The Korean Communications in Statistics
Vol. 13 No. 1, 2006, pp. 19-27

Weight Reduction Method for Outlier in Survey Sampling

Jin Kim[1)]

Abstract

Outliers in survey are a perennial problem for applied survey statisticians to estimate the total or mean of population. The influence of outliers is more increasing as they have large weights in survey sampling. Many techniques have been studied to lower the impact of outliers on sample survey estimates. Outliers can be downweighted by winsorization or reducing the weight of outliers. The weight reduction is more reasonable than replacing one outlier by one value of non-outliers, because it has at least one unit. In this paper, we suggest the square root transformation of weight as the weight reduction method. We show this method is efficient with real data, and it's also easy to apply in practical affairs.

Keywords : Outliers; Winsorization; Weight reduction; Transformation.

1. 서론

이상치(outlier)란 자료 값들이 주로 모여 있는 곳에서 멀리 떨어져 있는 관측치를 말하며 표본조사의 경우 이상치를 포함하는 표본으로부터 모집단의 평균이나 총계를 추정하는 과정에서 이상치로 인한 문제점에 자주 직면하게 된다. 이상치에 대한 문제점은 표본조사 뿐만 아니라 통계학의 전 분야에서 발생하며, 이상치에 대한 영향 및 처리에 대해 많은 방법론이 연구되었으나, 일반 통계학 분야에서 연구된 이상치 방법론은 표본조사에 적용하기가 어려우며 이에 대한 연구결과도 미비한 상태이다. 표본조사에서 이상치 방법론이 어려운 이유는 대개의 경우 분포에 대한 가정이 없고, 표본단위들이 서로 상관되어 있으며, 서로 다른 표본 가중치와 추출확률을 갖고, 대부분의 모집단의 분포가 치우친 형태이기 때문이다.

이상치는 합리적인 이상치(representative outlier)와 비합리적인 이상치(nonre-presentative outlier)로 구분할 수 있다. 비합리적인 이상치는 조사 및 코딩 에러로 인하여 발생하는 이상치를 말하며, 에디팅단계에서 추적(follow-up), 대체법(imputation)과 같은 방법에 의해 처리되어야 한다. 합리적 이상치는 정확하게 조사되어 입력된 이상치를 말하며 추정결과의 비편향성 및 정도의 제고를 위해 추정단계에서 처리되어야 한다. 본 논문에서 이상치는 합리적인 이상치만을 다루기로 한다.

표본조사의 자료에서 이상치를 포함한 채로 결과값을 추정하게 되면, 추정치가 편

1) Assistant Director, Regional Statistics & Sampling Division, Korea National Statistical Office, Korea. E-mail : jink@nso.go.kr.

향될 뿐만 아니라, 추정치의 정도가 떨어지게 되므로 적절한 방법에 의해 이상치를 처리함이 필요하다. 이상치를 처리하는 방법으로는 이상치를 제외하는(Trimming) 방법, 이상치를 다른 값으로 대치시키는 윈저화(Winsorization) 방법, 가중치를 조정하는 가중치 감소(Weight Reduction) 방법, 로버스트(Roburst) 기법을 적용하여 이상치의 값을 추정하는 방법 등이 있다.

이상치를 제외하는 방법은 추정치의 분산은 작아지나, 추정치를 실제보다 과소(또는 과대)추정하여 편향된 추정치를 얻게 되고, 이상치 또한 실제 조사된 수치이므로 이상치를 제외하는 것은 현실을 제대로 반영하는 방법으로 적절치 못하다.

윈저화 방법은 이상치를 제외한 나머지 값 중 최소값(또는 최대값)에 가까운 값으로 이상치의 값을 바꾸어 주는 방법으로 단순무작위추출시에 적합한 것으로 알려져 있다. 가중치의 감소방법은 이상치를 제외하거나 바꾸지 않고 해당 표본단위의 가중치를 조정하여 이상치의 영향을 다소 감소시키는 방법으로 표본조사에 적용하기에 적합하며 미국, 캐나다 등 통계조사에서 많이 쓰이는 방법이다. 로버스트 추정방법은 M-estimation 등의 로버스트 기법을 적용하여 이상치의 값을 추정하는 방법으로 모집단 분포에 로버스트한 성질을 갖고 있으며, 최근 많은 연구가 진행되고 있다. 로버스트 추정방법으로 M-estimation 기법을 이용한 비추정, 회귀추정이 주로 사용되며, 이 경우에 관심변수와 상관이 높은 보조변수의 선택이 중요하다.

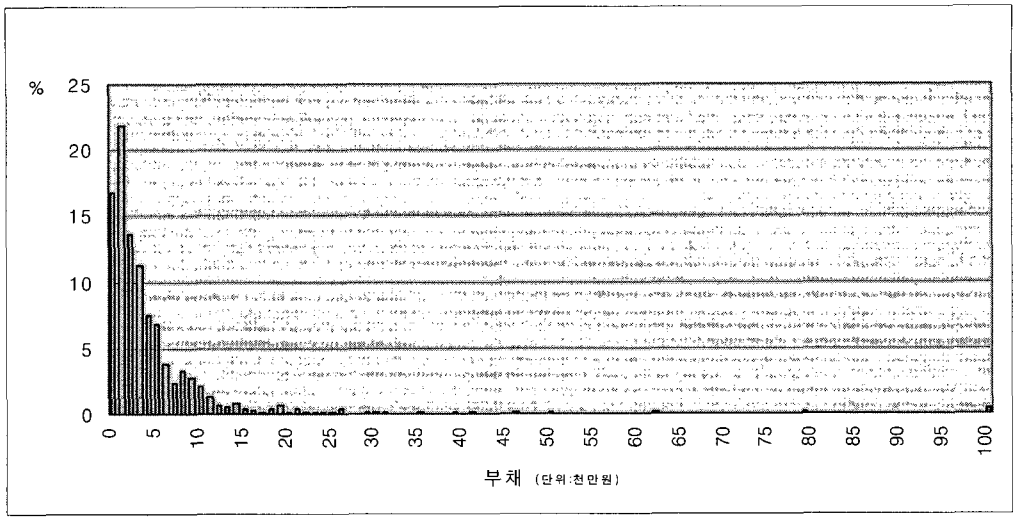
본 논문에서는 통계청에서 실제로 조사된 자료를 이용하여 윈저화방법과 가중치 감소방법으로 이상치를 처리한 결과를 비교하였다. 표본조사의 결과 얻어지는 이상치는 최소한 1개의 표본단위 이상에서 발견되는 값이므로 가중치를 감소시키는 것이 가장 합리적이라 판단된다. 따라서, 본 논문은 가중치 감소방법에 의해 조정된 가중치 w^*를 구하는 절차를 제시하고, 가중치 감소방법으로 가중치를 제곱근 변환하는 방법을 제안하였으며, 이 방법이 실제 업무에 효율적이고 적용이 용이한 방법임을 언급한다.

2. 이상치의 정의

통계청에서 실제 조사된 부채자료를 이용하여 이상치에 대한 분석을 수행하고 이에 대한 처리방법을 제안하고자 한다. 본 자료에서 부채변수는 <그림 1>과 같이 오른쪽으로 기울어진 분포(one-side right skewed distribution)로 값이 큰 이상치의 영향이 매우 심각하게 발생한다.

본 조사는 표본설계시 부채를 층화 및 주요특성변수로 사용하지 않았으므로, 각 표본단위의 가중치는 부채값을 대표하는 가중치로 보기 어려우며, 부채에 대한 이상치에 가중치가 다소 크게 부여되는 경우도 종종 발생한다. 이상치에 해당되는 표본단위의 가중치가 클수록 이상치의 영향은 더욱 심각하게 된다.

이상치를 처리하기 이전에 이상치에 대한 정의가 필요하다. 대개의 경우, 정렬된 관측치 $y_1, y_2, \cdots, y_n$가 허용범위(tolerance interval) $[m - c_l s,\ m + c_u s]$를 벗어나면 이상치라고 정의한다. 여기서, m은 위치 추정치(location estimator), s는 척도 추정치(scale estimator), c_l, c_u은 사전에 결정하는 절사점(cut-off value)이다.



<그림 1> 부채의 분포(Weighted)

많은 경우 위치와 척도 추정을 위해 표본평균과 표준편차를 이용하기도 하나, 평균과 표준편차는 이상치에 매우 민감한 통계량으로써 이상치가 존재하는 경우 이를 이용하는 것은 비효율적이다. 따라서, 평균과 표준편차에 로버스트한 통계량으로써 중위수와 MAD(Median Absolute Deviation)를 사용하기도 한다.

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

M-estimation을 기반으로한 새로운 기법들에서는 척도 추정치로 MAD를 사용하기도 하나, 대개의 표본조사에서는 MAD보다 사분위수 범위를 사용한 사분위수 방법(quartile method)에 의해 이상치를 정의하게 된다. 즉, $[q_{0.5} - c_l d_l,\ q_{0.5} + c_u d_u]$ 를 벗어나는 수치를 이상치로 정의한다. 여기서, $q_{0.5}$는 중위수 $q_{0.25}$, $q_{0.75}$는 1사분위수와 3사분위수이며, $d_l = q_{0.5} - q_{0.25}$, $d_u = q_{0.75} - q_{0.5}$는 사분위수 범위를 나타낸다. 사분위수 방법의 구간은 중위수에 대해 비대칭, 즉 $c_l d_l \neq c_u d_u$이다. 표본조사에서 치우친 분포인 경우, c_l나 c_u를 큰 값으로 주어 한쪽 구간만 만들고 나머지 구간은 최대값 또는 최소값으로 대신하기도 한다. 사분위수 방법과 유사하게 Tukey (1977)는 Box-plot의 상한(하한) 바깥쪽 울타리(outer fences)를 벗어나는 관측치를 이상치로 정의하였다. 즉, $[q_{0.25} - c_l IQ,\ q_{0.75} + c_u IQ]$를 벗어난 관측치가 이상치에 해당된다. 여기서, $IQ = q_{0.75} - q_{0.25}$이다.

본 논문에서는 Tukey의 방법을 적용하여 $[q_{0.25} - 3\,IQ,\ q_{0.75} + 3\,IQ]$를 벗어나는 수치를 이상치로 정의하였다. 이때, 이상치를 포함한 경우의 부채에 대한 가중평균이 41,468(천원), 이상치를 제거한 경우에는 29,954(천원)으로 나타났다.

3. 이상치의 처리

3.1 윈저화 방법의 적용

본 절에서는 이상치 처리에 대한 일반적인 방법인 윈저화 방법을 적용하여 그 결과를 비교하 여 보고자 한다. 어떠한 방법이 효율적인 이상치 처리방법인지를 판단하기 위한 기준이 필요하다. 이때, 부채에 대한 모집단의 정보(가중치, 이상치의 개수, 모평균, 모분산 등)를 알 수가 없으므로 정확하고 효율적인 방법을 선택하는 것은 어려운 작업이다. 따라서, 이상치 처리에 관한 문제에 있어서는 편의를 정확히 알 수 없으므로 효율성의 기준을 추정치의 변이계수(CV)을 줄이는 것으로 한다.

$y_1, y_2, \cdots, y_{n-k}, y_{n-k+1}, \cdots, y_n$ 은 k개의 이상치를 포함한 부채값을 오름차순으로 정렬한 값이다. 본 자료의 경우에는 부채의 값이 큰 경우만 이상치에 속한다. 윈저화 방법을 이용하여 다음과 같이 4가지 방법으로 이상치를 처리하였다.

$$① \overline{Y}_{w1} = \frac{\sum_{i=1}^{n-k} w_i y_i + \sum_{i=n-k+1}^{n} w_i y_{n-k}}{\sum_{i=1}^{n} w_i}$$

$$② \overline{Y}_{w2} = \frac{\sum_{i=1}^{n-k} w_i y_i + \sum_{i=n-k+1}^{n} w_i dt_u}{\sum_{i=1}^{n} w_i}, \quad (dt_u = q_{0.75} + 3\,IQ)$$

$$③ \overline{Y}_{w3} = \frac{\sum_{i=1}^{n-k} w_i y_i + \sum_{i=n-k+1}^{n} w_i y_{n-k+1}}{\sum_{i=1}^{n} w_i}$$

$$④ \overline{Y}_{w4} = \frac{\sum_{i=1}^{n-k} w_i y_i + \sum_{i=n-k+1}^{n} w_i \left(\frac{1}{w_i} * y_i + \frac{(w_i - 1)}{w_i} * dt_u \right)}{\sum_{i=1}^{n} w_i}.$$

첫 번째 방법은 이상치가 아닌 값들 중에 최대값을 이상치를 대치하였고, 두 번째 방법의 경우, 이상치를 정의하는 구간의 상한값 dt_u으로 이상치를 대치하였다. 세 번째 방법은 이상치 중에서 최소값으로 이상치를 대치하였다. 마지막으로, 네 번째 방법은 dt_u로 대치하되 dt_u로 완전대치하지 않고 이상치를 $1/w$ (w는 가중치)만큼 반영시켜 대치하는 방법을 제안하였다.

<표 1>은 위에서 제시한 방법에 의해 이상치를 처리한 결과이다.

<표 1> 윈저화 방법에 의한 이상치 처리결과(단위 : 천원)

method	평균	표준오차	변이계수	비고
$\overline{Y}$	41,468	4306.8	0.10439	이상치 무처리
$\overline{Y}_T$	29,954	1499.2	0.05005	이상치 제외
$\overline{Y}_{w1}$	34,432	1825.6	0.05302	
$\overline{Y}_{w2}$	34,455	1828.8	0.05308	
$\overline{Y}_{w3}$	34,508	1836.0	0.05321	
$\overline{Y}_{w4}$	34,537	1837.1	0.05319	

윈저화 방법은 이상치를 무처리한 원자료와 이상치를 제외한 방법의 중간정도의 결과인 평균 34,500 내외, CV 5.3% 정도의 비슷한 결과를 보인다. $\overline{Y}_{w1} \sim \overline{Y}_{w3}$은 모든 이상치의 값을 동일한 값으로 중복 대치하게 되어 분산을 과소추정하게 된다. 그러한 이유로, 동일한 값을 중복하여 대치하는 것보다 실제 관찰된 이상치를 $1/w$만큼 반영하여 처리하게 되면 이상치마다 서로 다른 수치들로 대치가 되고 실제값도 일정부분 반영이 되므로 $\overline{Y}_{w4}$가 더 합리적인 것으로 보인다.

3.2 가중치 감소 방법

본 절에서는 기존의 가중치를 조정하여 새로운 가중치 w_i^*를 얻은 후 이를 적용하는 가중치 감소방법에 의해 이상치를 처리하고 이에 대한 결과를 비교하고자 한다. 가중치 감소 방법을 적용한 가중평균은 다음과 같이 조정된 가중치 w_i^*에 의해 추정된다.

$$\overline{Y}_R = \sum_{i=1}^{n} w_i^* y_i \Big/ \sum_{i=1}^{n} w_i^* \quad . \tag{3.1}$$

식 (3.1)의 w_i^* 는 다음 단계에 의해 구할 수 있다.

단계1) 이상치에 해당되는w_i를 임의의 함수 $f(w_i)$를 이용하여 조정한다.

단계2) $\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_i^*$가 되도록 가중치 조정인자(weight reduction adjust factor) f_i를 구한다.

$$\cdot\ f_i = \begin{cases} 1 & (\text{for } outlier) \\ \dfrac{A+B}{B} & (\text{for } non-outlier) \end{cases}$$

· $A=\sum_{i=n-k+1}^{n} w_i - \sum_{i=n-k+1}^{n} f(w_i)$ (for *outlier*), $B=\sum_{i=1}^{n-k} w_i$ (for *non−outlier*)

단계 3) 실제 가중치에 가중치 조정인자 f_i를 곱하여 w_i^*를 구한다

$$w_i^* = \begin{cases} f(w_i) & (\text{for } outlier) \\ f_i \times w_i & (\text{for } non-outlier) \end{cases}$$

단계 4) $\overline{Y}_R = \dfrac{\sum_{i=1}^{n} w_i^* y_i}{\sum_{i=1}^{n} w_i^*} = \dfrac{\sum_{i=1}^{n-k} f_i \times w_i \times y_i + \sum_{i=n-k+1}^{n} f(w_i) \times y_i}{\sum_{i=1}^{n-k} f_i \times w_i + \sum_{i=n-k+1}^{n} f(w_i)}$

이때, 층화 표본설계된 경우에는 각 층에 대해서 $f(w_i)$와 f_i를 구해야 한다.

가중치 감소를 위해서 다음과 같은 $f(w_i)$를 적용하여 이상치를 처리하여 보자.

① $f(w_i)=1$

② $f(w_i)=\min(w_i)$

③ $f(w_i) = \begin{cases} w_i\left(1+\dfrac{k}{2n}\right) & (\text{for } non-outlier) \\ w_i\left(1-\dfrac{n+k}{2n}\right) & (\text{for } outlier) \end{cases}$

④ $f(w_i)=0.5w_i$

⑤ $f(w_i)=\sqrt{w_i}$

⑥ $f(w_i)=\log_{10}(w_i)$

⑦ $f(w_i)=\log(w_i)$

Rao (1971)와 Chinnappa (1976)는 ①과 같이 이상치에 대한 가중치를 1로 주어 이상치의 영향력을 감소하는 방법을 제시하였다. ②는 전체 가중치 중 최소의 가중치를 이상치에 대한 가중치로 주는 방법이며, ③은 Hidiroglou and Srinath (1981)에 의해 제시된 방법으로 이상치뿐만 아니라 전체 관측값에 대한 가중치를 조정하여 이상치에 대한 영향력을 감소시키는 방법이다. ④는 이상치의 가중치에 0.5배를 부여하는 방법으로 이상치 수가 적을 경우 Hidiroglou and Srinath의 방법과 유사하다. ⑤~⑦은 기존의 가중치를 감소시키기 위해서 변환(transformation) 중 가장 대표적인 제곱근 및 로그변환을 실행한 것이다.

<표 2>는 ①~⑦의 가중치 감소방법을 적용하여 구한 결과이다. <표 2>의 결과 가중치를 변환했을 때의 변이계수들이 적게 나타났으며 이중 가중치를 제곱근한 경우의 변이계수가 가장 작았다. 이에 Box-Cox의 변환식을 적용하여 가중치의 효율적인 변환식을 선택하여 보고자 한다.

<표 2> 가중치 감소방법에 의한 이상치 처리결과(단위 : 천원)

method	평균	표준오차	변이계수	비고
$\overline{Y}_{R1}$	30,204	1498.6	0.049616	$w=1$
$\overline{Y}_{R2}$	33,170	1627.3	0.049059	$\min(w)$
$\overline{Y}_{R3}$	35,506	2491.9	0.070184	
$\overline{Y}_{R4}$	35,760	2540.2	0.071035	$0.5*w$
$\overline{Y}_{R5}$	31,277	1533.4	0.049026	$\sqrt{(w)}$
$\overline{Y}_{R6}$	30,318	1499.4	0.049457	$\log10(w)$
$\overline{Y}_{R7}$	30,667	1503.9	0.049039	

3.3 가중치의 Box-Cox의 변환

Box-Cox의 변환식을 가중치 감소방법에 적용하기 위해 $f(w_i)$를 다음과 같이 정의한다.

$$f(w)=w^k$$

<표 3> 가중치의 Box-Cox 변환 적용결과

k	평균	표준오차	변이계수	비고
0.0	30,204	1498.62	0.049616	$w=1$
0.1	30,281	1498.92	0.049501	
0.2	30,397	1499.97	0.049346	
0.3	30,576	1503.03	0.049158	
0.4	30,851	1511.34	0.048989	
0.5	31,277	1533.36	0.049026	$=\sqrt{w}$
0.6	31,942	1590.91	0.049806	
0.7	32,987	1737.31	0.052667	
0.8	34,638	2087.24	0.060259	
0.9	37,264	2842.64	0.076283	
1.0	41,468	4306.78	0.103857	$=raw\ weight$

$k=0$은 $f(w_i)=w^0=1$이고 Rao (1971)와 Chinnappa (1976)가 제안한 모든 이상치에 대한 가중치를 1로 부여한 방법과 같다. $k=1$은 $f(w_i)=w^1=w$이고 가중치를 조정하지 않고 최초의 가중치를 그대로 사용한 경우이다. 가중치를 감소시키기 위한 k 값의 범위는 0과 1 사이이다. <표 3>은 가중치에 Box-Cox 변환식을 적용한 결과이다.

k값이 감소할수록 가중치 감소의 정도가 커져서 평균은 점차 작아진다. 이때 모집단의 정보 즉 부채의 모평균을 모르기 때문에 편의정도는 알 수 없으므로 추정방법의 효율성을 변이계수로 판단함이 바람직하다. 따라서, $k=0.4$ 또는 $k=0.5$일때 변이계수가 가장 적게 나타나므로 효율적인 추정결과를 위한 $f(w_i)$는 $\sqrt{w_i}$가 적당하다.

4. 결론

표본으로부터 모집단의 총계나 평균을 추정하는 과정에서 이상치는 추정값에 큰 영향을 미치게 된다. 특히, 가중치를 갖는 통계조사의 경우에는 이상치를 갖는 표본단위에 가중치가 다소 크게 부여됨에 따라 이상치의 영향은 더욱 커지게 된다. 따라서, 표본조사에서 이상치의 영향을 감소하기 위한 여러 가지 방법들이 연구되어 왔으며, 이들 방법들 중 윈저화 방법과 가중치 감소방법을 실제자료에 적용하여 그 결과를 비교하였다.

표본조사의 결과 얻어지는 이상치는 최소한 1개의 표본단위 이상에서 발견되는 값이므로 관측된 이상치의 값을 다른 수치로 변경하는 윈저화 방법보다 표본단위의 가중치를 감소시키는 것이 적합하다. 본 논문에서는 가중치를 감소하기 위해 Box-Cox 변환식을 적용하여 보았다. $k=1$은 이상치의 가중치를 그대로 사용하는 것이고, $k=0$은 이상치의 가중치를 1로 주는 경우이다. k값이 1에 가까울수록 가중치 감소효과가 적고, 보통 $0.3<k<0.6$에서 가중치 감소효과가 나타나게 되는데 본 자료분석 결과에서는 $k=0.4$ 또는 $k=0.5$가 적합하였다. 따라서, 본 논문에서는 비교적 효율적이고 실무에 적용하가 용이한 가중치 감소방법으로 가중치를 제곱근하여 조정하는 방법을 제안하였다. 향후 다양한 자료에 대한 수치실험으로 이에 대한 검증작업이 필요하며, 효율성에 대한 기준을 명확히 하여 이 기준에 부합한 효율적인 방법에 대한 심층적인 연구가 필요할 것이다.

참고문헌

[1] Chambers, R.L. (1986). Outliers robust finite population estimation. *Journal of Applied Statistics*, Vol. 81, 1063-1069.

[2] Cook, R.D. (1979). Influential observations in linear regression. *Journal of Applied Statistics*, Vol. 74, 169-174.

[3] Gwet, J.P. and Rivest, L.P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of Applied Statistics*, Vol. 87, 1174-1182.

[4] Hidiroglou, M.A. and Srinath, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of Applied Statistics*, Vol. 76, 690-695.

[5] Lee, H. (1995). Outliers in business surveys. *Business Survey Methods*. Chap. 26. 503-526.

[6] Smith, T.M.F. (1987). Influential observations in survey sampling. *Journal of Applied Statistics*, Vol. 14, 143-152.
[7] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley.

[Received August 2005, Accepted November 2005]