The Korean Journal of Applied Statistics (2019) DOI: http://dx.doi.org/10.5351/KJAS.2019.32.6.889
32(6), 889–898

Robust multiple imputation method for missings with boundary and outliers

Yousung Park[a] · Do Young Oh[a] · Tae Yeon Kwon[b,1]

[a]Department of Statistics, Korea University;
[b]Department of International Finance, Hankuk University of Foreign Studies

(Received October 2, 2019; Revised October 29, 2019; Accepted November 24, 2019)

Abstract

The problem of missing value imputation for variables in surveys that include item missing becomes complicated if outliers and logical boundary conditions between other survey items cannot be ignored. If there are outliers and boundaries in a variable including missing values, imputed values based on previous regression-based imputation methods are likely to be biased and not meet boundary conditions. In this paper, we approach these difficulties in imputation by combining various robust regression models and multiple imputation methods. Through a simulation study on various scenarios of outliers and boundaries, we find and discuss the optimal combination of robust regression and multiple imputation method.

Keywords: break-down point, robust regression, Bayesian multiple imputation

1. 서론

설문조사에서 발생한 항목 무응답(item missing)에 대한 대체(imputation)를 위하여 결측치를 포함한 변수에 대하여 충분한 설명력을 갖는 독립변수들이 존재한다면 회귀모형을 이용할 수 있다. 그러나 자료에 이상치(outlier)가 존재하는 경우 일반적인 회귀모형은 붕괴점(breakdown point)이 0%이기 때문에 회귀계수의 추정에 있어 편향(bias)의 문제가 발생하고 (Park 등, 2012) 이러한 편향된 회귀계수의 추정치는 대체값의 편향으로 이어질 수 있다.

또한 설문조사의 조사항목 중 결측치를 포함한 변수와 다른 변수들 간에 논리적 한계(logical boundary) 조건들이 존재하는 경우가 발생할 수 있다. 이러한 경우, 결측치 대체과정에서 이러한 한계조건들이 고려되어야 하기 때문에 추가적인 어려움이 따른다. 예를 들어, 건강 설문 조사에서 흡연기간은 흡연자의 나이보다 클 수 없으며 (Raghunathan 등, 2001), 가계소득조사에서 개인의 소득은 개인이 속한 가구의 소득보다 클 수 없고 (Schenker 등, 2006), 건강 관련 패널조사에서 청소년의 키는 그의 작년 키보다 작

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1C1B5043739), Korea University Grant (K1910711) and Hankuk University of Foreign Studies Research Fund.
[1]Corresponding author: Department of International Finance, Hankuk University of Foreign Studies, 81 Oedae-ro, Wangsan-ri, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Republic of Korea.
E-mail: tykwon@hufs.ac.kr

을 수 없다 (Geraci와 McLain, 2018). 이러한 결측치의 논리적 한계를 반영하기 위해 기존의 결측대체 방법에 기각/채택(rejection/acceptance) 절차를 추가하거나 새로운 대체방법을 적용해야 한다.

본 논문은 설문조사 자료에서 위와 같은 실질적 문제들이 발생하였을 때 이를 해결하기 위한 방법에 대하여 논의하고자 한다. 항목 무응답이 발생한 경우, 회귀모형에 기반을 두고 이들 결측치들을 대체를 함에 있어 이상치와 논리적 한계조건들이 자료에 존재하는 경우 다양한 로버스트 회귀모형과 베이지안 다중대체 방법의 조합을 통해 해결점을 모색하고자 한다.

먼저 이상치가 존재하는 경우 최소제곱법을 사용하는 일반적인 회귀모형(ordinary least square; OLS)에서 발생하는 추정된 모수의 편향을 해결하기 위한 방법으로 다양한 로버스트 회귀 추정량을 고려하였다. 본 논문에서는 높은 붕괴점을 갖는 MM-추정법 (Yohai, 1987), least trimmed square (LTS)-추정법 (Rousseeuw, 1984) 그리고 Park 등 (2012)의 data partition technique and M-estimation (DPM)-추정법을 기존의 OLS 대신 적용하였다.

두 번째로 결측치를 포함한 변수의 한계(boundary)가 있는 무응답 문제를 해결하기 위하여 기존의 회귀모형에 근거한 베이지안 다중대체 방법에 기각/채택 절차를 추가하는 방법을 먼저 적용하였다. 이를 위해 고려된 방법으로는 Little (1988)의 proportioned mean matching method (PMM) 그리고 Schenker과 Taylor (1996)의 local residual draw method (LRD)이다. Kwon과 Park (2015)은 한계가 있는 결측변수에 대하여 이러한 추가적인 절차 없이도 한계조건의 만족을 보장하는 새로운 다중대체방법인 boundary information matching proportioned residual draw method (BIM-PRD)을 제안하였다.

본 논문에서는 PMM, LRD, 그리고 BIM-PRD 방법들을 OLS 대신 앞서 제시된 로버스트 회귀모형 추정방법과 결합하여 새로운 결측치 대체방법으로 제안하고 이들의 최적의 조합을 찾고자 한다. 이를 위해 다양한 시나리오별로 생성된 자료에 대하여 모의실험을 실시하였다. 그 결과 로버스트 회귀모형의 추정 방법보다는 다중대체방법의 선택에 의해 성능차이가 크게 나타남을 확인하였고 DPM과 PMM 방법을 결합한 방법이 모든 시나리오에서 큰 차이 없이 가장 안정적인 성능을 나타내는 방법이지만 한계변수가 목적변수와 높은 상관관계를 가지고 있다면 PMM 대신 BIM-PRD 방법을 어떠한 로버스트 회귀모형 추정방법과 결합하여도 추정의 성능을 크게 향상시킬 수 있음을 확인하였다.

본 논문은 총 5장으로 구성되어 있다. 2장에서는 MM, LTS, 그리고 DPM 로버스트 회귀추정량들을 소개하고 3장에서는 PMM, LRD, 그리고 BIM-PRD 결측치 대체 방법들을 소개하였다. 4장에서는 모의실험을 통하여 이상치가 있고 결측치의 한계정보가 존재하는 경우, 로버스트 회귀모형과 결측치 대체방법의 조합들의 성능을 비교하였다. 마지막으로 5장은 결론으로 마무리 하였다.

2. 로버스트 회귀모형 추정법

로버스트 회귀추정량을 비교하는 척도인 붕괴점은 표본 중 어느 정도 부분의 관측치에 무한대 값을 넣었을 때 추정량의 발산하는지를 측정한 값이다. 예를 들어 OLS를 적합함에 있어서는 하나의 관측치만을 무한대로 보내도 회귀계수가 발산하기 때문에 OLS의 붕괴점이 0%이다. 본 논문에서는 높은 붕괴점을 갖는 MM 추정법 (Yohai, 1987), LTS 추정법 (Rousseeuw, 1984), 그리고 Park 등 (2012)의 DPM-추정법을 고려하였다. 이에 본 장에서는 다음의 회귀모형

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{2.1}$$

에서 MM, LTS, 그리고 DPM에 의한 회귀계수 추정법에 대해 간략하게 논의하고자 한다.

2.1. MM-추정량

먼저 MM-추정량은 M-추정량 (Huber, 1973)과 S-추정량 (Rousseeuw와 Yohai, 1984)을 결합한 형태로 S-추정법보다 더 효율적임이 알려져 있다 (Yohai, 1987). MM-추정량은 S-추정량 ($\hat{\beta}_S$)에 의해 계산된 표준화 잔차를 Tukey가 제안한 이중가중 목적함수 ($\rho(\cdot)$)에 적용하여 이를 최소화 하는 회귀계수 추정량(M-추정량)을 계산하는 2단계 과정으로 다음과 같이 구해진다.

$$\hat{\beta}_{\mathrm{MM}} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i\beta}{s_n}\right). \tag{2.2}$$

Tukey의 이중가중 목적함수의 조정상수로 1.548을 적용하였을 때 MM-추정량은 S-추정량과 마찬가지로 50%의 높은 붕괴점을 갖는다 (Stromberg, 1993). 본 논문에서는 R의 robustbase 라이브러리의 lmrob 함수를 이용하여 적합하였다 (Salibian-Barrera와 Yohai, 2006; Maronna와 Yohai, 2000).

2.2. LTS-추정량

LTS-추정량은 순위통계량의 선형결합을 기반으로 하며 다음과 같이 구해진다.

$$\hat{\beta}_{\mathrm{LTS}} = \arg\min_{\beta} \sum_{i=1}^{q} r_{(i)}\beta^2. \tag{2.3}$$

이때, $r_{(i)}(\beta) = y_{(i)} - x_{(i)}^T\beta$이고, $r_{(1)}(\beta)^2 \le r_{(2)}(\beta)^2 \le \cdots \le r_{(q)}(\beta)^2$, $q = [n(1-\alpha)+1]$, α는 절단비율이다. 만약 $q = n/2+1$이라면 붕괴점은 50%가 되며, $q = [n/2] + [(p+1)/2]$에서 최대 붕괴점 $[(n-p)/2]/n + 1/n$을 갖는다 (Rousseeuw, 1984). 이상치에 강건하지만 계산상의 문제로 일부 표본을 재표집(resampling)하는 FAST-LTS 알고리즘 (Rousseeuw와 Van Driessen, 2000)을 사용한다. 본 논문에서는 R-robustbase에서 제공하는 FAST-LRT알고리즘 함수 ltsReg 함수를 사용하여 적합하였다.

2.3. DPM-추정량

Park 등 (2012)은 LTS와 같은 기존의 로버스트 추정량이 재표집 방법을 사용한 것과 달리 전체자료를 작은 부분집합으로 분할한 후 각각의 부분집합에 OLS를 적용함으로써 이상치에 강건한 DPM-추정량을 새롭게 제안하였다. DPM-추정량은 다음과 같은 4단계에 걸쳐 구해진다.

- 단계 1. 자료를 예측변수와 설명변수에 의하여 $14(p+1)+1$개의 부분집합으로 분할한다. 이때 p는 독립변수의 수이다. 분할하는 방법은 다음과 같다.
  - 단계 1-1. 다음과 같은 4개의 부분집합을 만든다.

    $$\{i : y_i \ge y_{75}\}, \quad \{i : y_{50} \le y_i \le y_{75}\}, \quad \{i : y_{25} \le y_i \le y_{50}\}, \quad \{i : y_i < y_{25}\}.$$

    이때 y_{25}, y_{50}, 그리고 y_{75}는 y의 각각 25%, 50%, 그리고 75% 백분위 수이다. 위의 4개의 부분집합을 조합하여 전체데이터를 제외한 14개의 부분집합을 만든다.
  - 단계 1-2. 단계 1-1과 같이 4개의 부분집합을 만든다.

    $$\{i : y_i < \bar{y}_{lj}, x_{ij} < \bar{x}_j\}, \quad \{i : y_i \ge \bar{y}_{lj}, x_{ij} < \bar{x}_j\}, \quad \{i : y_i < \bar{y}_{uj}, x_{ij} \ge \bar{x}_j\}, \quad \{i : y_i \ge \bar{y}_{uj}, x_{ij} \ge \bar{x}_j\},$$

    이때 $\bar{x}_j$는 x_{ij}의 열 평균이고 $\bar{y}_{uj}$와 $\bar{y}_{lj}$는 각각 $x_{ij} \ge \bar{x}_j$와 $x_{ij} < \bar{x}_j$에서의 y의 평균이다. 단계 1-1과 마찬가지로 전체 자료를 제외한 부분집합을 만들면 $14p$개가 된다. 단계 1-1과 단계 1-2에서 만들어진 부분 집합과 전체데이터를 추가하면 $14(p+1)+1$개의 부분집합이 만들어진다.

- 단계 2. 다음과 같이 $\tilde{\beta}_n$을 정의한다.

$$\tilde{\beta}_n = \arg\min_{b_k} \sum_{|\tilde{r}_i(b_k)|<c} r_i^2(b_k) \tag{2.4}$$

  b_k는 각 부분집합에서의 OLS 추정량이며 $r_i(b_k) = y_i - x_i^T b_k$는 각 부분집합에 대한 잔차벡터이고, $\tilde{r}_i(b_k) = r(b_k)/s(b_k)$는 표준화 잔차이다. 이때, $s(b_k) = 0.6745^{-1}\text{median}(|r_i(b_k)|)$이다.
- 단계 3. $\tilde{\beta}_n$과 $s(\tilde{\beta}_n)$을 사용하여 M-추정을 하고 표준화 잔차를 구하여 표준화 잔차의 절대값이 고정된 상수 c보다 작은 관측치는 제거한다.
- 단계 4. 남은 자료들을 사용하여 단계 2와 단계 3을 다시 시행하고 이때 계산되는 회귀계수 $\hat{\beta}_{\text{DPM}}$으로 정의한다.

3. 베이지안 다중대체

결측치를 포함하는 변수 Y_i $(i = 1, \ldots, n)$는 관측치 Y_i^{obs} $(i = 1, \ldots, n_0)$와 결측치 Y_j^{mis} $(j = n_0 + 1, \ldots, n)$ 그룹으로 나눌 수 있다고 하자. 일반적인 회귀모형에 기반을 둔 베이지안 다중대체방법(Bayesian multiple imputation)의 시행은 아래의 사후분포에서 회귀계수 β^*와 σ^{*2}을 추출하는 것으로 시작된다.

$$\sigma^{*2} \sim \hat{\sigma}^2_{\text{OLS}} \frac{n_0 - p}{\chi^2_{n_0-1}}, \qquad \beta^* \sim N\left(\beta_{\text{OLS}}, \sigma^{*2}\left(X^T X\right)^{-1}\right). \tag{3.1}$$

이때 X는 결측치를 포함하지 않는 p개의 설명변수들의 행렬이고 $\hat{\beta}_{\text{OLS}}$, $\hat{\sigma}^2_{\text{OLS}}$는 각각 $Y^{\text{obs}}_{1:n_0}$로만 식 (2.1)에서의 회귀계수와 분산에 대한 OLS 추정량이다. 설명변수인 X행렬에는 결측이 없다고 가정하였기 때문에 모든 개체 $i = 1, \ldots, n_0, n_0+1, \ldots, n$에 대하여 추출된 β^*을 사용하여 $X_i^T\beta^*$ 즉 $\hat{Y}_i, Y$의 조건부 기댓값을 계산할 수 있다. 회귀계수와 분산을 식 (3.1)에 제시된 사후분포에서 M번 추출하면 M개의 조건부 기댓값을 계산할 수 있다. 이렇게 계산된 M개의 $\hat{Y}_i$에 근거하여 Y_j^{mis} $(j = n_0 + 1, \ldots, n)$에 대한 M번의 대체값을 찾는 베이지안 다중대체방법 중 본 논문에서는 PMM, LRD, 그리고 BIM-PRD 방법을 고려하였다. 위 세 방법은 각기 다른 방법으로 각 결측개에 대해 그와 유사성이 높다고 판단되는 도너집합(possible donor set)을 구성하고, 도너집합으로부터 M번 표집하여 결측치를 대체하는 핫덱 대체방법(hot-deck imputation)이 결합된 다중대체 방법으로 보다 일반적인 베이지안 다중 대체법에 대한 논의는 Rubin (1987)을 참조할 수 있다.

3.1. PMM

PMM (Little, 1988)은 $X_i^T\beta^*$ 즉 $\hat{Y}_i$을 모두 계산한 후, Y_j^{mis}에 대한 대체값으로 $\hat{Y}_j^{\text{mis}}$과 유사한 예측값 $\hat{Y}_i^{\text{obs}}$을 갖는 실제 관측치 Y_i^{obs}을 제안하는 방법이다. 즉 j번째 결측치 Y_j^{mis}를 대체하기 위하여 먼저 그의 예측치 $\hat{Y}_j^{\text{mis}}$와 모든 관측치의 예측치 $\hat{Y}_i^{\text{obs}}$ $(i = 1, \ldots, n_0)$간의 예측거리 δ_{ij}를 다음과 같이 계산한다.

$$\delta_{ij} = \hat{Y}_j^{\text{mis}} - \hat{Y}_i^{\text{obs}} \tag{3.2}$$

Y_i^{obs}들 중 δ_{ij}의 절대값의 크기가 작은 순서대로 정해진 크기의 도너집합을 구성한 후 이들 중 임의로 한 값을 추출하여 추출된 Y_i^{obs}로 Y_j^{mis}를 대체한다. 식 (3.1)에 제시된 β^*와 σ^{*2}의 추출부터의 모든 과정을 M번 반복하여 Y_j^{mis}에 대한 대체값을 각각 M개 즉 결측치가 대체된 M개의 자료(complete data set)을 구할 수 있다.

3.2. LRD

PMM이 결측치에 대한 예측치가 가장 가까운 예측치를 갖는 관측치 값 자체를 핫덱 대체하는 것이라면, LRD (Schenker와 Taylor, 1996)는 결측치에 대한 예측치가 가장 가까운 예측치 값을 갖는 관측치 값의 잔차를 핫덱 대체하는 방법이다. 잔차 즉, $e_i^{\text{obs}} = Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}}$들 중 식 (3.1)의 δ_{ij}의 절대값의 크기가 작은 순서대로 정해진 크기의 도너집합을 구성한 후 이들 중 임의로 한 값을 추출하여 추출된 $e_i^{*\text{obs}}$를 이용하여 다음과 같이 Y_j^{mis}를 다음의 $Y_j^{*\text{mis}}$로 대체한다.

$$Y_j^{*\text{mis}} = \hat{Y}_j^{\text{mis}} + e_i^{*\text{obs}} \tag{3.3}$$

식 (3.1)에 제시된 β^*와 σ^{*2}의 추출부터의 모든 과정을 M번 반복하여 Y_j^{mis}에 대한 대체값을 각각 M개 즉 결측치가 대체된 M개의 자료을 구할 수 있다.

3.3. BIM-PRD

앞서 제시된 PMM과 LRD 방법은 결측을 포함한 변수에 한계조건이 있는 경우, 결측치 대체 후 추가적인 기각/채택 절차가 필요하다. 즉 대체된 결측치 $Y_j^{*\text{mis}}$가 주어진 한계조건을 충족시키지 못한다면 버리고 한계조건을 충족시키는 대체치가 산출될 때까지 다시 추출하여야 한다. 이에 BIM-PRD 방법은 결측을 포함한 변수에 한계조건이 존재하는 경우 이를 결측치 대체과정에 포함하여 한계조건을 보장하는 대체치를 산출하도록 하는 방법으로 Park과 Kwon (2015)에 의해 제안되었다.

C_i는 모든 개체에 대해 알려진 값이고 결측을 포함하는 변수 Y_i값과 다음과 같은 관계에 있다고 가정하자.

$$Y_i \le C_i \tag{3.4}$$

BIM-PRD을 위해서는 먼저 다음과 같은 한계비례잔차(proportioned residual), $\tilde{r}_i$를 모든 관측치 $i = 1, \ldots, n_0$에 대해 구하고 이들을 $C_i - \hat{Y}_i^{\text{obs}}$의 부호에 따라 두 개의 집합 R^+와 R^-로 나눈다.

$$\tilde{r}_i = \frac{Y_i - \hat{Y}_i^{\text{obs}}}{C_i - \hat{Y}_i^{\text{obs}}}, \quad i = 1, \ldots, n_0, \tag{3.5}$$

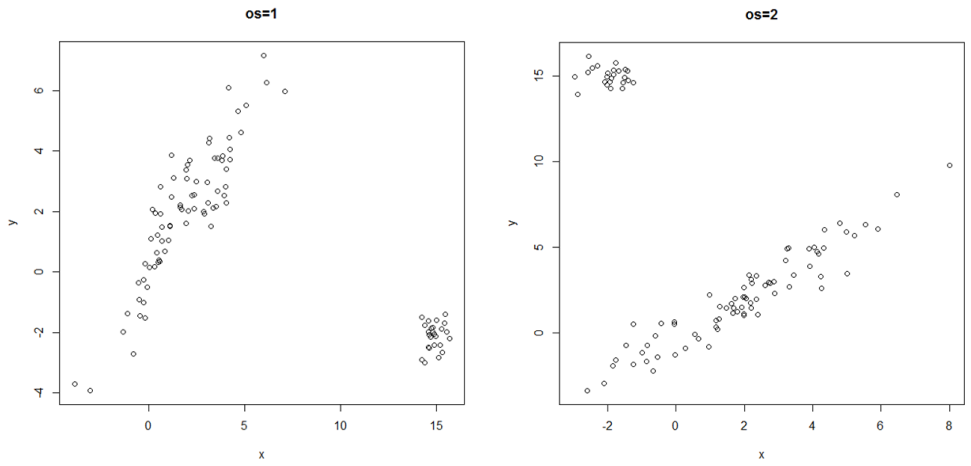
$$R^+ = \left\{ \tilde{r}_i | C_i - \hat{Y}_i^{\text{obs}} > 0 \right\}, \quad R^- = \left\{ \tilde{r}_i | C_i - \hat{Y}_i^{\text{obs}} < 0 \right\}. \tag{3.6}$$

만약 $C_j - \hat{Y}_j^{\text{mis}} > 0$이라면 $\tilde{r}_i \in R^+$들 중에서, 만약 $C_j - \hat{Y}_j^{\text{mis}} < 0$이라면 $\tilde{r}_i \in R^-$들 중에서, $|(C_j - \hat{Y}_j^{\text{obs}}) - (C_i - \hat{Y}_i^{\text{obs}})|$의 크기가 작은 순서대로 정해진 크기의 도너집합을 구성한 후 이들 중 임의로 한 값을 추출하여 추출된 $\tilde{r}_i^*$를 이용하여 다음과 같이 Y_j^{mis}를 다음의 $Y_j^{*\text{mis}}$로 대체한다.

$$Y_j^{*\text{mis}} = \hat{Y}_j^{\text{mis}} \tilde{r}_i^* \left(C_j - \hat{Y}_j^{\text{mis}} \right). \tag{3.7}$$

4. 모의실험

회귀모형 추정방법은 OLS, MM, LTS, 그리고 DPM-추정법을 그리고 결측치 대체방법으로는 PMM, LRD, 그리고 BIM-PRD 방법을 사용하여 이들 결합 중 어느 조합이 가장 우수한 결측치 대체 방법인지 다양한 자료의 시나리오별로 살펴보고자 한다. MM-추정량과 LTS-추정량은 각각 R에 내장되어 있는 robustbase 패키지의 lmrob, ltsReg함수를 사용하였다, DPM의 조정상수 $c = 2.5$로 설정하였다. MM-추정량과 DPM-추정량이 사용하는 M-추정량의 조정상수는 모두 동일하게 $c = 4.5$로 설정하였다.



(a) OS = 1 (b) OS = 2

Figure 4.1. Two outlier scenarios.

가상의 자료 생성은 아래의 모형으로부터 생성하였다.

$$Y_i = X_i + \epsilon_i, \quad Y_i < C_i \quad \text{for } i = 1, \ldots, n.$$

이때 X_i는 정규분포 $N(2, 2^2)$에서 생성하였다. $C_i = Y_i + |Z_i|$이고, 이때 Z_i는 정규분포 $N(0, \sigma_C^2)$에서 생성하였다. σ_C^2 값을 조정함으로써 이상치를 제외한 개체들에 대한 $\text{cor}(Y_i, C_i)$ 값을 약 0.6에서 0.9에 이르도록 조정하여 가상의 자료를 다양하게 생성하였다.

이상치를 제외한 개체들에 대한 $\text{cor}(Y_i, X_i)$값이 약 0.9가 되도록 오차항 ϵ_i의 분포는 $N(0, 1)$에서 생성하였다. X_i 혹은 Y_i 변수가 이상치인 경우에는 각각 $N(\mu_{x^{\text{out}}}, 0.5^2)$, $N(\mu_{y^{\text{out}}}, 0.5^2)$에서 생성하였다. 이때 $\mu_{x^{\text{out}}}, \mu_{y^{\text{out}}}$의 크기를 조정하여 다음의 두 가지 결측 시나리오(outlier scenario; OS)에 대한 이상치의 방향과 정도를 설정하였다.

$$\text{(a)} \quad \text{OS} = 1 : \mu_{x^{\text{out}}} = 15, \quad \mu_{y^{\text{out}}} = -2,$$
$$\text{(b)} \quad \text{OS} = 2 : \mu_{x^{\text{out}}} = -2, \quad \mu_{y^{\text{out}}} = 15$$

Figure 4.1에 두 가지 결측 시나리오에 따른 임의의 자료의 산점도를 제시하였다. 이상치의 비율(outlier rate; OR)은 10%, 25%, 그리고 40%의 경우에 대해 살펴보았다.

결측치는 주 모형에서만 발생시켰으며 결측비율은 30%로 하였으며 생성된 자료에 다음의 두 가지 형태의 결측 메커니즘에 따라 결측치를 생성하였다. 첫 번째는 j번째 Y변수의 결측 확률이 자료에 전혀 의존하지 않는 완전임의 결측(missing completely at random; MCAR)이며 두 번째 j번째 Y변수의 결측 확률이 X_j의 값에 의존하는 임의결측(missing at random; MAR)이다.

모의실험은 시나리오 별로 1,000번 반복하였고 다중대체에서의 대체횟수 $M = 5$번 그리고 도너의 사이즈는 5로 설정하였다 (Schenker와 Taylor, 1996). 대체방법을 표기할 때 적용한 회귀모형 추정방법을 앞에 대체방법을 뒷부분에 표기하였다. 예를 들면 DPM을 회귀모형 추정방법으로 결측치의 대체방법으로 BIM-PRD를 사용하였으면 DPM-BIM로 표기하였다.

Table 4.1. Simulation results: Root mean squared error (RMSE) when missing mechanism is MAR

$\text{cor}(Y_i, C_i)$	OS	OR	OBS	OLS LRD	OLS PMM	OLS BIM	MM LRD	MM PMM	MM BIM	LTS LRD	LTS PMM	LTS BIM	DPM LRD	DPM PMM	DPM BIM
0.9	1	10	0.323	0.065	0.064	0.067	0.067	0.064	0.050	0.067	0.064	0.049	0.067	0.064	0.049
		25	0.378	0.055	0.053	0.062	0.056	0.053	0.061	0.056	0.053	0.062	0.056	0.053	0.042
		40	0.351	0.045	0.043	0.054	0.045	0.043	0.054	0.046	0.044	0.054	0.046	0.043	0.040
	2	10	0.062	0.068	0.060	0.060	0.052	0.061	0.045	0.052	0.060	0.045	0.052	0.060	0.046
		25	0.431	0.079	0.051	0.050	0.045	0.051	0.042	0.045	0.052	0.042	0.046	0.052	0.042
		40	0.799	0.080	0.039	0.037	0.106	0.039	0.042	0.058	0.039	0.036	0.034	0.038	0.031
0.6	1	10	0.338	0.058	0.057	0.171	0.055	0.057	0.052	0.055	0.057	0.054	0.055	0.058	0.053
		25	0.367	0.046	0.045	0.149	0.046	0.044	0.139	0.046	0.045	0.136	0.043	0.045	0.044
		40	0.344	0.036	0.035	0.121	0.037	0.036	0.115	0.037	0.036	0.117	0.035	0.035	0.067
	2	10	0.066	0.059	0.055	0.100	0.053	0.055	0.054	0.052	0.056	0.054	0.052	0.056	0.054
		25	0.432	0.057	0.041	0.071	0.038	0.041	0.044	0.039	0.041	0.043	0.039	0.041	0.043
		40	0.795	0.062	0.037	0.067	0.077	0.036	0.120	0.046	0.036	0.057	0.033	0.037	0.036

OS = outlier scenario; OR = outlier rate; OBS = estimates with observed cases. In the case of other estimates, the applied robust regression estimation method (OLS, MM, LTS, or DPM) is shown at the beginning and the missing imputation method (LRD, PMM, or BIM) is shown at the end.

각 회귀모형에 기반을 둔 결측치 대체 방법에 대한 비교를 위하여 결측치 대체 이후 목적변수인 Y의 평균 추정치에 근거하여 $\bar{Y}$가 모수 μ_y를 추정하는데 있어 그 정확도(accuracy)와 효율성(efficiency)을 비교하였다. 이를 위해 평균제곱오차(root mean squared error; RMSE)와 95% 신뢰구간의 평균길이(average width of 95% confidence interval; AWCI)과 포함확률(coverage rate; CR)을 산출하여 Tables 4.1–4.3에 제시하였다. 이때 비교를 위하여 결측치들을 대체하지 않고 관측된 개체들만을 가지고 추정한 결과(OBS) 역시 제시하였다.

결측메커니즘이 MCAR인 경우의 결과는 관측된 개체들만을 가지고 추정한 결과(OBS)가 MAR에서 문제가 됨을 보이기 위함으로 공간의 제약 때문에 본문에는 제시하지 않았다. 결측 메커니즘이 MCAR이 아닌 MAR이기 때문에 관측치들 만으로 평균을 추정하면(OBS) 추정의 정확성이 좋지 않음을 확인할 수 있다. 이 경우 가장 넓은 95% 신뢰구간의 평균길이를 보임에도 불구하고 신뢰구간의 모수포함 확률이 40%에 미치지 못하는 경우도 발생함을 볼 수 있다. OBS를 제외하고 Table 4.2와 Table 4.3에 제시된 CR과 AWCI에는 로버스트 추정 및 대체 방법들 간 큰 차이가 확인되지 않기에 앞으로의 해석은 Table 4.1에 제시된 RMSE에 근거하여 하도록 하겠다.

로버스트 회귀모형의 추정방법에 따른 RMSE의 차이는 OS = 2인 경우 확연하게 나타났다. 이는 Y의 평균추정에 있어 추정된 회귀선의 위치가 미치는 영향력이 OS = 1의 경우보다 Y축 방향에서 큰 이상치들이 발생하는 OS = 2의 경우에 크기 때문인 것으로 판단된다. 로버스트 회귀모형의 추정방법에 따른 성능은 MM-LTS-DPM 순으로 나타나며 이는 붕괴점이 우수한 순서와 동일하다. 전체 자료에서 이상치들이 차지하는 비율이 10% 그리고 25%까지는 세 가지 로버스트 회귀모형의 추정방법간의 성능차이가 거의 없었다. 그러나 이상치의 비율이 40%에 다다르게 되면 MM 방법은 큰 문제를 나타냈다. 특히 MM-LRD 방법은 OLS 방법보다도 더 큰 RMSE를 보이기도 하였다.

결측치를 포함하는 목적변수인 Y변수와 그들의 한계변수인 C와의 상관계수가 높은 경우(즉 두변수간 상관관계가 0.9인 경우), 결측치 대체과정에서 한계변수를 고려하는 BIM-PRD 방법이 전반적으로 우수함을 확인할 수 있었다. 특히 로버스트 회귀모형의 추정방법 중 성능이 가장 우수한 DPM 방법과 함께 사용된 BIM-PRD 방법이 가장 작은 RMSE를 나타냈다. 또한 DPM 방법과 함께 사용하였을 때 결

Table 4.2. Simulation results: Coverage rate (CR) of 95% confidence interval when missing mechanism is MAR

$\text{cor}(Y_i, C_i)$	OS	OR	OBS	OLS LRD	OLS PMM	OLS BIM	MM LRD	MM PMM	MM BIM	LTS LRD	LTS PMM	LTS BIM	DPM LRD	DPM PMM	DPM BIM
		10	52.1	94.2	94.4	95.4	96.2	94.1	97.6	96.0	94.3	97.8	96.3	94.7	97.7
0.9	1	25	42.0	96.7	97.0	95.4	96.4	96.9	94.9	96.3	96.9	95.5	97.5	96.9	98.8
		40	39.6	96.8	97.0	95.8	96.8	97.2	96.7	96.2	96.9	96.3	97.2	97.0	98.1
		10	50.3	94.1	94.7	76.7	96.6	94.0	96.9	96.3	94.4	96.1	95.8	94.2	96.8
0.6	1	25	69.2	95.5	93.4	95.7	95.7	93.9	95.9	95.8	94.5	95.7	95.1	93.9	95.5
		40	39.4	98.4	98.5	85.6	98.3	98.5	86.8	98.3	98.6	85.3	98.6	98.3	94.6

OS = outlier scenario; OR = outlier rate; OBS = estimates with observed cases. In the case of other estimates, the applied robust regression estimation method (OLS, MM, LTS, or DPM) is shown at the beginning and the missing imputation method (LRD, PMM, or BIM) is shown at the end.

Table 4.3. Simulation results: Average width of 95% confidence interval (AWCI) when missing mechanism is MAR

$\text{cor}(Y_i, C_i)$	OS	OR	OBS	OLS LRD	OLS PMM	OLS BIM	MM LRD	MM PMM	MM BIM	LTS LRD	LTS PMM	LTS BIM	DPM LRD	DPM PMM	DPM BIM
		10	1.107	0.935	0.935	0.998	1.009	0.935	0.984	1.011	0.935	0.985	1.013	0.935	0.985
0.9	1	25	1.107	0.990	0.992	1.039	0.996	0.992	1.036	0.994	0.992	1.037	1.053	0.993	1.043
		40	1.049	0.996	0.999	1.031	0.995	0.999	1.030	0.996	0.999	1.032	1.032	0.999	1.041
		10	1.107	0.951	0.952	1.063	0.987	0.952	0.994	0.989	0.952	0.995	0.988	0.953	0.996
0.6	1	25	1.115	1.009	1.011	1.060	1.011	1.012	1.054	1.011	1.012	1.060	1.045	1.011	1.060
		40	1.049	1.013	1.015	1.019	1.012	1.014	1.018	1.011	1.014	1.017	1.030	1.015	1.043

OS = outlier scenario; OR = outlier rate; OBS = estimates with observed cases. In the case of other estimates, the applied robust regression estimation method (OLS, MM, LTS, or DPM) is shown at the beginning and the missing imputation method (LRD, PMM, or BIM) is shown at the end.

측치 대체 방법들 간에 (LRD, PMM, 그리고 BIM-PRD) 그리고 이상치 비율 및 이상치 위치 시나리오에 따른 RMSE 차이가 크게 변화 없이 그 성능이 유지됨을 확인할 수 있었다. 그러나 Y변수와 한계변수 C간의 상관관계가 낮은 경우(두 변수 간 상관관계가 0.6인 경우) 그리고 높은 이상치 비율(25%와 40%)문제가 함께 존재하는 경우에는 BIM-PRD 방법은 PMM과 LRD 방법에 비하여 그 정확도가 현저히 떨어짐을 확인 할 수 있었다.

5. 결론

본 논문은 회귀모형에 기반을 두고 결측치들을 대체를 함에 있어 이상치와 논리적 한계조건이 자료에 모두 존재하는 경우, 다양한 로버스트 회귀모형과 다중대체 방법의 최적의 조합을 다양한 시나리오별로 모의실험을 통하여 찾아보고 이에 대하여 논의하였다.

모의실험 결과에 따라 이상치와 한계가 있는 변수의 회귀모형에 근거한 대체를 함에 있어 다음과 같은 결론을 내릴 수 있다. 첫째, 이상치의 비율이 큰 경우 로버스트 방법의 선택이 중요하다. 그리고 한계치가 목적변수와 상관관계가 아주 높은 경우에만 결측치 대체를 위한 BIM-PRD이 성능이 좋으며 그렇지 않은 경우는 주의하여 사용하여야 한다.

둘째, 로버스트 방법의 선택에 비하여 대체방법의 선택에 따른 추정결과의 성능차이가 두드러진다. 가장 안정적인 성능을 나타내는 대체방법은 PMM 방법이다. PMM 방법과 결합되어 사용된 추정 결과들이 가장 낮은 RMSE를 보이지는 않으나 모든 경우 가장 안정적이고 크지 않은 RMSE를 보인다. 모든

경우 OLS-PMM과 MM-PMM, LTS-PMM, DPM-PMM간에 큰 성능차이가 나타나지 않음을 확인할 수 있었다.

마지막으로 DPM과 PMM 방법을 결합한 방법이 모든 시나리오에서 큰 차이 없이 가장 안정적인 성능을 나타내는 방법이지만 한계변수가 목적변수와 높은 상관관계를 가지고 있다면 PMM대신 BIM-PRD 방법을 어떠한 로버스트 회귀모형 추정방법과 결합하여도 추정의 성능을 크게 향상시킬 수 있다.

설문조사의 조사 항목 중 결측치를 포함한 변수와 그렇지 않은 다른 변수들 간에 논리적 한계 관계가 존재하는 경우가 발생할 수 있다. 패널조사의 경우 조사 내의 논리적 한계관계 뿐 아니라 조사 간의 논리적 한계관계 까지 더해져 문제는 더욱 복잡해 질 수 있다. 본 논문에 제시된 결측치의 로버스트 회귀 대체방법을 적용한다면 흡연자의 연령을 넘을 수 없는 흡연기간의 결측치 (Raghunathan, 등, 2001), 가구의 소득을 넘을 수 없는 개인의 소득의 결측치 (Schenker 등, 2006) 그리고 증가 혹은 유지되어야만 하는 청소년 성장(키)에 관한 패널 조사에서의 결측치 (Geraci와 McLain, 2018)에 대한 대체를 보다 정확하고 효과적으로 실행할 수 있을 것으로 기대하며 이를 후속 연구로 진행할 예정이다.

References

Geraci, M. and McLain, A. (2018). Multiple imputation for bounded variables, *Psychometrika*, **83**, 919–940.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo, *The Annals of Statistics*, **1**, 799–821.

Kwon, T. Y. and Park, Y. (2015). A new multiple imputation method for bounded missing values, *Statistics & Probability Letters*, **107**, 204–209.

Little, R. J. (1988). Missing-data adjustments in large surveys, *Journal of Business & Economic Statistics*, **6**, 287–296.

Maronna, R. A. and Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors, *Journal of Statistical Planning and Inference*, **89**, 197–214.

Park, Y., Kim, D., and Kim, S. (2012). Robust regression using data partitioning and M-estimation, *Communications in Statistics-Simulation and Computation*, **41**, 1282–1300.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, **27**, 85–96.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (pp. 256–272), Springer, New York.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P. J. and Van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. In *Data Analysis* (pp. 335–346), Springer, Berlin, Heidelberg.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys* (Vol. 81), John Wiley & Sons.

Salibian-Barrera, M. and Yohai, V. J. (2006). A fast algorithm for S-regression estimates, *Journal of computational and Graphical Statistics*, **15**, 414–427.

Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey, *Journal of the American Statistical Association*, **101**, 924–933.

Schenker, N. and Taylor, J. M. (1996). Partially parametric techniques for multiple imputation, *Computational Statistics & Data Analysis*, **22**, 425–446.

Stromberg, A. J. (1993). Computation of high breakdown nonlinear regression parameters, *Journal of the American Statistical Association*, **88**, 237–244.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression, *The Annals of Statistics*, **15**, 642—656.

한계와 이상치가 있는 결측치의 로버스트 다중대체 방법

박유성[a] · 오도영[a] · 권태연[b,1]

[a]고려대학교 통계학과, [b]한국외국어대학교 국제금융학과

(2019년 10월 2일 접수, 2019년 10월 29일 수정, 2019년 11월 24일 채택)

요 약

항목 무응답(item missing)이 발생한 설문조사에서 결측이 포함된 변수에 이상치(outlier)의 존재와 다른 설문문항 항목과의 논리적 한계(boundary) 조건들이 유의미하다면 결측치 대체문제는 매우 복잡해진다. 한계가 있는 결측값들을 포함한 변수에 이상치가 존재하는 경우, 기존의 회귀분석에 근거한 결측치 대체방법은 편향된 대체값 그리고 한계를 만족하지 않은 대체값을 제시할 가능성이 있다. 이에 본 논문은 회귀모형에 기반을 두고 결측치들을 대체를 함에 있어 이상치와 논리적 한계조건이 자료에 존재하는 경우, 다양한 로버스트 회귀모형과 다중대체 방법의 조합을 통해 해결점을 모색하고자 한다. 이를 위해 이들 방법들의 최적의 조합을 다양한 시나리오별로 모의실험을 통하여 찾아보고 이에 대하여 논의하였다.

주요용어: 붕괴점, 로버스트 회귀분석, 베이지안 다중대체

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1C1B5043739), Korea University Grant (K1910711) and Hankuk University of Foreign Studies Research Fund.

[1](17035) 경기도 용인시 처인구 외대로31, 한국외국어대학교 국제금융학과. E-mail: tykwon@hufs.ac.kr