applications, *Statistics in Medicine*, **28**, 1753–1773.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics*, **47**, 825–839.

Zhao, L. P., Prentice, R. L. and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model, *Journal of the Royal Statistical Society, Series B*, **54**, 805–811.