The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1347
29(7), 1347–1359

The EM algorithm for mixture regression with missing covariates

Hyungmin Kim[a] · Geonhee Ham[b] · Byungtae Seo[a,1]

[a]Department of Statistics, Sungkyunkwan University;
[b]Center for Public Opinion and Quantitative Research, The Asan Institute for Policy Studies

(Received August 31, 2016; Revised October 21, 2016; Accepted October 22, 2016)

Abstract

Finite mixtures of regression models provide an effective tool to explore a hidden functional relationship between a response variable and covariates. However, it is common in practice that data are not fully observed due to several reasons. In this paper, we derived an expectation-maximization (EM) algorithm to obtain the maximum likelihood estimator when some covariates are missing at random in the finite mixture of regression models. We conduct some simulation studies and we also provide some real data examples to show the validity of the derived EM algorithm.

Keywords: mixture models, missing covariates, mixture regression, EM algorithm

1. 서론

보통의 회귀분석은 설명변수와 반응변수가 하나의 함수식으로 표현될 수 있다는 가정에 기반한다. 이는 설명변수가 주어졌을 때 반응변수의 분포는 동일하다는 가정으로 해석될 수 있는데 현실적으로는 주어진 자료안에 서로 상이한 회귀식이 존재할 수 있으며 만약 일반적인 회귀분석에서 그러한 상이성을 고려하지 않을 경우 유의미한 결과를 기대하기 어렵다. 이렇게 서로 상이한 회귀식이 혼재한 경우는 설명변수가 주어졌을때 반응변수의 조건부 분포가 서로 섞여 있다고 볼 수 있는데 이를 통계적으로 모형화 한 것이 혼합회귀모형이다.

혼합회귀모형에서는 각 관측값이 두 개 혹은 그 이상의 서로 다른 회귀식을 가지는 여러 집단이 혼합된 집단에서 얻어졌다고 가정한다. 예를 들어 설명변수가 체질량지수이고 반응변수가 혈압이라고 하고 회귀분석을 시행했다고 하면 이는 관측값들이 모두 동질한 집단에서 나왔음을 가정하는 것이다. 하지만 현실적으로는 관측값들이 서로 다른 회귀식을 가지는 혼합된 모집단에서 얻어졌을 가능성이 높다. 예를 들어 관측되지 않은 성별이나 인종에 따라 서로 다른 회귀식을 가지는 혼합된 집단에서 관측값들이 얻어졌을 수 있는 것이다.

이러한 경우 혼합회귀모형은 두 개 혹은 그 이상의 숨겨진 회귀식을 얻어낼 수 있는 유용한 도구가 될 수 있다. 이를 통해 연구자들은 주어진 자료에서 숨겨진 그룹을 찾을 수 있을 뿐 아니라 각 그룹에서

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 Techonology (NRF-2013R1A1A2057715).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: seobt@skku.edu

설명변수와 반응변수간의 다른 선형식도 얻어낼 수 있다. 이러한 혼합회귀모형은 Quandt와 Ramsey (1978)에 의해 switching regression이라는 개념으로 처음 소개되었으며 이후 latent class regression (Bandeen 등, 1997), mixtures of experts (Jacobs 등, 1991) 등 다양한 형태로 확장 발전되어 오고 있다.

혼합회귀모형에서 모수의 최대우도추정량(maximum likelihood estimator; MLE)은 일반적으로 Expectation-Maximization(EM) 알고리즘 (Dempster 등, 1977)을 통해 얻어질 수 있는데 이는 각각의 관측치가 어느 그룹에서 왔는지에 대한 정보를 결측으로 가정한 후 EM 알고리즘을 적용한다. 보통의 일반적인 추정 알고리즘과 달리 EM 알고리즘은 각 단계에서 항상 우도함수를 증가시키는 성질을 가지고 있어 특히 혼합모형에서 매우 안정적으로 모수를 추정할 수 있는 가장 일반적인 도구로 이용되어지고 있다. 또한 정규혼합회귀모형을 비롯한 많은 혼합모형에서 E-step과 M-step이 수리적으로 계산 가능하다는 장점이 있다. 이를 처리하기 위한 소프트웨어로는 R에서 제공하는 Mixtools (Benaglia 등, 2009), FlexMix (Leisch, 2004) 등이 대표적이다.

실제 자료분석에서는 혼합회귀모형뿐만 아니라 일반적인 회귀모형 적합시 설명변수 중 일부가 관측되지 않는 경우가 매우 흔히 발생하게 되는데 만약 결측이 있는 일부 자료를 모두 제거하고 모형적합을 할 경우 분석의 효율성이 매우 떨어질 수 있다. 효율적인 통계적 분석을 위해서는 일부 결측이 있는 자료를 제거하지 않고 관측된 모든 정보를 충분히 사용한 통계 분석을 해야 하는데 이를 위해서는 먼저 결측이 일어나는 메커니즘에 대한 가정이 필요하다. 큰 틀에서 결측메커니즘은 Missing Completely At Random(MCAR), Missing At Random(MAR), Missing Not At Random(MNAR)로 구분되는데 이 중 가장 일반적으로 가정하는 결측유형은 MAR이며, MAR은 결측이 다른 결측값과는 무관하고, 오직 관측된 값에 의존하여서만 발생한다는 가정으로 본 논문에서는 혼합회귀모형에서 모수추정시 MAR 가정하에서 공변량에 결측이 있는 경우에서의 모수추정을 위한 EM 알고리즘을 제안하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 혼합모형과 혼합회귀모형을 소개하고 EM 알고리즘을 이용하여 MLE를 구하는 과정에 대해 설명하였다. 3장에서는 결측 공변량을 포함한 혼합회귀모형에서 EM 알고리즘을 이용하여 모수를 추정하는 방법에 대해 제안 하였고 4장의 모의 실험을 통해 제안한 EM 알고리즘의 효용성을 확인하였다. 또한, 5장의 사례연구에서는 제안하는 방법을 실제 자료에 적용해 보고 실제로 모수를 잘 추정하는지 확인하였다. 마지막으로 6장에서는 본 논문의 결론에 대해 다루었다.

2. 혼합회귀모형

2.1. 혼합모형

본 절에서는 혼합회귀모형에 대해 설명하기에 앞서 먼저 혼합모형에 대해 간략히 설명하고자 한다. 일반적으로 우리는 주어진 자료가 하나의 동질한 집단으로부터 나왔다고 생각을 하지만 실제로 자료들은 하나의 동질한 집단이 아닌 여러 집단으로 부터 나온 경우가 많다. 이러한 경우 집단의 분포를 추정하는데 있어서 사용할수 있는 통계적 모형이 혼합모형이다. 혼합모형은 1846년 영국의 수리통계학자 칼 피어슨(Karl Pearson)에 의해 처음 제안되었다. 칼 피어슨은 1,000마리의 게의 몸통 길이의 비율을 측정하여 히스토그램을 그리는 과정에서 비대칭의 종 모양을 발견하였다. 칼 피어슨은 자료 속에는 두 개의 서로 다른 종의 게들이 존재하며, 종에 따라 게의 몸통길이가 다르다고 판단하였다. 이에 하나가 아닌 두 개의 정규분포로 자료를 적합시켜 분석을 하였으며, 이것이 혼합모형에 대한 최초의 접근이였다. 이후 많은 현실 자료에서 잠재적 요인에 의해 여러 개의 분포를 따르는 자료의 분석에 있어서 혼합모형은 유용한 통계적 모형이 되었다.

Redner와 Walker (1984), Mclachlan과 Krishnan (1997)은 혼합모형에서 모수를 추정하는 방법으로 EM 알고리즘에 대해 소개하고 이를 이용하여 MLE를 구하는 방법에 대해 설명하였다. 혼합모형을 나

타내는 확률밀도함수는 일반적으로 다음의 식 (2.1)과 같이 표현될 수 있다.

$$f(x;\Theta)=\sum_{j=1}^{k}\pi_j f_j(x;\theta_j),\qquad \sum_{j=1}^{k}\pi_j=1,\ \ \pi_j>0. \tag{2.1}$$

여기서 $\Theta=(\pi_1,\ldots,\pi_k,\theta_1,\ldots,\theta_k)$이고, π_j는 혼합비율(mixing proportion), 그리고 f_j는 각 성분(component)의 확률밀도함수를 나타낸다. 일반적으로 k는 유한개이며, 따라서 식 (2.1)는 k개의 성분이 혼합된 유한 혼합모형(finite mixture model)을 나타낸다.

2.2. 혼합선형회귀모형

혼합선형회귀모형은 독립변수와 종속변수 사이의 관계를 선형관계로 나타낼 경우에 두 변수 사이의 관계가 일반적인 선형회귀모형처럼 하나의 선형관계로 표현되지 않고 여러개의 선형 관계가 존재하는 경우에 이용한다. 이러한 경우에 $Y_i=\boldsymbol{X}_i^T\boldsymbol{\beta}+\epsilon_i$, $i=1,\ldots,n$와 같은 보통의 선형회귀모형을 가정하는 대신 각 관측값은 π_j의 확률을 가지고 서로 다른 회귀계수 $\boldsymbol{\beta}_j$를 가지는 $Y_i=\boldsymbol{X}_i^T\boldsymbol{\beta}_j+\epsilon_{ij}$, $i=1,\ldots,n$, $j=1,\ldots,J$의 선형모형을 따른다고 가정한다. 여기서, $\boldsymbol{X}_i=(X_{i1},\ldots,X_{pi})^T$는 i번째 관측치의 공변량벡터이고 $\boldsymbol{\beta}_j=(\beta_{j1},\ldots,\beta_{jp})^T$는 j번째 성분의 회귀계수벡터이며 $\sum_{j=1}^{J}\pi_j=1$을 만족한다.

이때 각각의 주어진 j에 대하여 오차항 ϵ_{ij}가 평균이 0이고 분산이 σ_j^2인 정규분포를 따른다고 가정하면 $Y_i|\boldsymbol{X}_i$의 확률밀도 함수는 다음의 식 (2.2)와 같이 표현될 수 있다 (DeSarbo와 Cron, 1988).

$$f(y_i|\boldsymbol{x}_i)=\sum_{j=1}^{J}\pi_j\phi_1\left(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j,\sigma_j^2\right), \tag{2.2}$$

여기서 $\phi_k(\cdot;a,b)$는 평균이 a이고 분산이 b인 k차원 다변량 정규분포를 나타낸다. 이 모형은 공변량 $\boldsymbol{x}_i$를 확률변수가 아닌 고정된 값으로 간주한 모형으로 mixtures of regressions with fixed covariates(MRFC)라고 부른다.

MRFC는 $(\boldsymbol{X}_i^T,Y_i)$가 j번째 그룹에 속할 사후확률이

$$\frac{\pi_j\phi_1\left(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j,\sigma_j^2\right)}{\sum_{h=1}^{J}\pi_h\phi_1\left(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j,\sigma_j^2\right)}$$

과 같이 주어지는 모형으로 그 확률이 $Y_i|\boldsymbol{X}_i$의 분포에만 의존하는 모형을 나타내는데 Hennig (2000)은 이 확률이 $(\boldsymbol{X}_i^T,Y_i)$의 결합 확률분포에 의존하는 모형으로

$$f(y_i,\boldsymbol{x}_i)=\sum_{j=1}^{J}\pi_j\phi_1\left(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j,\sigma_j^2\right)\phi_p\left(\boldsymbol{x}_i,\boldsymbol{\mu}_j,\Sigma_j\right).$$

을 제시하였다. 이 모형은 Cluster-Weighted Models(CWMs)이라고 불리우며 Ingrassia 등 (2012, 2014), Subedi 등 (2013), Punzo (2014)에 의해 다양한 분포가정하에서 연구되었다.

또다른 모형으로 Jacobs 등 (1991)은 혼합확률 π_j가 설명변수에 의존하는 모형으로 다음과 같은 모형을 제시하였는데 이는 Mixture-of-Experts(ME)라고 불리우며 기계학습분야에서 많이 이용되고 있다.

$$f(y_i|\boldsymbol{x}_i)=\sum_{j=1}^{k}\pi\left(\boldsymbol{x}_i;\eta_j\right)\phi_1\left(y_i;\boldsymbol{x}_i^T\boldsymbol{\beta}_j,\sigma_j^2\right)$$

여기서, $\pi(\boldsymbol{x}_i;\eta_j)$는 설명변수에 의존하는 혼합확률에 관한 모형으로 모수 η_j를 가지는 모수적 모형이며 보통은 로지스틱이나 다항로지스틱 모형의 형태를 이용한다.

2.3. MRFC 모형에서의 EM 알고리즘

본 절에서는 혼합회귀모형 중 MRFC 모형에서 EM 알고리즘을 이용하여 모수를 추정하는 방법에 대해 간략하게 소개하고자 한다. 혼합모형에서의 모수에 대한 MLE는 일반적으로 EM 알고리즘을 통해 얻어질 수 있는데 혼합모형에서는 각각의 관측값이 어느 그룹에 속하는지에 대한 정보를 결측이라고 보고 일반적인 EM 알고리즘을 적용한다. 특히 혼합모형에서의 EM은 많은 경우에 E-step과 M-step이 계산 가능한 형태로 도출이 되어 매우 안정적으로 모수를 추정할 수 있다. MRFC 모형에서 EM 알고리즘을 이용하기 위해 우리는 먼저 각 관측값이 어떠한 그룹에 속하는지를 나타내는 지시함수 Z_{ij}를 다음과 같이 정의한다.

$$Z_{ij} = \begin{cases} 1, & Y_i|\boldsymbol{X}_i\text{가 } j\text{번째 성분에서 나온 경우,} \\ 0, & \text{그렇지 않은 경우.} \end{cases}.$$

이때, $\boldsymbol{X}_i$가 주어졌을 때 $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iJ})$와 Y_i의 결합확률 밀도함수는 다음과 같이 나타낼 수 있다.

$$f\left(y_i, \boldsymbol{z}_i|\boldsymbol{x}_i; \Theta\right) = \prod_{j=1}^{k} \left[\pi_j \phi_1\left(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2\right)\right]^{z_{ij}}$$

여기서, $\Theta = (\pi_1, \ldots, \pi_J, \beta_1, \ldots, \beta_J, \sigma_1^2, \ldots, \sigma_J^2)$이다. 주어진 자료 $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$과 위의 결합 확률밀도함수로부터 로그 우도 함수는 다음과 같이 표현될 수 있다.

$$\ell\left(\Theta\right) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\left(\log \pi_j + \log \phi_1\left(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2\right)\right). \tag{2.3}$$

위의 식 (2.3)에서 z_{ij}는 관측가능하지 않으므로 EM 알고리즘에서는 다음과 같은 E-step과 M-step을 반복하여 MLE를 구할 수 있다.

- E-step: 현재의 추정량을 $\Theta^{(t)}$라 할때, 전체우도함수의 조건부 기댓값인 Q는 다음과 같다.

$$\begin{aligned} Q\left(\Theta|\Theta^{(t)}\right) &= E\left[\ell\left(\Theta\,|\boldsymbol{X}_i, Y_i, \boldsymbol{Z}_i\right)|\,\boldsymbol{X}_i = \boldsymbol{x}_i, Y_i = y_i; \Theta^{(t)}\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \hat{z}_{ij}\left(\log \pi_j + \log \phi_1\left(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2\right)\right), \end{aligned} \tag{2.4}$$

  여기서,

$$\hat{z}_{ij} = E\left[Z_{ij}|\boldsymbol{X}_i = \boldsymbol{x}_i, Y_i = y_i; \Theta^{(t)}\right] = \frac{\pi_j^{(t)}\phi_1\left(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(t)}, \sigma_j^{2\,(t)}\right)}{\sum_{h=1}^{J}\pi_h^{(t)}\phi_1\left(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_h^{(t)}, \sigma_h^{2\,(t)}\right)} \tag{2.5}$$

- M-step: 위의 식 (2.4)을 최대화하는 단계이며, 그 결과는 다음과 같다.

$$\begin{aligned} \hat{\pi}_j^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ij}, \\ \hat{\beta}_j^{(t+1)} &= \left(X^T\hat{W}_jX\right)^{-1}X^T\hat{W}_jY, \\ \hat{\sigma}_j^{2(t+1)} &= \frac{\sum_{i=1}^{n}\hat{z}_{ij}\left(y_i - \left(\boldsymbol{x}_i^T\beta_j^{(t)}\right)\right)^2}{\sum_{i=1}^{n}\hat{z}_{ij}}. \end{aligned} \tag{2.6}$$

위의 결과는 $Q(\Theta|\Theta^{(t)})$를 각각의 모수로 미분하여 얻어지며, 식 (2.6)에서 X는 절편이 있는 보통의 선형회귀모형에서의 design matrix를 의미하며 $\hat{W}_j$는 $\hat{z}_{ij}$를 대각원소로 하는 $n \times n$ 대각행렬이다.

3. 결측공변량을 갖는 혼합회귀모형

본 장에서는 혼합회귀모형에서 공변량이 결측값을 가지는 경우에 있어서 모수를 추정하기 위한 EM 알고리즘을 MAR 가정하에서 제안하고자 한다. MAR 가정하에서 공변량의 일부 또는 전체가 결측인 경우에 EM 알고리즘을 적용하기 위해서는 공변량의 분포에 대한 가정이 있어야 하는데 본 논문에서는 MRFC 모형에서 공변량이 다변량정규분포를 따른다는 가정에서의 EM 알고리즘을 제안하고자 한다.

이를 위하여 먼저 $\boldsymbol{X}_i$가 평균이 $\boldsymbol{\mu}$이고 분산이 Σ인 p차원 다변량 정규분포를 따른다고 하자. 이 경우 $\boldsymbol{X}_i$, Y_i, $\boldsymbol{Z}_i$의 결합밀도함수로부터 로그우도함수는

$$\ell(\Theta) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\left(\log\pi_j + \log\phi_1\left(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2\right) + \log\phi_p\left(\boldsymbol{x}_i; \boldsymbol{\mu}, \Sigma\right)\right) \tag{3.1}$$

와 같이 주어진다. EM 알고리즘을 유도하기 위하여서는 $\boldsymbol{X}_i$에 결측이 존재하는 경우와 그렇지 않은 경우를 구분하여야 하는데 이를 위하여 인덱스 집합(index set) $\mathcal{O}$를 모든 공변량이 관측된 경우의 인덱스(index)의 집합으로 정의하고 $\mathcal{M}$을 그렇지 않은 경우의 인덱스의 집합이라 정의하자. 이경우에 $\mathcal{O}$와 $\mathcal{M}$은 $\mathcal{O}\cap\mathcal{M} = \emptyset$과 $\mathcal{O}\cup\mathcal{M} = \{1,\ldots,n\}$을 만족한다. 이러한 두 인덱스 집합을 이용하면 식 (3.1)은 다음과 같이 다시 표현될 수 있다.

$$\begin{aligned}\ell(\Theta) =& \sum_{i\in\mathcal{O}}\sum_{j=1}^{k} z_{ij}\left(\log\pi_j + \log\phi_1\left(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2\right) + \log\phi_p\left(\boldsymbol{x}_i; \boldsymbol{\mu}, \Sigma\right)\right)\\ &+ \sum_{i\in\mathcal{M}}\sum_{j=1}^{k} z_{ij}\left(\log\pi_j + \log\phi_1\left(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_j, \sigma_j^2\right) + \log\phi_p\left(\boldsymbol{x}_i; \boldsymbol{\mu}, \Sigma\right)\right).\end{aligned} \tag{3.2}$$

공변량에 결측이 있는 경우에서의 E-step은 공변량에 결측값이 존재하는 경우와 그렇지 않은 경우로 나누어서 계산하여야 하는데 공변량에 결측이 없는 경우(즉 $i\in\mathcal{O}$)의 계산은 식 (2.5)와 같이 $\hat{z}_{ij}^o = E(Z_{ij}|\boldsymbol{X}_i = \boldsymbol{x}_i, Y_i = y_i, \Theta)$만을 계산하면 되지만 공변량에 결측값이 있는 경우(즉, $i\in\mathcal{M}$)에서는 경우는 $\hat{z}_{ij}^m = E(Z_{ij}|\boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i, \Theta)$, $\widehat{z_{ij}\boldsymbol{x}_i} = E(Z_{ij}\boldsymbol{X}_i|\boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i, \Theta)$, $\widehat{z_{ij}\boldsymbol{x}_i\boldsymbol{x}_i^T} = E(Z_{ij}X_iX_i^T|\boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i, \Theta)$을 함께 계산하여야 한다. 여기서 X_i^o는 X_i에서 관측이 된 부분을 의미한다. 이 경우에 $\hat{z}_{ij}^m$은 X_i^O와 Y_i의 결합분포가 다변량 정규분포임을 이용하여 식 (2.5)와 유사하게 계산되어질수 있다. 하지만, $\widehat{z_{ij}\boldsymbol{x}_i}$과 $\widehat{z_{ij}\boldsymbol{x}_i\boldsymbol{x}_i^T}$의 계산은 관측된 값들이 주어졌을 때 $Z_{ij}\boldsymbol{X}_i$와 $Z_{ij}\boldsymbol{X}_i\boldsymbol{X}_i^T$의 조건부 분포를 구해야 하므로 그 계산이 다소 복잡할 수 있다. 하지만, 다음의 사실로부터 $E(X_i|\boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i, \Theta)$와 $E(X_iX_i^T|\boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i, \Theta)$의 계산만을 이용하여 $\widehat{z_{ij}\boldsymbol{x}_i}$과 $\widehat{z_{ij}\boldsymbol{x}_i\boldsymbol{x}_i^T}$을 쉽게 계산할 수 있다.

$$\begin{aligned}\widehat{z_{ij}\boldsymbol{x}_i} &= E\left[Z_{ij}\boldsymbol{X}_i|\boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i; \Theta\right]\\ &= E\left[E(Z_{ij}\boldsymbol{X}_i|Z_{ij}, \boldsymbol{X}_i^o, Y_i)|\boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i; \Theta\right]\\ &= E\left[Z_{ij}|\boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i; \Theta\right]E\left[X_i|Z_{ij} = 1, \boldsymbol{X}_i^o = \boldsymbol{x}_i^o, Y_i = y_i; \Theta\right]\\ &= \hat{Z}_{ij}^m\hat{X}_i,\end{aligned}$$

$$\begin{aligned}\widehat{z_{ij}\boldsymbol{x}_i\boldsymbol{x}_i^T} &= E\left[Z_{ij}\boldsymbol{X}_i^T\boldsymbol{X}_i|\boldsymbol{X}_i^o=\boldsymbol{x}_i^o, Y_i=y_i;\Theta\right]\\ &= E\left[E(Z_{ij}\boldsymbol{X}_i^T\boldsymbol{X}_i|Z_{ij},\boldsymbol{X}_i^o,Y_i)|\boldsymbol{X}_i^o=\boldsymbol{x}_i^o,Y_i=y_i;\Theta\right]\\ &= E\left[Z_{ij}|\boldsymbol{X}_i^o=\boldsymbol{x}_i^o,Y_i=y_i;\Theta\right]E\left[\boldsymbol{X}_i^T\boldsymbol{X}_i|Z_{ij}=1,\boldsymbol{X}_i^o=\boldsymbol{x}_i^o,Y_i=y_i;\Theta\right]\\ &= \hat{Z}_{ij}^m\widehat{X_i^TX_i}\end{aligned}$$

여기서, $\hat{\boldsymbol{x}}_i = E[\boldsymbol{X}_i|Z_{ij}=1,\boldsymbol{X}_i^o=\boldsymbol{x}_i^o,Y_i=y_i;\Theta]$이고 $\widehat{\boldsymbol{x}_i\boldsymbol{x}_i^T} = E[\boldsymbol{X}_i\boldsymbol{X}_i^T|Z_{ij}=1,\boldsymbol{X}_i^o=\boldsymbol{x}_i^o,Y_i=y_i;\Theta]$이며 이들은 $(\boldsymbol{X}_i^o,Y_i)$의 결합분포가 다시 다변량정규분포임을 이용하여 쉽게 계산할 수 있다. 따라서 식 (3.2)의 조건부 기댓값인 $Q(\Theta|\Theta^{(t)})$는 모든 관측값들과 $\hat{\boldsymbol{x}}_i$, $\widehat{\boldsymbol{x}_i\boldsymbol{x}_i^T}$, $\hat{z}_{ij}^o$, $\hat{z}_{ij}^m$의 식으로 표현할 수 있고 M-step에서는 이를 각 모수로 미분하여 $\Theta^{(t+1)}$를 다음과 같이 구할 수 있다.

$$\begin{aligned}\hat{\pi}_j^{(t+1)} &= \frac{1}{n}\sum_{i=1}^n \hat{z}_{ij},\\ \hat{\beta}_j^{(t+1)} &= \left(\sum_{i\in\mathcal{O}}\hat{z}_{ij}\boldsymbol{x}_i\boldsymbol{x}_i^T + \sum_{i\in\mathcal{M}}\hat{z}_{ij}\widehat{\boldsymbol{x}_i\boldsymbol{x}_i^T}\right)^{-1}\left(\sum_{i\in\mathcal{O}}\hat{z}_{ij}\boldsymbol{x}_iy_i + \sum_{i\in\mathcal{M}}\hat{z}_{ij}\hat{\boldsymbol{x}}_iy_i\right),\\ \hat{\sigma}_j^{2(t+1)} &= \frac{\sum_{i\in\mathcal{O}}\hat{z}_{ij}^o\left(y_i-\boldsymbol{x}_i^T\hat{\beta}_j^{(t)}\right)^2 + \sum_{i\in\mathcal{M}}\hat{z}_{ij}^m\left(y_i^2 - 2\hat{\boldsymbol{x}}_i^T\hat{\beta}_j^{(t)}y_i + \hat{\beta}_j^{(t)T}\widehat{\boldsymbol{x}_i\boldsymbol{x}_i^T}\hat{\beta}_j^{(t)}\right)}{\sum_{i\in\mathcal{O}}\hat{z}_{ij}^o + \sum_{i\in\mathcal{M}}\hat{z}_{ij}^m},\\ \hat{\mu}^{(t+1)} &= \frac{1}{n}\left(\sum_{i\in\mathcal{O}}\boldsymbol{x}_i + \sum_{i\in\mathcal{M}}\hat{\boldsymbol{x}}_i\right),\\ \hat{\Sigma}^{(t+1)} &= \frac{1}{n}\left(\sum_{i\in\mathcal{O}}\left(\boldsymbol{x}_i-\hat{\mu}^{(t)}\right)\left(\boldsymbol{x}_i-\hat{\mu}^{(t)}\right)^T + \sum_{i\in\mathcal{M}}\left(\widehat{\boldsymbol{x}_i\boldsymbol{x}_i^T} - 2\hat{\mu}^{(t)}\hat{\boldsymbol{x}}_i^T + \hat{\mu}^{(t)}\hat{\mu}^{(t)T}\right)\right).\end{aligned}$$

4. 모의 실험

본 장에서는 3장에서 유도된 EM 알고리즘을 이용하여 혼합회귀모형에서 공변량에 결측이 있는 경우에 대하여 모의 실험을 실시하였다. 모의실험에서는 공변량이 X_1, X_2 두 개인 경우에 대하여 다른 선형식을 가지는 두 집단이 있는 경우를 가정하였다. 첫 번째 모의 실험에서는 공변량에서의 결측이 X_1과 X_2에 동시에 일어나는 경우를 가정하였으며, 두 번째 모의실험은 결측이 오직 X_1에서만 발생하고, X_2는 항상 관측되는 경우로 설정하였다. 결측값을 생성하기 위하여 R을 결측이 있을 경우 1 그렇지 않은 경우 0을 나타내는 지시함수라고 정의하였을 때 결측이 일어날 확률은 다음의 모형으로 가정하였다.

$$P(R=1) = \frac{\exp(\alpha_0+\alpha_1 y)}{1+\exp(\alpha_0+\alpha_1 y)}. \tag{4.1}$$

이 결측메커니즘은 결측여부가 오직 항상 관측되는 반응변수 Y에 의해서 결정되므로 MAR을 만족하는 모형이 된다. 여기서 α_1은 결측확률이 관측된 다른 변수에 얼마나 의존하는지를 나타내는 값으로 $\alpha_1 = 0$일 경우 MCAR을 의미하게 된다. 본 모의 실험에서는 $\alpha_1 = 2$로 두었고 결측을 발생시켰다. α_0는 결측값이 발생하는 비율을 조정하는 모수로 본 모의실험에서는 결측비율이 30%와 50%가 되도록 α_0를 선택하였다. 모의 실험에서는 우리가 제안한 EM 알고리즘을 통하여 추정된 모수와 결측이 전혀 없는 관측된 자료로만으로 추정된 모수 비교하였다.

먼저 첫 번째 모의실험에서는 공변량 X_1과 X_2를 다음과 같은 다변량 정규분포로 부터 생성하였다.

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{x1}^2 & \rho \\ \rho & \sigma_{x2}^2 \end{pmatrix} \right).$$

이때 결측은 식 (4.1)의 확률로 X_1과 X_2에 동시에 일어난다고 가정하였다. 또한, 두 개의 선형회귀모형, $Y = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2 + \epsilon_j$, $j = 1, 2$를 이용하였으며 $\epsilon_j \sim N(0, \sigma_j^2)$을 가정하였다. 첫 번째 모의실험에서 이용한 모수의 참값으로는 $\sigma_{x1}^2 = \sigma_{x2}^2 = 1$, $\rho = 0.3$과

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \beta_{01} \\ \beta_{02} \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} \beta_{11} \\ \beta_{12} \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} \beta_{21} \\ \beta_{22} \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

을 사용하였으며 혼합비율은 $\pi_1 = 0.3$, $\pi_2 = 0.7$으로 두었다.

Table 4.1은 첫 번째 모의실험의 결과이며, CC는 결측이 있는 자료를 모두 제거하고 혼합회귀모형을 적합하여 얻은 결과이고 FD는 결측을 생성하기 전의 완전한 자료에 대하여 적합한 결과이며 MLE는 본 논문에서 제안한 방법에 의하여 얻어진 결과이다. 따라서 FD는 결측이 없는 경우에 비해 얼마나 제시한 추정방법이 효율적인지를 보기 위한 값이고 실제 상황에서는 구할 수 없는 값이다. 이 세 가지 방법으로 각각 100번의 반복시행을 통해 각 모수의 mean squared error(MSE)와 bias를 구하였으며 Table 4.1에서의 값들은 MSE에 1,000을 곱한 값들을 나타내며 괄호안의 값은 bias에 1,000을 곱한 값을 나타낸다. Table 4.1에서 CC와 MLE를 비교해 보면 MSE와 bias 모두에서 MLE가 CC에 비해 효율적인 추정을 함을 알 수 있다. 특히 missing rate가 증가함에 따라 CC의 경우 매우 큰 bias를 가짐에 반해 MLE는 여전히 신뢰할 수 있는 추정값을 주는 것을 알 수 있다.

두 번째 모의실험에서는 첫 번째 모의실험과 마찬가지로 식 (4.1)에 따라 결측을 발생시켰으나 오직 X_1에서만 결측을 발생시켰다. 두 번째 모의실험에서 이용한 모수의 참값은 $\sigma_{x1}^2 = \sigma_{x2}^2 = 1$, $\rho = 0.1$과

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \beta_{01} \\ \beta_{02} \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} \beta_{11} \\ \beta_{12} \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} \beta_{21} \\ \beta_{22} \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

을 사용하였으며 혼합비율은 $\pi_1 = 0.4$, $\pi_2 = 0.6$으로 두었다.

Table 4.2는 두 번째 모의실험에 대한 결과이다. 결과를 살펴보면 첫 번째 모의실험의 결과와 유사하게 동일한 30%의 결측 비율에서는 MLE 방법의 추정치들의 MSE 값이 대부분 CC 방법보다 작으며, FD 방법의 추정치와 비교했을 때도 큰 차이가 없으므로 좋은 성능을 보여주며, 동일한 결측비율에서 표본의 수를 250개에서 500개로 증가시키면, 전반적으로 MSE 값이 작아지는 것을 알수 있다. 또한, 결측 비율을 50%까지 증가시키면, MLE 방법과 CC 방법에서의 MSE 차이가 더욱 커짐을 알 수 있다.

5. 사례연구

사례연구에서는 UCI Machine Learning Repository(http://archive.ics.uci.edu/ml/)에서 제공하는 미국 위스콘신(Wisconsin) 의과대학의 유방암 진단 자료를 이용하여, MLE 방법이 실제 자료에서도 좋은 성능을 보이는지 알아보고자 한다. 자료의 수는 198개이며, 본 사례연구에서는 32개의 변수들 중 반응변수를 세포핵(cell nucleus)의 둘레(perimeter)로 두고 공변량을 세포핵의 부드러운 정도(smoothness)와 밀집된 정도(compactness)를 측정한 두 개의 변수로 설정하였다. 또한 자료에는 유방암의 재발여부에 대한 변수가 있는데 이를 잠재변수라고 간주하고, 성분(component)이 2개인 혼합회귀모형에 대해 모수의 추정을 실시하였다. 이 자료는 결측이 없는 완전한 자료이며, 제안한 방법의 성

Table 4.1. MSE×1000(Bias×1000) from the first simulation

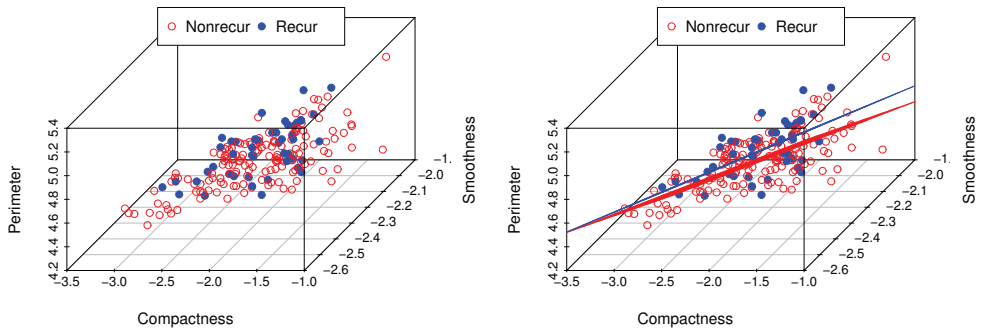
Missing rate(%)	n	Method	π	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}	σ_1^2	σ_2^2	μ_{x1}	μ_{x2}	σ_{x1}^2	σ_{x2}^2	ρ
30	250	FD	4.05	34.72	33.75	36.22	37.53	52.34	30.47	39.54	17.18	3.71	3.58	2.48	1.20	4.28
			(12.49)	(6.59)	(−14.17)	(12.45)	(−33.21)	(−32.76)	(−16.59)	(−34.05)	(−51.49)	(−13.30)	(−6.85)	(−2.48)	(−1.39)	(15.86)
		MLE	4.17	39.08	59.77	49.18	49.10	68.10	39.54	60.46	31.85	5.79	4.36	3.48	2.37	5.80
			(12.57)	(4.22)	(15.94)	(−6.81)	(−40.37)	(−28.92)	(−20.84)	(−69.61)	(−64.50)	(−14.63)	(−6.97)	(−2.49)	(−1.80)	(13.42)
		CC	33.64	165.36	47.04	133.48	474.93	162.37	499.30	92.78	72.84	9.18	69.85	30.75	77.30	21.77
			(−178.92)	(−341.52)	(−15.65)	(−247.55)	(−672.72)	(343.45)	(−681.36)	(−205.63)	(−248.23)	(−83.56)	(−260.78)	(−167.60)	(−275.15)	(−140.42)
	500	FD	1.62	16.49	15.88	21.47	17.17	18.84	18.51	13.24	6.41	1.85	1.81	0.80	1.01	2.36
			(−0.46)	(−22.52)	(11.24)	(20.19)	(−15.50)	(0.32)	(−10.94)	(−27.28)	(−7.55)	(−0.02)	(−3.36)	(−1.94)	(1.24)	(8.60)
		MLE	2.53	26.95	27.48	24.84	35.98	22.38	23.77	31.02	12.78	3.31	2.18	1.61	1.23	2.69
			(2.84)	(−36.93)	(2.98)	(−10.32)	(−32.49)	(5.98)	(−7.25)	(−16.18)	(−23.59)	(6.29)	(−0.61)	(0.34)	(4.41)	(12.13)
		CC	30.79	161.97	20.21	81.56	500.86	130.23	468.10	18.57	55.87	6.29	68.42	28.35	75.09	20.82
			(−173.57)	(−374.29)	(−51.68)	(−224.25)	(−696.49)	(338.81)	(−668.84)	(−140.40)	(−226.08)	(−68.86)	(−259.80)	(−165.29)	(−272.99)	(−141.55)
50	250	FD	3.91	37.31	37.29	40.89	38.85	52.64	34.98	43.09	18.45	4.20	3.23	1.97	1.46	4.86
			(12.59)	(16.34)	(−17.28)	(16.19)	(−28.32)	(−61.86)	(−6.75)	(−29.56)	(−56.13)	(−19.32)	(−2.25)	(0.96)	(−8.26)	(7.29)
		MLE	3.61	45.91	54.40	49.17	59.82	71.14	44.11	109.13	31.40	9.00	5.49	4.22	2.78	7.85
			(12.28)	(18.54)	(10.68)	(11.08)	(−48.99)	(−48.99)	(−15.70)	(−69.48)	(−71.45)	(−6.17)	(−0.37)	(3.20)	(−8.78)	(13.36)
		CC	52.36	486.40	76.11	158.02	2380.99	253.83	1141.83	91.51	183.71	8.36	93.35	97.59	174.27	37.22
			(−225.75)	(−596.96)	(−103.00)	(−286.47)	(−1530.01)	(445.73)	(−1046.99)	(−192.74)	(−411.41)	(−80.03)	(−302.99)	(−308.12)	(−415.67)	(−189.73)
	500	FD	2.16	15.72	18.37	27.82	15.18	21.03	20.42	16.56	6.29	2.15	2.04	1.13	0.94	1.90
			(−4.46)	(−5.31)	(20.06)	(20.97)	(−6.63)	(4.48)	(−22.57)	(−42.37)	(−10.75)	(−7.99)	(1.76)	(−3.88)	(−0.31)	(−1.31)
		MLE	3.27	55.49	50.19	32.14	41.95	46.21	32.41	77.70	19.45	5.40	3.16	2.16	1.88	3.56
			(6.51)	(−0.35)	(3.20)	(18.40)	(−12.29)	(−12.35)	(−21.90)	(−36.39)	(−28.46)	(−5.62)	(−1.87)	(−5.37)	(−2.20)	(−1.67)
		CC	51.93	475.69	39.45	158.85	2329.94	234.06	1144.83	19.74	148.76	6.29	94.18	99.03	168.66	37.80
			(−225.01)	(−647.42)	(−107.80)	(−296.32)	(−1520.58)	(449.91)	(−1053.52)	(−114.65)	(−374.79)	(−72.69)	(−305.88)	(−312.61)	(−409.75)	(−193.19)

Table 4.2. MSE × 1000(Bias × 1000) from the second simulation

Missing rate(%)	n	Method	π	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}	σ_1^2	σ_2^2	μ_{x1}	μ_{x2}	σ_{x1}^2	σ_{x2}^2	ρ
30	250	FD	13.46	51.62	29.12	29.60	21.90	23.80	21.00	28.48	13.50	3.65	3.51	2.56	1.21	3.98
			(19.17)	(−12.32)	(2.10)	(−20.89)	(21.78)	(21.46)	(−22.08)	(−57.29)	(−54.72)	(−12.93)	(−5.61)	(−4.25)	(−3.11)	(16.87)
		MLE	18.51	103.43	39.42	41.78	30.06	41.52	30.06	44.61	25.80	5.53	3.51	4.16	1.21	4.90
			(26.21)	(−10.95)	(15.24)	(−22.56)	(40.49)	(15.26)	(−25.25)	(−64.11)	(−76.34)	(−5.63)	(−5.61)	(−11.20)	(−3.11)	(20.80)
		CC	38.00	267.31	97.74	100.09	286.62	98.25	89.09	30.19	42.00	36.00	40.62	71.76	70.14	5.58
			(−179.19)	(−465.66)	(237.82)	(−243.08)	(−519.10)	(270.43)	(−270.03)	(−102.18)	(−183.87)	(183.45)	(−196.43)	(−264.28)	(−261.94)	(63.39)
	500	FD	4.94	18.98	15.33	15.99	11.77	7.97	6.96	14.14	4.95	1.84	1.80	0.84	1.06	2.12
			(−5.55)	(−32.92)	(−3.77)	(7.35)	(−3.88)	(−5.71)	(12.40)	(−43.91)	(−8.48)	(0.34)	(−3.45)	(−2.83)	(0.49)	(8.21)
		MLE	5.77	31.07	20.04	20.15	12.29	10.98	9.22	16.48	8.33	2.86	1.80	1.29	1.06	2.37
			(−9.38)	(−44.73)	(−3.95)	(−2.58)	(1.38)	(−7.77)	(17.77)	(−50.97)	(−10.04)	(1.33)	(−3.45)	(−3.95)	(0.49)	(5.50)
		CC	30.27	187.93	61.92	68.65	171.86	56.36	52.56	19.58	26.10	31.36	33.13	44.95	44.61	6.15
			(−158.40)	(−398.65)	(202.66)	(−220.65)	(−398.31)	(216.68)	(−211.86)	(−69.42)	(−141.77)	(173.36)	(−179.06)	(−210.12)	(−209.28)	(71.53)
50	250	FD	13.46	51.62	29.12	29.60	21.90	23.80	21.00	28.48	13.50	3.65	3.51	2.56	1.21	3.98
			(19.17)	(−12.32)	(2.10)	(−20.89)	(21.78)	(21.46)	(−22.08)	(−57.29)	(−54.72)	(−12.93)	(−5.61)	(−4.25)	(−3.11)	(16.87)
		MLE	18.95	107.32	33.90	48.77	47.81	41.83	37.70	38.26	54.97	13.18	3.51	6.16	1.21	6.00
			(33.44)	(−53.56)	(39.30)	(−54.29)	(40.44)	(−16.96)	(−32.98)	(−68.098)	(−131.60)	(−22.38)	(−5.61)	(−8.92)	(−3.11)	(8.04)
		CC	28.72	270.56	161.40	160.37	467.67	100.11	113.08	45.12	55.86	39.56	42.08	139.44	135.17	1.31
			(−117.75)	(−401.97)	(358.77)	(−354.05)	(−603.07)	(267.34)	(−293.85)	(−66.11)	(−213.75)	(194.78)	(−201.32)	(−371.17)	(−365.70)	(10.26)
	500	FD	4.94	18.98	15.33	15.99	11.77	7.97	6.96	14.14	4.95	1.84	1.80	0.84	1.06	2.12
			(−5.55)	(−32.92)	(−3.77)	(7.35)	(−3.88)	(−5.71)	(12.40)	(−43.91)	(−8.48)	(0.34)	(−3.45)	(−2.83)	(0.49)	(8.21)
		MLE	5.02	36.86	21.23	21.88	16.46	11.29	11.73	16.28	15.40	4.21	1.80	1.83	1.06	3.39
			(−7.21)	(−66.87)	(4.40)	(−14.10)	(12.04)	(−11.84)	(24.10)	(−60.13)	(−25.31)	(−1.43)	(−3.45)	(0.12)	(0.49)	(6.52)
		CC	60.15	445.66	77.88	87.03	604.50	89.20	86.91	34.16	32.28	41.35	43.94	96.38	97.48	2.33
			(−239.47)	(−638.07)	(222.90)	(−247.87)	(−768.12)	(281.12)	(−274.08)	(−136.77)	(−147.99)	(201.00)	(−207.39)	(−309.30)	(−310.91)	(38.63)

Table 5.1. Mechanism to generate 60 and 100 missing values

	Missing ID numbers	
	60 observations are missing	100 observations are missing
perimeter $\geq$ 150	All ID numbers	All ID numbers
120 $\leq$ perimeter $<$ 150	Multiples of 3	Even ID numbers
90 $\leq$ perimeter $<$ 120	Multiples of 5	Odd ID numbers
perimeter $<$ 90	85715, 844981, 855138, 855167, 856106, 863030, 927992, 937897	855138, 927992, 937897



(a) Scatter plot (b) Regression plane

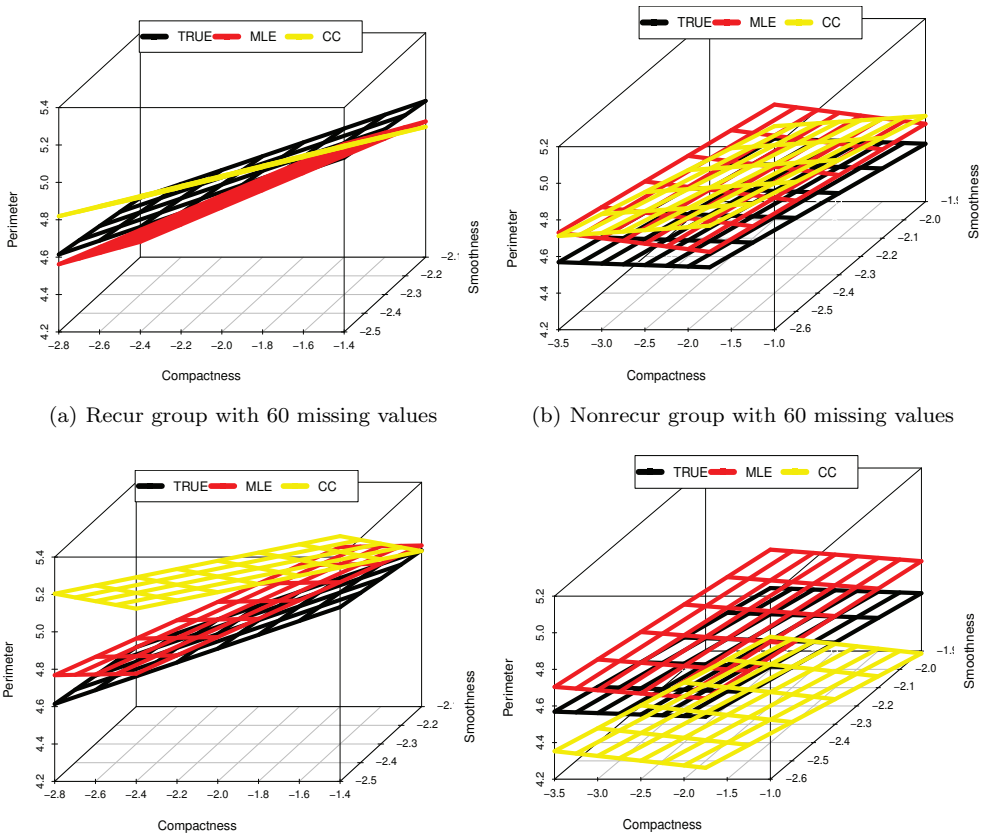
Figure 5.1. Scatter plot and regression plane.

Table 5.2. Parameter estimates from each method.

The number of missing values	Method	Parameter estimates													
		π	β_{01}	β_{11}	β_{21}	β_{02}	β_{12}	β_{22}	σ_1^2	σ_2^2	μ_{x1}	μ_{x2}	σ_{x1}^2	σ_{x2}^2	ρ
	FD	0.38	5.05	−0.24	0.37	2.81	−1.04	0.27	0.09	0.13	−2.28	−2.01	0.12	0.36	0.68
60	MLE	0.30	4.10	−0.70	0.46	2.72	−1.15	0.28	0.08	0.15	−2.29	−2.01	0.11	0.36	0.66
	CC	0.25	3.70	−0.75	0.27	3.16	−0.92	0.24	0.13	0.14	−1.59	−1.42	0.59	0.60	0.90
100	MLE	0.23	3.69	−0.98	0.49	2.92	−1.09	0.30	0.07	0.15	−2.28	−2.01	0.11	0.36	0.66
	CC	0.48	2.82	−1.15	0.22	2.29	−1.13	0.25	0.19	0.18	−1.12	−1.01	0.81	0.77	0.96

능을 알아보기 위해 모의 실험에서와 동일하게 MAR가정하에서 결측치 60개와 100개를 각각 생성하였다. Table (5.1)은 결측값을 생성한 방법과 생성된 결측치 번호를 보여주는 표이다. 또한, 자료가 정규분포를 따르지 않기 때문에 반응변수와 공변량에 로그변환을 한 다음에 분석을 실시하였다.

먼저 자료의 탐색을 위해 산점도를 그려보았으며 그 결과는 Figure 5.1(a)와 같다. 여기서 Recur는 유방암이 재발한 집단을 Nonrecur는 그렇지 않은 집단을 나타낸다. 산점도를 살펴보면 유방암이 재발한 집단과 재발하지 않은 집단 간의 뚜렷한 차이를 확인하기 어렵다. 따라서 두 집단에 대해 각각 회귀모형을 적용하여 비교해 보았으며 그 결과는 Figure 5.1(b)와 같다. 점선이 유방암이 재발한 집단에서의 회귀평면이고, 아래의 직선이 유방암이 재발하지 않은 집단에서의 회귀평면을 나타낸다. 이를통해 두 집단간에는 기울기 차이가 있음을 알수 있었다. 다음으로 FD, MLE, CC 방법을 통해 모수를 추정해 보았다. 그 결과는 Table 5.2이며, 결과를 살펴보면 MLE 방법이 CC 방법에 비해 FD의 추정치에 더욱 근접한 값을 가지는 것을 알 수 있다. 이로부터 MLE 방법이 실제 자료에서도 CC 방법에 비해 더 좋은 성능을 보여 준다는 것을 확인할 수 있었다.



(a) Recur group with 60 missing values

(b) Nonrecur group with 60 missing values

(c) Recur group with 100 missing values

(d) Nonrecur group with 100 missing values

Figure 5.2. Estimated regression plane for each group with different numbers of missing values.

다음으로 시각적인 비교를 위해 Figure 5.2에서는 유방암이 재발한 집단(Recur group)과 재발하지 않은 집단(Nonrecur group)으로 자료를 나누어 회귀분석을 적합시킨 실제 회귀 평면과 MLE 방법과 CC 방법으로 추정한 회귀 평면을 비교해 보았다. MLE로 구한 회귀 평면은 실제 회귀 평면과 약간의 차이는 있지만 상당히 유사하게 나타나는 반면 CC 방법의 경우에는 상당한 차이를 보인다. 또한, 결측값이 60개에서 100개로 증가하면 실제 회귀 평면과의 차이는 더욱 커진다는 것을 알 수 있다. 이와 더불어 두 추정회귀직선이 각각 참그룹을 얼마나 잘 반영하고 있는지 알아보기 위해 방법 별로 빈도표(confusion table)를 만들어 비교해 보았다. 그 결과 결측값이 60개인 경우와 100개인 경우 모두 MLE 방법이 CC 방법에 비해 참그룹을 더 잘 반영하고 있음을 알 수 있다.

6. 결론

기존의 연구에서는 결측 공변량을 포함한 혼합회귀모형에 대해 엄밀히 다룬 연구는 없었으며 본 논문에서는 공변량에 결측이 존재하는 경우에 공변량이 다변량 정규분포를 따른다는 가정하에서 EM 알고리

Table 5.3. Confusion table for the classification using FD method

		True group		
		Recur	Nonrecur	Total
Estimated group	Recur	23	54	77
	Nonrecur	24	97	121
	Total	47	151	198

Table 5.4. Confusion table from CC method with 60 missing values

		True group		
		Recur	Nonrecur	Total
Estimated group	Recur	5	30	35
	Nonrecur	29	74	103
	Total	34	104	138

Table 5.5. Confusion table from MLE method with 100 missing values

		True group		
		Recur	Nonrecur	Total
Estimated group	Recur	6	16	22
	Nonrecur	21	55	76
	Total	27	71	98

Table 5.6. Confusion table from CC method with 100 missing values

		True group		
		Recur	Nonrecur	Total
Estimated group	Recur	12	36	48
	Nonrecur	15	35	50
	Total	27	71	98

즘을 이용해 모수를 추정하는 방법에 대해 연구하였다. 일반적으로 결측된 자료가 많지 않은 경우에 있어서 혼합회귀분석을 시행할 때는 결측값을 포함하는 관측값 전체를 제외하고 모형적합을 하여도 큰 정보의 손실이 있지 않지만 모의실험에서 살펴본 바와 같이 결측된 관찰값들이 무시할 수 없을 정도로 많을 경우에는 이러한 모수추정 방법은 큰 편의를 보여주어서 신뢰할 수 있는 분석결과를 주지 못한다.

본 논문에서 다룬 혼합회귀모형은 MAR하에서 MRFC만을 다루었으나 CWMs와 ME의 경우도 비슷한 방법으로 EM 알고리즘을 유도할 수 있을 것이다. 또한, 비선형 회귀모형에 대해서도 EM 알고리즘이 유도될수 있을 것이다. 다만 이러한 경우에는 M-step이 계산가능하지 않은 형태가 되어 수치적 계산이 필요할 수 있다. 그외에 MAR이 아닌 결측 메커니즘하에서의 모수추정과 공변량의 분포가 다변량 정규분포가 아닌 경우에서의 모수추정 방법 등은 향후 보다 연구되어야 할 필요성이 있다.

References

Bandeen, R. K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes, *Journal of the American Statistical Association*, **92**, 1375–1386.

Benaglia, T., Chauveau, D., Hunter, D., and Young, D. (2009). mixtools: an R package for analyzing finite mixture models, *Journal of Statistical Software*, **32**, 1–29.

DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression, *Journal of Classification*, **5**, 249–282.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B (Methodological)*, **39**, 1–38.

Ingrassia, S., Minotti, S., and Vittadini, G. (2012). Local statistical modeling via the cluster-weighted approach with elliptical distributions, *Journal of Classification*, **29**, 363–401.

Ingrassia, S., Minotti, S., and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models, *Computational Statistics and Data Analysis*, **71**, 159–182.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression, *Journal of Classification*, **17**, 273–296.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts, *Neural Computation*, **3**, 79–87.

Leisch, F. (2004). FlexMix: a general framework for finite mixture models and latent glass regression in R, *Journal of Statistical Software*, **11**, 1–18.

Mclachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extension*, Wiley, New York.

Punzo, A. (2014). Flexible mixture modeling with the polynomial Gaussian cluster-weighted model, *Statistical Modelling*, **14**, 257–291.

Quandt, R. and Ramsey, J. (1978). Estimating mixtures of normal distributions and switching regressions, *Journal of the American Statistical Association*, **73**, 730–738.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, **26**, 195–239.

Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. (2013). Clustering and classification via cluster-weighted factor analyzers, *Advances in Data Analysis and Classification*, **7**, 5–40.

결측 공변량을 갖는 혼합회귀모형에서의 EM 알고리즘

김형민[a] · 함건희[b] · 서병태[a,1]

[a]성균관대학교 통계학과, [b]아산정책연구원

(2016년 8월 31일 접수, 2016년 10월 21일 수정, 2016년 10월 22일 채택)

요 약

혼합회귀모형은 반응 변수와 공변량 사이의 관계를 규명하는 유용한 통계적 모형으로 여러 분야에서 사용되어지고 있다. 하지만 실제로 혼합회귀모형을 이용하여 분석을 하는 과정에서 공변량이 결측값을 포함하는 문제는 흔하게 발생하며, 발생하는 결측의 유형 또한 다양하게 나타난다. 이러한 경우에 있어서 본 논문에서는 최대우도추정량을 구하기 위한 EM 알고리즘을 제안하고자 한다. 제안된 EM 알고리즘의 효용성을 모의실험을 통해 확인하였으며 또한 사례연구를 통해 제시된 방법이 어떻게 사용될수 있는지와 그 효용성을 함께 확인하였다.

주요용어: 혼합모형, 결측 공변량, 혼합회귀모형, EM 알고리즘

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 연구임(NRF-2013R1A1A2057715).

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: seobt@skku.edu