The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.7.1229
27(7), 1229–1241

A Multiple Imputation for Reducing Outlier Effect

Man-Gyeom Kim[a] · Key-Il Shin[a,1]

[a]Department of statistics, Hankuk University of Foreign Studies

(Received October 15, 2014; Revised December 05, 2014; Accepted December 05, 2014)

Abstract

Most of sampling surveys have outliers and non-response missing values simultaneously. In that case, due to the effect of outliers, the result of imputation is not good enough to meet a given precision. To overcome this situation, outlier treatment should be conducted before imputation. In this paper in order for reducing the effect of outlier, we study outlier imputation methods and outlier weight adjustment methods. For the outlier detection, the method suggested by She and Owen (2011) is used. A small simulation study is conducted and for real data analysis, Monthly Labor Statistic and Briquette Consumption Survey Data are used.

Keywords: Outlier detection, outlier imputation, outlier weight adjustment, penalized regression.

1. 서론

표본조사에서는 다양한 원인으로 비표본오차가 발생하고 있으며 이 중에서 항목 무응답과 자료에 발생하는 이상점은 비표본오차에 매우 큰 영향을 준다. 먼저 무응답으로 인해 발생한 결측자료는 자료의 분포를 변형시키고, 추정에 편향을 발생시켜 전체적인 비표본오차에 영향을 준다. 또한 이상점(outlier)은 비표본오차를 증가시키는 또 다른 중요한 요인으로 이상점을 적절히 처리하지 않고 분석할 경우 모수 추정 결과는 과대 또는 과소 추정될 수 있다. 따라서 표본조사의 정확성을 향상시키기 위해서는 이상점을 적절히 처리하고, 무응답 대체를 실시하여야 한다. 실제 표본조사에서는 이상점과 무응답이 동시에 발생하는 것이 일반적이다.

무응답으로 발생한 결측을 해결하기 위해 여러 통계적 기법이 개발되었으며 이 중에서 흔히 사용하는 방법이 대체법(imputation) 과 가중치 조정법(weight adjustment)이다. 특히 대체법의 성능은 결측을 제외한 완전자료(complete data)에 존재하는 이상점(outlier)에 큰 영향을 받게 된다. 물론 대체법 대신에 가중치 조정법을 사용하는 경우에도 이상점의 존재 유무에 따라 최종 가중치는 변하게 된다. 따라서 대체법 또는 가중치 조정법을 사용하기 전에 이상점은 적절히 처리되어야 한다.

이상점은 대표성이 있는 이상점(representative outlier)과 대표성이 없는 이상점(non-representative outlier)으로 나누어진다. 대표성이 없는 이상점은 값 자체가 잘못된 것이다. 따라서 재조사를 통해 자료를 바로 잡거나, 내용 검토를 통해 올바른 값으로 수정할 수 있다. 반면 대표성 있는 이상점은 그 값 자체가 참값이어서 재조사나 내용 검토를 통해서도 수정할 수 없는 값이다. 따라서 본 연구에서는 대표

The research was supported by Hankuk University of Foreign Studies research fund (2014).
[1]Corresponding author: Professor, Department of statistics, Hankuk University of Foreign Studies, Yongin, Gyeonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr

성 있는 이상점만을 다룬다. 결국 이상점의 경우에는 잘못된 값이 아니고 충분히 모집단에 있을 수 있는 참값이기 때문에 이상점이라 하여 무조건 제거하는 것은 좋은 방법이 아닐 수 있다. 이상점을 제거하는 대신에 이상점을 대체할 수 있으며 Ren과 Chamber (2004)는 이상점 대체법에 관하여 연구하였다.
이상점을 처리하기 위해서는 먼저 이상점을 탐지하여야 한다. 여러 이상점 탐지 방법이 제안되었으며 이때 탐지에 사용할 정보, 보조변수의 유무 등에 따라 이상점 탐지 방법은 달라진다. 보조변수는 전 시점에서 얻어질 수도 있고, 현 시점에서 얻어질 수도 있다. 전 시점의 자료를 이용한 이상점 탑지법이 제안되었으며 이에 관한 내용은 Hidiroglou과 Berthelot (1986)과 Belcher (2003)을 참고하기 바란다. 또한 회귀분석의 외표준화잔차를 이용하여 이상점을 탐지하는 방법이 있으며 이에 관한 내용은 McCullough과 Pennington (2009)를 참조하기 바란다. 위에서 설명한 두 내용은 Kim과 Shin (2013)에도 설명되었다. 본 연구에서는 최근에 제안된 방법인 벌점회귀법을 이용한 이상점 탐지법을 이용하여 이상점을 탐지하였다. 이에 관한 내용은 Park 등 (2013) 과 She과 Owen (2011)을 참조하기 바란다.
본 연구에서는 모의실험을 이용하여 임의의 자료를 생성한 후 자료 값에 발생하는 이상점의 대체법과 이상점 가중치 보정법을 이용하여 이상점을 적절히 처리한 후 항목 무응답 처리를 위한 대체법에 따른 다중 대체법의 성능을 비교하였다. 이때 무응답은 MAR(missing at random)을 따른다고 가정하였으며 이에 관한 내용은 Rubin (1987)을 살펴보기 바란다. SAS/MI를 이용하여 다중 대체가 이루어졌고 또한 실제 자료분석이 수행되었으며 이를 통하여 모의실험에서 얻어진 결과의 타당성을 확인하였다. 실제 자료분석에는 매월노동통계와 연탄소비실태조사 자료가 사용되었다.
논문의 구성은 다음과 같다. 2절에서는 이상점을 탐지하는 방법으로 회귀모형을 이용하는 방법과 전 시점 자료만을 이용하는 방법인 Hidiroglou-Berthelot 방법을 설명하였다. 또한 She과 Owen (2011)이 제안한 $\boldsymbol{\Theta}$-IPOD(iterative procedure for outlier detection) 방법을 설명하였다. 3절에서는 이상점의 영향력을 줄이기 위한 처리 방법으로 이상점 가중치 보정방법과 이상점 대체법을 설명하였으며 4절에서는 다중 대체법을 설명하였다. 5절에서는 모의실험을 통하여 결측값 대체를 위한 이상점 처리 효과를 살펴보았으며 6절에서는 실제 자료분석이 수행되었다. 7절에 결론이 있다.

2. 이상점 탐지

이상점 탐지법 중에서 회귀분석의 외표준화잔차(studentized deleted residual)를 이용하는 방법과 Hidiroglou와 Berthelot (1986)이 제안한 방법을 간단히 설명하였다. 이에 관한 자세한 내용은 Kim과 Shin (2013)을 살펴보기 바란다. 또한 She과 Owen (2011)의 이상점 탐지 방법을 간단히 설명하였으며 본 연구에서는 이상점 탐지 방법으로 She과 Owen (2011) 방법을 사용하였다.

2.1. 외표준화잔차(externally strudentized residual, studentized deleted residual)

특정한 $\boldsymbol{i}$번째 자료가 이상점일 경우 잔차 $\boldsymbol{r}_i$가 커지게 되지만 동시에 $\boldsymbol{i}$ 번째 자료가 분산 추정에도 영향을 미쳐 분산이 커지게 된다. 따라서 외표준화잔차를 사용하여 이상점을 찾아내는 것은 매우 타당하다. 다음이 외표준화잔차의 정의이다.

$$\boldsymbol{t}_i = \frac{\boldsymbol{r}_i}{\boldsymbol{s(d_i)}} = \frac{\boldsymbol{r}_i}{\boldsymbol{s(i)}\sqrt{\boldsymbol{(1-h_{ii})}}}.$$

여기서 $\boldsymbol{r}_i = \boldsymbol{y}_i - \hat{\boldsymbol{y}_{(i)}}, \boldsymbol{y}_i$는 관측값 그리고 $\hat{\boldsymbol{y}_{(i)}}$는 $\boldsymbol{i}$번째 관측값을 제거한 후에 얻어진 예측값을 의미한다. 또한 $\boldsymbol{h_{ii}}$는 지렛값 또는 레버리지이다. $\boldsymbol{t}_i$의 분포는 우리가 잘 알고 있는 자유도 $\boldsymbol{(n-p-1)}$인 $\boldsymbol{t}$분포를 따르는 것으로 알려져 있으며 외표준화잔차는 SAS/PROC REG의 출력결과에서 쉽게 얻을 수 있다.

2.2. Hidiroglou-Berthelot 방법

이 방법은 전 시점 자료와 현 시점 자료의 비를 이용하여 이상점을 탐지한다. 먼저 $\boldsymbol{i}$번째 자료의 $\boldsymbol{t-1}$ 시점 자료를 $\boldsymbol{x_i(t-1)}$라 하고 $\boldsymbol{t}$시점 자료를 $\boldsymbol{x_i(t)}$라 할 때 두 자료의 비 $\boldsymbol{r_i} = \frac{\boldsymbol{x_i(t)}}{\boldsymbol{x_i(t-1)}}$을 구한다. 다음으로 계산의 편리를 위하여 변환된 자료인 $\boldsymbol{s_i}$를 다음과 같이 구한다.

$$\boldsymbol{s_i} = \begin{cases} \boldsymbol{1 - \frac{r_M}{r_i}}, & \boldsymbol{0 < r_i < r_M}, \\ \boldsymbol{\frac{r_i}{r_M} - 1}, & \boldsymbol{r_M \leq r_i}. \end{cases}$$

여기서 $\boldsymbol{r_M}$은 $\boldsymbol{r_i}$의 중앙값이다. 다음으로 자료의 크기를 반영하기 위해 다음과 같은 $\boldsymbol{E_i}$를 구한다.

$$\boldsymbol{E_i = s_i \times \left\{ max(x_i(t), x_i(t-1)) \right\}^U}.$$

이제 $\boldsymbol{E_M}$을 $\boldsymbol{E_i}$의 중앙값이라 하고, $\boldsymbol{d_{Q1} = E_M - E_{Q1}}$, $\boldsymbol{d_{Q3} = E_{Q3} - E_M}$ 그리고 $\boldsymbol{E_{Q1}}$, $\boldsymbol{E_{Q3}}$를 각각 제 1사분위수와 제 3사분위수라 하면 탐지 구간은 다음과 같다.

$$\boldsymbol{(E_M - Cd_{Q1}, E_M + Cd_{Q3})}.$$

여기서 $\boldsymbol{U}$는 자료값의 영향력을 나타내는 모수이고, $\boldsymbol{C}$는 이상점 탐지 구간의 길이를 결정하는 모수가 된다. 적당한 $\boldsymbol{C}$와 $\boldsymbol{U}$를 사용하여 이상점을 탐지한다.

2.3. Θ-IPOD(iterative procedure of outlier detection)

She와 Owen (2011)이 제안한 비볼록(nonconvex) 노음(norm) 벌점함수를 이용한 이상점 탐지 방법은 가면효과(masking effect)와 수렁효과(swamping effect)의 영향을 받지 않는 우수한 방법으로 알려져 있다. 이 방법을 간단히 설명하면 다음과 같다. 먼저 흔히 사용하는 선형 회귀모형에 추가적으로 관측치마다 개별 절편을 갖는 다음의 모형을 고려하자.

$$\boldsymbol{y_i = \beta_0 + x_i{}^T \beta + \gamma_i + \epsilon_i, i = 1, \cdots, n}.$$

여기서 $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$, $\boldsymbol{x_i} = (x_{1i}, x_{2i}, \cdots, x_{pi})^T$이다. 이 식에서 모든 i에 대해 $\boldsymbol{\gamma_i} = 0$이면 기존의 선형 회귀모형과 동일해진다. 만약 특정 $\boldsymbol{i}$ 의 $\boldsymbol{\gamma_i}$값이 "0"이 아니라면 이는 $\boldsymbol{x_i}$에서의 조건부 평균값과 $\boldsymbol{\gamma_i}$만큼 차이가 난다는 것을 의미하고 이는 회귀분석에서 이상치의 일반적 정의와 동일하게 된다. 따라서 이 모형은 관측치마다 이상치 여부를 탐색하고 또한 추정할 수 있는 모형이다. 또한 복수개의 $\boldsymbol{\gamma_i}$ 값이 "0"이 아닌 것도 판단할 수 있으며 이는 여러 개의 이상치를 동시에 탐지할 수 있다는 것을 의미한다. 이상점 검출 및 효과가 우수한 반면에 추정될 모수의 개수는 $\boldsymbol{n + p}$로 자료의 수 $\boldsymbol{n}$보다 항상 크게 되어 축소추정방법이 사용된다. 이 방법에서는 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_n)$에 대한 비볼록(nonconvex) 노음 벌점함수를 사용한다. 즉 $\boldsymbol{\Theta}$-IPOD 추정량은 다음에 정의되는 목적 함수를 최소로 하는 해 $\boldsymbol{(\beta, \gamma)}$ 로 정의된다.

$$\boldsymbol{L(\beta, \gamma) = \sum_{i=1}^{n} \left(y_i - x_i{}^T \beta - \gamma_i \right)^2 + \sum_{i=1}^{n} P(\gamma_i; \lambda)}.$$

여기서 $\boldsymbol{P(\gamma_i; \lambda)}$는 $\boldsymbol{\gamma_i}$와 조율모수(tuning parameter) $\boldsymbol{\lambda}$ 에 특징지어지는 비볼록 벌점함수이다. 현재 $\boldsymbol{R}$ 에 이 방법을 적용할 수 있는 함수가 나와 있어 쉽게 사용이 가능하다. 결과 창에는 추정된 γ_i 값이 "0"과 "1"로 나오게 되며 γ_i 값이 "0"으로 추정된 자료 값은 정상으로 "1"로 추정된 자료 값은 이상점으로 결론을 내리면 된다.

3. 이상점 처리

3.1. 가중치 보정방법

가중치 보정방법은 무응답과 이상점의 영향력을 줄이고 벤치마킹 정보를 이용하기 위해 흔히 사용된다. 일반적인 방법은 표본설계 시에 정해진 설계 가중치에 각각의 요인에 해당되는 보정인자를 구한 후 이 보정 인자를 곱하여 최종 가중치를 얻는다. 본 연구에서는 이상점 처리를 위한 보정을 고려한다. 따라서 최종 가중치 $\boldsymbol{w}^f$는 설계 가중치를 $\boldsymbol{w}$라 하고 이상점 보정인자를 $\boldsymbol{f}$라 하였을 때 $\boldsymbol{w}^f = \boldsymbol{w} \times \boldsymbol{f}$로 구해진다. 다음이 $\boldsymbol{n}$개의 자료에서 $\boldsymbol{k}$개의 이상점이 존재할 경우 본 연구에서 고려한 가중치 보정방법이다. 여기서 하나의 층을 고려하기 때문에 각 자료에 배정된 가중치는 동일하게 된다. 이 방법들은 Kim과 Shin (2013)에서도 사용되었다.

방법 1. 이상점 보정인자 $\boldsymbol{f} = 0$으로 한다. 따라서 이상점의 최종 가중치 $\boldsymbol{w}^f = 0$이 된다. 이 방법은 이상점을 제거한 자료를 만드는 효과를 준다. 따라서 최종 가중치는 다음과 같다.

$$\boldsymbol{w}^f = 0 : \text{이상점인 경우}$$

$$\boldsymbol{w}^f = \boldsymbol{w}\left(\frac{\boldsymbol{n}}{\boldsymbol{n}-\boldsymbol{k}}\right) = \boldsymbol{w}\left(1+\frac{\boldsymbol{k}}{\boldsymbol{n}-\boldsymbol{k}}\right) : \text{정상자료인 경우.}$$

방법 2. 이상점 보정인자 $\boldsymbol{f} = \frac{1}{\boldsymbol{w}}$로 한다. 따라서 이상점의 최종 가중치 $\boldsymbol{w}^f = \boldsymbol{1}$이다. 이 방법은 BLS에서 흔히 사용하는 방법으로 이에 관한 내용은 Lee와 Shin (2008)을 살펴보기 바란다. 따라서 최종 가중치는 다음과 같다.

$$\boldsymbol{w}^f = \boldsymbol{w} \times \frac{\boldsymbol{1}}{\boldsymbol{w}} = \boldsymbol{1} : \text{이상점인 경우,}$$

$$\boldsymbol{w}^f = \boldsymbol{w}\left(1+\frac{\boldsymbol{k}(\boldsymbol{w}-\boldsymbol{1})}{\boldsymbol{w}(\boldsymbol{n}-\boldsymbol{k})}\right) : \text{정상자료인 경우.}$$

3.2. 이상점 대체법

Ren and Chamber (2004)은 이상점으로 탐지된 경우에 사용할 수 있는 이상점 대체법을 연구하였다. 이 연구에서는 총계가 알려진 경우에 사용할 수 있는 reverse calibration imputation 방법이 제안되었으며 또한 알려진 총계가 없는 경우에 사용할 수 있는 회귀대체 방법이 제안되었다. 국내 현실을 반영하면 총계가 알려진 경우가 거의 없기 때문에 본 연구에서는 총계에 관한 정보가 존재하지 않는 경우에 사용할 수 있는 회귀대체방법을 이용하여 이상점의 영향력을 제거하였다. 이상점 회귀대체법은 먼저 정상자료를 이용한 회귀분석을 통하여 회귀계수를 찾는다. 이 회귀계수를 $\hat{\boldsymbol{\beta}}_0^*$라 하면 회귀분석에 의한 대체값은 다음과 같다.

$$\boldsymbol{y}_k^* = \hat{\boldsymbol{\beta}}_0^* \boldsymbol{x}_k + \boldsymbol{\delta}_k, \boldsymbol{k} \in \boldsymbol{s}_1,$$

이제 $\boldsymbol{\delta}_k$를 결정하는 방법에 따라 결정적 대체법과 확률적 대체법으로 나누어진다. 여기서 s_1은 이상점이 아닌 자료 세트이다.

3.2.1. 결정적 대체법 결정적 대체법은 $\boldsymbol{\delta}_k = sign(\boldsymbol{y}_k - \hat{\boldsymbol{\beta}}_0^* \boldsymbol{x}_k)\boldsymbol{z}_{1-\alpha/2}\hat{\boldsymbol{\sigma}}_0^2$을 사용한다. 즉 이상점이 아닌 자료를 이용하여 신뢰구간을 구한 후 상한 또는 하한 값으로 대체값을 결정한다. 여기서 $\hat{\sigma}_0^2$은 추정된 오차의 분산이고 $\boldsymbol{z}_{1-\alpha/2}$는 표준정규분포에서 얻는다.

3.2.2. 확률적 대체법 확률적 대체법은 $\boldsymbol{\delta}_k = sign(\boldsymbol{y}_k - \hat{\boldsymbol{\beta}}_0^* \boldsymbol{x}_k)|\boldsymbol{z}_k|\hat{\boldsymbol{\sigma}}_0^2$를 사용한다. 즉 표준정규분포에서 랜덤으로 난수 $\boldsymbol{z}_k$를 생성한 후 이를 사용한다. 따라서 $\boldsymbol{E}(|\boldsymbol{z_k}|) = \sqrt{2/\pi}$이 성립하므로 많은 경우 결정적 대체에 비해 회귀선에 가까운 값으로 대체값이 결정된다.

4. 다중 대체법

단일 대체법은 결측이 있는 불완전한 자료의 결측에 하나의 값으로 대체값을 결정하는 방법이다. 즉 각각의 결측값에 예측되는 분포에서 자료를 생성한 후 하나의 적절한 값을 대체하여 완전자료를 만드는 방법이다. 그러나 단일 대체법의 경우 대체된 값의 분포를 고려하지 않기 때문에 단일 대체법으로 만들어진 완전자료를 이용하여 분석할 경우 분산을 과소추정하게 된다. 이러한 문제를 해결하기 위한 방법으로 다중 대체법이 사용된다.

본 연구에서 사용된 다중 대체법은 SAS/PROC MI에서 사용하는 여러 대체 방법 중에서 회귀 분석법이 사용되었다. 일반적으로 회귀대체법에서는 가중치를 고려하지 않지만 본 연구에서는 이상점이 존재하는 경우를 고려하기 때문에 이상점의 가중치와 정상 자료의 가중치를 각각 대체 모형에 적용하여야 한다. 이를 위해서 PROC/MI를 사용하기 전에 자료에 미리 제곱근 가중치를 곱하여 새로운 자료를 만든 후 이 자료에 다중 대체법을 적용하였다. 다중 대체법을 이용하여 얻어진 최종 대체값에 제곱근 가중치의 역수를 곱하면 원하는 대체 값이 얻어진다. 이 내용을 설명하면 다음과 같다. 먼저 다음의 중회귀모형을 고려하자.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon.$$

이제 양변에 $W^{1/2}$를 곱하면 다음의 모형이 된다.

$$\boldsymbol{W^{1/2}y} = \boldsymbol{W^{1/2}X}\boldsymbol{\beta} + \boldsymbol{W^{1/2}}\epsilon.$$

여기서, $\boldsymbol{W^{1/2}y} = \boldsymbol{y^*}$, $\boldsymbol{W^{1/2}X} = \boldsymbol{X^*}$ 그리고 $\boldsymbol{W^{1/2}}\epsilon = \epsilon^*$ 라 하면 다음의 모형이 얻어진다.

$$\boldsymbol{y^*} = \boldsymbol{X^*}\boldsymbol{\beta^*} + \epsilon^*.$$

그러면 최소제곱추정량은 $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*'}\boldsymbol{X^*})^{-1}\boldsymbol{X}^{*'}\boldsymbol{y^*} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$가 된다. 따라서 변환된 자료의 대체값은 $\hat{\boldsymbol{y^*}} = \boldsymbol{X^*}\hat{\boldsymbol{\beta}}^*$이 된다. 이제 양변에 $W^{-1/2}$를 곱하면 원하는 다음의 최종 대체값이 얻어진다.

$$\boldsymbol{W^{-1/2}}\hat{\boldsymbol{y^*}} = \boldsymbol{W^{-1/2}X^*}\hat{\boldsymbol{\beta}}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}^*.$$

결국 원하는 $\boldsymbol{i}$번째 결측값의 대체값은 예측값인 $\hat{\boldsymbol{y}}_i = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}^*$이 된다. 이 방법은 단일 대체법에 관한 설명이지만 같은 방법으로 다중 대체법이 얻어진다. 본 논문에서는 SAS 결과에서 얻어진 5개의 다중 대체 자료 세트를 사용하였다. 만약 이상점 회귀대체에 의해 이상점의 영향이 제거되었다면 $\boldsymbol{W} = \boldsymbol{I}$를 사용한다.

5. 모의실험

이상점의 영향력을 줄인 후 얻은 다중 대체 결과의 각 방법별 우수성을 살펴보기 위해 모의실험이 수행되었다. 모의실험을 위한 자료의 생성과정은 Lee 등 (1995)에서 사용한 동일한 방법을 사용하였다. 먼저 크기 $\boldsymbol{N} = 10,000$인 모집단을 다음과 같이 생성하였다. 보조자료 $\boldsymbol{x}_k$는 평균 48이고 분산 768을 갖는 감마분포로부터 생성하였다. 주어진 $\boldsymbol{x}_k$ 값에 대하여 모두 네 종류의 조사자료 $\boldsymbol{y}_k$ 값을 발생시켰

Table 5.1. Coefficients for the simulation

조사변수형태	a	b	c	d	g
비례형	0	1.50	0.00	5.13	0.50
선형	20	1.50	0.00	13.79	0.25
볼록형	0	0.25	0.01	4.91	0.50
오목형	0	3.00	-0.01	5.60	0.50

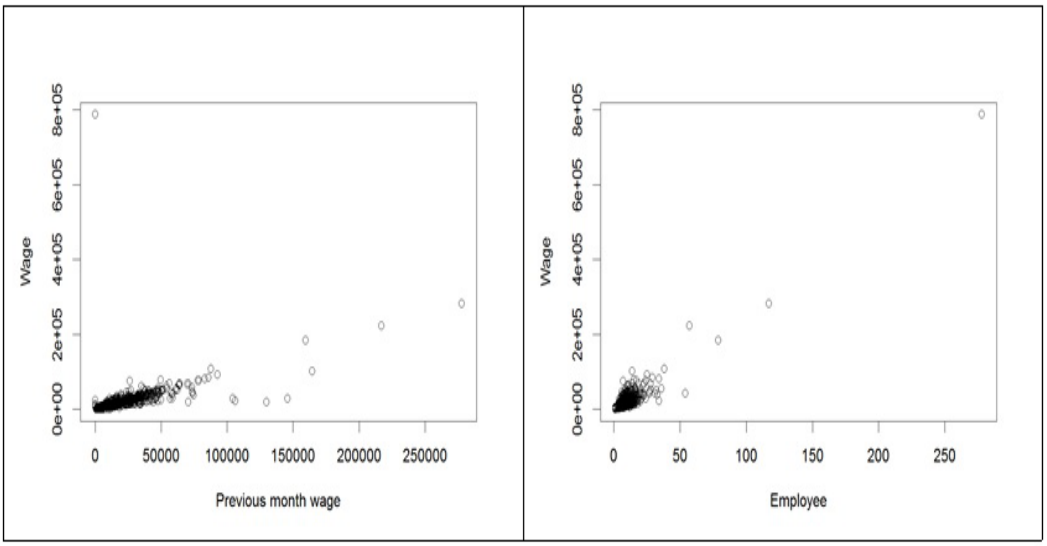

Figure 5.1. Scatter plots of wages of May and April in the left panel and wage and employee of May in the right panel for data set 1

으며 각각 평균 $\boldsymbol{\mu(x)} = \boldsymbol{a} + \boldsymbol{bx} + \boldsymbol{cx}^2$ 값이고 분산 $\boldsymbol{\sigma}^2(\boldsymbol{x}) = \boldsymbol{d}^2\boldsymbol{x}^{2g}$를 갖는 감마분포를 사용하였다. Table 5.1 은 선택된 상수 $\boldsymbol{a, b, c, d, g}$의 값을 나타낸다. 첫 번째로 생성된 자료는 보조변수와의 관계가 원점을 지나는 비례적(ratio) 형태이고, 두 번째 자료는 양의 절편 값을 갖는 선형관계(regression)를 갖도록 하였다. 또한 오목형과 볼록형의 관계가 성립하도록 상수를 조정한 후 자료를 생성하였다.

본 연구에서 사용한 비교 통계량은 절대편향(absolute bias)와 평균제곱오차(mean squared error)이다. 정의는 다음과 같다.

$$\text{Absolute bias} : AB = \frac{1}{R}\frac{1}{n}\frac{1}{5}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{j=1}^{5}|\hat{\boldsymbol{y}}_{i,j}^{(r)} - \boldsymbol{y}_i|,$$

$$\text{Mean squared error} : MSE = \frac{1}{R}\frac{1}{n}\frac{1}{5}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{j=1}^{5}(\hat{\boldsymbol{y}}_{i,j}^{(r)} - \boldsymbol{y}_i)^2.$$

모의실험에서 $R = 1,000$과 $n = 200$을 사용하였으며 랜덤으로 $m = 10$, 20 그리고 40개의 자료를 결측 처리 하였다. 여기서 추정을 위해 사용하는 설계 가중치(design weight)는 50이 되어야 하나, 본 모의실험은 가중치를 고려한 대체법의 우수성을 비교하기 위한 것이기 때문에 설계 가중치 보다 큰 가중치 $w = 400$을 사용하였다. 표본으로 추출된 자료를 R-code로 되어있는 $\boldsymbol{\Theta}$-IPOD 방법으로 이상점을

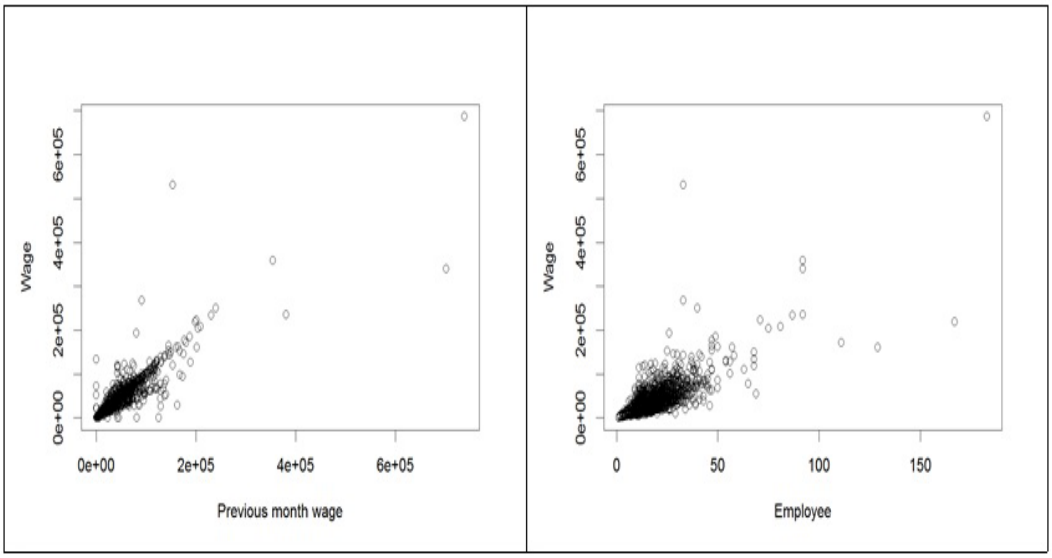

Figure 5.2. Scatter plots of wages of May and April in the left panel and wage and employee of May in the right panel for data set 2

탐지하였으며 이상점 대체는 SAS/Proc MI를 사용하였으며 5개의 자료 세트가 얻어졌다. 따라서 R과 SAS를 동시에 사용하기 때문에 많은 수의 모의실험을 하기에는 어려움이 있다. 모의실험 결과는 Table 5.2와 Table 5.3에 수록하였다. 이후 Table에서 M1은 이상점을 처리하지 않은 결과이고, M2, M3는 이상점인 경우 최종 가중치를 각각 0과 1로 두는 가중치 조정 방법을 사용한 결과이며, M4, M5는 이상점에 각각 결정적 대체법과 확률적 대체법을 시행한 결과이다. 모의실험 결과를 살펴보면 이상점 처리를 실시한 M2, M3, M4, M5 방법 모두 이상점 처리를 하지 않은 M1 결과보다 좋은 것을 알 수 있다. 이 중에서도 이상점인 경우 최종 가중치를 1로 두는 M3 방법이 가장 우수한 것으로 나타났다. 그 다음으로 이상점인 경우 최종 가중치를 0 으로 두는 즉, 이상점을 삭제하는 M2 방법이 우수한 결과를 주고 있다. 이에 반해 상대적으로 이상점 대체를 실시한 방법은 성능이 떨어지는 것을 확인 할 수 있다. 결론적으로 이상점과 무응답이 동시에 존재하는 경우에는 이상점을 적절히 처리한 후 무응답 대체법을 사용하면 우수한 결과를 얻을 수 있음을 알 수 있었으며 이 모의실험에서는 그 중 이상점인 경우 최종 가중치를 1로 두는 M3 방법을 사용하면 가장 우수한 결과를 얻을 수 있음을 확인할 수 있다.

6. 실제 자료분석

6.1. 매월노동통계 자료분석

이 절에서는 2007년 매월노동통계 표본 자료 7,066개 자료 중에서 사업체 규모 기준으로 규모 1인 1,664개 자료 세트 1과 규모 2 인 1,976개 자료 세트 2에서 $\boldsymbol{t}$월의 총임금과 종사자 수 그리고 $\boldsymbol{t-1}$월의 총임금을 선택하여 자료분석을 실시하였다. 이상점 탐지 방법인 $\boldsymbol{\Theta}$-IPOD 방법을 위해 $\boldsymbol{t-1}$월의 총임금 자료와 $\boldsymbol{t}$월의 종사자 수가 보조변수로 $\boldsymbol{t}$월의 총임금 자료가 관심변수로 사용되었다. 다중 대체에서도 두 개의 보조변수 즉 $\boldsymbol{t}$월의 종사자 수와 $\boldsymbol{t-1}$월의 총임금 자료가 사용되었으며 관심변수는 $\boldsymbol{t}$월의 총임금 자료가 사용되었다. 이를 그림으로 나타낸 것이 각각 Figure 6.1과 Figure 6.2 이며 직관적으로 이상점이 있음을 쉽게 알 수 있다. 자료 세트의 모든 자료가 분석에 사용되었으며 무응답이 존재하지 않아

Table 5.2. Simulation results for absolute bias

가중치 / 결측개수	조사변수형태	Absolute Bias				
		$M1$	$M2$	$M3$	$M4$	$M5$
400 / 10	비례형	6.43	6.23	6.04	6.30	6.28
	선형	6.54	6.49	6.44	6.47	6.51
	볼록형	6.37	5.11	4.79	6.11	6.14
	오목형	6.66	6.61	6.61	6.64	6.67
400 / 20	비례형	6.43	6.20	6.01	6.30	6.31
	선형	6.54	6.46	6.37	6.47	6.47
	볼록형	6.40	5.17	4.80	6.17	6.20
	오목형	6.67	6.63	6.55	6.60	6.62
400 / 40	비례형	6.46	6.21	6.02	6.31	6.31
	선형	6.55	6.46	6.43	6.48	6.49
	볼록형	6.38	5.17	4.80	6.19	6.17
	오목형	6.66	6.63	6.57	6.62	6.63

Table 5.3. Simulation results for MSE results

가중치 / 결측개수	조사변수형태	MSE				
		$M1$	$M2$	$M3$	$M4$	$M5$
400 / 10	비례형	7.88	7.63	7.37	7.71	7.70
	선형	8.02	7.95	7.90	7.95	7.98
	볼록형	7.82	6.68	6.15	7.47	7.51
	오목형	8.15	8.10	8.08	8.11	8.15
400 / 20	비례형	7.92	7.63	7.39	7.76	7.74
	선형	8.05	7.95	7.84	7.96	7.96
	볼록형	7.88	6.80	6.18	7.57	7.60
	오목형	8.21	8.15	8.06	8.12	8.14
400 / 40	비례형	7.97	7.66	7.40	7.77	7.76
	선형	8.07	7.97	7.92	7.99	8.00
	볼록형	7.88	6.81	6.21	7.61	7.58
	오목형	8.21	8.16	8.09	8.16	8.17

자료 세트 1과 2에 대해 10, 50, 100, 200개의 자료를 각각 랜덤으로 결측 처리하였고 사용된 가중치는 각각 200, 400 그리고 1,000이다. 결과는 Table 6.1에서 Table 6.4에 수록하였다.

매월노동통계 자료분석 결과를 살펴보면 이상점 처리를 한 M2, M3, M4, M5 방법 모두 이상점 처리를 하지 않은 M1 결과보다 좋은 것을 알 수 있고 이상점인 경우 최종 가중치를 '1'로 두는 M3 방법이 가장 우수한 것을 확인할 수 있다. 그 다음으로 이상점인 경우 최종 가중치를 '0' 으로 두는 즉 이상점을 삭제하는 M2 방법과 이상점 대체법이 유사한 결과를 주고 있다. 결론적으로 실제 자료 분석에서도 이상점과 무응답이 동시에 존재하는 경우에는 이상점을 적절히 처리한 후 무응답 대체법을 사용하는 것이 타당하다는 것을 확인하였다.

Table 6.5와 Table 6.6에 분석 결과를 수록하였으며 결과를 살펴보면 이상점 처리를 한 M2, M3 방법이 이상점 처리를 하지 않은 M1 결과보다 좋은 것을 알 수 있고 이상점인 경우 최종 가중치를 '1'로 두는 M3 방법이 가장 우수한 것으로 나타났다. 그 다음으로 이상점인 경우 최종 가중치를 '0'으로 두는 즉, 이상점을 삭제하는 M2 방법이 우수한 결과를 주고 있다. 그러나 M4, M5 방법은 이상점 처리 후 오히려 대체 결과가 나빠진 것을 확인 할 수 있다.

Table 6.1. Absolute bias results for data set 1

가중치	결측개수	Absolute Bias				
		$M1$	$M2$	$M3$	$M4$	$M5$
200	10	5612.73	3172.92	3134.67	3183.47	3159.05
	50	5636.90	3222.62	3180.77	3226.77	3226.70
	100	5633.45	3210.61	3171.98	3216.76	3215.13
	200	5630.70	3213.27	3174.52	3219.16	3217.16
400	10	5640.82	3235.17	3181.17	3229.81	3234.34
	50	5615.01	3177.95	3127.71	3185.14	3189.49
	100	5628.94	3203.79	3152.81	3207.75	3207.27
	200	5627.99	3219.24	3171.62	3227.16	3223.35
1000	10	5665.17	3209.80	3153.24	3205.83	3215.27
	50	5633.33	3213.07	3152.25	3218.24	3207.98
	100	5637.27	3219.13	3163.14	3222.06	3227.24
	200	5620.79	3228.82	3174.51	3237.94	3236.12

Table 6.2. Absolute bias results for data set 2

가중치	결측개수	Absolute Bias				
		$M1$	$M2$	$M3$	$M4$	$M5$
200	10	15235.33	7767.12	7719.10	7801.60	7804.22
	50	15089.36	7712.15	7647.50	7734.86	7722.52
	100	15123.22	7762.27	7693.60	7780.07	7776.62
	200	15066.81	7737.93	7668.95	7744.69	7737.33
400	10	15254.60	7845.24	7753.32	7875.52	7894.70
	50	15131.94	7795.88	7678.22	7785.12	7789.00
	100	15113.80	7703.52	7600.17	7709.63	7711.23
	200	15119.54	7758.20	7655.93	7777.51	7772.27
1000	10	15156.52	7729.94	7626.89	7744.53	7739.73
	50	15102.26	7695.57	7582.63	7717.08	7721.01
	100	15129.57	7803.69	7689.80	7827.91	7814.84
	200	15059.98	7721.52	7594.25	7740.13	7734.46

Table 6.3. MSE results for data set 1

가중치	결측개수	MSE				
		$M1$	$M2$	$M3$	$M4$	$M5$
200	10	7367.49	4988.32	4948.47	5002.84	4971.84
	50	7777.17	6143.19	6113.45	6147.53	6150.50
	100	7865.29	6471.00	6440.73	6474.26	6469.78
	200	7926.68	6694.40	6667.29	6702.20	6700.07
400	10	7476.60	5208.68	5170.43	5207.45	5214.95
	50	7754.34	6051.93	6026.46	6059.59	6059.59
	100	7855.68	6428.73	6398.96	6433.23	6429.94
	200	7931.04	6741.30	6713.92	6743.53	6742.33
1000	10	7463.44	5152.03	5093.32	5144.48	5146.73
	50	7745.61	6021.82	5990.44	6027.13	6019.11
	100	7878.09	6452.01	6426.45	6451.54	6458.68
	200	7895.72	6734.92	6714.00	6739.21	6743.82

Table 6.4. MSE results for data set 2

가중치	결측개수	MSE				
		$M1$	$M2$	$M3$	$M4$	$M5$
200	10	20250.53	12464.13	12397.81	12495.74	12499.50
	50	21158.73	14624.31	14559.88	14639.24	14642.43
	100	21823.82	15666.45	15606.60	15677.96	15678.14
	200	22153.14	16348.75	16302.04	16364.05	16352.12
400	10	20235.98	12499.31	12406.91	12519.67	12540.11
	50	21345.80	14819.32	14735.91	14809.55	14807.24
	100	21533.42	15289.86	15212.19	15288.20	15293.97
	200	22266.29	16380.37	16307.81	16388.92	16385.20
1000	10	19953.40	12184.19	12090.63	12202.90	12173.71
	50	21343.72	14643.56	14563.45	14651.87	14672.09
	100	22113.75	16051.04	15981.66	16075.20	16061.01
	200	22075.40	16413.33	16331.17	16422.25	16416.62

Table 6.5. Absolute Bias results for Briquette data

가중치	결측개수	Absolute Bias				
		$M1$	$M2$	$M3$	$M4$	$M5$
200	10	220066.10	197727.76	194482.34	226537.02	227040.35
	50	220674.22	198744.73	194212.30	226234.49	226579.37
	100	220401.43	198796.40	193790.75	226015.31	226142.74
	200	219542.68	198132.84	193031.06	225735.60	225684.07
400	10	219357.63	198451.42	193122.20	225821.18	226164.68
	50	219607.63	198697.55	193444.79	225955.03	226101.99
	100	219650.91	198530.21	193427.52	226050.39	225866.87
	200	220099.09	198839.81	194032.84	225824.38	226007.46
1000	10	221968.51	199459.66	195386.83	228290.41	228968.39
	50	220320.74	199656.53	194179.52	226776.38	226431.14
	100	219877.78	198420.73	194015.78	225789.22	225877.72
	200	219757.04	198200.69	192931.87	225723.31	225795.69

Table 6.6. MSE results for Briquette data

가중치	결측개수	MSE				
		$M1$	$M2$	$M3$	$M4$	$M5$
200	10	288752.69	273407.84	268852.47	296929.42	297435.86
	50	301232.27	286973.19	282064.86	307271.95	307961.67
	100	304215.42	290963.10	285294.14	310450.96	310596.91
	200	305371.51	292248.56	286475.91	311905.72	311804.90
400	10	286110.01	273015.73	266697.33	294382.96	294435.99
	50	301868.66	288950.83	282889.27	308592.39	308438.35
	100	302379.58	289879.31	283818.28	309394.66	309111.94
	200	306776.24	294126.12	288845.29	312920.40	313147.14
1000	10	290955.46	274958.00	269483.18	296936.10	298945.79
	50	300386.74	287886.38	281847.69	307800.03	307350.53
	100	303366.29	290218.92	285267.99	309897.91	309794.87
	200	306154.09	292977.01	287425.68	312563.10	312573.72

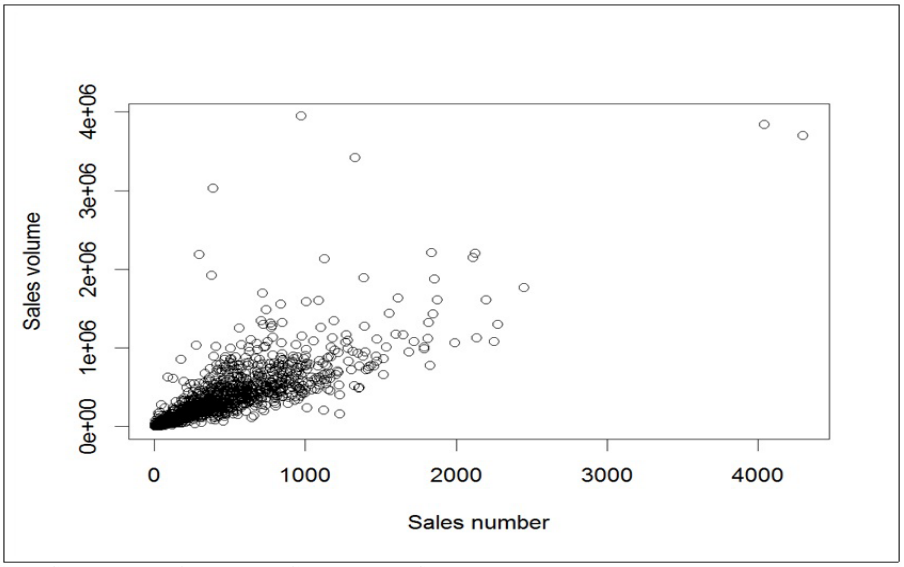

Figure 6.1. Scatter plot of Briquette Consumption Survey data

본 자료분석에서도 이상점과 무응답이 동시에 존재하는 경우에는 이상점을 적절히 처리한 후 무응답 대체법을 사용하면 우수한 결과를 얻을 수 있음을 확인하였으며 제안된 방법 중에서 이상점인 경우 최종 가중치를 '1'로 두는 M3 방법을 사용하면 가장 우수한 결과를 얻을 수 있음을 알 수 있다.

7. 결론

본 논문에서는 이상점 탐지를 위해 **Θ**-IPOD 방법을 이용하였다. 탐지된 이상점은 본 논문에서 연구된 방법을 이용하여 이상점 대체와 이상점 가중치 보정이 이루어졌다. 모의실험 결과 이상점 처리 후 모든 대체법에서 이상점 처리를 하지 않은 대체법에 비해 우수한 것을 확인하였다. 그 중에서 본 논문에서 연구된 M3 방법이 가장 우수한 것을 확인하였다.

그러나 실제 자료분석 결과를 살펴보면 매월노동통계 자료 분석에서는 본 논문에서 사용된 모든 이상점 처리 방법 사용 후의 무응답 대체가 우수한 결과를 주었으나 연탄실태자료분석 결과를 살펴보면 이상점 대체를 이용한 무응답 대체 결과는 이상점 처리를 실시하지 않고 무응답 대체를 실시한 결과보다 좋지 않게 나왔다. 따라서 자료에 따라 이상점 처리 방법을 신중히 선택해야 하거나 다른 추가적인 처리가 필요한 것으로 판단된다.

References

Belcher, R (2003). Application of the Hidiroglou-Berthelot method of outlier detection for periodic business surveys, *SSC Annual Meeting, Proceeding of the Survey Method Section.*

Hidiroglou, M. A. and Berthelot, J.-M. (1986). Statistical editing and imputation for Periodic Business Surveys, *Survey Methodology*, **12**, 73-83.

Kim, J.-Y. and Shin, K.-I. (2013). Multiple Imputation reducing outlier effect using weight adjustment methods, *The Korean Journal of Applied Statistics,* **26**, 635-647.

Lee, H., Rancourt, E. and Sarndal, C.-E.(1995). Experiment with variance estimation from survey data with imputed value, *Journal of Official Statistics*, **10**, 231-243.

Lee, S.-J. and Shin, K.-I. (2008). A Study on the sensitivity of the BLS Methods, *Communications of the Korean Statistical Society,* **15**, No. 6, 843-858.

McCullough, M. and Pennington, T. L. (2009). identifying outliers when creating an imputation based for the Qualterly Financial Report,*JSM, Section on Survey Research Methods.*

Park, D.-I., Kang, H., Han S.-T. and Choi, H. (2013). Comparison study of outlier detection methods in a regression model, *Journal of the Korean Data Analysis Society,* **15**, 177-186.

Ren, R. and Chamber, R. L. (2004). Outlier robust imputation of survey data, *Proceeding of ASA Section on Survey Research Methods.*

Rubin, D. B. (1987). Multiple imputation for Nonresponse in Survey, New York.

She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression, *Journal of the American Statistical Association,* **106**, 626-639.

이상점 영향력 축소를 통한 무응답 대체법

김만겸[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2014년 10월 15일 접수, 2014년 12월 05일 수정, 2014년 12월 05일 채택)

요 약

이상점과 무응답이 동시에 존재하는 경우에는 무응답만 있는 경우에 비해 무응답 대체의 성능이 떨어지게 된다. 이러한 경우에는 먼저 이상점을 탐지하고, 탐지된 이상점의 영향력을 축소한 후 무응답 대체를 실시하여야 한다. 본 논문에서는 이상점의 영향력을 축소하여 무응답 대체법의 성능을 향상시키는 방법을 연구하였다. 이를 위해 She and Owen (2011)이 제안한 이상점 탐지법을 살펴보았고, 탐지된 이상점의 영향력을 줄이기 위한 방법으로 흔히 사용되는 가중치 조정법과 이상점 대체법을 살펴보았다. 또한 이상점 처리 방법을 적용한 무응답 대체법을 살펴보았으며 모의실험과 사례분석을 통하여 이상점 영향력 축소 효과를 살펴보았다.

주요용어: 이상점 탐지, 이상점 대체, 이상점 가중치 조정, 벌점화 회귀.

이 논문은 2014년 한국외국어대학교 학술연구지원에 의해 수행되었음.

[1]교신저자: (449-791) 경기도 용인시 처인구 모현면 외대로 81, 한국외국어대학교 통계학과 교수. E-Mail: keyshin@hufs.ac.kr