Multiple imputation reducing outlier effect by weight adjustment method

Jinyoung Kim*
Hankuk University of Foreign Studies, Seoul, Republic of Korea,
suwonfis@naver.com
Key-Il Shin
Hankuk University of Foreign Studies, Seoul, Republic of Korea,
keyshin@hufs.ac.kr

Abstract

Imputation is the most used method for handling missing values in survey. In this paper we investigate multiple imputation methods that use outlier weight adjustment method in order to reduce the effect of outlier. A regression method in PROC/MI in SAS is used for the multiple imputation method and obtained final adjusted weight is used as a weight variable to obtain the imputed values. Simulation studies are conducted to compare the performance of various weight adjustment methods and outlier detection methods. The Korea monthly labor statistics are used for real data analysis.

Key Words : Business survey, non-response, outlier, regression method

1. Introduction

It is well known that survey results are affected by errors arising from several sources. Among them, one crucial effect could be introduced by unit and item non-responses with the consequence that the last could produce bias and distortions of distributions. Another problem for misleading analysis result is the existence of outlier. Sometimes the effect of outlier may be great enough not to interpret correctly the result. That is, without proper treatment of outliers, the estimate of total may be under or over estimated.

Practically in survey data, both non-response and outliers occur simultaneously and these two factors are crucial to produce non-sampling errors. So in this paper, we investigate simple and appropriate imputation methods which reduce this effect of outlier using the weight adjustment.

Outlier detection methods usually depend on the selected model to detect outliers. Since a regression method is widely used for imputation, it is proper to use a detection method based on regression model. Also for the panel type data, a detection method to use the information of previous wave should be used. Therefore in this study, two methods are used to detect outliers. The first one is to use the studentized deleted residual statistic widely used for regression model and the second one is Hidiroglou-Berthelot (1986) outlier detection method which is usually used for panel type data.

After detection of outliers, we assign a weight to each value according to outliers or non-outliers. In this paper we consider three weight adjustment methods which reduce weights assigned to outliers. And then imputation is conducted with transformed variables applied by the final weight and we re-transform the obtained imputed values to get final imputed values. Small simulation studies are conducted to compare the superiority of results obtained by two outlier detection methods and three weight adjustment methods. For real data analysis, the Korea monthly labor statistics are used and the regression method SAS/MI is used.

In this paper we assume that the missing pattern is missing at random(MAR). Section 2 and section 3 presents briefly outlier detection method and weight adjustment methods respectively. In section 4, we illustrate the multiple imputation method which uses the transformation with finally adjusted weight. In section 5, we conduct a small

Monte-Carlo simulation to compare the performance of weight adjustment methods and outlier detection methods. Finally summary and discussions are in section 6.

2. Outlier detection

For outlier detection, studentized deleted residuals and Hidiroglou-Berthelot method are used in this study. Brief reviews are followings.

1) Studentized deleted residual method

In this section we consider studentized deleted residual statistic, which is easily obtained and printed out in SAS/REG

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

Where $d_i = y_i - \hat{y}_i$ and $\hat{y}_i$ denote the observed response for the i-th observation and h_{ii} is the leverage.

2) Hidiroglou-Berthlot method

Hidiroglou-Berthlot (1986) suggested a method using the ratio of the last wave and current wave. The simple procedure is followings.

1. Let $x_i(t-1)$ be the i-th variable value in $(t-1)$-th wave and $x_i(t)$ be the i-th variable value in t-th wave. Then calculate the ration r_i and s_i defined by

$$r_i = \frac{x_i(t)}{x_i(t-1)} \text{ and } s_i = \begin{cases} 1 - \frac{r_M}{r_i}, if\ 0 < r_i < r_M \\ \frac{r_i}{r_M} - 1, \qquad if\ r_M \leqq r_i \end{cases}$$

where r_M is the median of r_i

2. In order to consider the magnitude of values in each waves, we calculate E_i

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

and then the acceptance region is defined by

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

where U, $0 \leqq U \leqq 1$, is a parameter to count on the effect of the magnitude of values and in this study we used $U = 0.4$. Also E_M is the median of E_i, E_{Q1} and E_{Q3} are the first and third quartiles and $d_{Q1} = E_M - E_{Q1}$ and $d_{Q3} = E_{Q3} - E_M$. In this study we used $C = 1.5$.

3. Weight adjustment method

Let w be the design weight assigned to values and f be a outlier weight adjustment factor. Even the final weight, w^f, is obtained by $w^f = w \times f$. In this study following weight adjustment factors are used.

Method 1. For the outliers, we use $f = 0$. Then the final weight is

$$w^f = \begin{cases} w^f = 0, & for\ outliers \\ w\left(\frac{n}{n-k}\right) = w\left(1 + \frac{k}{n-k}\right) & for\ non-outliers \end{cases}$$

Method 2. For the outliers, we use $f = \frac{1}{w}$. Then the final weight is

$$w^f = \begin{cases} w \times \frac{1}{w} = 1 & for\ outliers \\ w\left(1 + \frac{k(w-1)}{w(n-k)}\right) & for\ non-outliers \end{cases}$$

Method 3. Following weight adjustment is suggested by Hidiroglou and Srinath (1981)

$$w^f = \begin{cases} w\left(1 - \frac{n-k}{2n}\right) & for\ outliers \\ w\left(1 + \frac{k}{2n}\right) & for\ non-outliers \end{cases}$$

4. Multiple imputation method

Multiple imputation method is conducted with the final adjusted weight assigned to each value. For this, first consider multiple regression model defined by

$$y = \mathrm{X}\beta + \epsilon$$

where y is a interesting variable vector, X is a design matrix of auxiliary variables and ϵ is an error vector. Also let W be a final weight matrix. Then multipling W to the model, we have $y^* = \mathrm{X}^*\beta^* + \epsilon^*$. Then the weighted least squares estimates of β is obtained by $\widehat{\beta^*} = (\mathrm{X}'W^2\mathrm{X})^{-1}\mathrm{X}'W^2y$. Using these estimates, and usual multiple imputation method, we obtain multiple imputed values. For this, we use SAS/MI with regression method. Finally using re-transformation by multiplication of W^{-1}, we have the final imputed values. That is we have

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

and the i-th imputed value is obtained by $\hat{y}_i = x_i\hat{\beta}$.

5. Simulation Study

For simulation study the population data is generated using the same steps as those in Lee et al. (1995). The population of auxiliary variable x_i, with size N=5,000 is generated with mean 48 and variance 768 from gamma distribution. Given x_i, we generate the interesting variable y_i with mean function $\mu = \mathrm{a} + \mathrm{b}x + cx^2$ and variance $\sigma^2 = d^2x^{2g}$ from gamma distribution. The constants used in this study can be found in Lee et al. (1995).

Let $y^{(2)}$be the second wave variable and x be the auxiliary variable. Then the first wave variable $y^{(1)}$ is generated by adding disturbance from $Uni(-10,10)$ to $y^{(2)}$. If $y^{(1)}$is less than zero, we replace them with 0.01.

For generating outliers, we randomly select 0% and 1% samples and multiply 5 to $y_i^{(2)}$. Also we use two well-known comparison statistics, Absolute bias and Root MSE. In this study we use replication number R=1,000. Also we generate 5 imputed data sets.

In this simulation, studentized deleted residual method is used under the assumption that we have two auxiliary variables, the $t-1$-th wave variable $y^{(1)}$and independent variable x. On the other hand, Hidiroglou-Berthlot method is used under the assumption of existence of the only one auxiliary variable $y^{(1)}$. Therefore the direct comparison of two methods should be avoided.

Table 5.1 and Table 5.2 show the results of number of non-responses 60. M1 through M3 stand for the weight adjustment methods explained in section 3 respectively. For example, M1 means method 1. Also M4 stands for the result of no consideration of outler adjustment.

Table 5.1 Studentized deleted residual method

Number of outliers	Variable type	Estimator							
		M1		M2		M3		M4	
		Abias	MSE	Abias	MSE	Abias	MSE	Abias	MSE
0%	Ratio	6.56	12.01	6.60	12.06	6.98	12.75	7.73	15.23
	Linear	6.65	8.76	6.64	8.76	6.69	8.83	6.85	9.10
	Concave	6.48	9.14	6.42	9.09	6.71	9.62	7.34	11.26
	Convex	5.17	7.67	5.17	7.64	5.29	7.83	6.26	8.86
1%	Ratio	10.02	45.20	11.35	45.62	18.27	49.93	31.23	63.90
	Linear	10.18	43.43	11.55	43.83	19.40	48.40	32.46	61.70
	Concave	11.77	58.32	14.12	58.96	25.85	65.29	44.66	83.53
	Convex	6.97	30.65	7.49	30.83	10.22	32.79	17.32	41.20

Table 5.2 Hidiroglou-Berthlot method

Number of outliers	Variable type	Estimator							
		M1		M2		M3		M4	
		Abias	MSE	Abias	MSE	Abias	MSE	Abias	MSE
0%	Ratio	6.33	12.63	5.91	12.29	6.56	13.14	7.67	15.12
	Linear	6.19	7.92	5.82	7.43	6.13	7.84	6.82	8.83
	Concave	6.39	15.83	6.12	15.66	7.34	17.96	9.03	22.99
	Convex	5.48	7.10	5.03	6.39	5.18	6.63	6.03	7.48
1%	Ratio	9.17	37.98	9.53	38.09	17.86	43.09	29.11	54.74
	Linear	9.98	44.01	10.51	44.17	19.42	48.93	30.86	60.15
	Concave	10.47	53.62	11.58	53.90	23.96	60.77	39.14	76.45
	Convex	6.56	21.63	6.36	21.52	9.55	24.50	15.56	32.19

The results of Table 5.1 show that outlier adjustment methods improve the precision of imputed values. Especially M1 and M2 methods have a great improvement regardless of the four population types. Also as the number of outliers increases, the effect of outlier adjustment increases. Table 5.2 shows the similar results to Table 5.1

6. Real data analysis

Total wage and number of employers of the Korea monthly labor statistic are used for comparison of superiority of weight adjustment methods after treating outliers. $t-1$-th month total wage and t-th month number of employers are used as auxiliary variables and t-th total wage is the interesting variable.

Among about 1,670 data we make n=100,200 missing values and 3 weight adjustment methods and 2 outlier detecting methods are used. Multiple imputation method is used for imputation and 5 data sets are obtained. The design weight is about 400.

Table 6.1 Comparison results of weight adjustment methods

Detecting method	number of missing values	Method							
		M1		M2		M3		M4	
		Abias	RMSE	Abias	RMSE	Abias	RMSE	Abias	RMSE
Studentized deleted residual	100	3666	6542	3624	6523	4283	6850	5605	7914
	200	3672	6535	3638	6513	4292	6850	5630	7943
H-B method	100	2193	7032	2180	7044	4146	6781	5629	7889
	200	2145	6773	2133	6783	4096	6623	5611	7884

This result shows that all outlier adjustment methods are superior to M4. Using the studentized deleted residual method, M1 and M2 show better results than M3. However using H-B method and RMSE criterion, M3 is better than M1 and M2. Also if we want to have minimum Abias, then it is better to use H-B method and M2 and if we want to have minimum RMSE then it is better to use studentized deleted residual method and M2.

7. Conclusion and Summary

In this study we investigate multiple imputation using three weight adjustment methods that reduce the influence of outliers. Two outlier detecting methods, studentized deleted residual method and Hidiroglou-Berthlot method are used and compared. Using small simulation studies, the three weight adjustment methods for multiple imputation method are compared and the results shows that the superiority of weight adjustment methods do not depend on the underlying population types and outlier detection methods. Therefore if outlier exists in data set, in order to obtain more accurate imputed values, it is better to treat them properly.

References

[1] Hidiroglou, M. A. and Srinath, K. P. (1981) "some estimators of a population total from simple random samples constraining large units," *Journal of Applied statistics*, Vol. 76, 690-695

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

[3] 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.

[4] SAS, SAS Institute Inc., Cary, NC, USA.