Journal of the Korean Data & Information Science Society
2016, **27**(1), 131–142

http://dx.doi.org/10.7465/jkdi.2016.27.1.131
한국데이터정보과학회지

경제조사에서의 이상치 탐지와 처리방법

주영선[1] · 조교영[2]

[12]경북대학교 통계학과

접수 2016년 1월 4일, 수정 2016년 1월 12일, 게재확정 2016년 1월 13일

요 약

통계조사에서 이상치는 총계추정에 큰 영향을 줄 수 있다. 통계조사에서 보고된 값은 극단적이 아니지만 그것의 가중치 (weight)가 커서 추정값에 큰 영향을 주거나, 극단값이라 해도 그것이 작은 가중치를 가질 때 추정에 큰 영향을 주지 않는 경우도 있다. 이러한 극단값이나 추정에 영향을 주는 값들은 표본조사에서 민감하다. 일반적으로 치우친 분포를 가진 모집단에서 추출된 표본으로 조사를 하는 사업체 조사에서는 특별히 더 큰 영향을 준다. 본 연구에서는, 우리는 이상치를 판별하고 처리하는 방법에 대해서 다루고자 한다. 이상치 판별은 분위수에 기초해서 판정하였으며, 판정된 이상치는 여러 가지 다양한 방법을 적용해 보았다. 연구에서는 2가지 winsorised 방법과 세가지 cut-off 방법에 대하여 적용하였다. 그리고 시뮬레이션에서는 4가지 방법의 가중치를 각각 적용하여 진행하였다. 여러 가지 이상치 처리방법들을 비교해 본 결과 type I 윈저화 방법보다는 type II 윈저화 방법이 효율적인 결과값을 보여주었으며, 가중치 변환방법들 중에서는 제곱근 변환을 통한 가중치 감소방법이 다른 처리방법에 비해 좋은 결과값을 보여주었다.

주요용어: 가중치 축소, 윈저화 방법, 이상치 탐지, 이상치 처리.

1. 서 론

이상치 (outlier)란 일부 자료값들이 대다수의 다른 표본에 비해 매우 크거나 작은 극단적인 값 (extreme observation)을 갖는 것을 말한다. 표본조사에서이상치를 포함한 경우 모집단의 평균이나 총합을 추정하는 과정에서 이상치로인한 문제점에 직면하게 된다. 이상치에 대한 문제점은 표본조사 뿐만 아니라 통계학의 전 분야에서 발생하고 있으며, 이상치에 대한영향 및 처리방안에 대해 많은 방법론들을 연구하고 있으나, 일반 통계학 분야에서 연구된 이상치 방법론은 표본조사에 적용하기가 어려우며 이에 대한 연구결과도 미비한 상태이다. 표본조사에서 이상치에 대한 연구는 Eltinge 와 Cantwell (2006), Ishikawa 등 (2010), Lee (1995), Mattews와 Bernard (2002), Kim (2014), Shon과 Shin (2012)에 의하여 이루어졌다. Kim (2006)은 표본조사에서 가중치 감소 방법에 대하여 연구하였고, Song 등 (2011)은 R을 사용하여 이상치를 탐지하는 알로리즘을 구현하였다.

표본조사에서 이상치 방법론이 어려운 이유는 대개의 경우 분포에 대한 가정이 없고 표본단위들이 서로 상관되어 있으며, 서로 다른 표본 가중치와 추출확률을 갖기 때문이다.

이상치는 목표모집단에서 얻은 유효한 관찰값에 해당하는 대표적 이상치 (repr-esentative outlier)와 추정과정에서 유입된 오차에 해당하는 비대표적 이상치 (non-representative outlier)의 2개 그룹으로 분류할 수 있다. 조사 및 코딩 에러로 인하여 발생하는 이상치를 비대표적 이상치라고 하며,보통 추

[1] (702-701) 대구광역시 북구 대학로 80, 경북대학교 통계학과, 석사.
[2] 교신저자: (702-701) 대구광역시 북구 대학로 80, 경북대학교 통계학과, 교수. E-mail: gycho@knu.ac.kr

적(follow-up)이나 대체 (imputation)와 같은 방법에 의해 에디팅단계에서 처리되어야 한다. 대표적 이상치는 정확하게 조사되어 입력된 이상치를말하며 추정결과의 비편의성 및 정도의 제고를 위해 추정단계에서 처리되어야 한다.

본 연구에서 비대표적 이상치는 없는 것으로 간주하고 대표적 이상치 만을 다루기로 한다. 유한모집단에서 대표적인 이상치를 어떻게 처리하는가의 문제는 몇 개의 표본들이 평균값보다 상당히 큰 값을 가지게 되는 경제조사에서는 특히나 더 중요하다고 할 수 있다. 사업체 표본조사의 자료에서 이상치를 포함한 채로 결과값을 추정하게 되면, 추정치가 편의될 뿐만 아니라 추정치의 정도가 떨어지게 되므로 적절한 방법에 의한 이상치 처리가 필요하다.

본 연구에서는 이상치를 탐지하는 방법 및 탐지된 이상치를 처리하는 방법들을 검토한 후 사례분석을 통해 이상치 처리방법들을 비교 검토하였다. 사업체 자료는 대부분 왜도가 높은 치우친 분포 (skewed distribution)를 가지고있으므로 본 연구에서는 하한의 이상치는 고려하지 않고 상한을 벗어나는 이상치만을 고려하였다.

2. 이상치의 탐지

이상치의 탐지는 주로 양적 판단에 의존하는데, 양적판단을 위해 각 자료의벗어남의 정도를 나타내는 지표가 필요하다. 일반적으로 이상치는 자료의 중심으로부터의 상대적인 거리를 사용해서 탐지된다. $y_1, y_2, \cdots, y_n$가 정렬된 자료라고 한다면 허용범위 (tolerance interval) $[m - c_l s, m + c_u s]$를 벗어나는 자료에 대해 이상치라고 정의한다. 이 때, m은 위치추정치, s는 척도추정치이며, c_l과 c_u는 상수로 보통 3을 사용하나 경우에 따라 달라진다. 자료가 치우친 분포인 경우 c_l과 c_u는 다른 값을 가지게 된다. 많은 경우 위치와 척도 추정을 위해 표본평균과 표준편차를 이용하기도 하지만, 평균과 표준편차는 이상치에 매우 민감한 통계량으로서 이상치가 존재할경우 이를 이용하는 것은 매우 비효율적이다. 따라서 평균과 표준편차에 로버스트한 통계량으로서 중위수 (median)와 MAD (median absolute deviation)을 사용하기도 한다.

$$MAD = median_i \left\{ |y_i - median_j(y_j)| \right\}.$$

M-estimation을 기반으로 한 새로운 기법들에서는 척도 추정치로 MAD를 사용하기도 하지만, 대개의 표본조사에서는 MAD보다는 사분위수 범위를 척도 추정치로 사용한다.

사분위수 방법 (quartile method)은 이상치 탐지에 로버스트한 비모수적 방법이라고 할 수 있으며, 사분위수 방법에서의 허용범위는 다음과 같이 나타낼 수 있다.

$$[q_m - c_l d_l, q_m + c_u d_u]$$

$$\text{여기서, } q_m: \text{ 중앙값}(median),\ q_1, q_3: \text{ 제1, 3 사분위수, } d_l : q_m - q_1,\ d_u : q_3 - q_m.$$

자료가 치우친 분포인 경우는 c_l나 c_u를 큰 값으로 주어 한쪽 구간만 만들고 나머지 구간은 최대값 또는 최소값으로 대신하기도 한다.

또한 사분위수 방법과 유사한 것으로 Tukey (1977) 방법이 있는데, box-plot의 상한이나 하한 즉 바깥쪽 울타리 (outer fences)를 벗어나는 관측치를 이상치로정의하는 방법이다. Turkey (1977)방법의 허용범위는 $[q_1 - c_l IQR, q_3 + c_u IQR]$이며, 여기서 $IQR = q_3 - q_1$이다.

추세자료 (trend data)의 경우 비율 (ratio)을 이용하여 이상치를 탐지하기도 한다. 전년 (월, 분기) 조사 실적이 있는 경우를 기준으로 하여 비율값에 의해 이상치를 판단한다.

$$r_i = \frac{\text{현 조사 } (t)\text{에서 표본 } i\text{의 값}}{\text{이전조사 } (t-1)\text{에서 표본 } i\text{의 값}}$$

비율을 이용한 이상치 탐지시 비율 (r)의 범위가 0에서 무한까지이므로 왼쪽에서 이상치를 탐지하기가 힘들고, 또한 사업체 조사는 일반적으로 사업체 규모가 작을수록 비율의 변동 폭이 크기 때문에 주의를 해야 한다. 비율을 이용한 이상치 탐지시 규모가 큰 사업체보다 작은 사업체에서만 이상치가 탐지되는 경향을 "크기 마스크 효과 (size masking effect)"라고 한다.

Hidiroglou와 Berthelot (1986)은 크기 마스크 효과 (size masking effect)를 없애기 위한 방안으로사분위수방법과 비율방법을 결합하여 이상치를 탐지하는 수정사분위 방법을 제시하였는데 다음과 같은 과정을 통해 허용범위를 구할 수 있다.

(Step 1) 두 시점 $(t-1, t)$간의 조사자료의 비 (ratio)를 구한다.

$$r_i = \frac{y_i(t)}{y_i(t-1)}$$

(Step 2) 조사자료의 비를 다음과 같이 변환한다.

$$s_i = \begin{cases} 1 - \dfrac{r_m}{r_i} & \text{if } \ 0 < r_i < r_m \\ \dfrac{r_i}{r_m} - 1 & \text{if } \ r \geq r_m \end{cases} \quad , \ r_m : r_i\text{의 중앙값},$$

$$E_i = s_i \left\{ Max \left(y_i \left(t \right), y_i \left(t-1 \right) \right) \right\}^u , \ 0 \leq u \leq 1$$

(Step 3) 허용범위의 상·하한값 산출

$$D_l = \max \left(E_m - E_{q_1}, |AE_m| \right)$$
$$D_u = \max \left(E_{q_3} - E_m, |AE_m| \right)$$

여기서, $A = 0.05$

$E_m : E_i$의 중앙값

$E_{q_1} : E_i$의 제1사분위수

$E_{q_3} : E_i$의 제3사분위수

Hidiroglou와 Berthelot (1986) 방법의 허용범위는 $(E_m - cD_l, E_m + cD_u)$과 같으며 여기서 c는 상수이다.

이상치를 탐지하는 여러 가지 방법들을 살펴보았는데, 본 연구에서는 사분위수 방법을 적용하였다. 또한 경제조사에서는 자료값들이 일반적으로 양의 값을 가지고 극단값들은 큰 값을 가지는 형태가 일반적이므로 본 연구에서는 하한은 최소값 0을 적용하고 상한은 $q_m + 23\left(q_3 - q_m\right)$을 적용하여 이상치를 탐지하였다.

3. 이상치 처리방법

탐지된 이상치를 처리하는 방법은 크게 세 가지로 구분할 수 있다. 먼저 이상치를 제외하는 방법 (trimming)과 이상치의 값을 변경시키거나 가중치를 조정하여 이상치의 영향력을 감소시키는 방법 그리고 로버스트 (robust) 기법을 적용하여 이상치의 값을 추정하는 방법 등이 있다.

3.1. 이상치를 제외하는 방법 (trimming)

이상치로 판단되는 관측값을 제외하고 추정하는 방법으로, 추정치의 분산은작아지지만 실제보다 과소 (또는 과대) 추정되어 편의가 발생한다. 이상치 자료도실제 조사된 수치이므로 이상치를 제외하는 것은 현실을 제대로 반영하는 방법으로 적절하지 않다.

3.2. 윈저화 방법 (winsorization)

윈저화 방법은 가중치를 조정하거나 관측값 자체를 다른 값으로 대체하는 방식으로 이상치를 처리한다. Kokic 과 Bell (1994), Chambers 등 (2000)이 윈저화 방법에 대하여 연구하였다.

• **관측값 변경 (value modification)**

이상치를 제외한 나머지 값 중 최대값 또는 최소값에 가까운 값으로 이상치를대체하는 방법이다. $y_1, y_2, \cdots, y_{n-k}, y_{n-k+1}, \cdots, y_n$을 h층에서 k개의 이상치를 포함한 정렬된 매출액 조사 자료라고 했을 때, 이상치 값을 변경한 후의 변경된 자료값은 다음과 같이 나타낼 수 있다.

$$y_{hi} = \begin{cases} y_{hi} & \text{if } \ y_{hi} < K_h \\ f_h y_{hi} + (1 - f_h) K_h & \text{otherwise} \end{cases}$$

여기서, $0 \le f_h \le 1$, h : 층, K_h : 절사값 (winsoring cutoff) 이때, $f_h = 0$인 경우의 총합추정량을 winsorized type I estimator이라 하고, $f_h = \frac{n_h}{N_h}$인 경우의 총합추정량을 winsorized type II estimator라고 한다.

$$\text{총합추정량} : \widehat{Y} = \sum_h \sum_{i=1}^{n} w_h y_{hi}$$

$$\text{winsorized type I estimator} \ : \widehat{Y} = \sum_h \left\{ \sum_{i=1}^{n-k} w_h y_{hi} + \sum_{i=n-k+1}^{n} w_h K_h \right\}$$

$$\text{winsorized type II estimator} : \widehat{Y} = \sum_h \left\{ \sum_{i=1}^{n-k} w_h y_{hi} + \sum_{i=n-k+1}^{n} w_h \left\{ f_h y_{hi} + (1 - f_h) K_h \right\} \right\}$$

관측값을 변경하여 이상치를 처리하고자 할 때 가장 중요한 것이 절사값 (winsoring cutoff) k의 선택이다. 절사값이 클수록 총합추정시 이상치의 자료값을 많이 반영하게 되고 절사값이 작을수록 추정시에 이상치의 영향력을 감소시킬 수 있다.

• **가중치 조정 (weight modification method)**

가중치 조정방법은 이상치 값 자체를 바꾸거나 제외하지 않고 가중치를 조정함으로써 이상치의 영향을 감소시키는 방법으로 표본조사에 적용하기에 적합하다. 가중치를 조정하기 전의 총합추정량은 다음과 같이 나타낼 수 있다.

$$\widehat{Y} = \sum_h \left(w_h \sum_{i=1}^{n_h} y_{hi} \right), \text{여기서 } n_h : \text{층 } h\text{에서 표본수}, w_h : \text{층 } h\text{에서 가중치}$$

만일 $s_1 \in$ 이상치 그룹

$s_2 \in$ 이상치가 아닌 그룹

$n_{h1} : s_1$그룹에서의 표본수

$n_{h2} : s_2$그룹에서의 표본수 $(= n_h - n_{h1})$

$N_h = \sum_{i=1}^{n_h} w_{hi} : h$층에서의 모집단수

라고 했을 때, 이상치로 탐색된 자료의 가중치를 감소시킨 후의 총합추정치는 다음과 같이 나타낼 수 있다.

$$\widehat{Y} = \sum_h \left\{ f(w_h) \sum_{i \in s_1} y_{hi} + \frac{N_h - n_{h1} f(w_h)}{n_{h2}} \sum_{i \in s_2} y_{hi} \right\}.$$

3.3. 로버스트 추정방법

로버스트 추정방법은 M-estimation 등의 로버스트 기법을 적용하여 이상치의값을 추정하는 방법으로 모집단 분포에 로버스트한 성질을 갖고 있으며, 최근 많은 연구가 진행되고 있다. 여러 가지 통계모형이 제안되었지만, 구조가 너무 복잡하여 사업체 조사에서는 거의 사용되지 않고 있다.

4. 모의실험

본 절에서는 2010년 경제총조사의 산업대분류 M (전문, 과학 및 기술 서비스업)자료를 모집단으로 하여 층화 랜덤 추출 (Stratified random sampling)방법으로 1000번 반복 추출한 자료를 이용하여 이상치탐지와 처리 방법들을 비교 검토하였다.

이상치 처리의 효율성을 검토하기 위해 이상치 처리 후 총합추정량의 상대편의 (relative bias), 상대표준 오차 (relative standard error)를 구하여 각 방법들의 효율성을 비교하였다. R번 반복시 총합추정량 및 상대편의 (RB), 상대표준오차 (RSB)의 값은 다음과 같이 구하였다.

$$\text{relavive bias (RB)} : RB(\%) = \frac{1}{R}\sum_{r=1}^{R} \frac{\bar{\hat{t}}_y - t_y}{t}_y \times 100$$

$$\text{relative standard error (RSE)} : RSE(\%) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\frac{\hat{t}_y - \bar{\hat{t}}_y}{t_y}\right]} \times 100$$

여기서, $\bar{\hat{t}}_y = \frac{1}{R}\sum_{r=1}^{R}\hat{t}_y(r)$, t_y : 모총합, $\hat{t}_y$: 총합추정량.

Table 4.1 Population sales by industrial classification 2-digit (KSIC) (Unit: million won)

KSIC2	SALES						
	TOTAL	MIN	MEDIAN	MEAN	MAX	SKEWNESS	STANDARD DEVIATION
70	34,137,319	1	342	8,590.2	5,951,275	39.3	117,073.1
71	53,845,838	1	186	1,660.3	928,566	28.8	15,432.2
72	25,217,483	1	226	1,433.1	723,550	38.4	10,759.1
73	4,477,213	1	58	278.7	86,569	29.4	1,730.5

모집단 자료의 분포를 살펴보면 왜도가 높은 오른쪽으로 긴 꼬리를 가지는 치우친분포를 하고있음을 알 수 있으므로, 이상치 탐지시 하한은 최소값 0으로 정하고 상한은사분위수 방법을 이용하여 허용범위 $[0, q_m + 23(q_3 - q_m)]$를 벗어나는 자료를 이상치로 정의하였다.

이상치로 탐지된 자료는 자료값 변경 (value modification) 방식인 윈저화방법과 가중치 감소 (weight reduction) 방법을 각각 적용하여 처리한 후 총합추정량을 비교하였다. 윈저화방법 적용시 type I winsorization 방법과 type II winsorization 방법을 각각 적용하였고, 이 때 윈저화 절사값 (winsoring cut-off) k는 자료의 분포값을 이용하여 각각 이상치탐지의 허용범위 상한값인 $q_m + 23(q_3 - q_m)$과99백분위수 (Percentile), 95백분위수 3개의 값을 k값으로 선택하였다.

이상치 자료의 가중치 감소를 위해서 가중치 변환함수 (weight transformation function) $f(w_i)$는 다음과 같은 4가지 경우를 적용하였다.

- $(w_i) = 1$
- $f(w_i) = \sqrt{w_i}$
- $f(w_i) = \log(w_i)$
- $f(w_i) = \frac{1}{2}w_i$

위의 이상치 처리방법을 적용한 후의 총합추정량 및 추정량의 RB, RSE의 값을 살펴보면 다음과 같다.

Table 4.2 Total estimator after outlier treatment (Unit: hundred millin won)

Treatment Method				KSIC2			
			Treatmentmark	70	71	72	73
Parameter				34,137	53,846	25,217	4,477
No Treatment				34,109	54,036	25,118	4,482
value modification	type I winsorizaion	K1	A	25,875	51,228	22,088	4,291
		K2	B	42,373	54,808	24,670	4,442
		K3	C	23,308	50,868	21,223	4,272
	type II winsorizaion	K1	D	33,650	52,845	24,190	4,427
		K2	E	34,192	54,060	25,429	4,465
		K3	F	33,812	52,608	24,062	4,392
weight reduction	$f(w_i) = 1$		G	33,418	52,083	23,711	4,357
	$f(w_i) = \sqrt{w_i}$		H	33,616	52,639	24,061	4,392
	$f(w_i) = \log(w_i)$		I	20,758	49,674	20,735	4,219
	$f(w_i) = \frac{1}{2}w_i$		J	27,299	51,499	22,720	4,328

여기서 $K1 = q_m + 23(q_3 - q_m)$, $K2 = 99\ percentile$, $K3 = 95\ percentile$

Table 4.3 RB(%) of total estimator after outlier treatment

Treatment Method				KSIC2			
			Treatmentmark	70	71	72	73
No Treatment				-0.08	0.35	-0.39	0.10
value modification	type I winsorizaion	K1	A	-24.20	-4.86	-12.41	-4.16
		K2	B	24.13	1.79	-2.17	-0.79
		K3	C	-31.72	-5.53	-15.84	-4.59
	type II winsorizaion	K1	D	-1.43	-1.86	-4.07	-1.13
		K2	E	0.16	0.40	0.84	-0.27
		K3	F	-0.95	-2.30	-4.58	-1.91
weight reduction	$f(w_i) = 1$		G	-2.11	-3.27	-5.97	-2.69
	$f(w_i) = \sqrt{w_i}$		H	-1.53	-2.24	-4.59	-1.90
	$f(w_i) = \log(w_i)$		I	-39.19	-7.75	-17.78	-5.76
	$f(w_i) = \frac{1}{2}w_i$		J	-20.03	-4.36	-9.90	-3.33

Table 4.4 RSE(%) of total estimator after outlier treatment

Treatment Method				KSIC2			
			Treatmentmark	70	71	72	73
No Treatment				5.79	3.15	3.30	2.64
value modification	type I winsorizaion	K1	A	5.53	2.35	2.08	2.43
		K2	B	5.89	3.48	4.39	2.80
		K3	C	5.63	2.40	2.19	2.40
	type II winsorizaion	K1	D	5.54	2.42	2.17	2.44
		K2	E	5.87	3.43	4.23	2.75
		K3	F	5.64	2.47	2.25	2.40
weight reduction	$f(w_i) = 1$		G	5.51	2.31	1.98	2.36
	$f(w_i) = \sqrt{w_i}$		H	5.55	2.44	2.16	2.37
	$f(w_i) = \log(w_i)$		I	5.53	2.34	2.04	2.36
	$f(w_i) = \frac{1}{2}w_i$		J	5.57	2.49	2.33	2.39

1000번 반복 시행한 자료의 이상치 처리방법별 총합추정량 비교는 산업중분류별로 상자그림 (boxplot)을 통하여 살펴보았다. 각 처리방법별 구분은 표 4.2의 처리기호에 따라 나타내었다.

수평상자그림을 통해 통계추정치의 분포를 살펴보면 type II 윈저화 방법 (처리D, E, F)과 가중치 변환시 가중치를 1로 두는 경우 (처리 G)와 제곱근 변환을 한 경우(처리 H)가 가장 모수에 근접한 값을 가짐을 알 수 있었다.

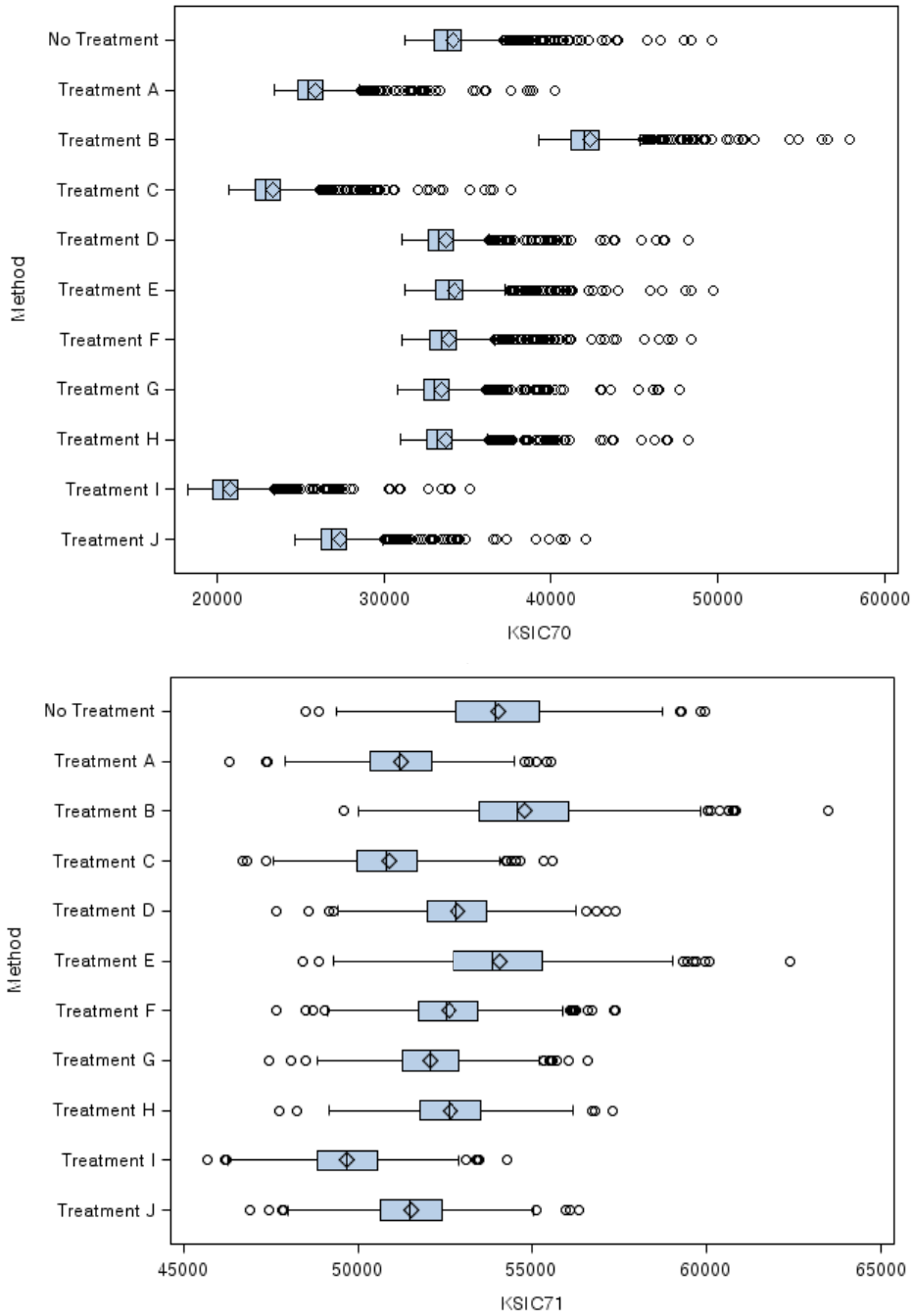


Figure 4.1 Horizontal box-plot of total estimator by KSIC70 & KSIC71/ treatment method

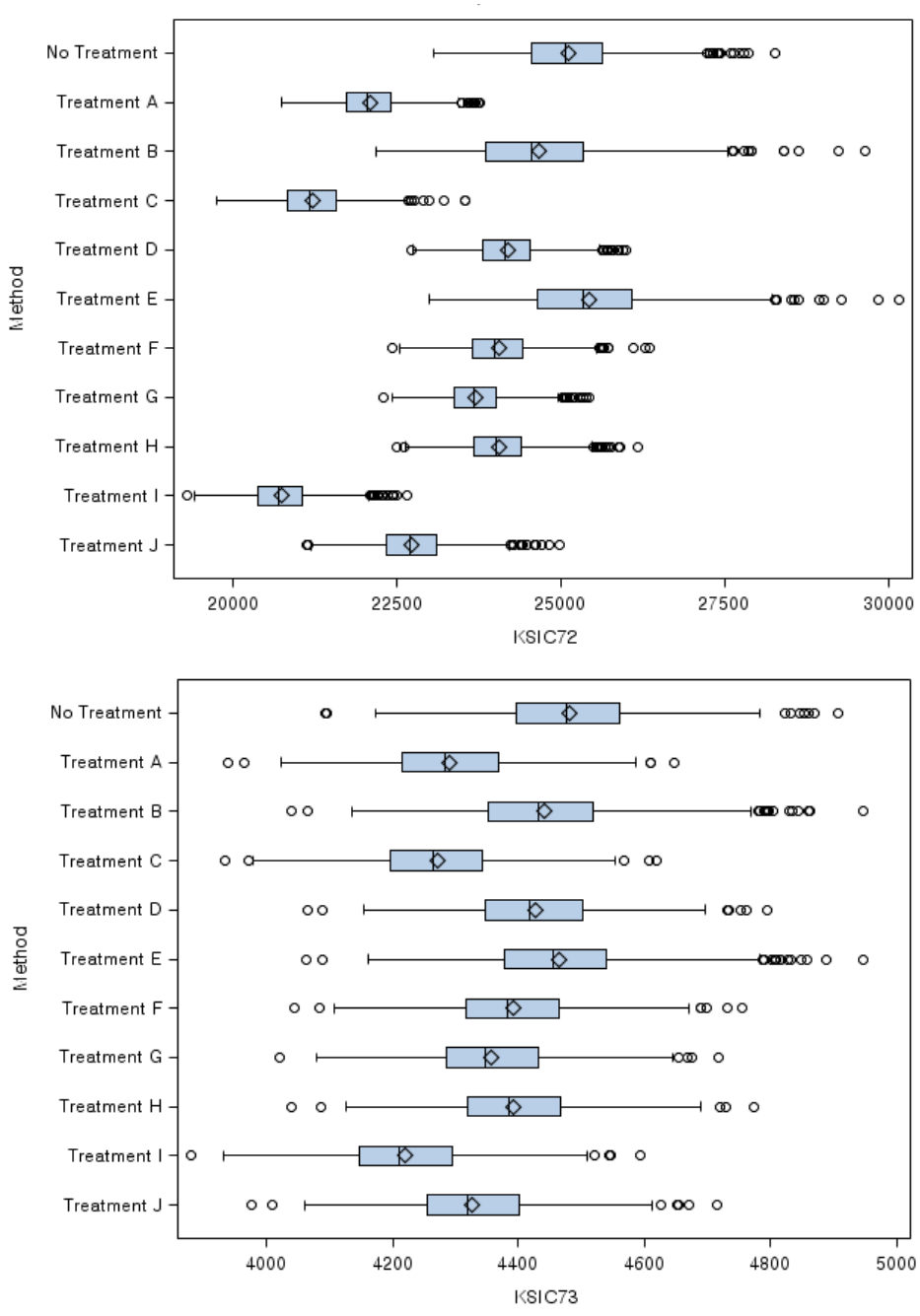
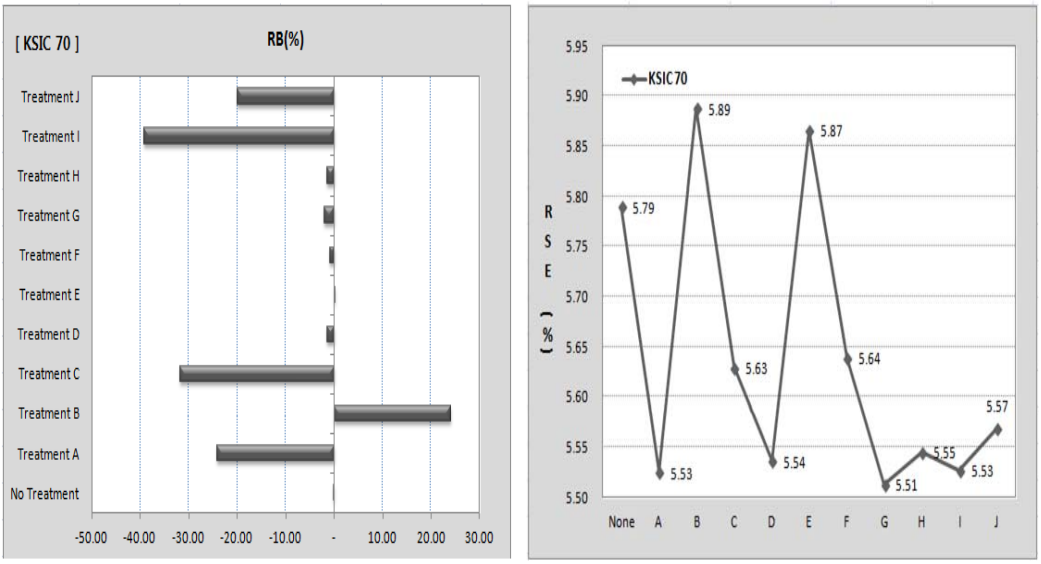
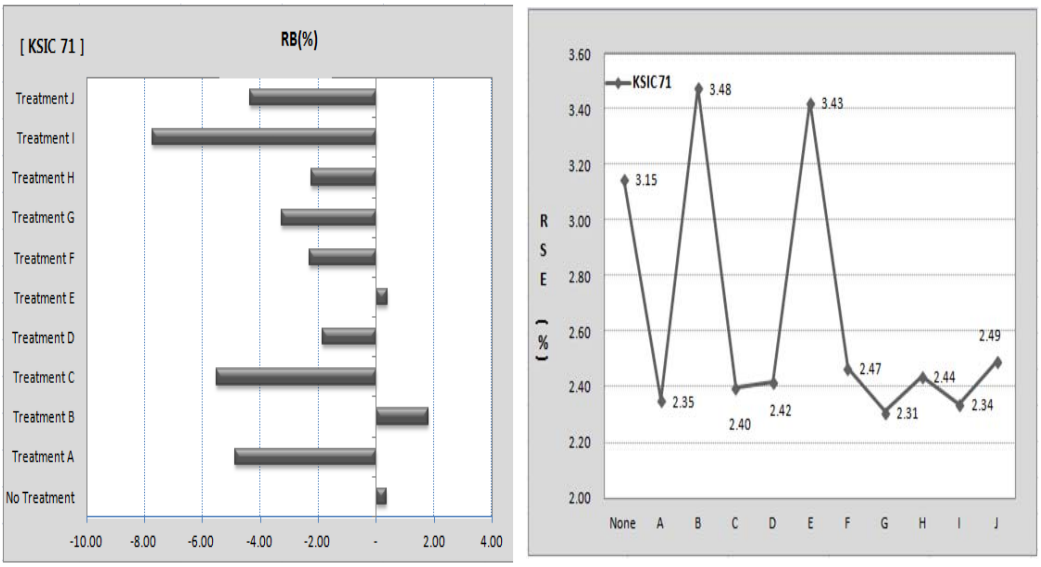


Figure 4.2 Horizontal box-plot of total estimator by KSIC72 & KSIC73/ /treatment method

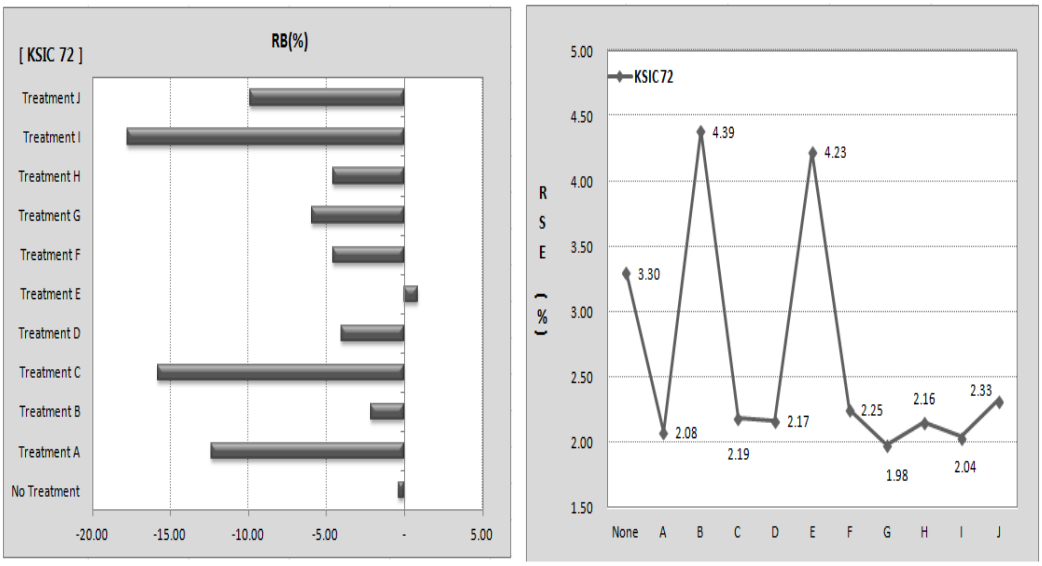


KSIC70

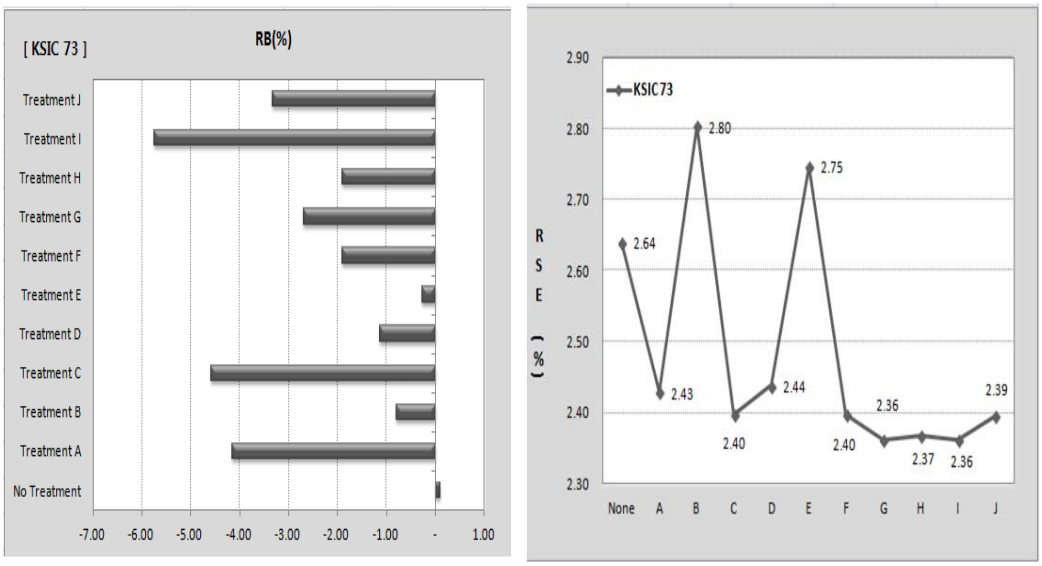


KSIC71

Figure 4.3 RB(%), RSE(%) of total estimator by KSIC70 & KSIC71



KSIC72



KSIC73

Figure 4.4 RB(%), RSE(%) of total estimator by KSIC72 & KSIC73

각 처리방법별로 살펴보면 type II 윈저화방법인 처리 D, E와 가중치를 제곱근 변환한 처리H가 다른 처리방법에 비해 편의 (RB)이 작게 나타났다. 처리E의 경우는 편의 (RB)는 작게 나타났으나 상대표준오차 (RSE)가 처리 D, H에 비해 다소 크게 나타나는 결과를 보였다.

여러 가지 이상치 처리방법들을 비교해 본 결과 type I 윈저화 방법보다는 type II 윈저화 방법이 효율적인 결과값을 보여주었으며, 가중치 변환방법들 중에서는 제곱근 변환을 통한 가중치 감소방법이 다른 처리방법에 비해 좋은 결과값을 보여주었다.

5. 결론

표본조사에서 모집단의 총합이나 평균을 추정할 때 이상치는 큰 영향을 미치게 되므로, 이상치의 영향을 감소하기 위한 여러 가지 방법들이 연구되어 왔다. 여러 가지 이상치 처리 방법들 중에서 조사값 제거를 통한 처리방법은 이상치 자료도 일단 조사된 자료값이므로 제거하는 것은 바람직하지 않다고 생각된다. 그러나 이상치가 추정값에 미치는 영향이 작지 않으므로 이상치의 영향을 감소시키기 위한 방법들은 필요하다고 생각된다.

이상치 영향을 감소시키기 위한 방법들 중 윈저화방법과 가중치 감소 방법을 모의실험을 통해 비교검토 한 결과 type II 윈저화방법을 통한 이상치 처리 방법과 제곱근변환을 통한 가중치감소 처리방법이 효율적인 결과값을 보여주었다. 다만, type II 윈저화방법 적용시 윈저화 절사값 (winsoring cut-off) k값에 따라 추정치의 효율이 많이 좌우되는 결과를 보였다. 따라서 Type II 윈저화방법을 적용하기 위해서는 절사값 k에 대한 좀 더 많은 연구와 검토가 필요할 것으로 생각된다.

여러 가지 이상치 처리 방법 중에서 제곱근 변환을 통한 가중치 감소방법을 통해서 가장 안정적이고 효율적인 추정값을 얻을 수 있었다

References

Chambers, R., Kokic, P., Smith, P. and Cruddas, M. (2000). Winsorization for identifying and treating outliers in business surveys. *Proceedings of the Second International Conference on Establishment Surveys*, 717-726, American Statistical Association Alexandria, Virginia.

Eltinge, J. L. and Cantwell, P. J. (2006). *Outliers and influential observations in establishment surveys*, Federal Economic Statistics Advisory Committee, `http://www.bls.gov/bls/fesacp3060906.pdf`.

Hidiroglou, M. A. and Berthelot, J. M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, **12**, 73-83.

Ishikawa, A., Endo, S. and Shiratori, T. (2010). *Treatment of outliers in business surveys : The case of short-term economic survey of enterprises in Japan (Tankan)*, 10-E-8, Bank of Japan, Japan.

Kim, J. (2006). Weight reduction method for ourlier in survey sampling. *Communications for Statistical Applications and Methods*, **13**, 19-27.

Kim, J. T. (2014). Lowness and outlier analysis of biological oxygen demand on Nakdong main stream river. *Journal of the Korean Data & information Science Society*, **25**, 119-130.

Kokic, P. N. and Bell, P. A. (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, **10**, 419-435.

Lee, H. (1995). *Outliers in business surveys. Chapter 26 in Business Survey Methods (B. Cox et al., eds.)*, Wiley, New York.

Mattews, S. and Berard, H. (2002). The outlier detection and treatment strategy for the monthly wholesale and retail trade survey of statistics Canada. in *Proceedings of the Survey Methods Section*, 63-68, Statistical Society of Canada.

Sohn, K. C. and Shin, I. H. (2012). Outlier detection using Grubb and Cochran test in clinical data. *Journal of the Korean Data & information Science Society*, **23**, 657-663.

Song, G. M., Moon, J. E. and Park, C. (2011). Realization of an outlier detection algorithm using R. *Journal of the Korean Data & information Science Society*, **22**, 449-458.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, California.

Journal of the Korean Data & Information Science Society
2016, **27**(1), 131–142

http://dx.doi.org/10.7465/jkdi.2016.27.1.131
한국데이터정보과학회지

Outlier detection and treatment in industrial sampling survey

Young Sun Joo[1] · Gyo-Young Cho[2]

[12]Department of Statistics, Kyungpook National University

Received 4 January 2016, revised 12 January 2016, accepted 13 January 2016

Abstract

Outliers in surveys can have a large effect on estimates of totals. This is especially true in business surveys where the populations are drawn are typically skewed. In this paper, we discussed the practical development and implementation of methods to identify and deal with outliers. A detection method is based on quartile method and detected outlier is processed in various ways. The study examines two versions of winsorised estimators with three different cut-off thresholds for each one. For the simulation study, four types of weight transformation function have been considered.

Keywords: Outlier detection, outlier treatment, winsorization, weight reduction.

[1] Graduate student, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea.
[2] Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: gycho@knu.ac.kr