

Multiple Imputation Reducing Outlier Effect using Weight Adjustment Methods

Jin-Young Kim^a · Key-Il Shin^{a,1}

^aDepartment of Statistics, Hankuk University of Foreign Studies

(Received April 27, 2013; Revised June 17, 2013; Accepted June 18, 2013)

Abstract

Imputation is a commonly used method to handle missing survey data. The performance of the imputation method is influenced by various factors, especially an outlier. The removal of the outlier in a data set is a simple and effective approach to reduce the effect of an outlier. In this paper in order to improve the precision of multiple imputation, we study a imputation method which reduces the effect of outlier using various weight adjustment methods that include the removal of an outlier method. The regression method in PROC/MI in SAS is used for multiple imputation and the obtained final adjusted weight is used as a weight variable to obtain the imputed values. Simulation studies compared the performance of various weight adjustment methods and Monthly Labor Statistic data is used for real data analysis.

Keywords: Outlier detection, Hidroglou-Berthelot method, studentized deleted residual.

1. 서론

표본조사에서는 다양한 원인으로 인하여 오차가 발생하고 있다. 여러 오차 중에서 무응답으로 인해 발생한 결측자료는 자료의 분포를 변형시키거나 편향을 발생시키는 등 전체 비표본오차에 매우 큰 영향을 주고 있다. 쉽게 생각할 수 있는 방법이 결측 자료를 제거하여 완전자료(complete data)를 만든 후 분석하는 것이다. 그러나 결측 자료를 제거한 후 얻어진 완전자료는 모집단 분포와 다른 형태의 분포가 될 수도 있어 자료 분석 후 편향이 발생되거나 자료 수가 줄어들게 되어 분석의 정확도가 떨어질 수 있다. 결측으로 인한 비표본오차를 줄이기 위한 여러 연구가 활발히 진행되었으며 가장 흔하면서도 효과적인 방법이 대체법(imputation)이다. 대체법은 단일 대체법과 다중 대체법으로 나누어지며 최근 다중 대체법을 이용한 대체가 통계적으로 우수한 결과를 주는 것으로 알려져 있다.

또한 표본조사에서 무응답 문제와 더불어 이상점(outlier)의 존재는 비표본오차를 증가시키는 중요한 요인이다. 이상점을 적절히 처리하지 않고 분석 할 경우 모수 추정 결과가 과대 또는 과소 추정될 수 있다. 이러한 과대 또는 과소 추정을 막기 위해서는 이상점을 판별하고 판별된 이상점을 적절히 처리하여야 한다. 실제 표본조사에서는 이상점과 무응답이 동시에 발생하게 된다. 이 경우 이상점을 적절히 처리하지 않고 무응답 대체법을 사용하게 되면 대체법의 성능이 떨어지게 된다. 따라서 무응답 대체를 실시하기

This research was supported by the research fund of Hankuk University of Foreign Studies(2013).

¹Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin, Kyonggi 449-791, Korea. E-mail: keyshin@hufs.ac.kr

전에 이상점을 판별하고 적절한 처리를 한 후 무응답 대체를 실시하여야 한다. 이에 본 연구에서는 이상점을 판별하고 판별된 이상점의 가중치를 보정하여 이상점의 영향력을 줄인 후 다중 대체를 실시하여 대체의 성능을 향상시키는 방법을 연구하였다.

이상점 판별 방법은 판별에 사용할 정보, 독립변수의 유무 등에 따라 달라진다. 본 연구에서는 패널 자료와 같이 전 시점에서 얻어진 정보만을 이용하여 이상점을 판별하는 방법과 전 시점 자료에 추가로 횡시점 자료에 독립변수가 존재하는 경우에 회귀모형을 사용하여 이상점을 판별하는 방법을 살펴보았다. 전 시점 자료의 정보만을 이용하여 이상점을 판별하는 방법으로는 Hidiroglou-Berthelot (1986) 방법을 사용하였으며 이에 관한 내용은 Hidiroglou와 Berthelot (1986)과 Belcher (2003)을 참고하기 바란다. 또한 회귀분석의 외표준화잔차를 이용하여 이상점을 판별하는 방법에 관한 내용은 McCullough와 Pennington (2009)를 참조하기 바란다.

이상점이 판별된 후에는 이상점의 영향력을 줄이기 위해 이상점의 가중치를 줄일 필요가 있다. 이상점의 가중치를 줄이는 여러 가중치 보정 방법이 이미 개발되었으므로 이러한 가중치 보정방법을 이용하여 이상점이나 판별된 자료의 가중치를 보정한 후 다중 대체를 실시하였다. 또한 모의실험을 이용하여 임의의 자료를 생성한 후 자료의 특성과 가중치 보정 방법에 따른 다중 대체법의 성능을 비교하였다. 이때 무응답은 MAR(missing at random)을 따른다고 가정하였다. 이에 관한 내용은 Rubin (1987)을 살펴보기 바란다. SAS/MI를 이용하여 다중 대체가 이루어졌으며 실제 자료 분석과 모의실험을 통하여 본 연구에서 제안한 방법의 우수성을 확인하였다. 실제 자료 분석에는 매월노동통계 자료가 사용되었다.

논문의 구성은 다음과 같다. 2절에서는 이상점을 판별하는 두 가지 방법, 즉 회귀모형을 이용하는 방법과 전 시점 자료만을 이용하는 방법인 Hidiroglou-Berthelot 방법을 설명하였다. 3절에서는 이상점의 효과를 줄이기 위한 처리 방법으로 여러 이상점 가중치 보정 방법에 관하여 설명하였으며 4절에서는 보정된 최종 가중치를 이용한 대체법을 설명하였다. 5절에서는 모의실험을 통하여 가중치를 이용한 이상점 가중치 보정 효과를 살펴보았으며 6절에서는 실제 자료 분석이 수행되었다. 7절에 결론이 있다.

2. 이상점 판별

본 연구에서는 회귀모형에서 사용되는 이상점 판별법 중에서 외표준화잔차(studentized deleted residual)를 기준으로 한 판별법이 사용되었다. 또한 Hidiroglou-Berthelot (1986)이 제안한 방법도 본 연구에서 사용되었다. 전 시점 자료와 횡시점 자료에서 이상점 판별을 위한 독립변수가 얻어진 경우에는 외표준화잔차를 사용하고 오직 전 시점 자료가 이상점 판별에 사용될 경우에는 Hidiroglou-Berthelot 방법을 사용하였다. 이 두 방법을 간단히 살펴보면 다음과 같다.

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

특정한 i 번째 자료가 이상점일 경우 잔차 r_i 가 커지게 되지만 동시에 i 번째 자료가 분산 추정에도 영향을 미쳐 분산이 커지게 된다. 따라서 i 번째 자료를 제거한 후 분산을 추정한 외표준화잔차를 사용하여 이상점을 찾아내는 것은 매우 타당하다. 다음이 외표준화잔차의 정의이다.

$$t_i = \frac{r_i}{s(d_i)} = \frac{r_i}{s(i)/\sqrt{(1-h_{ii})}},$$

여기서 $r_i = y_i - \hat{y}(i)$, y_i 는 관측값 그리고 $\hat{y}(i)$ 는 i 번째 관측값을 제거한 후에 얻어진 예측값을 의미한다. 또한 h_{ii} 는 지렛값 또는 레버리지이다. t_i 의 분포는 우리가 잘 알고 있는 자유도 $(n-p-1)$ 인 t 분포를 따르는 것으로 알려져 있다. 또한 외표준화잔차는 SAS/Proc REG의 출력결과(Rstudent)에서 쉽게 얻을 수 있는 결과이다.

2.2. Hidiroglou-Berthelot 방법

Hidiroglou-Berthelot 방법은 이상점을 전 시점($t-1$) 자료와 현 시점(t) 자료의 비율을 이용하여 이상점을 판단한다. 이를 위한 판단 방법을 구체적으로 설명하면 다음과 같다.

1. i 번째 자료의 $t-1$ 시점 자료를 $x_i(t-1)$ 이라 하고 t 시점 자료를 $x_i(t)$ 라 할 때 두 자료의 비율 $r_i = x_i(t)/x_i(t-1)$ 을 구한다.
2. 계산의 편리를 위하여 변환된 자료인 s_i 를 다음과 같이 구한다.

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

여기서 r_M 은 r_i 의 중앙값이다.

3. 자료의 크기를 반영하기 위하여 다음과 같은 E_i 를 구한다.

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

여기서 U 는 자료의 크기 효과를 반영하기 위한 모수로 $0 \leq U \leq 1$ 이다. 만약 $U = 0$ 이면 자료의 크기를 전혀 반영하지 않는 것이며 결과적으로 단지 E_i 는 s_i 에만 의존하게 된다. 반면 $U = 1$ 인 경우에는 자료의 크기가 E_i 에 매우 큰 영향을 미치게 된다. 실제 자료 분석에서는 $0.3 \leq U \leq 0.5$ 가 흔히 사용되며 본 연구에서는 $U = 0.4$ 를 사용하였다. 이에 관한 내용은 Belcher (2003)를 보기 바란다.

4. 이상점 판별을 위하여 하한과 상한을 이용한 이상점 판별 구간을 정한다. 먼저 E_M 을 E_i 의 중앙값이라 하자. 그리고 $d_{Q1} = E_M - E_{Q1}$, $d_{Q3} = E_{Q3} - E_M$, E_{Q1} , E_{Q3} 을 각각 제 1사분위수와 제 3사분위수라고 하자. 그러면 판별 구간은 다음과 같다.

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

여기서 C 는 이상점 판별 구간의 길이를 결정하는 모수가 되며 본 연구에서는 $C = 1.5$ 를 사용하였다. 따라서 이 구간 밖에 있는 자료는 이상점으로 판별된다.

3. 가중치 보정방법

가중치 보정방법은 무응답과 이상점의 영향력을 줄이고 벤치마킹을 이용하기 위해 흔히 사용된다. 일반적인 방법은 표본설계 시에 정해진 설계 가중치에 각각의 요인에 해당되는 보정인자를 구한 후 이 보정인자를 곱하여 최종 가중치를 얻는다. 가중치 보정 방법에 관한 내용은 Lee와 Shin (2008)을 참조하기 바란다. 본 연구에서는 무응답 대체가 주된 목적이므로 무응답 보정과 벤치마킹 보정방법을 고려하지 않고 다만 이상점 보정만을 고려한다. 따라서 최종 가중치 w^f 는 설계 가중치를 w 라 하고 이상점 보정인자를 f 라 하였을 때 $w^f = w \times f$ 로 정해진다. 다음이 n 개의 자료에서 k 개의 이상점이 존재할 경우 본 연구에서 고려한 가중치 보정방법이다. 이 방법들은 Kim (2006)에서도 연구되었다.

- 방법1. 이상점 보정인자 $f = 0$ 으로 한다. 따라서 이상점의 최종 가중치 $w^f = 0$ 이 된다. 이 방법은 이상점을 제거한 자료를 만드는 효과를 주게 되므로 다중 대체 시 이상점은 제거된다. 최종 가중치는 다음과 같다.

$$w^f = 0 : \text{이상점인 경우}$$

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

방법2. 이상점 보정인자 $f = 1/w$ 로 한다. 따라서 이상점의 최종 가중치 $w^f = 1$ 이다. 이 방법은 BLS에서 흔히 사용하는 방법으로 이에 관한 내용은 Lee와 Shin (2008)을 살펴보기 바란다. 따라서 최종 가중치는 다음과 같다.

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

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

방법3. 이상점인 경우 보정인자 $f = (1 - (n-k)/2n)$ 로 하고 이상점이 아닌 정상점은 보정인자 $f = (1 + k/2n)$ 을 사용한다. 이 방법은 Hidioglou와 Srinath (1981)가 제안한 방법으로 최종 가중치는 다음과 같다.

$$w^f = w \left(1 - \frac{n-k}{2n} \right) : \text{이상점인 경우}$$

$$w^f = w \left(1 + \frac{k}{2n} \right) : \text{정상자료인 경우}$$

따라서 세 방법 모두 최종 가중치를 합치면 nw 가 되는 것을 확인할 수 있다. 또한 이상점 보정의 효과를 모의실험을 통하여 비교하기 위해 이상점 보정을 실시하지 않은 경우도 살펴보았다.

4. 다중 대체법

단일 대체법은 결측이 있는 불완전한 자료의 결측에 하나의 적절한 값을 대체하여 완전자료를 만드는 방법이다. 그러나 단일 대체법의 경우 대체된 값의 분포를 고려하지 않기 때문에 단일 대체법을 통하여 만들어진 완전자료 분석 시 분산을 과소추정하게 된다. 이러한 문제를 해결하기 위한 방법으로 다중 대체법이 제안되었다. 본 연구에서 사용된 다중 대체법은 SAS/PROC MI에서 사용하는 여러 대체 방법 중에서 회귀 분석법이 사용되었다. 일반적으로 회귀대체법에서는 가중치를 고려하지 않지만 본 연구에서는 이상점의 영향력을 줄이기 위해 가중치 보정 방법을 사용하기 때문에 보정된 이상점의 가중치와 정상자료의 가중치를 대체 모형에 적용하여야 한다. 이를 위해서 PROC/MI를 사용하기 전에 자료에 미리 가중치를 곱하여 새로운 자료를 만든 후 이 자료에 다중 대체법을 적용하였다. 다중 대체법을 이용하여 얻어진 최종 대체값에 가중치의 역수를 곱하면 원하는 대체 값이 얻어진다. 예를 들어 다음의 단일 대체법을 살펴보자. 이를 위한 중회귀 모형은 다음과 같다.

$$y = X\beta + \epsilon.$$

이제 최종 가중치 행렬을 W 라 하면, 모형의 양변에 W 를 곱하여 새로운 모형을 만든다.

$$y^* = X^*\beta^* + \epsilon^*.$$

그러면 가중최소제곱추정량은 $\hat{\beta}^* = (X'^*W^2X)^{-1}X'^*W^2y^*$ 가 된다. 이 값을 이용하여 예측값이 얻어지며 얻어진 예측값에 가중치의 역수를 곱하게 되면 최종 예측값이 얻어진다. 즉 최종 예측값은 다음과 같다.

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

따라서 원하는 i 번째 결측값의 대체값은 예측값인 $\hat{y}_i = x_i\hat{\beta}^*$ 이 된다. 같은 방법으로 다중 대체법을 사용한 대체값이 얻어진다.

Table 5.1. Coefficients for simulated population

조사변수형태	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
비례형	0	1.50	0.00	5.13	0.50
선형	20	1.50	0.00	13.79	0.25
오목형	0	0.25	0.01	4.91	0.50
블록형	0	300	-0.01	5.60	0.50

5. 모의실험

이상점의 영향력을 줄여 얻은 대체 결과의 우수성을 살펴보기 위하여 모의실험이 수행되었다. 전차 자료만을 이용하여 대체법을 실시하는 경우를 고려하여 독립변수의 수가 1개인 경우를 고려하였으며 전차 자료와 횡시점의 독립변수가 모두 사용 가능한 경우를 고려하여 독립변수의 수가 2인 경우를 고려하였다. 물론 외표준화잔차는 독립변수의 수와 무관하게 구해지게 되며 SAS/PROC MI도 독립변수의 수와 무관하게 사용된다. 또한 가중치 조정법은 종속변수의 가중치만을 조정하기 때문에 독립변수와 무관하다. 따라서 본 연구에서 제안된 방법은 독립변수의 수와 무관하게 사용될 수 있다.

모의실험을 위한 자료의 생성과정은 Lee 등 (1995)에서 사용한 동일한 방법을 사용하였다. 먼저 크기 $N = 5,000$ 인 모집단을 다음과 같이 생성하였다. 보조자료 x_k 는 평균 48이고 분산 768을 갖는 감마분포로부터 생성하였다. 주어진 x_k 값에 대하여 모두 네 종류의 조사자료 y_k 를 발생시켰는데, 각각 평균 $\mu(x) = a + bx + cx^2$ 이고 분산 $\sigma^2(x) = d^2x^{2g}$ 을 갖는 감마분포를 가정하였다. Table 5.1은 선택된 상수 a, b, c, d 의 값을 나타낸다. 첫 번째로 생성된 자료는 보조변수와의 관계가 원점을 지나는 비례적 형태(ratio)이고, 두 번째 자료는 양의 절편 값을 갖는 선형관계(regression)를 갖도록 하였다. 또한 오목형과 블록형의 관계가 성립하도록 상수를 조정한 후 자료를 생성하였다.

이렇게 얻어진 자료를 $y^{(2)}$ 라 하면 이 자료가 2차년도 자료가 된다. 또한 1차년도 자료 $y^{(1)}$ 은 2차년도 자료 $y^{(2)}$ 에 $\text{Uni}(-10, 10)$ 에서 랜덤으로 얻어진 값을 더하여 생성하였다. 만약 $y^{(1)}$ 에 음수가 나오면 이 값은 '0.01'로 대체하였다. 모의실험은 다음의 두 방법으로 수행되었다.

모의실험 1 :

먼저 이상점은 2차년도 모집단 자료에서 랜덤으로 0과 50개 자료를 각각 선택한 후 선택된 각각의 자료에 5를 곱하여 생성되었다. 다음으로 표본 규모를 200 그리고 400으로 하여 랜덤으로 표본을 추출한 후 얻어진 표본자료에 Hidiroglou-Berthelot 방법을 적용하여 이상점을 판별하였다. 이 때 결측 자료 수는 20과 40을 사용하였으며 다중 대체를 위해 5세트의 자료가 생성되었다. 다중 대체를 위해서는 전 시점의 자료만 이용하였으며 따라서 대체에 사용된 회귀모형은

$$y_i^{(2)} = \beta_0 + \beta_1 y_i^{(1)} + \epsilon_i$$

이다. 이 방법은 Hidiroglou-Berthelot 방법을 이상점 판별법으로 사용하였을 때 각 이상점 보정 방법의 성능을 비교하기 위해 사용되었다.

모의실험 2 :

1차년도 자료, $y^{(1)}$ 과 x 를 독립변수로 한 후 2차년도 모집단 자료 $y^{(2)}$ 중에서 외표준화편차가 2이상인 자료에 5를 곱하여 이상점을 생성한다. 그리고 이상점의 수는 50개와 200개를 사용하였다. 또한 이상점을 만들지 않은 경우도 살펴보았다. 다음으로 표본 규모 200과 400을 사용해 표본을 추출한 후 표본 자료에 외표준화잔차와 Hidiroglou-Berthelot 방법을 각각 적용해 이상점을 판별하였다. 다중 대체를

Table 5.2. Results of H-B method with one independent variable and $n = 200$

이상점수	결측개수	자료형태	추정량							
			M1		M2		M3		M4	
			Abias	RMSE	Abias	RMSE	Abias	RMSE	Abias	RMSE
0	20	비례형	6.14	7.77	5.73	7.19	6.06	7.66	6.84	8.72
		선형	6.29	11.21	5.94	10.88	6.49	11.62	7.42	13.19
		오목형	6.27	12.75	5.94	12.48	7.02	14.63	8.42	18.82
		블록형	5.44	7.31	5.03	6.66	5.23	6.98	6.24	8.08
	40	비례형	6.25	11.12	5.85	10.75	6.46	11.61	7.52	13.48
		선형	6.21	7.95	5.82	7.43	6.15	7.86	6.80	8.75
		오목형	6.26	9.35	5.87	8.91	6.52	9.92	7.54	11.90
		블록형	5.48	7.24	5.03	6.54	5.30	6.98	6.32	8.21
50	20	비례형	8.86	37.41	8.79	37.39	16.95	41.98	27.25	52.45
		선형	10.02	46.57	10.25	46.63	22.19	52.98	35.65	65.72
		오목형	11.16	57.44	11.59	57.48	25.73	64.31	41.05	78.07
		블록형	6.62	20.91	6.32	20.71	10.19	23.90	16.65	31.44
	40	비례형	9.79	45.44	10.07	45.50	19.93	50.74	32.67	63.08
		선형	9.96	44.05	10.33	44.14	21.07	49.80	33.67	61.73
		오목형	11.06	54.42	11.70	54.59	24.85	61.28	39.80	75.13
		블록형	6.90	25.59	6.65	25.47	10.30	28.32	16.86	35.93

Table 5.3. Results of H-B method with one independent variable and $n = 400$

이상점수	결측개수	자료형태	추정량							
			M1		M2		M3		M4	
			Abias	RMSE	Abias	RMSE	Abias	RMSE	Abias	RMSE
0	20	비례형	6.09	7.49	5.64	6.87	5.95	7.31	6.59	8.15
		선형	6.12	7.48	5.77	6.98	6.13	7.50	6.84	8.51
		오목형	6.15	7.70	5.79	7.20	6.09	7.62	6.74	8.49
		블록형	5.45	7.12	5.04	6.48	5.20	6.72	6.09	7.65
	40	비례형	6.27	16.52	5.92	16.29	7.46	18.18	9.60	22.35
		선형	6.16	8.04	5.81	7.59	6.23	8.17	7.02	9.30
		오목형	6.16	7.60	5.78	7.07	6.07	7.48	6.68	8.30
		블록형	5.41	7.00	5.01	6.36	5.16	6.60	6.02	7.46
50	20	비례형	8.87	34.81	8.73	34.75	17.64	39.20	28.48	48.52
		선형	9.66	43.61	9.81	43.68	22.89	49.98	37.39	62.36
		오목형	10.41	48.69	10.61	48.75	24.37	54.91	39.44	67.17
		블록형	7.12	28.20	6.90	28.12	13.28	32.09	23.67	41.52
	40	비례형	8.99	34.89	8.91	34.88	18.00	39.45	29.18	48.95
		선형	9.82	44.42	10.07	44.49	22.97	50.69	37.39	63.11
		오목형	10.86	52.26	11.19	52.31	25.73	58.60	41.55	71.25
		블록형	7.06	25.29	6.81	25.17	11.88	28.60	20.52	36.81

위해 사용된 독립변수는 이상점 판별 방법과 무관하게 1차년도 자료 $y^{(1)}$ 과 x 이다. 따라서 회귀모형은

$$y_i^{(2)} = \beta_0 + \beta_1 y_i^{(1)} + \beta_2 x + \epsilon_i$$

이다. 이 때 결측 자료 수는 20, 40 그리고 60을 사용하였다. 또한 다중 대체를 위해 5세트의 자료가 생성되었다. 다음으로 본 연구에서 사용한 비교 통계량은 절대편향(absolute bias)과 RMSE이다. 각각의

Table 5.4. Results of H-B method with two independent variables and $n = 200$

이상점수	결측개수	자료형태	추정량							
			M1		M2		M3		M4	
			Abias	RMSE	Abias	RMSE	Abias	RMSE	Abias	RMSE
0	20	비례형	6.04	7.53	5.64	6.94	5.97	7.43	6.77	8.54
		선형	6.22	9.24	5.88	8.85	6.79	10.86	8.13	14.68
		오목형	6.29	10.73	5.91	10.37	6.72	11.70	7.88	14.33
		블록형	5.47	8.90	5.11	8.43	5.62	9.47	7.02	12.17
	40	비례형	6.17	8.19	5.71	7.60	6.11	8.15	6.93	9.33
		선형	6.25	8.80	5.85	8.33	6.33	9.00	7.20	10.46
		오목형	6.20	9.03	5.87	8.64	6.61	10.10	7.75	13.01
		블록형	5.45	7.22	5.09	6.62	5.24	6.85	6.09	7.70
	60	비례형	6.15	8.29	5.72	7.74	6.11	8.32	7.01	9.70
		선형	6.17	7.60	5.79	7.06	6.04	7.42	6.67	8.27
		오목형	6.35	12.08	5.99	11.77	6.68	12.88	7.78	15.29
		블록형	5.45	7.65	5.09	7.09	5.37	7.57	6.46	9.04
50	20	비례형	8.91	37.56	8.90	37.54	17.90	42.78	28.74	53.76
		선형	10.22	49.43	10.56	49.51	23.58	56.03	38.11	69.78
		오목형	10.62	55.35	11.25	55.46	27.35	63.77	45.03	80.86
		블록형	6.73	22.75	6.46	22.59	10.18	25.36	16.54	32.29
	40	비례형	8.96	36.39	9.03	36.42	16.95	41.08	27.07	51.38
		선형	9.27	38.80	9.51	38.85	19.13	44.35	30.46	55.78
		오목형	10.29	47.39	10.82	47.51	23.15	54.21	37.25	68.12
		블록형	6.92	23.54	6.77	23.44	10.49	26.18	17.18	33.30
	60	비례형	8.50	31.66	8.57	31.71	14.41	35.47	22.59	44.35
		선형	9.85	43.24	10.40	43.40	19.61	48.63	31.11	60.20
		오목형	10.72	50.12	11.70	50.35	23.93	56.95	38.60	71.53
		블록형	7.22	28.50	7.21	28.48	11.08	31.43	18.33	39.67
200	20	비례형	15.52	63.49	16.23	63.57	35.84	71.93	58.67	91.23
		선형	20.34	86.35	22.12	86.55	50.98	98.23	82.39	124.00
		오목형	24.13	100.54	26.16	100.73	58.63	112.99	92.23	139.51
		블록형	11.98	56.81	12.26	56.84	24.94	62.41	44.51	77.63
	40	비례형	17.72	76.10	19.29	76.32	41.12	85.12	68.37	107.31
		선형	19.85	80.60	21.88	80.85	47.26	91.24	76.20	115.15
		오목형	24.38	105.62	27.50	106.00	59.55	118.13	95.67	146.56
		블록형	12.83	58.86	13.59	58.98	27.80	65.67	51.08	84.67
	60	비례형	17.70	74.35	19.88	74.73	40.01	83.39	66.11	104.97
		선형	20.38	84.60	23.57	85.07	48.83	95.63	79.76	121.49
		오목형	23.78	101.67	27.85	102.18	57.61	114.14	93.10	142.76
		블록형	11.99	60.30	12.95	60.53	24.54	65.92	44.72	82.68

정의는 다음과 같다. 여기서 k 는 결측 개수이며 반복수 $R = 1,000$ 을 사용하였다.

$$\text{Absolute bias} : \frac{1}{1000} \frac{1}{5k} \sum_{r=1}^{1000} \sum_{i=1}^k \sum_{j=1}^5 \left| y_{i,j}^{(r)} - \hat{y}_i^{(r)} \right|,$$

$$\text{RMSE} : \left(\frac{1}{1000} \frac{1}{5k} \sum_{r=1}^{1000} \sum_{i=1}^k \sum_{j=1}^5 \left(y_{i,j}^{(r)} - \hat{y}_i^{(r)} \right)^2 \right)^{\frac{1}{2}}.$$

Table 5.5. Results of H-B method with two independent variables and $n = 400$

이상점수	결측개수	자료형태	추정량							
			M1		M2		M3		M4	
			Abias	RMSE	Abias	RMSE	Abias	RMSE	Abias	RMSE
0	20	비례형	6.04	9.32	5.65	8.92	6.49	10.05	7.84	12.34
		선형	6.15	8.32	5.79	7.85	6.23	8.44	6.97	9.43
		오목형	6.18	8.02	5.82	7.54	6.16	7.98	6.84	8.88
		블록형	5.37	7.21	5.03	6.65	5.21	6.94	6.09	7.79
	40	비례형	6.06	7.69	5.65	7.13	5.99	7.61	6.79	8.66
		선형	6.22	10.17	5.86	9.80	6.51	10.58	7.59	12.16
		오목형	6.24	9.91	5.88	9.54	6.70	10.65	7.91	12.76
		블록형	5.48	12.38	5.20	12.12	6.00	13.27	7.85	16.24
	60	비례형	6.16	12.48	5.80	12.21	6.92	13.77	8.63	17.20
		선형	6.16	8.71	5.81	8.30	6.28	8.89	7.18	10.17
		오목형	6.34	14.59	6.00	14.37	6.96	15.52	8.37	18.00
		블록형	5.39	7.68	5.06	7.16	5.31	7.52	6.32	8.59
50	20	비례형	8.73	36.33	8.70	36.29	18.97	41.90	31.19	52.95
		선형	10.05	44.69	10.09	44.71	21.56	49.95	34.17	59.92
		오목형	10.50	49.40	10.78	49.49	26.40	56.85	42.94	70.87
		블록형	6.81	28.70	6.63	28.61	11.71	31.97	20.36	40.54
	40	비례형	9.18	41.63	9.17	41.62	18.77	46.20	30.55	55.95
		선형	9.81	44.40	10.11	44.47	23.70	51.18	38.80	64.84
		오목형	10.18	49.10	10.59	49.19	25.65	56.48	42.13	70.77
		블록형	6.86	27.89	6.73	27.82	11.84	31.40	20.62	40.45
	60	비례형	8.75	35.01	8.81	35.04	18.14	40.22	29.97	51.35
		선형	9.97	46.47	10.36	46.56	23.41	52.80	38.29	65.58
		오목형	10.61	53.13	11.26	53.28	27.12	60.84	44.68	76.25
		블록형	6.58	21.66	6.36	21.52	9.73	23.84	15.76	29.45
200	20	비례형	17.18	71.47	17.96	71.54	40.95	80.89	67.17	101.53
		선형	20.53	86.15	21.86	86.31	50.29	97.05	80.60	120.70
		오목형	25.05	105.98	26.91	106.13	60.76	117.97	95.13	143.46
		블록형	10.79	45.73	10.78	45.68	21.56	50.45	37.70	62.16
	40	비례형	17.41	72.80	18.39	72.95	40.85	81.84	66.76	102.05
		선형	20.92	89.56	22.69	89.82	52.10	100.95	83.86	125.36
		오목형	24.07	103.34	26.24	103.59	59.48	115.70	94.00	141.80
		블록형	12.58	58.94	12.95	58.99	27.37	64.77	49.57	80.69
	60	비례형	16.88	69.50	17.87	69.62	38.41	77.70	62.74	96.53
		선형	21.09	89.76	23.11	90.02	51.47	100.89	82.75	124.88
		오목형	24.67	105.02	27.23	105.26	60.06	117.19	95.41	144.09
		블록형	13.48	73.22	14.38	73.37	31.68	80.29	59.42	100.84

Table 5.2와 Table 5.3은 Hidioglou-Berthelot 방법을 이용하여 이상점을 판별한 후 전 시점의 자료를 독립변수로 하여 대체값을 구한 결과이다. 표에서 M1은 2절에서 설명한 방법 1을 의미하며 M2 그리고 M3은 각각 방법 2와 3을 의미한다. M4는 무응답 가중치 보정을 실시하지 않은 것을 의미한다.

모의실험 1의 결과인 Table 5.2와 Table 5.3을 살펴보면 다음과 같다. 먼저 Table 5.2에서 이상점을 만들지 않은 경우를 살펴보면 이상점 판별과 이상점 보정을 하지 않은 M4에 비해 이상점 판별을 실시하고 이상점 가중치 보정을 실시한 모든 방법에서 매우 우수한 결과를 주고 있음을 확인할 수 있다. 이상

Table 5.6. Results of SDR method with two independent variables and $n = 200$

이상점수	결측개수	자료 형태	추정량							
			M1		M2		M3		M4	
			Abias	RMSE	Abias	RMSE	Abias	RMSE	Abias	RMSE
0	20	비례형	6.46	8.73	6.47	8.73	6.66	9.01	7.06	9.86
		선형	6.62	8.92	6.62	8.95	6.89	9.38	7.32	10.58
		오목형	6.51	8.58	6.43	8.48	6.59	8.70	6.98	9.30
		블록형	5.12	6.66	5.13	6.62	5.20	6.73	6.00	7.48
	40	비례형	6.47	8.24	6.45	8.22	6.50	8.28	6.68	8.56
		선형	6.67	9.26	6.66	9.25	6.81	9.49	7.15	10.31
		오목형	6.45	8.14	6.35	8.02	6.47	8.16	6.75	8.55
		블록형	5.16	8.42	5.19	8.43	5.51	9.07	6.87	11.64
	60	비례형	6.56	12.01	6.60	12.06	6.98	12.75	7.73	15.23
		선형	6.65	8.76	6.64	8.76	6.69	8.83	6.85	9.10
		오목형	6.48	9.14	6.42	9.09	6.71	9.62	7.34	11.26
		블록형	5.17	7.67	5.17	7.64	5.29	7.83	6.26	8.86
50	20	비례형	9.59	41.56	10.31	41.80	17.28	45.50	29.57	57.15
		선형	10.33	41.95	11.07	42.10	19.24	46.37	32.21	58.21
		오목형	12.11	61.61	13.56	61.92	26.05	67.57	44.76	83.25
		블록형	7.36	32.29	7.94	32.52	12.20	35.64	22.60	47.52
	40	비례형	9.45	40.14	10.47	40.49	17.43	44.86	30.16	59.04
		선형	10.25	41.66	11.20	41.92	18.83	46.10	31.43	58.06
		오목형	11.33	52.44	12.86	52.80	23.60	58.41	40.25	74.23
		블록형	7.28	32.48	7.85	32.67	11.43	35.11	20.41	44.80
	60	비례형	10.02	45.20	11.35	45.62	18.27	49.93	31.23	63.90
		선형	10.18	43.43	11.55	43.83	19.40	48.40	32.46	61.70
		오목형	11.77	58.32	14.12	58.96	25.85	65.29	44.66	83.53
		블록형	6.97	30.65	7.49	30.83	10.22	32.79	17.32	41.20
200	20	비례형	20.57	68.47	21.94	68.78	36.42	75.10	61.61	94.83
		선형	26.78	92.26	28.91	92.64	49.98	101.07	82.76	126.08
		오목형	27.78	107.08	31.16	107.62	59.25	118.59	99.77	150.14
		블록형	11.83	48.65	12.99	48.92	21.95	53.22	41.74	68.97
	40	비례형	21.78	78.73	23.64	79.20	38.36	85.50	65.17	106.06
		선형	24.92	86.21	27.61	86.83	47.53	95.15	79.37	120.41
		오목형	26.86	104.72	30.57	105.33	55.36	115.17	92.56	143.37
		블록형	14.28	75.96	15.91	76.38	25.83	80.66	49.03	96.87
	60	비례형	20.83	73.23	23.18	73.86	37.08	80.24	63.46	101.49
		선형	25.34	89.37	28.72	90.18	48.06	98.54	80.61	123.99
		오목형	28.84	112.04	34.29	113.06	60.40	123.72	101.40	155.18
		블록형	14.32	60.48	16.43	60.98	26.29	65.44	49.72	83.59

점 보정 방법 중에서는 M3의 결과가 가장 좋지 않게 나왔으며 M1과 M2는 유사한 성능을 보이고 있다. 이러한 결과는 결측 개수 및 모집단 자료 형태와 무관하게 나타나고 있다. 또한 이상점의 수가 50개인 경우에도 유사한 결과를 주고 있다. 다만 이상점 수가 많기 때문에 전체적으로 정도는 나빠진다. 또한 표본수가 400인 경우의 결과인 Table 5.3에서도 Table 5.2와 유사한 결과가 얻어졌다. 따라서 독립변수가 1개인 경우 Hidiroglou-Berthelot 방법을 사용하여 이상점을 판별하고 이상점 보정법을 사용하여 다중대체를 실시한다면 우수한 대체 결과를 얻을 수 있을 것으로 판단된다.

Table 5.7. Results of SDR method with two independent variables and $n = 400$

이상점수	결측개수	자료형태	추정량							
			M1		M2		M3		M4	
			Abias	RMSE	Abias	RMSE	Abias	RMSE	Abias	RMSE
0	20	비례형	6.55	11.98	6.57	11.98	6.98	12.53	7.82	14.11
		선형	6.58	8.11	6.57	8.11	6.89	8.62	7.47	10.07
		오목형	6.48	10.19	6.45	10.18	6.88	10.74	7.75	12.38
		블록형	5.10	6.81	5.12	6.78	5.19	6.89	6.10	7.77
	40	비례형	6.38	8.05	6.39	8.06	6.45	8.14	6.69	8.45
		선형	6.59	9.19	6.60	9.20	6.84	9.55	7.33	10.57
		오목형	6.46	11.00	6.46	11.00	7.13	12.01	8.33	14.85
		블록형	5.08	6.71	5.07	6.66	5.19	6.83	6.11	7.73
	60	비례형	6.46	9.03	6.47	9.05	6.68	9.33	7.19	10.24
		선형	6.59	8.34	6.57	8.32	6.62	8.37	6.72	8.52
		오목형	6.58	15.21	6.62	15.24	7.70	16.81	9.51	21.33
		블록형	5.22	13.89	5.31	13.92	6.04	15.06	8.28	19.68
50	20	비례형	9.44	39.41	9.97	39.57	16.43	42.60	27.83	51.86
		선형	11.28	52.04	12.33	52.33	24.14	57.69	42.27	72.27
		오목형	11.00	55.74	12.26	56.04	24.88	61.69	43.88	77.17
		블록형	6.91	27.23	7.33	27.40	11.40	30.00	21.32	39.57
	40	비례형	9.43	36.09	10.09	36.28	17.07	39.77	29.64	50.62
		선형	10.40	45.25	11.32	45.47	21.07	50.01	36.37	62.72
		오목형	11.96	60.12	13.48	60.45	27.31	66.38	47.99	82.43
		블록형	6.88	23.68	7.24	23.82	10.65	26.03	19.38	34.12
	60	비례형	8.64	27.71	9.11	27.89	14.10	30.80	23.36	39.12
		선형	10.38	45.54	11.45	45.84	21.38	50.55	37.22	63.78
		오목형	10.79	47.86	12.03	48.16	23.26	53.50	40.61	67.95
		블록형	6.68	26.11	7.10	26.29	10.61	28.44	19.54	36.50
200	20	비례형	20.15	70.31	21.51	70.55	37.07	77.03	63.82	97.36
		선형	25.18	86.55	26.72	86.79	46.48	94.53	77.55	117.55
		오목형	27.25	108.28	30.51	108.81	59.28	119.11	99.57	149.63
		블록형	12.15	49.71	13.26	49.93	23.19	54.47	45.10	71.13
	40	비례형	19.31	65.10	20.63	65.42	35.37	71.79	60.36	91.05
		선형	24.90	87.85	26.95	88.28	47.85	96.60	80.33	120.92
		오목형	28.43	113.07	32.00	113.73	60.62	124.11	102.09	153.83
		블록형	13.29	55.57	14.68	55.90	25.80	61.03	51.24	80.93
	60	비례형	22.07	78.49	24.02	78.97	41.30	85.96	71.31	108.13
		선형	23.50	81.27	25.58	81.64	44.77	89.22	74.51	111.86
		오목형	28.54	108.85	32.01	109.47	58.75	119.59	98.44	148.60
		블록형	13.54	56.72	14.94	57.01	25.09	61.30	48.11	77.48

Table 5.4에서 Table 5.7은 독립변수가 두 개인 모의실험 2의 결과를 정리한 것이다. 이중에서 Table 5.4와 Table 5.5는 Hidiroglou-Berthelot 방법을 사용하여 이상점을 판별한 후 얻은 이상점 보정 방법의 성능비교 결과이다. 또한 Table 5.6과 Table 5.7은 회귀모형의 외표준화잔차로 이상점을 판별한 후 얻은 이상점 보정 방법의 성능비교 결과이다.

Hidiroglou-Berthelot 방법을 이용하여 이상점을 판별을 하고, 독립변수로 전 시점 자료와 횡시점 자료를 사용하여 다중 대체를 실시한 결과인 Table 5.4와 Table 5.5의 결과를 살펴보면 다음과 같다. 먼저

Table 6.1. Comparison results of real data analysis using Monthly Labor Statistics

이상점 판별 방법	결측개수	Method							
		M1		M2		M3		M4	
		Abias	RMSE	Abias	RMSE	Abias	RMSE	Abias	RMSE
외표준화 잔차	100	3666	6542	3624	6523	4283	6850	5605	7914
	200	3672	6535	3638	6513	4292	6850	5630	7943
H-B방법	100	2193	7032	2180	7044	4146	6781	5629	7889
	200	2145	6773	2133	6783	4096	6623	5611	7884

Table 5.4에서 이상점의 수에 상관없이 이상점 판별과 이상점 가중치 보정법을 이용할 경우 대체의 정도가 우수한 것을 확인할 수 있다. 즉 M4에 비해 다른 모든 이상점 보정 방법에서 우수한 결과를 주고 있다. 이 중에서 M3의 결과가 가장 좋지 않게 나타났다. 또한 M1과 M2의 비교를 살펴보면 미미하지만 이상점의 수가 작은 경우에는 M2 방법이 우수하고 이상점 수가 많은 경우에는 M1 방법이 우수한 것을 확인할 수 있다. 이러한 결과는 결측개수 및 모집단 형태와 무관하게 유사한 결과를 주고 있으며 표본 수가 400인 경우에도 유사한 결과를 주고 있다. 또한 이 결과는 독립변수가 하나인 경우인 Table 5.2와 Table 5.3의 결과와도 일치한다.

Table 5.6과 Table 5.7은 외표준화잔차를 이상점 판별 방법으로 사용한 후 전 시점자료와 횡시점 자료를 이용하여 다중 대체한 결과이다. 먼저 Table 5.6의 결과를 살펴보면 이상점의 수와 결측 개수에 상관없이 이상점 보정을 실시한 결과인 M1, M2, M3가 이상점 보정을 실시하지 않은 M4에 비해 모두 우수한 결과를 주고 있음을 확인할 수 있다. 이상점의 수가 0인 경우에는 M2 방법이 미미하지만 우수한 결과를 주고 있으나 이상점의 수가 50, 200인 경우에는 M1 방법이 가장 우수한 결과를 주고 있다. 이러한 결과는 Table 5.7에서도 확인할 수 있다.

결론적으로 본 연구에서 설명한 이상점 판별 방법을 이용하여 이상점 판별을 실시하고, 이상점 보정방법으로 M1 또는 M2 방법을 사용한다면 이상점의 영향력을 줄여 매우 우수한 대체 결과를 얻을 수 있을 것으로 판단된다.

6. 매월노동통계 자료 분석

이상점의 영향력을 줄인 후 대체 결과의 우수성을 확인하기 위한 실제 자료 분석을 위해 2007년 매월노동통계 자료 중 t 월의 총임금과 종사자 수 그리고 $t-1$ 월의 총임금 자료 약 7,000개 중 규모 1인 자료 1,600 여개가 사용되었다. 이상점 판별 방법에서 Hidiroglou-Berthelot 방법은 $t-1$ 월의 총임금 자료와 t 월의 총임금 자료가 사용되었으며 외표준화잔차를 이용한 이상점 판별에서는 두 독립변수가 모두 사용되었다. 다중 대체를 위해 이상점 판별 방법과 상관없이 두 개의 독립변수 즉 t 월의 종사자수와 $t-1$ 월의 총임금이 사용되었다. 자료에는 무응답이 존재하지 않아 분석 대상 자료 1660여개에서 100와 200개의 자료를 결측 처리하였으며 사용된 가중치는 400이다. 또한 5개의 완전자료가 비교 통계량 산출에 사용되었으며 반복수는 1,000이다. 결과는 Table 6.1에 수록되었다.

자료분석 결과를 살펴보면 이상점 처리를 하지 않은 M4에 비해 모든 방법이 우수한 것을 확인할 수 있다. 외표준화잔차를 사용한 결과를 살펴보면 M1과 M2가 M3에 비해 우수한 결과를 주고 있다. 또한 미미하지만 M2가 M1에 비해 우수한 결과를 주고 있다. 또한 Hidiroglou-Berthelot 방법을 사용한 결과를 살펴보면 M1과 M2가 편향을 기준으로 우수한 결과를 주고 있는 반면 RMSE 기준으로는 M3 방법이 우수한 결과를 주고 있음을 확인할 수 있다. 이제 두 이상점 판별법과 가중치 보정법을 종합하면 RMSE 기준으로는 외표준화잔차를 사용하고, M2를 사용하는 것이 가장 우수한 결과를 주고 있으며

편향을 기준으로 한다면 Hidiroglou-Berthelot 방법을 사용하고 M2 방법을 사용하는 것이 가장 우수한 결과를 주는 것을 확인 할 수 있다.

7. 결론

본 논문에서는 이상점 판별을 위해 회귀분석에서 흔히 사용하는 이상점 판별법인 외표준화잔차와 전 시점 자료만을 이용하여 이상점을 판별하는 Hidiroglou-Berthelot 방법을 사용하였다. 외표준화잔차는 SAS PROC/REG를 이용하여 쉽게 얻을 수 있으나 Hidiroglou-Berthelot 방법은 이상점 판별 구간을 정하기 위하여 몇 개의 모수를 추정해야 하는 등의 번거로움이 있다. 모의실험 결과 Hidiroglou-Berthelot 방법 또는 외표준화잔차를 사용하여 이상점을 판별한 후 본 논문에서 연구한 가중치 보정 방법을 사용하면 다중 대체 결과의 성능이 크게 향상되는 것을 확인할 수 있다. 특히 이상점의 수가 많은 경우에는 M1 방법이 그리고 이상점의 수가 적은 경우에는 M2 방법이 미미하지만 우수한 결과를 주는 것을 확인하였다. 또한 이러한 결과는 실제 자료 분석에서도 확인되었다. 결론적으로 이상점과 무응답이 동시에 존재하는 경우에는 이상점을 적절히 처리한 후 무응답 대체법을 사용하면 우수한 결과를 얻을 수 있음이 확인되었다. 다만 본 논문에서 사용된 Hidiroglou-Berthelot 방법에서 모수 U 와 C 의 값을 고정하여 모의실험이 수행되었기 때문에 본 연구에서 얻어진 Hidiroglou-Berthelot 방법의 결과를 일반화하는 것은 주의해야 한다.

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.
- Hidiroglou, M. A. Srinath, K. P. (1981). Some estimators of a population total from simple random samples constraining large units, *Journal of Applied Statistics*, **76**, 690–695.
- Kim, J. (2006). Weight reduction method for outlier in survey sampling, *The Korean Communications in Statistics*, **13**, 19–27.
- 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**, 843–858.
- McCullough, M. and Pennington, T. L. (2009). Identifying outliers when creating an imputation based for the Quarterly Financial Report, *JSM, Section on Survey Research Methods*.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Survey*, New York.

가중치 보정을 이용한 다중대체법

김진영^a · 신기일^{a,1}

^a한국외국어대학교 통계학과

(2013년 4월 27일 접수, 2013년 6월 17일 수정, 2013년 6월 18일 채택)

요약

다중 대체법은 표본조사에서 결측값이 발생하였을 때 가장 흔히 사용하는 방법이다. 이 방법은 여러 요인에 의해 그 성능이 좌우되며 특히 이상점의 영향을 많이 받는다. 본 연구에서는 가중치 보정법을 이용하여 이상점의 영향력을 줄여 다중 대체법의 성능을 향상시키는 방법을 연구하였다. 가중치 보정법을 이용하여 얻어진 최종 가중치를 다중 대체에 사용하였으며 SAS의 PROC MI가 다중 대체를 위해 사용되었다. 모의실험과 매월노동통계 자료를 이용한 실제 자료 분석을 통하여 제안된 방법의 우수성을 확인하였다.

주요용어: 이상점 판별법, Hidioglou-Berthelot 방법, 외표준화잔차.

이 논문은 2013년도 한국외국어대학교 학술연구비 지원에 의해 이루어진 것임.

¹교신저자: (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 통계학과, 교수.

E-mail: keyshin@hufs.ac.kr