단위 무응답 하에서 사후층화와 보정에 관하여

A Post Stratification and Calibration under the Unit Nonresponse

손창균*, 홍기학**, 이기성***

<요 약>

단위 무응답 상황에서의 사후층화 추정, 회귀 추정, 보정 추정 또는 일반화 래킹 추정을 포함하는 다양한 추정 방법에 대해 살펴보았다. 이러한 모든 방법들은 무응답 하에서 보정 추정의 일반적인 형태로 나타나며, 또한 이들은 주어진 범주형 보조변수 하에서의 사후층화 이론에 기초를 두고 있다.

<Abstract>

In this paper we consider a various estimation methods including the post-stratification estimation, regression estimation and calibration estimation or a generalized raking estimation under a unit nonresponse. All of them have a common type of calibration estimation based on the post-stratification for a categorical auxiliary variables.

Ⅰ. 서 론

모집단에서 단순임의표본을 추출하여 이들을 관찰한 다음 각 단위의 특성에 따라

* 동신대학교 컴퓨터 응용학과군, (520-714) 전남 나주시 대호동 (☎ 061-330-3359)
e-mail : ckson85@blue.dongshinu.ac.kr

** 동신대학교 컴퓨터 학과, (520-714) 전남 나주시 대호동 (☎ 061-330-3353)
e-mail : khhong@blue.dongshinu.ac.kr

*** 우석대학교 전산통계학과, (565-701) 전북 완주군 삼례읍 후정리 (☎ 063-290-1522)
e-mail : gisung@core.woosuk.ac.kr

층화하는 방법을 사후층화라 하며, 이를 근거로 이루어진 추정을 사후층화 추정이라 한다. 예를 들어, 여론조사에서 단순임의표본을 추출한 다음 성별, 연령별 찬성율이나 지지율을 추정하거나 소득조사에서 단순임의표본을 추출한 다음 연령별, 직업별 소득액 등을 추정하는 것을 사후층화 추정이라 할 수 있다.

대부분의 경우에 있어서 인구학적인 특성 요인인 성, 지역, 나이, 교육, 소득 수준에 따라 사후층화를 하게 되는 데, 이 때 모집단 비율과 표본 비율을 맞추기 위하여 가중치를 이용하여 자료를 보정하게 되며, 이러한 과정을 통해 모집단 비율을 추정하는 것을 보정 추정이라 한다.

많은 학자들이 사후층화 추정과 보정 추정에 대한 연구를 하였으며, 특히 Holt와 Smith(1979)는 사후층화 추정, Bethlehem과 Wouter(1987)는 회귀 추정, Deville과 Sarndal(1992)은 보정 추정에 대한 연구를 하였다.

그리고, Deville 외(1993)는 처음으로 일반화된 보정추정량을 도출했으며, 보조변수가 범주형일 때 보정추정량의 특별한 경우로서 사후층화추정량과 래킹 추정량을 설명하였다. 그 밖에도 많은 표준적인 방법들을 이용하여 범주형 보조정보에 따라 모집단에 대한 구조를 추론하는 연구들이 이루어졌다. 일반적으로 보조정보를 이용하게 되면 첫째, 관심변수와 높은 상관 관계가 있는 보조정보를 사용할 수 있을 경우에 변수들의 표본분산을 줄일 수 있으며, 둘째, 무응답(nonresponse)과 비 포괄성(non-coverage)에 의한 편향을 감소시킬 수 있으므로 조사의 질을 높일 수 있다.

한편, Smith(1990)는 보다 폭넓은 정의로서 사후층화는 표본을 관찰한 후에 유사한 집단에 속하는 단위들로 형성된 자료를 분류하는 방법으로 일반적으로 추가적인 모집단 보조정보뿐만 아니라 표본에 대한 보조정보를 이용할 수 있다고 하였다.

최근에 Zhang(2000)은 사후층화와 보정은 모집단에 대한 구조를 파악하는 반면, 사후층화추정량과 보정추정량은 추정을 목적으로 하고 있다고 언급하고 있다. 그리고, 사후층화는 구조화된 선형모형과 유사하며, 보정은 구조화된 선형모형에서 일부 또는 모든 상호작용 효과를 제거하여 축소된 선형모형의 형태를 갖는다고 사후층화와 보정의 차이를 설명하고 있다. 또한, 보조변수가 범주형일 때 가중치 조정을 해 본 결과 사후층화추정량의 특별한 형태로서 보정추정량이 된다는 사실을 보였다. 이는 Deville 외(1993)의 결과와는 상반되는 측면에서 접근방법이다.

본 논문에서는 Zhang의 관점에서 완전응답 하에서 사후층화와 사후층화 추정 및 보정 추정에 대하여 살펴보고, 사후층화와 보정의 조건부 편향에 대해서도 다루어 보고자 한다. 그리고, 무응답 상황 하에서 무응답으로 인한 편향을 감소시키기 위해 사용된 보조정보의 형태가 범주형일 경우에 대해 고찰하고, 이와 관련하여 사후층화추정량과 보정추정량과의 연관성에 대해 살펴보고자 한다.

Ⅱ. 완전응답 하에서 사후층화, 사후층화 추정 그리고 보정 추정

이 장에서는 Zhang의 관점에서 완전응답 하에서 사후층화와 사후층화 추정에 대하여 살펴보고, 사후층화와 보정 추정에 대하여 다루어 보고자 한다,

2.1 사후층화와 사후층화 추정

완전응답 하에서의 사후층화와 사후층화 추정을 구분해 보면, 전자는 보조정보에 따른 모집단의 구조를 파악하는 것이며, 후자는 이러한 모집단의 구조를 특정한 추정방법으로 추정하는 것이다.

조사에 있어서 y를 관심변수라 하고, x를 보조변수라 하며, $U=\{1,2,\cdots,N\}$를 모집단이라 정의하자. 사후층화는 표본을 추출한 후에 모집단을 x의 값에 따라 H개의 상호 배반이고 포괄적인 모집단 사후층으로 분할하는 것으로 $U=\cup_{h=1}^{H}U_h$로 표현할 수 있으며, $h\neq g$에 대해 $U_h\bigcap U_g=\phi$이 성립되어야 한다. 한편, 표본 s에 대한 사후층화는 표본 사후층인 $s_1, s_2, \cdots, s_H$로 분할된다. 사후층 h에 대해 보조변수 x에 대응되는 사후층은 (s, U)에서 $\{(s_1, U_1), \cdots, (s_H, U_H)\}$로 구조적인 분할이 된다. 즉, 동질적인 부차모집단 U_h로부터 추출된 표본 s_h로 간주할 수 있다.

사후층화로부터 모집단 총합 $Y=\sum_{k\in U}y_k=\sum_h(\sum_{k\in U_h}y_k)=\sum_{h=1}^{H}Y_h$를 얻게 된다. 모집단 사후층의 주변비율(marginal proportion)의 정보인 N_h/N를 알고 있고, 표본 사후층의 크기가 0인 곳이 없으면 Y에 대한 사후층화추정량은 $\hat{Y}_{post}=\sum_h \hat{Y}_h$의 형태가 된다. 이 때, N_h는 U_h의 크기이고, $\hat{Y}_h$을 구하는데 필요하며, 실제 층내 추정량 $\hat{Y}_h$는 포함확률 π_k가 각각의 U_h에서 일정한지, 그렇지 않은 지에 따라 다르다(Smith, 1991). 그리고, $k\in U_h$에 대해 포함확률이 $\pi_k=\pi_h$인 경우 Y_h는 단순한 확장으로 추정가능하며, 이 추정량을 단순사후층화추정량(simple post-stratification estimator)이라 하며, 다음과 같이 나타낼 수 있다.

$$\hat{Y}_{post}=\sum_h\sum_{k\in s_h}y_k/f_h\,. \qquad <2.1>$$

여기서, $f_h=n_h/N_h$로서 실현된 층내 추출률이며, n_h는 s_h의 크기를 나탄낸다.

복합설계 하에서 π_k가 층화나 집락과 같은 각각의 사후층내에서 서로 다를 때,

Y_h의 비편향추정량은 사후층내에서 Horvitz-Thompson 추정량 $\overline{Y}_h = \sum_{k \in s_h} y_k / \pi_k$

로 주어진다. Simth(1991)는 Hajek 추정량을 이용하여 다음과 같은 추정량을 제시하였다.

$$\hat{Y}_h = N_h(\sum_{k \in s_h} y_k / \pi_k)/(\sum_{k \in s_h} 1/\pi_k) \qquad <2.2>$$

이 추정량은 각각의 모집단 사후층인 U_h내에서 비 추정량을 응용한 형태이며, $k \in s_h$에 대해 가중치는 $(N_h/\pi_k)(\sum_{k \in s_h} 1/\pi_k)$가 된다.

사후층화 추정과 연관된 주된 이론적 관심사는 조건부 추론에 있다. Holt와 Smith(1979)는 단순임의추출의 경우에 있어서 사후층화추정량의 성질들은 사후층의 실현된 표본 $(n_1, n_2, \cdots, n_H)$에 대해 조건부를 취해야만 얻을 수 있음을 언급하였다. 한편, Rao(1985)는 복합설계의 경우 n_h만 조건부를 취할 때 $\{\pi_k : k \in s_h\}$가 고정이 아니기 때문에 분포를 쉽게 구할 수 없다는 점을 지적하였다.

한 예로서 층화임의 추출설계를 고려하고, 이 때 사후층화가 층에 교차하여 나타났다고 하자. 만일 사후층화를 모집단에 대한 주민등록표에 기초하였다면, 원칙적으로 뽑힌 표본과 이 등록표를 결합함으로서 사후층화가 가능하다. 말하자면, 사후층화는 층화지표를 보조변수로서 사용하여 확장이 가능하며, 따라서 결합된 등록표는 필연적으로 주변총합 N_h를 제공한다. 그러므로 일반적인 경우에 사후층화는 추가적인 보조정보로서 포함확률 π_k를 포함한다. 사후층화 추정뿐만 아니라 이러한 방법의 문제점은 크기가 0인 표본 사후층이 존재한다는 것이다. 이러한 문제의 다른 측면은 모집단 사후층의 총합들을 항상 이용가능 하거나, 신뢰할 만한 것이 아니라는 것이다. 크기가 0인 표본 사후층을 무시한 사후층화 추정은 크기가 0이 아닌 y_k에 대해 과소 편향된다(Jager, 1986). 다소 직관적인 방법으로 사후층을 붕괴(결합)함으로서 크기가 0인 셀을 줄이거나 없앨 수 있으며, 이에 대해 보정 추정은 일반적인 대체방법이 될 것이다(Deville과 Sarndal, 1992 ; Deville외, 1993).

2.2 사후층화와 보정 추정

주어진 표본에 대한 가중치 $\{d_k = 1/\pi_k : k \in s\}$가 기지인 보조변수의 모집단 총합을 이용하여 재조정된다면, 모집단 총합의 관점에서 보정된다고 말할 수 있다. 보정된 총합추정량을 $\hat{Y}_w = \sum_s w_k y_k$이라 하면, 다음과 같은 기지의 값으로부터 구할 수 있다.

$$X = \sum_{h \in R} N_h, \quad R \subseteq \{1, 2, \cdots, H\} \qquad <2.3>$$

모집단의 주변총합과 관련된 보정을 일반화 래킹(generalized raking)이라 한다.(Deville외, 1993). 래킹의 방법을 이용하면, 크기가 0인 표본 사후층의 경우에 비록 래킹이 불안정하거나, 수렴성을 보장하지 못한다 할 지라도(Oh와 Scheuren, 1987), 어느 정도는 사후층이 붕괴되는 것을 방지할 수 있는 장점이 있다. 크기가 0인 표본 사후층이 존재할 때, 래킹은 사후층에 대한 추정량을 얻을 수 있는 반면에 보정은 불가능하며, 또한 $\sum_{k \in s} w_k y_k$의 형태의 선형추정량의 경우에도 불가능하다.

모집단 사후층에 대한 보정은 다음과 같은 보정방정식으로 표현될 수 있다.

$$\begin{aligned} X = (N_1, N_2, \cdots, N_H) &= \sum_{k \in s} w_k \boldsymbol{x}_k \\ &= \sum_h x_h \left(\sum_{k \in s_h} w_k \right) \end{aligned} \qquad <2.4>$$

일반적으로 임의의 보조변수에 대해 보정에 대한 더미 변수화를 수행하여 보정방정식 $X = \sum_{k \in s} w_k \boldsymbol{x}_k$의 형태를 가지는 0과 1로 이루어진 보조변수 벡터로 나열함으로서 $\sum_{k \in s} w_k \boldsymbol{x}_k = (N_1, N_2, \cdots, N_H)$, $\sum_{k \in s_h} w_k = N_h$와 같이 표현할 수 있다.

다음으로 선형 보정과 회귀 추정에 대하여 살펴보기로 하자. 보정방정식 하나 만으로는 가중치를 결정하는 데에는 불충분하므로 2가지 이상의 가정이 필요한데 (a) $d_k = 1/\pi_k$가 되는 초기가중치의 집합 $\{d_k : k \in s\}$에 대한 가정과 (b) G로 표현되는 $\{d_k : k \in s\}$와 보정된 가중치 $\{w_k : k \in s\}$와의 거리를 측정하는 거리함수에 대한 가정이다. Deville 외(1993)는 G에 대해 $r_k = w_k / d_k$를 선택하여 전체 표본에 대한 거리 측도로서 $\sum_{k \in s} d_k G(r_k)$를 이용하였다. 여기서의 개념은 보정방정식에 대해 초기가중치 $\{d_k\}$와 최소의 거리가 되는 새로운 가중치 $\{w_k\}$를 찾는 것이다.

$g = \partial G / \partial r$을 1차 편미분이라 하자. $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_J)'$를 라그란쥐 승수라 하면 $\{w_k : k \in s\}$에 대한 해를 구하면 된다.

$$\frac{\partial \left[\sum_{k \in s} d_k G(r_k) - \left(\sum_{k \in s} w_k \boldsymbol{x}_k - X \right) \boldsymbol{\lambda} \right]}{\partial w_k} = g(r_k) - \boldsymbol{x}_k' \boldsymbol{\lambda} = 0 \qquad <2.5>$$

$F(u) = g^{-1}(u)$를 g의 역함수라 하자. 보정된 가중치는 $w_k = d_k F(\boldsymbol{x}_k' \boldsymbol{\lambda})$로 주어지며, 여기서 라그란쥐 승수 $\boldsymbol{\lambda}$는 보정방정식 $X = \sum_{k \in s} d_k F(\boldsymbol{x}_k' \boldsymbol{\lambda}) \boldsymbol{x}_k$을 만족한다.

Deville 외(1993)가 제시한 여러 가지 거리함수들 중에서 특별히 $G=(r-1)^2/2$일 경우에는 $g=r-1$이고, $F(u)=1+u$인 선형방법으로서 보정된 가중치는 다음과 같다.

$$\begin{aligned} w_k &= d_k(1+ \boldsymbol{x}_k' \boldsymbol{\lambda}) \\ &= d_k g_k \end{aligned} \qquad <2.6>$$

여기서, $g_k=\left[1+(X-\sum_{k\in s} d_k \boldsymbol{x}_k)(\sum_{k\in s} d_k \boldsymbol{x}_k' \boldsymbol{x}_k)^{-1} \boldsymbol{x}_k\right]$이다.

이 값은 특별한 경우에 $\{d_k: k\in s\}$에 기초한 회귀추정치와 동일하다(Bethleham과 Wouter, 1987).

관심변수의 모집단 벡터 $\boldsymbol{y}=(y_1, y_2, \cdots, y_N)'$를 보조변수벡터와 동일한 차원의 열벡터 $\boldsymbol{\beta}$를 통해 $\varepsilon_k=y_k-\boldsymbol{x}_k'\boldsymbol{\beta}$를 변환할 수 있으며, 특별히 모집단에 기초한 최소제곱 적합은 $\boldsymbol{\beta}=(\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}$로 정의된다. 여기서, $\boldsymbol{x}$는 k번째 행에 대해 $\boldsymbol{x}_k$인 모집단 보조행렬이다. 표본에 근거하여 $\boldsymbol{\beta}$는 $\boldsymbol{x}'\boldsymbol{x}$와 $\boldsymbol{x}'\boldsymbol{y}$의 추정치의 결합으로 추정할 수 있다. 특히 동일한 가중치 $\{d_k: k\in s\}$가 $\boldsymbol{x}'\boldsymbol{x}$와 $\boldsymbol{x}'\boldsymbol{y}$에 사용되었다면, 이미 앞에서 언급한 선형 보정추정량과 같은 가중치를 얻게 된다.

Deville과 Sarndal(1992)은 선형 보정추정량 $\hat{Y}_{greg}$을 다음과 같이 표현하였다.

$$\begin{aligned} \hat{Y}_{greg} &= X\tilde{\boldsymbol{\beta}}+\tilde{\varepsilon} \\ &= X\tilde{\boldsymbol{\beta}}+\sum_{k\in s} d_k(y_k \boldsymbol{x}_k \tilde{\boldsymbol{\beta}}) \\ &= (\sum_{k\in s} d_k y_k)+(X-\sum_{k\in s} d_k \boldsymbol{x}_k)\tilde{\boldsymbol{\beta}} \\ &= \sum_{k\in s} w_k y_k \end{aligned} \qquad <2.7>$$

여기서, $\tilde{\boldsymbol{\beta}}=(\sum_{k\in s} d_k \boldsymbol{x}_k' \boldsymbol{x}_k)^{-1}(\sum_{k\in s} d_k \boldsymbol{x}_k' y_k)$가 된다.

일반화 회귀추정에서는 보정방정식 $X=\sum_{k\in s} w_k \boldsymbol{x}_k$를 만족하는 가중치를 구한 후에 $\{d_k: k\in s\}$에 근거한 선형 보정으로 간주할 수 있다. 또한, 극단적인 경우로서 사후층화 추정은 사후층에 대한 지시자로 더미변수화 하여 얻을 수 있다(Sarndal 외, 1992). 이러한 방법으로 사후층화는 보조변수들 간의 모든 "상호작용"을 포함한 구조화된 선형모형으로 표현되며, 보정은 주효과 모형과 같이 포화모형의 축소를 나타낸다.

Deville과 Sarndal은 가중된 전체 거리식 G에 개별적인 계수 $1/q_k$가 첨가된 보다 일반적인 거리함수의 집합을 고려하였다. 이들은 특별히 선형방법이 점근적으로 모든 보정추정량들에 대해 1차 근사함을 보였고, 이와 동시에 반복적인 과정이 필요 없기 때문에 가장 수렴속도가 빠르다고 언급하였다. 가중치가 보정된 추정치는 적용한 방법에 따라 다소 차이가 있을 수 있지만, 선형방법은 표본수가 작을 경우 음의 가중치를 초래할 수 있다는 단점이 있다. 이에 대해 Jayasuriya와 Valliant(1996)은 반복 알고리즘을 적용하여 가중치의 범위를 제한하여 회귀추정량을 전개하였다. 기본적으로 보정된 가중치의 상한과 하한을 결정하는 것으로 가중치의 비율 w_k/d_k이 3 또는 4

를 초과하면 크다고 고려된다. 각각의 반복과정 이후에 이 범위를 벗어나는 가중치들을 버리고, 나머지 표본들에 대해서 적합 알고리즘을 다시 시작하는 방법을 사용하였다.

식<2.6>으로부터, 선형 보정가중치들의 부호가 $\sum_{k\in s} d_k \boldsymbol{x}_k' \boldsymbol{x}_k$의 역행렬에 의해 좌우됨을 알 수 있다. 이에 대해 Chambers(1996)는 소위 능형-회귀를 적용하여 동일한 차원의 양의 대각행렬 $\boldsymbol{D}$를 보정된 가중치의 식으로부터 $\left(\sum_{k\in s} d_k \boldsymbol{x}_k' \boldsymbol{x}_k\right)^{-1}$에 $\left(\boldsymbol{D}^{-1}+\sum_{k\in s} d_k \boldsymbol{x}_k' \boldsymbol{x}_k\right)^{-1}$를 대입하였다.

Ⅲ. 완전응답 하에서 사후층화와 보정 추정량의 조건부 편향

이 장에서는 Zhang의 관점에서 완전응답 하에서 사후층화와 보정의 조건부 편향에 대하여 크기가 0인 표본 사후층이 없는 경우와 있는 경우로 나누어 살펴보고자 한다.

3.1 크기가 0인 표본 사후층이 없는 경우

우선 표본 사후층들이 모두 크기가 0이 아니라고 가정하면, $1 \le h \le H$에 대해 $n_h > 0$이 성립한다. 식<2.7>의 일반화 회귀추정량과 식<2.1>의 단순 사후층화추정량은 둘 다 선형추정량에 속하므로 이들 각각의 가중치에 대해 둘 간의 차이를 $v_k = w_k - q_k$로 표현할 수 있으며, 여기서 w_k는 일반화 회귀추정방법 하에서의 가중치이며, $k \in s_h$에 대해 $q_k = 1/f_h$이다.

보정된 모집단 총합의 식<2.7>로부터, 가중치 w_k와 q_k의 관계는 $\sum_{k\in s} v_k \boldsymbol{x}_k = \sum_{k\in s} w_k \boldsymbol{x}_k - \sum_{k\in s} q_k \boldsymbol{x}_k = 0$가 성립한다.

임의의 고정된 β에 대해, $\varepsilon_k = \varepsilon_k(\beta) = y_k - x_k'\beta$라 하면, 일반화 회귀추정량은 다음과 같이 다시 표현할 수 있다.

$$\begin{aligned} \hat{Y}_{greg} &= \sum_{k\in s} w_k y_k = \sum_{k\in s} q_k y_k + \sum_{k\in s} v_k y_k \\ &= \hat{Y}_{post} + \sum_{k\in s} v_k(x_k\beta + \varepsilon_k) \\ &= \hat{Y}_{post} + \sum_{k\in s} v_k \varepsilon_k \end{aligned} \qquad <3.1>$$

여기서, $\sum_{k\in s} v_k \varepsilon_k(\beta) = \sum_{k\in s} v_k \tilde{\varepsilon}_k(\tilde{\beta})$가 되도록 해야 하며, $\tilde{\varepsilon}_k$는 일반화 회귀추정방법에 의해 추정된 ε_k이다.

만일, $n = (n_1, n_2, \cdots, n_H)$가 주어진 조건하에서 (a) $k\in U_h$에 대해 $\pi_k = \pi_h$이고, (b) $k\in U_h$에 대해 $w_k = w_h$가 성립하여 $v_k = v_h$이면, 일반화 회귀추정량의 조건부 편향은 간단히 다음과 같다.

$$\begin{aligned} E[\hat{Y}_{greg} - Y \mid n] &= \sum_h v_h E[\sum_{k\in s_h} e_k \mid n] \\ &= \sum_h n_h v_h (\sum_{k\in U_h} \varepsilon_k / N_h) \end{aligned} \qquad <3.2>$$

만일 다음의 방정식 <3.3>이 β에 대한 해가 존재하면, 초기가중치에 관계없이 조건부 비편향이 된다.

$$\sum_{k\in U_h} \varepsilon_k(\beta) = 0 \Leftrightarrow Y_h = X_h\beta \qquad <3.3>$$

여기서, $1 \le h \le H$이며, $X_h = \sum_{k\in U_h} x_k$이다.

일반화 회귀추정에 근거한 변환에서 β는 주어진 모집단에 대해 $\sum_{k\in U} \varepsilon_k^2$을 최소로 한다. 방정식 <3.3>으로부터 $\sum_{k\in U} x_k\varepsilon_k = 0$가 성립한다. 즉, 각각의 주변 수(marginal count)에 대해 잔차의 합은 0이 성립한다. 그러나 이 조건은 식<3.3>에서는 필요 없다. 만일 n에 대한 조건부로부터 층화임의추출과 $k\in U_h$에 대해 $w_k = w_h$가

된다면, 다음이 성립한다.

$$var(\hat{Y}_{greg}|\boldsymbol{n}) = \sum_h n_h(1-f_h)\, w_h^2\, \sigma_h^2 \qquad <3.4>$$

여기서, $\sigma_h^2 = \sum_{k\in U_h}(y_i - \overline{Y})^2/(N_h-1)$ 이다.

위의 조건은 만일 (c) $k\in U_h$에 대해 $d_k = d_h$ 이면, $k\in U_h$에 대해 $w_k = w_h$ 를 만족한다. 이는 보정방정식 $\sum_h x_h(\sum_{k\in s_h} w_k) = X$ 의 조건하에서 거리함수를 최소로 한다.

3.2 크기가 0인 표본 사후층을 가진 경우

$R_0 \cup R_0^c = \{1, \cdots, H\}$이고 $R_0 \cap R_0^c = \phi$라 하며, $h\in R_0$에 대해 $n_0 = 0$이고 $h\in R_0^c$에 대해 $n_h > 0$이라 하자. 또한, $X_0 = \sum_{h\in R_0}\sum_{k\in U_h} \boldsymbol{x}_k$와 $\sum_{k\in s} w_k\, \boldsymbol{x}_k = X$라 하자. 즉, $\sum_{k\in s} q_k\, \boldsymbol{x}_k = X - X_0$이고, $\sum_{k\in s} v_k\, \boldsymbol{x}_k = X_0$이면 다음이 성립한다.

$$\begin{aligned}\hat{Y}_{greg} &= \sum_{k\in s} w_k y_k \\ &= \sum_{k\in R_0^c}(\sum_{k\in s_h} y_k/f_h) + \sum_{k\in s} v_k y_k \\ &= \sum_{k\in R_0^c}(\sum_{k\in s_h} y_k/f_h) + X_0\beta + \sum_{k\in s} v_k \varepsilon_k\end{aligned} \qquad <3.6>$$

앞의 조건 (a)와 (b) 하에서 일반화 회귀추정량의 조건부 편향은 다음과 같이 축소된다.

$$\begin{aligned}E[\ \hat{Y}_{greg} - Y \mid \boldsymbol{n}] &= -\sum_{h\in R_0}\sum_{k\in U_h} y_k + X_0\beta + E_{R^c} \\ &= -(X_0\beta + \sum_{h\in R_0}\sum_{k\in U_h}\varepsilon_k) + X_0\beta + E_{R^c}\end{aligned} \qquad <3.7>$$

여기서, $E_{R^c} = \sum_{h\in R_0^c} n_h v_h(\sum_{k\in U_h}\varepsilon_k/N_h)$ 이다.

이는 다음과 같은 두 방정식에 대해 $\boldsymbol{\beta}$에 대한 해가 존재한다면 초기가중치에 관계없이 조건부 비편향이다.

$$\begin{aligned}Y_0 &= X_0\beta \\ Y_h &= X_h\beta\end{aligned} \qquad <3.8>$$

여기서, $Y_0 = \sum_{h\in R_0}\sum_{k\in U_h} y_k$이고 $h\in R_0^c$이다.

조건부 분산에 대한 식은 $h\in R_0^c$에 대한 합으로서 층화임의추출의 조건과 $k\in U_h$에 대해 $w_k = w_h$의 조건하에서 앞의 식<3.4>와 같다. 식<3.7>은 식<3.6>으로부터 얻어지기 때문에 모집단에 대해 일반화 회귀추정량의 비편향성은 크기가 0인 표본층에 대해 제외된다.

Ⅳ. 단위 무응답 하에서의 사후 추정과 보정 추정

2.1절의 관심변수와 모집단에 대한 가정과 더불어 크기 m인 응답집합 r을 정의하고, 다음과 같이 표본에 포함된 단위 k와 l에 대한 응답확률을 가정하자.

$$\Pr(k\in r \mid s) = \theta_k,\ \Pr(k\&l\in r \mid s) = \theta_{kl}$$

Lundstrom과 Srandal(1999)에 따르면 Deville과 Srandal(1992)이 제안한 선형 거리함수와 보정방정식의 조건하에서 보정추정량은 각각 다음과 같이 구해진다. 즉, 존재하는 보조변수의 수준이 모집단인 경우에 대해 총합추정량은 다음과 같다.

$$\hat{Y}_{wU} = \sum_r d_k v_{Uk} y_k \qquad <4.1>$$

여기서, $v_{Uk} = 1 + q_k(\sum_U \boldsymbol{x}_k - \sum_r d_k \boldsymbol{x}_k)'(\sum_r d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k')^{-1}\boldsymbol{x}_k$이다.

또한, 보조변수의 수준이 표본인 경우에 대한 총합추정량은 다음과 같다.

$$\hat{Y}_{ws} = \sum_r d_k v_{sk} y_k \qquad <4.2>$$

여기서, $v_{sk} = 1 + q_k(\sum_s d_k \boldsymbol{x}_k - \sum_r d_k \boldsymbol{x}_k)'(\sum_r d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k')^{-1}\boldsymbol{x}_k$이다.

4.1 단일 범주형 보조변수

보조변수가 범주형인 경우 사후층화를 위한 보조변수 벡터는 $\boldsymbol{x}_k = \Gamma_k$로서 $\Gamma_k = (\gamma_{1k}, \gamma_{2k}, \cdots, \gamma_{Pk})'$이며, $p = 1, \cdots, P$에 대해 다음과 같다.

$$\gamma_{pk} = \begin{cases} 1, & \text{if } k\in p \text{ group} \\ 0, & \text{otherwise} \end{cases}$$

표본 s의 일부가 그룹 p에 속할 때, s_p라하고, 모집단 U의 일부가 그룹 p에 속할 때, U_p라 정의하자. 그룹들은 상호 배반적이며, 포괄적이라 가정하고, 표본과 모집단에 대해 $s = \cup_{p=1}^{P} s_p$, $U = \cup_{p=1}^{P} U_p$라 하자. 주요 벡터 총합의 원소들을 다음과 같이 정의하자.

$$\Sigma_U \boldsymbol{x}_k = (N_1, \cdots, N_p, \cdots, N_P)', \Sigma_s \boldsymbol{x}_k = (n_1, \cdots, n_p, \cdots, n_P)', \Sigma_r \boldsymbol{x}_k = (m_1, \cdots, m_p, \cdots, m_P)' \quad \text{<4.3>}$$

표본 보조정보 하에서 $\boldsymbol{x}_k$는 각각의 $k \in s$에 대해 기지이므로, 다음을 계산할 수 있다.

$$\Sigma_s d_k \boldsymbol{x}_k = (\hat{N}_1, \cdots, \hat{N}_p, \cdots, \hat{N}_P)' \quad \text{<4.4>}$$

여기서, $\hat{N}_p = \Sigma_{s_p} d_k = N n_p / n$이다.

또한, 모집단 보조정보 하에서는 $\Sigma_U \boldsymbol{x}_k = (N_1, \cdots, N_p, \cdots, N_P)'$임을 알 수 있다.

표본정보에 대해 단순임의추출을 가정하면 보정추정량의 식<4.2>에서 $v_{sk} = n_p / m_p$가 됨으로 이를 대입하여 정리하면 다음과 같이 가중계급 추정량(weighting class estimator)으로 축소된다.

$$\hat{Y}_{ws} = \hat{Y}_{WCL} = \frac{N}{n} \Sigma_{p=1}^{P} n_p \bar{y}_{r_p} \quad \text{<4.5>}$$

여기서, $\bar{y}_{r_p} = \Sigma_{r_p} y_k / m_p$이다.

또한, 모집단 정보에 대해 보정추정량의 식<4.3>에서 $v_{Uk} = N_p n / m_p N$이 되어 이를 대입하여 정리하면 다음과 같이 사후층화추정량으로 축소된다.

$$\hat{Y}_{wU} = \hat{Y}_{post} = \Sigma_{p=1}^{P} N_p \bar{y}_{r_p} \quad \text{<4.6>}$$

사후추정량으로 표현되는 경우 추정량에 나타난 형태는 추출과정이 1단계가 아니라 2단계(추출단계와 응답단계)라는 사실이 숨어 있기 때문에 이상적인 형태가 아니다.

4.2 단일 범주형 보조변수와 수치변수

보조정보에 대한 앞의 가정과 더불어 보다 폭넓은 보조정보를 가정하여 수치적 보조변수인 x를 이용할 수 있다고 하자. 그러면, 보조벡터는 $\boldsymbol{x}_k = (\Gamma_k', x_k \Gamma_k')'$로 정의된다. 표본정보 하에서 보정방정식은 다음과 같이 표현된다.

$$\Sigma_r w_k \boldsymbol{x}_k = \Sigma_s d_k \boldsymbol{x}_k = (\hat{N}_1, \cdots, \hat{N}_p, \cdots, \hat{N}_P, \hat{X}_1, \cdots, \hat{X}_p, \cdots, \hat{X}_P)' \quad <4.7>$$

여기서, $\hat{N}_p = Nn_p/n$이고, $\hat{X}_p = N\Sigma_{s_p} \boldsymbol{x}_k/n$이다.
또한, 모집단 정보 하에서 보정방정식은 다음과 같이 표현된다.

$$\Sigma_r w_k \boldsymbol{x}_k = X = (N_1, \cdots, N_p, \cdots, N_P, X_1, \cdots, X_p, \cdots, X_P)' \quad <4.8>$$

여기서, $X_p = \Sigma_{U_p} \boldsymbol{x}_k$이다.
그러면, $\hat{Y}_{ws}$는 다음과 같다.

$$\hat{Y}_{ws} = \Sigma_{p=1}^{P} \hat{N}_p [\bar{y}_{r_p} + (\bar{x}_{s_p} - \bar{x}_{r_p}) B_p] \quad <4.9>$$

여기서, $\bar{x}_{s_p} = 1/n_p \Sigma_{s_p} \boldsymbol{x}_k$, $\bar{x}_{r_p} = 1/m_p \Sigma_{r_p} \boldsymbol{x}_k$,
$B_p = \Sigma_{r_p}(\boldsymbol{x}_k - \bar{x}_{r_p})(y_k - \bar{y}_{r_p}) [\Sigma_{r_p}(\boldsymbol{x}_k - \bar{x}_{r_p})^2]^{-1}$이다.
또한, $\hat{Y}_{wU}$ 는 다음과 같다.

$$\hat{Y}_{wU} = \Sigma_{p=1}^{P} N_p [\bar{y}_{r_p} + (\bar{X}_p - \bar{x}_{r_p}) B_p] \quad <4.10>$$

여기서, $\bar{X}_p = \Sigma_{U_p} \boldsymbol{x}_k / N_p$이다.

4.3 2개 이상의 범주형 보조변수

사실상 2개 이상의 범주형 보조변수를 가진 경우가 일반적이다. 여기서는 가능한 문제를 다룬다는 의미에서 2원 분류의 경우를 살펴보기로 한다.

2차원의 교차분류를 고려하자. 이들 중 첫 번째를 $p = 1, \cdots, P$인 첨자로서 일단의 그룹을 정의하자. 다음으로 $h = 1, 2, \cdots, H$에 대해 δ_h를 단위 k가 그룹 h에 속하면 1, 그렇지 않으면 0을 가지는 지시변수로 정의하자. 만일 모든 셀의 응답 수가 적절히 커서, 해당 셀에 대해 정보가 존재하게 되면, $\boldsymbol{x}_k$를 PH차원의 벡터로서 정의할 수 있으며, 이 때 단위 k가 셀에 포함됨을 나타내는 $PH-1$개의 0과 나머지 한

개의 1로 된 원소들로 구성된다. 따라서, 보조변수 벡터는 $\boldsymbol{x}_k=(\Gamma_k', \Delta_k')'$로 표현되며, $\Delta_k=(\delta_{1k},\cdots,\delta_{hk},\cdots,\delta_{H-1,k})'$이다. 이 때, $\sum_r d_k q_k \boldsymbol{x}_k \boldsymbol{x}_k'$의 역행렬이 존재하지 않는 경우를 방지하기 위해, 마지막 셀에 대해 h그룹은 제외한다. 이러한 정보를 이용한 경우에 대해 대체방법이 래킹 비 추정량이다(Oh와 Scheuren, 1987). 래킹 비 추정과정은 필연적으로 반복기법을 적용해야 하며, 이 때 반복과정이 수렴한다는 조건하에서는 구해진 가중치는 항상 양의 값을 갖는다.

V. 결 론

Zhang(2000)은 Deville 외(1993)의 주장과는 달리 사후층화의 특별한 경우를 보정으로 설명하고 있으며, 또한 사후층화추정량의 특별한 형태를 보정추정량으로 다루고 있다.

따라서, 본 논문에서는 우선 Zhang의 관점에서 완전응답의 경우 사후층화와 사후층화 추정 및 보정 추정의 연관성에 대하여 다루어 보았고, 사후층화와 보정의 조건부 편향에 대해서도 살펴보았다.

다음으로 Zhang의 관점에서 사후층화 추정과 보정 추정의 연관성을 단위 무응답이 존재하는 경우에 대해 살펴보았으며, 존재하는 보조정보의 수준(모집단/표본)에 따라 사후층화추정량과 가중계급 추정량의 형태로 나타나며, 특히 범주형 보조변수인 경우에 보조변수의 차원에 따라 단일 보조변수인 경우와 2개 이상의 보조변수인 경우에 대해 살펴보았다. 그 결과, 단위 무응답이 존재하는 경우 사후층화와 보정 그리고 사후층화 추정과 보정 추정과의 관계는 먼저 범주형 보조변수를 이용한 사후층화 후에 가중치를 조정하는 방법을 적용하였으므로 보다 폭넓은 의미에서 보정은 사후층화의 특별한 경우이며, 또한 사후층화추정량은 보정추정량의 보다 일반적인 형태라는 사실을 알 수 있었다.

<참고문헌>

1. Bethlehem, J. G., and Wouter, J. K. 1987. "Linear Weighting of Sample Survey." *Journal of Official Statistics* 3 : 141-153.
2. Chambers, R. L. 1996. "Robust Case-Weighting for Multipurpose Establishment Surveys." *Journal of Official Statistics* 12 : 3-32.

3. Deville, J. C., and Sarndal, C. E. 1992. "Calibration Estimators in Survey Sampling." *Journal of American Statistical Association* 87 : 376-382.
4. Deville, J. C., Sarndal, C. E., and Sautory, O. 1993. "Generalized Raking Procedure in Survey Sampling." *Journal of American Statistical Association* 88 : 1013-1020.
5. Fuller, W. 1966. "Estimation Employing Post Strata." Journal of American Statistical Association, 61 : 1171-1183.
6. Holt, D., and Smith, T. M. F. 1979. "Post Stratification." *Journal of Royal Statistical Society* Ser. A 142 : 33-46.
7. Jager, P. 1986. "Post Stratification Against Bias in Sampling," *International Statistics Review* 54 : 159-167.
8. Jayasuriya, B. R., and Valliant, R. 1996. "An Application of Restricted Regression Estimation in a Household Survey." *Survey Methodology* 22 : 127-137.
9. Lundstrom, S., and Sarndal, C. E. 1999. "Calibration as a Standard Method for Treatment of Nonresponse." *Journal of Official Statistics* 15 : 305-327.
10. Oh, H. L., and Scheuren, F. 1987. "Modified Raking Ratio Estimation." *Survey Methodology* 13 : 209-219.
11. Rao, J. N. K. 1985. "Conditional Inference in Survey Sampling." *Survey Methodology* 11 : 15-31.
12. Sarndal, C. E., Swensson, B., and Wretman, J. 1992. *Model Assisted Survey Sampling*. New York : Springer-Verlag.
13. Singh, A. C., and Rao, J. N. K. 1997. "Ridge-Shrinkage Methods for Range-Restricted Weight Calibration in Survey Sampling." *in Proceedings of the Section on Survey Research Methods* Alexandria, VA : American Statistical Association.
14. Smith, T. M. F. 1990. "Post-Stratifiaction." *The Statistician* 40 : 315-323.
15. Zahng, L. C. 2000. "Post-Stratification and Calibration- A Synthesis." *The American Statistician* 54 : 178-184.