응용통계연구 제 **21**권 3호, *2008*년, *pp. 413 427*

표본조사에 따른 추정방법 비교: 가중치 조정기법을 중심으로*

이상은[1)]

요 약

표본설계는 대표성있는 표본추출과 이에 따른 적절한 추정식을 산출해 주어야 한다. 이때 조사결과가 모집단의 대표성을 갖기 위해서는 적절한 가중치가 부여되어야 한다. 일반적으로 표본설계시 제시된 기본가중치는 추출확률에의해 계산된다. 그러나 조사 후 무응답등과 같은 요인이 발생된 경우 모집단의 대표성을 유지하기 위해 기본가중치의 보정이 필요하다. 본 논문은 조사 후 가중치 보정방법으로 기존에 흔히 사용하고 있는 갈퀴법(Raking)과 현재 BLS에서 적용하고 있는 방법(BLS) 그리고 최근 표본조사의 추정식으로 관심을 끌고 있는 일반화회귀식(GREG)에 의한 방법을 비교하였다. 자료분석은 산업단지공단내의 산업분류별 총 종사자수가 사용되었으며 MSE, Coverage, CV, Large Error(LE), Non-Centrality(NC) 등의 비교통계량을 이용하여 그 결과를 비교하였다.

주요용어: 갈퀴법, BLS방식, 일반화회귀식, CV, large error, non-centrality.

1. 서론

표본설계에 포함된 총계추정식에는 표본추출 방법에 관한 내용이 반드시 반영되어야 한다. 표본설계는 모집단으로부터 표본을 추출하는 과정을 말하며, 각각의 표본에 대하여 그 표본이 추출될 확률을 할당함으로써 결정된다. 기본가중치는 각 표본이 선택될 확률로써 반영되어 각각의 표본마다 계산되며 개별 표본자료의 기록과 관련된 수치적 통계측도와 표본의 불균형을 고려하여 조정되어야 한다. 따라서 표본조사 결과에 따른 최종가중치는 표본추출단위와 관련된 실제 표본의 선택과정을 반영한 기본가중치와 무응답으로 인하여 발생한 원표본(Original sample)과의 불균형을 조정하기 위한 가중 층 보정(Weighting class adjustment: WCA)값 그리고 중요한 의미를 지닌 특정변수에 대한 모집단 분포와 표본분포의 불균형을 수정하기 위한 사후층화 보정(PCA)값을 계산한 후 이들을 곱한 값으로 표현된다. 즉, 최종가중치의 일반적인 표현은 다음과 같다.

$$w = w_1 \times f_2 \times f_3, \tag{1.1}$$

여기서,

w_1: 불균등 선택 확률의 가중치(기본가중치),

f_2: 표본의 무응답 조정을 위한 보정값,

f_3: 모집단의 무응답 및 비포함(non-coverage) 조정을 위한 사후층화 보정값

* 본 논문은 2007학년도 경기대학교 학술연구비(일반연구과제)지원에 의하여 수행되었음.

1) (443-760) 경기도 수원시 영통구 이의동 산 94-3, 경기대학교 응용정보통계학과, 부교수.
E-mail: sanglee62@kyonggi.ac.kr

따라서 본 논문 2장에는 표본조사의 보정방법으로 이용되는 갈퀴법(Raking)과 BLS방법 그리고 일반화회귀방법(GREG) 등 3가지 방법에 대한 이론적 근거가 제시되며 3장에는 산업단지공단 자료를 이용한 사례연구 결과가 제시된다. 마지막으로 4장에는 3장의 결과를 근거로한 각 방법의 장단점을 제시하였다.

2. 추정방법

2.1. 갈퀴법(Ranking method)

표본조사 후 사용되는 사후 조정기법의 대표적인 방법이 갈퀴법(Raking Method)이다. 갈퀴법은 범주형 자료에서 주변분포(marginal distribution)는 알려져 있지만 결합분포가 알려져 있지 않을 경우, 주변합(marginal total, N_{i+}, N_{+j})과 알려진 주변분포를 이용하여 미지의 결합분포를 추정하는 사후층화 방법이다. 이는 층화변수가 2개 이상일 때 이 변수들의 조합에 대한 모든 결합층 대신 주변층의 정보만을 일치하도록 하는 조정이다. 즉, 표본설계시 얻어진 기본가중치(N_{i+}^R, N_{+j}^R)와 주변합을 이용하여 $\mid \widehat{N_{i+}^R} - N_{i+} \mid < \varepsilon$와 $\mid \widehat{N_{+j}^R} - N_{j+} \mid < \varepsilon$을 만족하는 최종가중치, $\widehat{N_{ij}^{(R)}}$를 구한다. 이때 모집단의 주변합을 벤치마크(Benchmark)로 하여 가중치가 사후 보정되며 사후 보정된 가중치를, $w_{k,R} = \widehat{N_{ij}^{(R)}}$이라 하면 갈퀴법에의한 추정량, $\hat{y_R}$은 다음과 같다.

$$\hat{y_R} = \sum_{k=1}^{n} w_{k,R} y_k^*,$$

여기서 y_k^*는 조사 후 관측값이다.

갈퀴법은 Deming과 Stephan (1940)에 의해 소개되었으며 Oh와 Scheuren (1987)의 무응답조정에 사용되어지기도 했다. 자세한 내용은 신민웅과 이상은 (2001)을 참조하기 바란다.

2.2. BLS 방법

BLS 방법은 미국 노동통계청(Bureau of Labor Statistics)의 사업체조사에서 주로 사용하는 방법이다. 본 논문에서는 BLS 가중치 보정 방법중에서 대표적인 보정 단계인 다음의 4단계를 다루도록한다. 즉 첫 번째는 집계 후 보정(Re-aggregation Adjustment)이고 두 번째는 무응답 보정(Nonresponse Adjustment), 세 번째는 이상값 보정(Outlier Adjustment), 마지막 네 번째는 벤치마크 보정(Benchmark Adjustment)이다. 이와 같은 4단계는 대부분의 경우 표본조사의 총계 추정에 적용되어져야 하는 요인이며 그 외 조사결과에 따라 "0"값의 보정단계등 다양한 보정단계가 조사의 특성에 따라 적용될 수도 있다. 다음의 표 2.1은 BLS방법을 다루기 위한 자료의 형태이다.

여기서 기본가중치는 표본설계 당시에 제시된 추출율의 역수를 의미하며 표본설계시값은 표본설계에 이용한 모집단 자료의 값을 그리고 관측값은 조사 후 얻은 값을 말한다. 그리고 상태는 조사에서 나타난 조사상태로 무응답(nonresponse), 폐업(oob: out of business) 등을 나타낸다. 다음은 BLS의 4단계에 대한 보정요인(adjustment factor)의 이론적 배경을 보기로 한다.

표 2.1: BLS 자료형태

사업체	기본가중치	표본설계시값	관측값	상태
1	w_1	y_1	y_1^*	reag
2	w_2	y_2	y_2^*	nonresponse
3	w_3	y_3	y_3^*	outlier
⋮	⋮	⋮	⋮	
n-1	w_{n-1}	y_{n-1}	y_{n-1}^*	collected
n	w_n	y_n	y_n^*	collected

a) 집계 후 보정(Re-aggregation Adjustment: REAG)

REAG는 y_i^*의 값이 표본설계시의 값인 y_i보다 매우 크거나 작게 관측된 경우로 예를 들면 표 2.1에서 사업체 1의 경우 조사표의 종사자수를 기업체 단위로 작성을 한 경우라 한다면 이 값은 사업체 단위의 조사에서 위배되는 값이 되기 때문에 대표성을 갖는 표본으로 관측값의 보정이 필요하게 된다. 또한 조사 사업체가 조사시점에서 일시적으로 종사자수가 증가한 경우라면 이 역시 표본 사업체의 관측값으로 보정이 필요하다. 이를 집계 후 보정요인(reaggreation factor)이라 하며 이 때 REAG에 대한 보정은 다음과 같다.

$$\text{REAG factor} = f_i^{REAG} = \frac{y_i}{y_i^*}, \tag{2.1}$$

여기서 y_i는 표본설계시의 값이고 y_i^*는 관측된 값이다.

b) 무응답 보정(Nonresponse Adjustment: NRA)

NRA는 최근 가장 많은 관심을 갖는 보정요인으로 BLS의 무응답 보정방법은 다음과 같다. 먼저 표본설계에서 가중치의 합은 모집단 수가 된다. 즉, $\sum_{i=1}^{n} w_i = N$이 된다. 이제 w_i^*를 보정된 가중치라 하자. 그리고 표본설계시 모집단과 조사시점에서의 모집단이 동일하다고 하면 $\sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n-NR} w_i^* y_i$를 만족해야한다. 이 때 무응답 보정요인, f_i^{NRA}을 이용하면 $w_i^* = w_i f_i^{NRA}$로 쓸 수 있으므로 $\sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n-NR} w_i f_i^{NRA} y_i$을 만족하기위한 f_i^{NRA}를 구하면 된다. 여기서 $n - \text{NR}$은 무응답으로 처리된 자료를 모집단에서 제외한 수이다. 이제 $f_i^{NRA} = f^{NRA}$이므로

$$\sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n-NR} w_i^* y_i = \sum_{i=1}^{n-NR} w_i f_i^{NRA} y_i = \left\{ \sum_{i=1}^{n-NR} w_i y_i \right\} f^{NRA}$$

이 된다. 따라서 무응답 보정결과는 다음과 같다.

$$\text{NRA factor} = f_i^{NRA} = f^{NRA} = \frac{\sum_{i}^{n} w_i y_i}{\sum_{i=1}^{n-NR} w_i y_i}. \tag{2.2}$$

c) 이상값 보정(Outlier Adjustment Factor: OAF)

OAF의 경우는 관측된 값이 표본설계 당시의 값보다 매우 크거나 작은 경우이다. 이때 조사 된 값이 정확하다면 이 역시 표본의 대표성을 위해 보정이 필요하다. 이 방법은 REAG와는 달리 해당 사업체의 조사값을 유지하되 가중치를 "1"로 환원한다. 즉, 해당 사업체를 전수조사 한 것과 같은 개념이다. 그 후 이상값으로 지정된 사업체를 제외한 나머지 사업체들의 가중치를 보정해준다. 보정방법은 NRA의 보정방법과 같은 개념을 사용한다. 우선 이상값인 사업체의 가중치를 "1"하기 위한 보정요인은 다음과 같다.

$$\text{outlier factor} = f_i^{OAF} = \frac{1}{w_i}. \tag{2.3}$$

이상값으로 판명된 값에 가중치를 "1"로 환원한 후에 이상값이 갖고 있던 나머지 가중치를 이상값이 아닌 다른 자료에 분배하여 전체 가중치의 합을 유지시켜야한다. 이는 모집단의 수는 정해진/알려진 것으로 전체 가중치의 합은 모집단의 수와 같아야 하므로 이상값에서 환원된 가중치는 이상값이 아닌 나머지 자료에 NRA와 같은 방법으로 배분하여야 한다. 즉, 이상값이 아닌 나머지 사업체들의 보정요인은 $\sum_{i=1}^{n} w_i y_i = \sum_{i \in outlier} y_i + \sum_{i \notin outlier} w_i f_i^{OAF(no)} y_i$을 만족하는 $f_i^{OAF(no)}$를 구해야 한다. 먼저

$$\sum_{i=1}^{n} w_i y_i = \sum_{i \in outlier} y_i + \sum_{i \notin outlier} w_i f_i^{OAF(no)} y_i$$

이므로 따라서

$$\sum_{i=1}^{n} w_i y_i - \sum_{i \in outlier} y_i = \sum_{i \notin outlier} w_i f_i^{OAF(no)} y_i$$

이 된다. 이제 $f_i^{OAF(no)} = f^{OAF(no)}$ 이므로

$$\sum_{i=1}^{n} w_i y_i - \sum_{i \in outlier} y_i = \sum_{i \notin outlier} w_i f^{OAF(no)} y_i = \left\{ \sum_{i \notin outlier} w_i y_i \right\} f^{OAF(no)}$$

이 되어 NRA의 경우와 같이

$$f_i^{OAF(no)} = f^{OAF(no)} = \frac{\sum_{i} w_i y_i - \sum_{i \in outlier} y_i}{\sum_{i \notin outlier} w_i y_i} = \frac{\sum_{i} w_i y_i - \sum_{i \in outlier} y_i}{\sum_{i}^{n} w_i y_i - \sum_{i \in outlier} w_i y_i}$$

이 된다. 따라서 이상값 보정 결과는 다음과 같다.

$$\text{non\_outlier factor} = f_i^{OAF(no)} = \frac{\sum_{i}^{n} w_i y_i - \sum_{i \in outlier} y_i}{\sum_{i}^{n} w_i y_i - \sum_{i \in outlier} w_i y_i}. \tag{2.4}$$

d) 벤치마크 보정(Benchmark Adjustment Factor: BMF)

BMF는 가중치 보정 단계 중 마지막 단계로 조사 시점에 알려진 최신 정보를 기준으로 가중치를 보정한다. 일반적으로 벤치마크에 따라 가중치의 값이 크게 좌우되는 경우가 많기 때문에 벤치마크의 설정은 신중해야 하며 많은 경우 센서스나 행정자료를 이용한다. 벤치마크의 보정요인, f_i^{BMF}은 다음과 같다.

$$\text{Benchmark factor} = f_i^{BMF} = \frac{\text{Target value}_i(\text{benchmark value}_i)}{\text{reported value}_i} = f^{BMF}. \tag{2.5}$$

표본자료에서 얻어진 값, y^*와 위의 3단계에서 얻어진 보정값을 이용하여 가중치를 구한 후 reported value를 구한다. Target value 혹은 Benchmark value가 주어지면 reported value를 이용하여 보정요인을 구한다.

e) 최종가중치(Final Weight)

이와 같이 4가지 단계를 마친 후 최종가중치는 다음과 같다.

$$W_i^{Final} = w_i \times f_i^{REAG} \times f_i^{NRF} \times f_i^{OAF} \times f_i^{BMF}. \tag{2.6}$$

식 (2.6)에 의해 최종가중치가 정해지면 BLS에 의한 추정량, $\hat{y}_B$은 다음과 같다,

$$\hat{y_B} = \sum_{i=1}^{n_i} W_i^{Final} y_i^*.$$

2.3. 일반화회귀(Generalized Regression: GREG) 추정방법

갈퀴법과 BLS방법이 표본조사에서 얻어진 자료만을 이용한 설계기반(design-based) 추정방법이라면 모형기반(model-based) 추정방법에는 일반화회귀추정방법(GREG)이 있으며 GREG방법은 최근 관심이 증가되고 있다. 자세한 내용은 Sarndal 등 (1992)을 살펴보기 바란다. 먼저 변수 Y의 총계, $t_y = \sum_U y_i$를 추정한다고 하자. 이때 종속변수 Y와 설명변수인 $X(x_{jk}$: j 변수의 k번째 관측값)를 이용하여 회귀식을 구하고 y_k와 가장 가까운 예측값을 y_k^0으로 표시하면 y_k의 예측값인 y_k^0은 다음과 같이 표현할 수 있다.

$$y_k^0 \doteq \sum_{j=1}^{J} A_j x_{jk} = A' X_k.$$

여기서 이해를 돕기위해 $x_1 = x$, $A_1 = 1$이라 가정하면 $y_k^0 = x_k$가 되며 이는 곧

$$y_k \doteq y_k^0 = x_k$$

로 표현할 수 있다. 위의 식들을 이용하여 추정하고자 하는 총계를 다시 표현하면

$$t_y = \sum_U y_k = \sum_U y_k^0 + \sum_U (y_k - y_k^0)$$

이 된다. 이제 π_k는 표본 추출률이고 $D_k = y_k - y_k^0$라 하면 $\sum_U D_k \doteq \sum_S D_k/\pi_k$으로 표현될 수 있다. 따라서 x, y의 선형관계를 이용한 총계 추정량은 다음과 같이 표현될 수 있다.

$$\hat{t}_y \doteq \sum_U y_k^0 + \sum_S \frac{y_k - y_k^0}{\pi_k},$$

$$\mathrm{Var}(\hat{t}_y) \doteq -\frac{1}{2}\sum\sum_s \Delta_{kl}(D_k - D_l)^2,$$

여기서 $\Delta_{ij} = (\pi_{ij} - \pi_i\pi_j)/\pi_{ij}$이고 π_{ij}는 i와 j가 동시에 표본으로 추출될 확률이다. 이제 S를 표본 집합, U는 전체 집합이라 할 때 (y_k, x_k), $k \in S$로부터 회귀분석을 이용하여 계수의 추정값 $\hat{B} = (\hat{B}_1, \ldots, \hat{B}_J)$을 얻을 수 있으며 이를 이용하면 GREG 추정식은 다음과 같다.

$$\hat{y}_{GR} \doteq \hat{t}_{y\pi} + \sum_{j=1}^{J} \hat{B}_j(t_{x_j} - \hat{t}_{x_j\pi})$$

여기서,

$$\hat{t}_{y\pi} = \sum_S \frac{y_k}{\pi_k}, \qquad \hat{t}_{x_j\pi} = \sum_S \frac{x_{jk}}{\pi_k}, \qquad t_{x_j} = \sum_U x_{jk},$$

$$\hat{B} = \left(\hat{B}_1, \ldots, \hat{B}_J\right) = \left(\sum_S \frac{x_k x_k'}{\sigma_k^2 \pi_k}\right)^{-1} \sum_S \frac{x_k y_k}{\sigma_k^2 \pi_k}$$

이다. 또한 π_k는 표본 추출률이고 σ_k^2는 k번째 자료의 분산이다. 이와 같이 GREG 추정량은 설계기반 추정량과 회귀추정량의 합으로 이루어진다. 이때 GREG 추정량의 분산은 다음과 같다.

$$\mathrm{Var}(\hat{y}_{GR}) = N^2\left(\frac{\bar{x}_U}{\hat{N}\hat{x}_s}\right)\frac{1-f}{n}\frac{\sum_s (y_k - \hat{B}x_k)^2}{n-1}.$$

이에 관한 자세한 내용은 Särndal 등 (1992)을 살펴보기 바란다.

3. 모의실험

3.1. 자료소개

본 논문에서는 전국 30개 산업공업단지에서 32,658개 사업체를 대상으로 매출액, 생산액 등을 전수 조사한 자료에서 종사자수가 5명에서 300명 미만이고 산업중분류가 17, 24, 28, 29, 33에 해당하는 사업체 6,333개만을 자료로 사용하였다. 이는 중소기업규모로 연구자의 관심 대상 산업분류만을 고려했기 때문이다. 여기서 표본은 단지별, 산업분류별, 종사자수를 기준으로 500개(unit)를 계통추출하여 선정하였다. 이 때 종사자수는 (1: 5명~19명, 2: 20명~49명, 3: 50명~99명, 4: 100명~299명) 등 총 4개의 그룹으로 만들었다. 그리고 위에서 언급한 3가지의 방법, Raking, BLS, GREG를 이용하여 총 종사자수를 추정하였다. 모집단 자료의 형태는 표 3.1과 같다.

표 3.1: 자료의 형태

사업체	단지	산업분류	매출액	생산액	종사자수
01	서울디지털산업단지	24	3500	3500	15

표 3.2: 산업분류별 사업체수

산업분류	17	24	28	29	33	계
사업체수	360	763	2228	2570	412	6333

표 3.3: 종사자 그룹별 사업체수

종사자그룹	5~19명	20~49명	50~99명	100~299명	계
사업체수	4142	1463	460	268	6333

조사 후 관측값에 집계 후 보정(REAG), 무응답 보정(NRA), 이상값 보정(OAF) 효과를 보기 위해, 3%, 10%, 20%의 보정 요소들이 적용되도록 자료를 새롭게 구성하였다. 예를 들어 원 자료에 3%, 10%, 20%의 이상점을 발생시켜 관측값으로 활용하였다. 같은 방법으로 무응답 보정과 집계 후 보정이 적용될 수 있도록 자료를 준비하였다. 일반적으로 보정요인(=기본가중치/최종가중치)이 0.5에서 2사이에 있을 때 보정이 의미가 있는 것으로 알려져 있기 때문에 이를 유지하기 위해 약 3%, 5%, 10%의 비율을 사용하였다. 이는 너무 많은 이상점 또는 집계 후 보정이 이루어져 최종 보정요인이 0.5에서 2사이를 벗어나면 이는 모집단이 다른 분포로 변화된 것으로 판단하는 것이 좋을 것이다. 끝으로 벤치마크 보정을 위한 벤치마크 값은 원 모집단의 값을 사용하였다.

3.2. 추정방법 적용

3%, 10%, 20%의 보정요인이 적용된 모집단에서 500개의 표본을 계통추출 하였으며 1,000번의 반복시행을 하였다.

3.2.1. Raking 방법

갈퀴법에서는 알려진 값인 각 산업분류별, 종사자 그룹별 주변합을 보정의 기준값으로 활용하여 최종가중치을 구하게 되며 산업분류별, 종사자 그룹별 주변합은 표 3.2와 3.3에 나와있다.

표본설계시 얻어진 기본가중치와 주변합을 이용하여 $|\widehat{N_{i+}^R} - N_{i+}| < \varepsilon$와 $|\widehat{N_{+j}^R} - N_{j+}| < \varepsilon$을 만족하는 최종가중치, $\widehat{N_{ij}^{(R)}}$를 구한다. 최종가중치 결과는 표 3.4과 같다.

위의 가중치를 이용하여 총계 추정치가 구해진다. 최종 추정식은 최종가중치와 관측값의 곱으로 이루어진다.

$$\widehat{y_R} = \sum_{j=1}^{4}\sum_{i=1}^{5} w_{ij.R} y_{ij}, \qquad i = 1,2,3,4,5; \quad j = 1,2,3,4,$$

표 3.4: 갈퀴법에 의한 최종 가중치

산업중분류 \ 종사자그룹	5명~19명	20명~49명	50명~99명	100명~299명	계
17	164.8557	101.4850	59.7583	33.90093	360
24	369.8370	222.7984	81.0070	89.35762	763
28	1549.4670	480.8501	132.7233	64.95957	2228
29	1761.3780	584.8341	159.1537	64.63380	2570
33	296.4619	73.0324	27.3576	15.14807	412
계	4142	1463	460	268	6333

여기서,

i는 산업 중분류,

j는 종사자수 그룹,

w_{ij}는 i산업 j 종사자 수 그룹에 해당하는 가중치,

y_{ij}는 i산업 j 종사자 수 그룹에 해당하는 종사자수

이다.

3.2.2. BLS 방법

3.1절에서도 언급하였듯이 추출된 500개의 자료에 이상점 보정, 집계 후 보정, 무응답 보정 등을 위한 선행 자료처리가 행해졌다. 이에 관한 예제 표본이 표 3.5에 나와 있다. 이상점은 1번 unit에, 폐업(Out of Business: Oob)은 4번과 497번, 무응답은 500번 그리고 집계 후 보정은 7번 Unit에 적용되었다. BLS의 경우는 갈퀴법과 달리 각 사업체 마다 가중치가 적용된다.

1 단계. 집계 후 보정: 식 (2.1)을 이용하여 REAG에 해당되는 unit의 보정인자인 $f_i^{REAG} = 0.7$이 되고 나머지 unit들은 1이 된다.

2 단계. 무응답 보정: 식 (2.2)를 이용하여 NRA요인이 보정된다. 이때 unit 4와 497등은 note가 Oob인 패업인 상태이다. 즉 사업체가 존재하지 않는 상태이므로 무응답 보정 요인에서 Oob에 해당하는 사업체를 제외한다. 따라서

$$\text{NRA factor} = f_i^{NRA} = \frac{\sum_{i \in h} w_i y_i}{\sum_{i \in h-NR} w_i y_i} = 1.016, \tag{3.1}$$

여기서 h는 모집단에서 폐업에 해당하는 사업체를 제외한 집단이다.

3 단계. 이상값 보정: 식 (2.3)와 (2.4)을 이용하면 OAF에 해당되는 사업체의 보정값은 $f_i^{OAF} = 0.077$으로 다음에 의해 구해진다.

$$f_i^{OAF} = \frac{1}{w_i \times f_i^{REAG} \times f_i^{NRA}}. \tag{3.2}$$

표 3.5: BLS를 위한 표본 자료 형태

Unit	Orig_wei	Asgn_jong	Asgn_wei_jong	Col_Status	Notes	Rep_jong
1	12.666	13	164.658	collect	**out**	58
2	12.666	19	240.654	collect		18
3	12.666	9	113.994	collect		13
4	12.666	12	151.992	**Oob**		
5	12.666	5	63.33	collect		13
6	12.666	8	101.328	collect		18
7	12.666	8	101.328	collect	**rea**	22
8	12.666	17	215.322	collect		8
9	12.666	5	63.33	collect		17
10	12.666	13	164.658	collect		12
⋮	⋮	⋮	⋮	⋮	⋮	⋮
496	12.666	13	164.658	collect		13
497	12.666	22	278.652	**Oob**		
498	12.666	22	278.652	collect		40
499	12.666	21	265.986	collect		30
500	12.666	46	582.636	**Refusal**		

여기서, Unit: 사업체 번호, Orig_wei: 기본가중치, Asgn_jong: 표본설계시 사업체 종사자수,
Asgn_wei_jong: Orig_wei× Asgn_jong, Col_status: 조사상태(Oob: 폐업, Refusal: 응답거부),
Note: 조사 후 자료상태; 이상값(Outlier: out) 보정값(REAG: rea),
Rep_jong: 조사에서 얻은 관측값; 사업체 종사자수

또한 이상값이 아닌 보정인자는 $f_i^{OAF(no)} = 1.02$가 되며 이는 다음식에서 구해진다.

$$f_i^{OAF(no)} = \frac{\sum_i^n w_i y_i - \sum_{i \in outlier} y_i \times w_i \times f^{REAG}}{\sum_i^n w_i y_i - \sum_{i \in outlier} w_i y_i \times w_i \times f^{REAG} \times f^{NRA}}$$

$$= \frac{\sum_i^n w_i y_i - \sum_{i \in outlier} y_i}{\sum_i w_i y_i - \sum_{i \in outlier} w_i y_i \times f^{NRA}}. \tag{3.3}$$

식 (3.2)와 (3.3)은 식 (2.3)과 (2.4)와 차이가 있어 보인다. 순수하게 이상점 보정만이 필요한 경우 또는 $f_{REAG} = 1, f_{NRA} = 1$인 경우가 식 (2.3)과 (2.4)이다. 만약 $f_{REAG} = 1, f_{NRA} = 1$이 아니면 REAG와 NRA를 보정하게 되며 이 경우 식 (3.2)와 (3.3)을 사용한다.

4 단계. 벤치마크 보정: 일반적으로 가중치 보정의 마지막 단계가 벤치마크 보정이며 식 (2.5)의 Target value를 이용한다.

표 3.6: 최종가중치

Unit	Original weight	REAG Factor	NRA Factor	Outlier Factor	BMF	Fanal
1	12.666	1	1.016	0.077	1.18	1.173
2	12.666	1	1.016	1.02	1.18	15.549
3	12.666	1	1.016	1.02	1.18	15.549
4	12.666	1				
5	12.666	1	1.016	1.02	1.18	15.549
6	12.666	1	1.016	1.02	1.18	15.549
7	12.666	0.7	1.016	1.02	1.18	7.774
8	12.666	1	1.016	1.02	1.18	15.549
9	12.666	1	1.016	1.02	1.18	15.549
10	12.666	1	1.016	1.02	1.18	15.549
⋮	⋮	⋮	⋮	⋮	⋮	⋮
496	12.666	1	1.016	1.02	1.07	14.100
497	12.666	1				
498	12.666	1	1.016	1.02	1.07	14.100
499	12.666	1	1.016	1.02	1.07	14.100
500	12.666	1				

5 단계. 최종가중치: 사업체별 최종가중치는 위에서 얻어진 각 보정인자를 곱하여 얻어진다.

즉 최종가중치는

$$W_i^{Final} = w_i \times f_i^{REAG} \times f_i^{NRF} \times f_i^{OAF} \times f_i^{BMF}$$

이며 결과는 표 3.6에 나와 있다.

여기서 BLS에 의한 총계 추정치는 최종가중치와 조사되어진 값을 이용하여 구해지며 결과는 다음과 같다.

$$\hat{Y}_B = \sum_{i=1}^{500} W_i^{Final} y_i^*,$$

여기서 y_i^*은 조사에서 얻은 자료값이다.

3.2.3. GREG

주어진 설명변수(생산액)과 종속변수(종사자수)를 SAS의 GLM을 이용하여 5개의 산업분류별로 모형을 설정하였다. 선형모형은 다음과 같다.

$$\hat{y}_{ijk} = \hat{\beta}_{ij} x_{ijk}, \quad i = 1, \ldots, 5; \quad j = 1, \ldots, 4; \quad k = 1, \ldots, n_{ij},$$

여기서,

y_{ijk} : i 산업분류, j번째 종사자수 그룹에 속하는 k번째 사업체의 종사자수,

β_{ij} : 회귀계수,

x_{ijk} : i 산업분류, j번째 종사자수 그룹에 속하는 k번째 사업체의 생산액,

n_{ij} : i 산업분류, j번째 종사자수 그룹에 속하는 사업체수

표 3.7: GREG의 추정된 회귀계수와 R^2

산업분류	종사자그룹	R^2	$\hat{B}$	P-value	모집단
17	1	0.3193	0.00198	0.0001	176
	2	0.5818	0.00408	0.0001	96
	3	0.5977	0.0034	0.0001	57
	4	0.727	0.00418	0.0001	35
24	1	0.01	0.00001974	0.0497	384
	2	0.1704	0.00027594	0.0001	212
	3	0.2241	0.00019166	0.0001	85
	4	0.0576	0.00002048	0.026	86
28	1	0.1284	0.00071374	0.0001	1543
	2	0.0543	0.00021585	0.0001	495
	3	0.5366	0.00256	0.0001	127
	4	0.1996	0.00043328	0.0001	67
29	1	0.0159	0.00006198	0.0001	1746
	2	0.3869	0.00188	0.0001	592
	3	0.293	0.00093894	0.0001	168
	4	0.5823	0.00167	0.0001	68
33	1	0.4269	0.00313	0.0001	298
	2	0.4798	0.00283	0.0001	73
	3	0.6908	0.00413	0.0001	28
	4	0.7672	0.00488	0.0001	17

이다.

계산한 선형모형의 추정된 회귀계수 $\hat{B}$와 R^2은 표 3.7과 같다.

따라서 GREG의 총계 추정량은 다음과 같다.

$$\hat{Y}_{GR} = \hat{t}_{y\pi} + \sum_{j=1}^{J} \hat{B}_j (t_{x_j} - \hat{t}_{x_j\pi}),$$

여기서,

$w_{ij} = 1/\pi_{ij}$, 기본가중치(추출율의 역수),

$\hat{t}_{y_\pi} = w_{ij} \sum_{k\in S} y_{ijk}$, 기본 가중치와 관측값의 곱,

$\hat{t}_{x_j\pi} = w_{ij} \sum_{k\in S} x_{ijk}$, 기본 가중치와 표본 사업체의 생산액의 곱,

$\hat{B} = (\hat{\beta}_1, \cdots, \hat{\beta}_J) = (\sum_S x_k x_k'/\sigma_k^2\pi_k)^{-1} \sum_S x_k y_k/\sigma_k^2\pi_k$, GLM에서 얻은 회귀계수

이다.

3.3. 비교 통계량 및 최종 결과

3.3.1. 비교 통계량

Raking, BLS, GREG 세가지 추정방법에 의한 결과를 MSE, Coverage(CVR), CV, Non-centrality(NC) 그리고 Large-Error(LE)를 이용하여 비교하였다. 일반적으로 간편하면서도

절대적인 비교 방법을 제시하기는 매우 어렵다. 그러나 MSE, CVR, CV 통계량은 일반적으로 많이 사용되며 CV, CVR는 흔히 MSE를 뒷받침해주는 통계량이다. 이들 통계량의 정의는 다음과 같다.

$\mathrm{MSE}=1/R\sum_R(y-\hat{y})^2$이며, $\mathrm{CV}=\sqrt{\mathrm{Var}(\hat{y})}/\hat{y}$이다. 또한 CVR은 추정치의 값이 95% 신뢰구간($\pm 1.96\sqrt{\mathrm{Var}(\hat{y})}$) 안에 포함되는 확률/비율로 계산 되어진다. 추정량의 비교에 MSE는 매우 신뢰할 수 있는 추정량임에 틀림없으나 본 논문에서는 NC, LE와 같은 비교 통계량을 추가하여 비교하였다.

NC는 추정치의 값이 $\pm 0.675 \times \sqrt{\mathrm{Var}(\hat{y})}$에서 벗어나는 비율을 의미하며 여기서 0.675는 자료가 정규분포일 경우 75백분위수를 의미하므로 NC는 자료가 중앙에서 얼머나 벗어나는지를 알 수 있게 해 주는 통계량이다.

LE는 max | 95%percentile – 5%percentile | /ture total로 계산되며 이는 조사오차(survey error) 즉, 추정치와 참값을 차이, 손실함수(loss function)에서 최대최소요건(minimax criteria)의 측도이다. 이 척도는 CV 보다 쉽게 이해할 수 있는 척도로 특히 공식통계에서 매우 중요하게 사용될 것으로 생각된다.

3.3.2. 비교 결과

산업중분류(17, 24, 28, 29, 33)와 종사자수 그룹(1, 2, 3, 4)에 각 3%, 10%, 20%의 집계 후 보정, 이상점 보정, 무응답 및 폐업 보정이 적용 가능하도록 자료를 만든 후 1,000번을 반복 시행하였다. 결과 중에서 Case1은 위의 모든 보정이 가능하도록 자료를 처리하였으며 Case2의 경우에는 자료에 각 3%, 10%, 20%의 무응답(refusal) 및 폐업(oob)보정만을 적용하여 1,000번의 모의실험을 실시하였다. 이는 BLS의 경우에만 집계 후 보정과 이상점 보정의 처리가 가중치에 적용되기 때문에 Case1의 경우 당연히 BLS가 가장 우수한 결과를 줄 것으로 예상되기 때문이다. 따라서 주어진 세 방법의 우수성을 다양한 조건에서 비교하기 위해 집계 후 보정과 이상점 보정이 적용되지 않은 경우도 살펴보았다. 비교 통계량은 MSE, CVR, CV, LE, NC이 사용되었으며 결과는 표 3.8에 나와있다.

4. 결론

표 3.8의 5개 비교통계량에서 MSE가 자료의 분포보다는 추정값의 차이를 비교하기 위해 사용되는 반면 그 외의 CV, LE NC, CVR은 자료의 분포가 고려된 통계량이다. 우선 MSE를 보면 BLS에서는 case1과 case2의 결과에 크게 변함이 없게 나타났다. 그러나 BLS를 제외한 두 방법은 case1보다 case2의 추정값의 정도가 크게 향상되었다. case2는 outlier와 REAG요인이 고려되지 않은 경우로, raking과 GREG는 outlier와 REAG요인을 보정하는 기능이 없기 때문에 MSE값이 작아지는 것은 타당한 결과이다. MSE의 결과만을 기준으로한다면 outlier, REAG의 보정요인이 있는 경우에는 BLS 보정방법이 바람직하며 그렇지 않은 경우에는 Raking나 GREG가 추전할 만한 방법이다. 특히 보정요인의 비중이 많은 경우, 즉 보정요인이 3%에서 20%인 경우들의 MSE값을 보면 GREG방법이 효과적임을 알 수 있다. 이는 GREG의 방법은 BLS와 Raking과는 달리 무응답과 같은 경우에 모형을 통한 대체 효과를 볼

표 3.8: 각 추정량의 최종 비교 결과

비교 통계량	추정방법	3%		10%		20%	
		case1	case2	case1	case2	case1	case2
MSE	Raking	2475.08	1589.60	4718.09	1750.71	7809.98	2268.76
	BLS	1902.06	1866.92	2633.85	2697.80	3860.30	4459.11
	GREG	2593.34	1746.89	4471.94	1753.06	6950.56	2143.44
CV	Raking	11.85	8.63	17.45	9.27	18.73	9.65
	BLS	9.47	8.42	12.22	8.50	15.92	7.92
	GREG	3.62	10.59	9.78	3.90	3.51	13.80
LE	Raking	0.10	0.07	0.15	0.07	0.15	0.07
	BLS2	0.08	0.07	0.11	0.07	0.14	0.07
	GREG	0.01	0.10	0.09	0.02	0.02	0.14
NC	Raking	0.89	0.84	0.93	0.84	0.93	0.85
	BLS	0.85	0.87	0.84	0.92	0.86	0.85
	GREG	0.90	0.87	0.91	0.83	0.90	0.82
CVR	Raking	0.47	0.45	0.43	0.45	0.30	0.42
	BLS	0.41	0.37	0.43	0.29	0.39	0.18
	GREG	0.44	0.39	0.44	0.44	0.37	0.48

수 있기 때문이다.

이제 분포도가 고려된 CV, LE, NC, CVR 통계량을 살펴보기로 하자. 우선 CVR은 자료의 정규성을 볼 수 있는 척도로 모든 경우에 매우 낮은 커버리지를 보이고 있어 자료의 정규성은 약하다고 볼 수 있다. 마찬가지로 NC값 역시 자료가 중앙에서 벗어나 있음을 말해주고 있고 이 역시 약한 정규성을 보여주고 있다. LE 기준을 보면 전반적으로 10% 이하로 안정적이다. LE는 CV와 비교해서 볼 수 있다. 본 자료와 같이 약한 정규성을 띠고 있는 경우에는 CV보다 LE값을 판단 기준으로 사용되는것이 바람직하다. 여기서 NC와 LE를 동시에 고려해 보면 전체적으로 치우져 있는 분포로 판단되며 각 그룹 안에서의 자료분포도는 자료 전체의 분포도보다 더 정규성을 가질 것으로 예측할 수 있다.

다음으로 case별 결과를 살펴보기로 하자. 대부분의 경우 case2가 case1에서 보다 좋은 추정치를 주고 있으며 case2는 보정의 필요성이 작기때문에 이 결과는 타당해 보인다. 또한 보정요인의 비중정도가 3%, 10%, 20%인 경우들을 case별로 보면 같은 모집단을 이용한 것이기 때문에 NC, CVR의 값은 거의 같은 값을 보여준다.

결론적으로 자료의 추정에 앞서 outlier와 REAG의 보정이 필요한 경우 BLS의 가중치 방법이 가장 좋을 수 있다. 그리고 갈퀴법은 case2의 경우에 좋은 추정값을 주고 있다. 그러나 outlier와 REAG의 경우를 포함한 자료에서는 3가지(BLS, raking, GREG)중 가장 나쁜 추정값을 내고 있다. 이는 갈퀴법이 알려진 주변정보의 값, 즉 벤치마크값에 매우 의존적이기 때문이다. 따라서 주변정보가 없는 경우 갈퀴법을 적용하는 것은 바람직하지 않다. GREG의 경우에서는 우선 BLS와 갈퀴법과는 달리 벤치마크 값이 없는 경우에도 적용가능하다. 또한 GREG는 case2의 경우 refusal과 oob의 자료형태에 모형에서 얻은 예측값을 대체해 주는 효과를 냄으로써 좋은 추정값을 줄 수 있다. 그러나 case1의 경우 outlier와 REAG 요인을 모형에 추가한 후 보정하기 위해서는 추가적인 연구가 필요해 보인다. 따라서 적절한 모형의 선택

이 가능한 경우 GREG의 적용은 매우 효과적일 것으로 판단된다.

참고문헌

신민웅, 이상은 (2001). <표본조사를 위한 표본설계>, 교우사.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *The Annals of Mathematical Statistics*, **11**, 427–444.

Oh, H. L. and Scheuren, F. (1987). Modified raking ratio estimation, *Survey Methodology*, **13**, 209–219.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer, New York.

[2008년 4월 접수, 2008년 5월 채택]

Comparison of Estimation on Sample Survey: Focusing on Weight Adjustment*

Sang Eun Lee[1)]

ABSTRACT

In sample design, it is usually planned by purpose and the range of the announcing statistics from the survey. After survey, getting a proper and decent statistics, applying the proper weights on the results of survey is very important and necessary. Therefore in this study, three estimation methods which are raking, BLS and general linear regression method are compared with MSE, Coverage, CV, LE and NC.

Keywords: Raking method, BLS method, GREG, large error, non-centrality.

* This work was supported by Kyonggi University Research Grant.

1) Associate Professor, Dept. of Applied and Information Statistics, San 94-3, Yongtong, Suwon, Kyonggi, Korea.
E-mail: sanglee62@kyonggi.ac.kr