The Korean Journal of Applied Statistics
2021, Vol. 34, No. 3, 427–438

DOI: https://doi.org/10.5351/KJAS.2021.34.3.427
Print ISSN1225-066X / Online ISSN 2383-5818

Integrated calibration weighting using complex auxiliary information

Inho Park[1,a], Sujin Kim[b]

[a]Department of Statistics, Pukyong National University;
[b]Health Insurance Research Institute, National Health Insurance Service

Abstract

Two-stage sampling allows us to estimate population characteristics by both unit and cluster level together. Given a complex auxiliary information, integrated calibration weighting would better reflect the level-wise characteristics as well as multivariate characteristics between levels. This paper explored the integrated calibration weighting methods by Estevao and Särndal (2006) and Kim (2019) through a simulation study, where the efficiency of those weighting methods was compared using an artificial population data. Two weighting methods among others are shown efficient: single step calibration at the unit level with stacked individualized auxiliary information and iterative integrated calibration at each level. Under both methods, cluster calibrated weights are defined as the average of the calibrated weights of the unit(s) within cluster. Both were very good in terms of the goodness-of-fit of estimating the population totals of mutual auxiliary information between clusters and units, and showed small relative bias and relative mean square root errors for estimating the population totals of survey variables that are not included in calibration adjustments.

Keywords: two-stage sampling, integrated weighting, calibration estimation, complex auxiliary information, American community survey

1. 서론

조사연구를 위한 표본설계는 흔히 다단추출방식을 채택하며, 이로 인해 개체(unit)는 물론 집락(cluster) 단위 수준의 모수추정을 가능하게 한다. 보조정보(auxiliary information)가 모집단과 표본단위별로 존재하면 칼리브레이션 추정(calibration estimation) 혹은 칼리브레이션 가중치(calibration weight)를 통해 정도수준과 포함률을 개선할 수 있다 (Särndal 등, 1992). 칼리브레이션 가중치란 보조정보의 가중표본총합이 모총합과 같아지도록 조정전 가중치를 조정하는 것을 말하며, 칼리브레이션 추정이란 칼리브레이션 가중치를 사용한 가중표본추정을 일컫는다. 만약 개체 및 집락에 대한 보조정보, 즉 복합보조정보(complex auxiliary information)가 존재한다면 이를 효율적으로 이용하여 개체와 집락 단위에 대한 통합 칼리브레이션 추정을 함께 고려할 수 있게 된다.

본 연구는 Estevao와 Särndal (2006)와 Kim (2019)에서 고려한 복합보조정보를 이용한 통합 칼리브레이션 가중치 산출방법들에 대해 살펴본다. 가상모집단을 활용한 모의실험을 통해 통합 칼리브레이션 가중치 산출방법의 효율성을 비교한다. 2절에서는 이단추출의 표본설계와 보조정보의 조건들에 대해 논의하고 관련한

This paper is based upon the second author's master thesis from the Pukyong National University.

[1] Corresponding author: Department of Statistics, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 45813, Republic of Korea. E-mail: ipark@pknu.ac.kr

Published 30 June 2021 / journal homepage: http://kjas.or.kr
© 2021 The Korean Statistical Society. All rights reserved.

칼리브레이션 추정에 대해 간단히 논의한다. 3절에서는 Estevao와 Särndal (2006)이 이단추출과 관련하여 논의한 두 가지의 가중치 통합옵션과 네 가지의 통합 칼리브레이션 가중치 산출방법에 대해 살펴본다. 더불어, Kim (2019)이 제안한 단위수준별 반복적합을 이용한 통합 칼리브레이션 가중치 산출방법에 대해 살펴본다. 4절에서는 미국사회조사(American Community Survey; ACS) 자료를 이용한 모의실험 설계와 다섯가지 통합 칼리브레이션 가중치 산출방법의 적용에 대한 모의실험의 결과에 대해 논의한다. 마지막으로 5절에서는 연구논의를 포함한다.

2. 이단추출을 위한 칼리브레이션 추정

2.1. 이단추출

이단추출(two-stage sampling)은 모집단의 집락구조를 활용하여 집락모집단에서 표본집락을 먼저 추출한 뒤 추출된 표본집락에서 표본개체를 단계별로 추출하는 표집방법을 일컫는다. 개체모집단 $U = \{1, \ldots, k, \ldots, N\}$는 N_j개의 개체를 갖는 M개의 집락 $U_j = \{1, \ldots, k, \ldots, N_j\}$ $(j = 1, \ldots, M)$으로 나뉘며 $N = \sum_{j=1}^{M} N_j$이다. 집락모집단은 $U_I = \{1, \ldots, j, \ldots, M\}$으로 표기하고 개체모집단은 $U = \cup_{j \in U_I} U_j$으로 표기할 수 있다. 일단계에서는 집락모집단 U_I에서 표본집락 $s_I = \{1, \ldots, j, \ldots, m\}$을 포함확률(inclusion probability) π_{Ij} $(j \in U_I)$로 추출하고, 이단계에서는 추출된 표본집락 U_j로부터 표본개체 $s_j = \{1, \ldots, k, \ldots, n_j\}$을 조건부 포함확률(conditional inclusion probability) $\pi_{k|j}$ $(k \in U_j)$로 추출한다. 여기서 m은 표본집락수를 나타내며 n_j는 j번째 표본집락에서 추출한 표본개체수를 나타낸다. 따라서 일단계의 집락 설계가중치(design weight)는 $d_{Ij} = 1/\pi_{Ij}$ $(j \in U_I)$이고, 이단계의 개체 조건부 설계가중치는 $d_{k|j} = 1/\pi_{k|j}$ $(k \in U_j)$이다. 개체표본은 $s = \cup_{j \in s_I} s_j$으로 나타내며 k번째 표본개체의 설계가중치는 $d_k = d_{Ij} d_{k|j}$으로 정의된다.

2.2. 칼리브레이션 추정

앞서 언급한 바와 같이 이단추출에서는 개체모집단의 특성 추정을 위한 개체통계량(unit statistics)과 집락모집단의 특성 추정을 위한 집락통계량(cluster statistics)의 산출이 모두 가능하다. 복합보조정보 및 단위수준별 표본추정을 표기하기 위해 이후 논의에서 Estevao와 Särndal (2006)의 기호를 사용고자 한다.

개체특성 $y_{(u)}$에 대한 표본개체값 $y_{(u)k}$ $(k \in s)$이 관측된다면, 개체모총합 $Y = \sum_{k \in U} y_{(u)k}$을 추정할 때 개체 보조정보를 이용하여 설계가중추정량 $\hat{Y}_\pi = \sum_{k \in s} d_k y_{(u)k} = \sum_{j \in s_I} d_{Ij} \sum_{k \in s_j} d_{k|j} y_{(u)k}$의 정도수준과 포함률을 개선할 수 있다. 또한 집락특성 $y_{(c)}$에 대한 표본집락값 $y_{(c)j}$ $(j \in s_I)$이 관측된다면, 집락모총합 $Y_I = \sum_{j \in U_I} y_{(c)j}$을 추정할 때 집락 보조정보를 이용하여 설계가중추정량 $\hat{Y}_{I\pi} = \sum_{j \in s_I} d_{Ij} y_{(c)j}$의 정도수준과 포함률을 개선할 수 있다. 집락과 개체의 예로 가구와 가구원, 지역과 가구, 학교와 학생, 기업과 종업원 등을 고려할 수 있다. 만약 가구와 가구원의 경우라면, $y_{(c)j}$는 가구소득, $y_{(u)k}$는 가구원 학력을 나타낼 수 있다.

개체 k의 보조정보(값 혹은 벡터)는 $x_{(u)k}$이고 집락 j의 보조정보는 $x_{(c)j}$으로 표기한다면, 조사연구자가 고려할 수 있는 보조정보의 주어진 조건들은 다음과 같이 정리할 수 있다.

(A1) 집락 모총합 $\sum_{j \in U_I} x_{(c)j}$이 주어짐.

(A2) 표본집락 $j \in s_I$에 대해 $x_{(c)j}$값이 주어짐.

(A3) 개체 모총합 $\sum_{k \in U} x_{(u)k}$이 주어짐.

(A4) 표본개체 $k \in s$에 대해 $x_{(u)k}$값이 주어짐.

모든 집락 j $(j \in U_I)$의 보조정보 $x_{(c)j}$가 주어진 상황이라면, A1와 A2의 조건을 만족한다. 국가승인통계를 위한 표본설계를 수행할 때 통계청의 인구주택총조사의 조사구 명부를 사용할 수 있는데, 조사구 내 주택유형별 가구수, 연령대별 인구수 등의 정보를 알 수 있다. 하지만 모든 개체 k $(k \in U)$의 보조정보 $x_{(u)k}$가 주어지는

상황은 그리 흔치 않다. 이는 이단추출의 표본설계를 고려하는 근본적인 이유이기도 하다. 예로, 통계청 국가통계포털을 통해 인구총계를 알 수 있고, 표본가구를 대상으로 가구원조사를 수행하여 표본개체들 보조정보 $x_{(u)k}$만을 파악하는 일반적인 경우는 A3와 A4의 조건을 만족한다고 할 수 있다.

단위수준별 칼리브레이션 총합추정은 우선 해당 설계가중치 d_{Ij}와 d_k을 수정하여 다음의 칼리브레이션 조건식을 만족시키는 칼리브레이션 가중치 w_{Ij}와 w_k을 산출한다.

$$\sum_{j\in s_I} w_{Ij}x_{(c)j} = \sum_{j\in U_I} x_{(c)j}, \tag{2.1}$$

$$\sum_{k\in s} w_k x_{(u)k} = \sum_{k\in U} x_{(u)k}. \tag{2.2}$$

이를 바탕으로한 집락총합 칼리브레이션 추정량 $\hat{Y}_{I,\mathrm{CAL}} = \sum_{j\in s_I} w_{Ij}y_{(c)j}$와 개체총합 칼리브레이션 추정량 $\hat{Y}_{\mathrm{CAL}} = \sum_{k\in s} w_k y_{(u)k}$을 사용한다.

3. 통합가중치 산출방안

3.1. Estevao와 Särndal (2006) 방법

Estevao와 Särndal (2006)는 집락과 개체 간 단위 및 보조정보 구성을 종합적으로 반영할 수 있는 통합 칼리브레이션 추정을 논하기 위해서 두 가지의 가중치 통합옵션을 다음과 같이 고려하였다.

(I1) $\sum_{k\in s_j} w_k = N_j w_{Ij}$ $(j \in s_I$, 집락크기 N_j는 주어짐).

(I2) $w_k = d_{k|j}w_{Ij}$ $(k \in s_j,\ j \in s_I)$.

두 옵션은 모두 칼리브레이션 조건식 (2.1)과 (2.2)을 만족시킴은 물론 통합가중치의 산출을 편리하고 단순하게 하는 역할을 한다. 통합옵션 (I1)하에서 집락 칼리브레이션 가중치는 $w_{Ij} = \sum_{k\in s_j} w_k/N_j$이 되며, 이로 인해 수준별 가중치의 선택과 상관없이 집락(집단)내 추정단위총수를 같게 해준다. 반면, 통합옵션 (I2)를 사용하면 표본개체의 칼리브레이션 가중치를 산출할 때 집락내 조건부 설계가중치 $d_{k|j}$를 그대로 유지할 수 있다. 일단집락추출(one-stage cluster sampling)인 경우에는 설계가중치가 1이므로 통합옵션 (I2)하에서는 집락 칼리브레이션 가중치(w_{Ij})와 개체 칼리브레이션 가중치(w_k)가 같아진다 (Lemaitre와 Dufour, 1987).

논의의 단순화를 위해 기호 $d \overset{x}{\Longrightarrow} w$는 보조정보 x를 사용하여 설계가중치 d로 하여금 칼리브레이션 조건식을 만족시키도록 하는 칼리브레이션 가중치 w를 산출하는 절차를 표기하기로 한다. 예로, 식 (2.1)과 (2.2)는 각각 $d_j \overset{x_{(c)j}}{\Longrightarrow} w_{Ij}$와 $d_k \overset{x_{(u)k}}{\Longrightarrow} w_k$로 표기할 수 있다. Estevao와 Särndal (2006)는 수준별 칼리브레이션 추정량 $\hat{Y}_{I,CAL}$와 $\hat{Y}_{CAL}$의 가중치를 산출할 수 있는 다음과 같은 네 가지의 가중치 산출방법을 고려하였다.

(W1) **비통합 칼리브레이션** 집락과 개체 각각의 보조정보를 사용하여 수준별 칼리브레이션 가중치를 독립적으로 다음과 같이 산출한다(통합옵션 (I1) 과 (I2)는 적용되지 않는다.)

(1) $d_{Ij} \overset{x_{(c)j}}{\Longrightarrow} w_{Ij}$.

(2) $d_k \overset{x_{(u)k}}{\Longrightarrow} w_k$.

(W2) **통합옵션 (I1)을 이용한 개체화 통합 칼리브레이션** 집락수준 보조정보를 개체별로 다음과 같이 균등하게 분배한 후,

$$x_{(c)k} = \frac{x_{(c)j}}{N_j}, \tag{3.1}$$

개체수준의 쌓여진 보조정보 $x_{(cu)k} = (x_{(c)k}, x_{(u)k})'$를 이용하여 개체 칼리브레이션 가중치를 산출하고, 통합옵션 (I1)에 의해 집락 칼리브레이션 가중치를 산출한다. 즉,

(1) $d_k \overset{x_{(cu)k}}{\Longrightarrow} w_k$와 $w_{Ij} := \sum_{k\in s_j} w_k/N_j$.

(W3) **통합옵션 (I2)를 이용한 집락화 통합 칼리브레이션** 개체 보조정보를 이용하여 집락총합 X_j을 다음과 같이 Horvitz-Thompson 추정량으로 추정한 후,

$$\hat{x}_{(u)j\pi} = \sum_{k\in s_j} d_{k|j} x_{(u)k}, \tag{3.2}$$

집락수준의 쌓여진 보조정보 $x_{(cu)j} = (x_{(c)j}, \hat{x}_{(u)j\pi})'$를 이용하여 집락 칼리브레이션 가중치를 산출하고, 통합옵션 (I2)를 이용하여 개체 칼리브레이션 가중치를 산출한다. 즉,

(1) $d_{Ij} \overset{x_{(cu)j}}{\Longrightarrow} w_{Ij}$와 $w_k := d_{k|j} w_{Ij}$.

(W4) **통합옵션 (I1)을 이용한 순차적 칼리브레이션** 집락과 개체 칼리브레이션 가중치를 순차적이고 독립적으로 산출하되, 개체 칼리브레이션 가중치 산출시 통합옵션 (I1)을 만족하도록 한다. 즉,

(1) $d_{Ij} \overset{x_{(c)j}}{\Longrightarrow} w_{Ij}$.

(2) $d_k \overset{x_{(u)k}}{\Longrightarrow} w_k$와 $\sum_{k\in s_j} w_k = N_j w_{Ij}$.

3.2. 수준별 반복적합에 의한 통합가중치 산출방안 및 방법 비교

Kim (2019)은 통합 칼리브레이션 가중치 산출의 문제를 집락과 개체의 수준별 칼리브레이션 가중치 산출을 반복적으로 수행하는 다음과 같은 방법을 제안하였다.

(W5) **통합옵션 (I1)을 이용한 수준별 반복적합 칼리브레이션** 개체 칼리브레이션 가중치와 집락 칼리브레이션 가중치를 반복적으로 산출하되 통합옵션 (I1)을 이용한다. 즉,

(1) $r = 1$와 $w_k^{(r-1)} = d_k$,

(2) $w_k^{(r-1)} \overset{x_{(u)k}}{\Longrightarrow} w_k^{*(r)}$와 $w_{Ij}^{*(r)} := \sum_{s_j} \frac{w_k^{*(r)}}{N_j}$,

(3) $w_{Ij}^{*(r)} \overset{x_{(c)j}}{\Longrightarrow} w_{Ij}^{(r)}$와 $w_k^{(r)} := c_j^{(r)} w_k^{*(r)}$, 여기서 $c_j^{(r)} := \frac{w_{Ij}^{(r)}}{w_{Ij}^{*(r)}}$,

(4) $r := r + 1$로 놓고 단계 2와 3을 반복하여 최종 수렴값 w_k와 w_{Ij}을 얻는다.

W1은 단위수준별로 칼리브레이션 조정을 독립적으로 수행하므로 단위수준간 다변량적 관계가 반영되지 못할 수 있다. W2는 집락 보조정보를 개체화하여 단일 단계의 개체 칼리브레이션 조정만을 수행함으로 절차가 매우 단순하며, 단위수준간 다변량적 관계가 반영될 수 있을 것으로 판단된다. W3은 개체 보조정보로 표본총합을 추정하는 집락화를 통해 단일 단계의 집락 칼리브레이션 조정만을 수행함으로 절차가 매우 단순하지만 개체수준의 통계추정이 다소 부정확할 가능성이 존재할 수 있다. W4는 집락과 개체 통합 보조정보를 독립적으로 적합하지만 W1와는 달리 두 단위간 통합옵션 (I1)을 만족하게 함으로 단위수준간 총수 관계는 유지될 수 있지만 다변량 특성을 잘 반영하지 못할 가능성이 존재한다. W5는 W2와 마찬가지로 통합옵션 (I1)을 이용하고 있지만 집락 보조정보를 개체화하는 대신에 수준별 반복적합을 통해 간접적으로 단위수준간 다변량적 관계를 보존할 수 있을 것으로 판단된다. 다음 절에서는 간단한 모의실험을 통해 앞서 논의한 다섯가지 통합가중치 산출방법의 효율성에 대해 비교하고 평가한다.

4. 모의실험

4.1. 모의실험 설계

본 연구를 위해 2012년 미국사회조사 자료를 이용하여 3절에서 기술한 통합 칼리브레이션 가중치 산출방법을 적용한 결과를 비교하였다. 모의실험의 기본설정은 Kolenikov와 Hammer (2015)를 준용하였다. 우선 미국사회조사 자료를 가상 모집단으로 설정하고 가구(집락)을 추출하고 가구내 모든 성인가구원을 조사하는 일단추출(이단추출의 특수형태)의 표본설계를 적용한 후 응답성향모형(response propensity model)을 통해 최종적 접촉성공 가구 및 응답 가구원 자료을 구성한 후 통합 칼리브레이션 가중치를 산출하였다.

미국사회조사 자료는 미국센서스국(U.S. Census Bureau)이 주관하며 매년 2백만여 가구가 참여한다. 표본으로 선정된 가구에는 조사참여 의무가 부여되며 대략 97% 이상의 응답률을 보이고 있고, 50여개의 인구, 사회 및 경제 등과 관련된 특성이 조사된다. 2012년 미국사회조사 자료는 IPUM.org 싸이트를 통해 입수하였고 총 1,207,415개 가구의 2,294,898명의 만 18세 이상 성인을 포함한다. 미국사회조사 자료에서 고려한 주요 변수는 Table 1에 나열하였다.

가상모집단인 미국사회조사 자료로부터 5,000개 가구를 단순확률로 추출하고 가구와 가구내 모든 성인들에 대해 로지스틱 회귀 응답모형을 순차적으로 적용하여 가구접촉성공 및 가구내 개인 응답여부를 결정하였다. 먼저, 5,000개 가구는 다음과 같이 정의되는 가구접촉성공확률 $\phi_j = \phi(x_j)$을 적용하였다.

$$\log\left(\frac{\phi_j}{1-\phi_j}\right) = x_j\beta \quad (j = 1, 2., \ldots, m).$$

여기서 β는 가구특성 x_j에 대한 로지스틱 회귀계수를 나타내며, m는 총 표본가구수를 나타낸다. 접촉성공한 가구내 개인은 다음과 같이 정의되는 개인응답확률 $\phi_{k|j} = \phi(x_{jk})$을 적용하였다.

$$\log\left(\frac{\phi_{k|j}}{1-\phi_{k|j}}\right) = x_{jk}\gamma \quad (k = 1, 2., \ldots, N_j).$$

여기서 γ는 개인특성 x_{jk}에 대한 로지스틱 회귀계수를 나타내며, N_j는 j번째 가구내 성인 가구원수를 나타낸다. Table 2와 Table 3은 각각 가구와 개인에 대한 로지스틱 회귀계수를 정리하고 있다. 가구소득(HH income)을 제외한 다른 모든 특성변수는 지시자이고 기준범주(reference category)의 로지스틱 회귀계수는 0으로 설정하였다. 가구응답확률은 평균 0.8875이고 0.8172에서 0.9394의 범위를 갖는다. 또한 개인응답확률은 평균 0.6205이고 0.4750에서 0.7408의 범위를 갖는다.

Table 1은 미국사회조사 자료 가상모집단의 가구 및 성인들의 특성별 모집단 총수 및 비율은 물론 $R = 200$개의 확표본들에 대한 응답 가구 및 성인들의 특성별 총수 평균 및 총수 평균 비율을 정리하고 있다. 평균적 응답가구수는 2,132.5개이고 응답자수는 2,970.6으로 조사특성별로 모집단 가구수 및 성인수 비율과는 다소 상이함을 알 수 있다. 예로, 1인 가구의 모집단 비율은 32.2%인 반면, 응답가구 기준의 평균적 총수 비율은 17.8%로 매우 낮은 것을 알 수 있다. 성인 남성의 모집단 비율은 47.3%인 반면, 응답자 기준의 평균적 총수 비율은 45.2%으로 약간 낮게 나타났다.

200개 (응답)표본가구 및 (응답)가구원별로 3절의 다섯가지 가중치 산출방안 W1–W5을 각각 적용하여 통합 칼리브레이션 가중치를 산출하였다. 칼리브레이션 가중치는 R의 surveys 패키지에 있는 calibrate함수 중 일반화선형회귀추정(generalized regression estimation)을 이용하여 산출하였다. 칼리브레이션 추정을 위해 사용한 보조변수는 Table 1에 진하게 표기하고 있는데, 가구수준에서 가구크기(가구원수, 4분류)와 가구소득(연속형, 달러)이고 개인수준에서 성별(2분류), 인종(3분류), 나이(5분류), 교육(5분류)이다.

Table 1: Population and sample distribution

Variable	Category	Population		Sample	
		Size	%	(ave) size	(ave)%
Households		1,207,415	100.0	2,132.5	100.0
Household Size	**1**	**388,470**	32.2	379.4	17.8
	2	**629,353**	52.1	1314.3	61.6
	3	**131,901**	10.9	306.0	14.3
	≥ 4	**57,791**	4.8	132.8	6.2
Household Income	Under $20k	224,677	18.6	295.7	13.9
	$20k-under $40k	252,356	20.9	411.4	19.3
	$40k-under $650k	249,978	20.7	447.5	21.0
	$65k-under $100k	219,408	18.2	424.3	19.9
	$100k-above	260,996	21.6	553.6	26.0
	Mean **income ($)**	71,456.0	-	79,904.4	-
Lingusitically	isolated	47,061	3.9	79.3	3.7
	not isolated	1,160,354	96.1	2053.2	97.3
Individual	Category	2,294,898	100.0	2,970.6	100.0
Sex	**Male**	**1,085,531**	47.3	1,342.0	45.2
	Female	**1,209,367**	52.7	1,628.6	54.8
Race	**White only**	**1,814,707**	79.1	2,424.2	81.6
	BLack/African American	**227,826**	9.9	260.2	8.8
	Others	**252,365**	11.0	286.2	9.6
Age	**18–29**	**395,250**	17.2	498.9	16.8
	30–44	**528,792**	23.0	610.1	20.5
	45–54	**437,672**	19.1	584.8	19.7
	55–64	**428,807**	18.7	586.5	19.7
	65 or above	**504,377**	22.0	690.4	23.2
Education	**Below high school**	**299,730**	13.1	377.5	12.7
	High school	**656,608**	28.6	855.0	28.8
	Some college	**697,947**	30.4	899.1	30.3
	Bachelor's degree	**399,943**	17.4	521.6	17.6
	Graduate/professional	**240,670**	10.5	317.5	10.7
Married	Married	1,260,721	54.9	1,746.0	58.8
	Others	1,034,177	45.1	1,224.6	41.2
Employment	Employed	1,342,689	58.5	1,713.8	57.7
	Unemployed	122,905	5.4	150.7	5.1
	Not in labor force	829,304	36.1	1,106.1	37.2

Table 2: Logistic regression model for the household contact probability

Variable	Category/Transform	Logistic regression coefficient
Intercept		0.05
Household size (one)	2	0.60
	3	0.50
	4 or more	0.25
Household Income	ln(HH income+20,000)	0.10

Table 3: Logistic regression model for the individual participation probability

Variable	Category	Logistic regression coefficient by household size			
		1	2	3	4 or more
Intercept		0.10	0.20	0.20	0.15
Sex (male)	Female	0.15	0.10	0.20	0.20
	30–44yr	–0.10	–0.35	–0.20	–0.15
(18–29yrs)	45–54yr	0.05	0.00	0.10	0.05
Age	55–64yr	0.20	0.10	0.25	0.20
	65 or above	0.10	–0.02	0.15	0.10
Race	Black	–0.10	–0.25	–0.15	–0.10
(white)	Others	–0.20	-0.10	–0.15	–0.20
Education	High school	0.15	0.30	0.25	0.20
	Some college	0.15	0.30	0.25	0.25
(below high)	Bachelor	0.50	0.30	0.40	0.45
	Graduate	0.40	0.30	0.35	0.40

Table 4: Median of summary statistics for calibration weights

Level	Method	Mean	CV	DEFF	(d, w)	Min	Max
	W1	565.4	037.9	1.143	668,216.7	364.2	1,036.9
	W2	565.4	048.1	1.231	793,448.1	133.9	1,347.3
Household	W3	565.4	120.8	2.459	2,530,101.4	–1,340.0	4,094.5
	W4	565.4	037.9	1.143	668,216.7	364.2	1,036.9
	W5	565.4	048.3	1.234	798,259.8	131.8	1,457.8
	W1	771.5	12.0	1.014	1,785,531.3	615.5	1,092.1
	W2	771.5	019.3	1.037	1,868,994.6	452.0	1,358.1
Individual	W3	771.5	095.9	1.920	5,114,130.8	–1340.0	4,094.5
	W4	772.2	051.6	1.266	2,708,971.9	–1848.1	3,326.6
	W5	771.7	019.1	1.037	1,864,445.4	505.1	1,459.8

4.2. 모의실험 결과

Table 4는 다섯가지 가중치 산출방법으로 얻은 200개 표본들의 가중치 분포를 여섯가지 기술통계량의 중간값(median)으로 정리하고 있다. 고려된 통계량은 평균(mean), 변동계수(coefficient of variation; CV), 분산증가분(design effect; DEFF), 거리측도($G(w, d)$), 최소값(min)과 최대값(max)이다. 여기서 분산증가분(DEFF) 이란 가중치를 사용하지 않는 단순추정과 비교하여 불균등 가중치의 사용으로 인해 증가하는 평균추정량의 분산 크기를 나타내며 다음과 같이 정의된다 (Park과 Lee, 2002).

$$\text{DEFF} = 1 + \text{CV}^2.$$

칼리브레이션 가중치 $\{w_i\}_{i\in s}$는 조정전 가중치 $\{d_i\}_{i\in s}$으로 하여금 칼리브레이션 조건식 (2.1) 혹은 (2.2)을 만족되는 시키도록 하되 조정 전과 후 가중치 간의 거리 $G(d, w)$를 최소화하도록 한다 (Deville와 Särndal, 1992). 다양한 거리측도가 선택될 수 있지만 본 연구의 모의실험에서는 일반화선형회귀추정의 방식으로 사용하였으므로 이에 해당하는 거리함수인 다음의 최소제곱거리(least-squares distance function)을 고려하여 모의실험의 결과를 평가하였다.

$$G(d, w) = \sum_{i\in s} \frac{(w_i - d_i)^2}{d_i}.$$

W1 단위수준별로 칼리브레이션 가중치를 산출하므로 다른 방법들에 비해 CV, DEFF, $G(d, w)$의 가중치 변동량이 가장 작게 나타나고 있다. W3은 집락화를 통한 통합 칼리브레이션 가중치를 산출하므로 식 (3.2)의 집락별 총합추정량이 칼리브레이션 조건식을 만족시키므로 관련된 표본추정의 변동성이 추가되어 가구수준은 물론 개인수준의 통합가중치 변동량 모두 다른 방법에 비해 가중치 변동량인 CV, DEFF, $G(d, w)$가 가장 크게 나타났다. W4는 W1과 마찬가지로 단위수준별로 칼리브레이션 가중치를 산출하되 개인수준 통합가중치가 통합옵션 (I2)를 추가적으로 만족시켜야한다. 따라서 이러한 제약식은 W1과 비교하여 개인 통합가중치의 변동량, 즉 CV, DEFF, $G(d, w)$값을 많이 증가시켰고 개인 통합가중치의 평균값도 다른 방법들의 평균값과 미세하게 다르게 나타났다. W2은 개체화를 통한 통합 칼리브레이션 가중치를 산출하므로 W1에 비해 통합가중치의 변동량을 다소 증가시켰다. W5는 단위수준별 칼리브리에션 가중치 산출을 순차적으로 수행하므로 W1에 비해 통합가중치의 변동량을 다소 증가시켰다. W5 방법은 W2 방법과 거의 유사한 결과를 보여주고 있는데, 가구 통합 가중치의 변동량은 미세하게 커지나 개인 통합 가중치의 변동량은 미세하게 작아지는 경향을 보였다.

칼리브레이션 가중치 산출을 위해 고려된 보조정보의 총합추정값과 모총합간의 차이는 다음과 같은 적합도를 표현하는 카이제곱 통계량으로 평가할 수 있다.

$$\chi^2 = \sum_{g=1}^{G} \frac{\left(\hat{X}_g - X_g\right)^2}{X_g},$$

여기서 G는 평가에서 고려하는 수준 총갯수를 나타내며, X_g와 $\hat{X}_g$는 수준 g의 모총합과 가중표본총합을 나타낸다. Table 5는 가구특성과 개인특성 분류 조합에 대한 다섯가지 가중치 방법의 결과로 얻은 200개 표본에 대해 가중치 산출결과의 가구와 개인의 다변량적 적합도 통계 χ^2의 분포를 정리하고 있다. Figure 1은 적합도 통계 χ^2에 대한 바이올린 그림이다.

바이올린 그림은 적합도 통계 χ^2의 커널 밀도 그림을 좌우대칭의 곡선으로 둘러싸인 영역으로 표현하고 있으며, 그 중앙에는 적합도 통계 χ^2의 상자그림과 함께 평균값은 다이아몬드로 나타내었다. 적합도 통계 χ^2는 통합 칼리브레이션 가중치 산출을 위해 고려한 가구특성인 가구크기(4분류) 및 가구소득분류(5분류)과 개인특성인 성별(2분류), 연령대(5분류), 교육수준(5분류)에 대해 산출하였다. χ^2의 평균값을 기준으로 보면, W2와 W5는 모든 조합에서 가장 작은 것으로 나타났고, W4는 그 다음으로 작게 나타났다. 하지만 가구크기와 개인연령 조합의 경우에는 W4의 χ^2의 평균값이 W2와 W5의 평균값에 비해 많이 높게 나타났다. W1과 W3은 W2, W4, W5에 비해 χ^2의 평균값이 크게 나타났다. W1은 W3에 비해 가구크기와 네가지의 개인 특성간 조합 모두에서 큰 χ^2의 값을 갖으며, W3은 W1에 비해 가구소득에 대해서 네가지 개인 특성간의 조합 모두에서 χ^2 값이 크게 나타난다.

추정량의 상대편향(relative bias; RB)과 상대 평균 제곱근 오차(relative root mean squared error; RRMSE)은 각각 다음의 식을 통해 산출할 수 있다.

$$\text{RB}\ (\%) = 100 \times \frac{1}{Y}\left(\frac{1}{R}\sum_{r=1}^{R}\hat{Y}_r - Y\right),$$

$$\text{RRMSE}\ (\%) = 100 \times \frac{1}{Y}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{Y}_r - Y\right)^2},$$

여기서 Y는 조사특성 y의 모총합이고 $\hat{Y}_r$은 $r(r = 1, \ldots, R)$번째 표본의 총합추정량을 나타낸다. Table 6은 칼리브레이션 조정에 고려하지 않은 가구변수인 가구의 언어적 고립(linguistically isolated)과 개인 변수인 기혼자(married), 고용상태(employed), 실업상태(unemployed), 비경제활동상태(not in labor force)에 대한 상대편향과 상대 평균 제곱근 오차에 대해 정리하고 있다. W2와 W5는 거의 유사한 결과를 보여주고 있는 다른 방법에 비해 상대편향은 매우 작은데, –0.02%에서 –0.57%의 수준으로 나타났다.

Table 5: Goodness-of-fit statistics by calibration weighting method

Cluster	Unit	Statistics of χ^2's	Weighting method				
			W1	W2	W3	W4	W5
Household size	Sex	Mean	40,425.9	2,045.5	15,772.4	4,005.7	2,296.0
		cv(%)	25.9	79.8	50.0	84.4	77.5
		Min	16,534.5	15.2	2,314.6	101.4	125.5
		Max	72,508.3	7,566.6	41 635.1	19,896.8	8,374.4
		IQR	12,747.8	1,956.0	11 597.5	3,912.5	2,047.1
	Age	Mean	49,478.5	12,467.7	42 456.5	29,921.8	12,278.8
		Cv(%)	23.2	39.1	43.0	48.7	38.1
		Min	21,215.2	2,598.7	11,374.8	4,881.5	2,788.0
		Max	83,898.9	33,904.1	144,142.5	109,521.7	31,051.1
		IQR	14,392.3	6,286.7	20,893.1	16,537.3	5,806.3
	Race	Mean	47,780.8	10,129.7	41,792.6	13,069.2	9,948.5
		cv(%)	24.1	52.8	53.9	54.3	54.0
		Min	20,186.2	817.7	5,605.6	1,442.8	1,048.7
		Max	78,324.1	30,899.6	130,727.9	46,009.2	32,342.7
		IQR	16,180.5	6,621.2	28,013.6	8,961.4	6,284.5
	Educ.	Mean	48,520.3	11,759.0	42,900.1	15,259.9	11,620.6
		Cv(%)	22.2	42.0	44.1	42.0	41.8
		Min	21,886.3	2,465.8	13,804.7	2,624.6	2,534.7
		Max	81,658.7	25,716.0	148,785.2	45,129.5	26,884.6
		IQR	15,889.1	6,281.7	22,386.4	8,275.0	6,448.4
Household income	Sex	Mean	12,845.2	7,108.1	17,189.8	8,297.7	7,062.4
		cv(%)	42.5	51.9	60.9	46.0	51.7
		Min	1,341.4	1,132.7	1,820.0	1,720.1	1,223.5
		Max	36,941.3	21,366.0	57,002.1	24,307.1	19,804.3
		IQR	7,876.5	4,675.2	13,515.0	5,110.7	4,572.2
	Age	Mean	26,682.8	20,146.2	46,589.5	22,044.7	20,056.5
		Cv(%)	29.8	34.2	33.8	33.5	34.8
		Min	8,678.0	6,288.3	7,481.2	6,381.9	6,655.2
		Max	58,061.8	47,866.7	97,533.3	48,024.5	49,960.6
		IQR	9,500.5	8,355.2	18,649.4	9,001.1	8,258.3
	Race	Mean	22,427.1	16,660.8	51,003.0	16,736.4	16,541.9
		Cv(%)	34.5	38.7	40.9	41.0	39.0
		Min	8,884.5	4,381.5	17,057.8	3,692.2	5,061.3
		Max	57,330.6	44,395.9	158,791.3	38,087.7	46,287.3
		IQR	10,584.5	9,368.7	25,054.0	10,194.5	8,817.9
	Educ.	Mean	25,786.6	19,835.4	49,909.8	21,892.8	19,756.2
		Cv(%)	29.2	32.6	39.8	31.1	32.7
		Min	6,658.5	5,467.9	19,241.8	7,215.8	5,497.4
		Max	54,674.2	41,732.6	147,719.5	47,619.3	41,548.9
		IQR	9,168.6	8,342.9	24,300.4	7,776.8	8,125.5

5. 논의

본 연구는 가상모집단으로부터의 이단추출에 대한 간단한 모의실험을 통해 집락과 개체 수준 간 복합보조정보를 함께 고려하여 수준별 칼리브레이션 가중치를 통합적으로 산출하는 방법을 비교하였다. 고려된 방법들

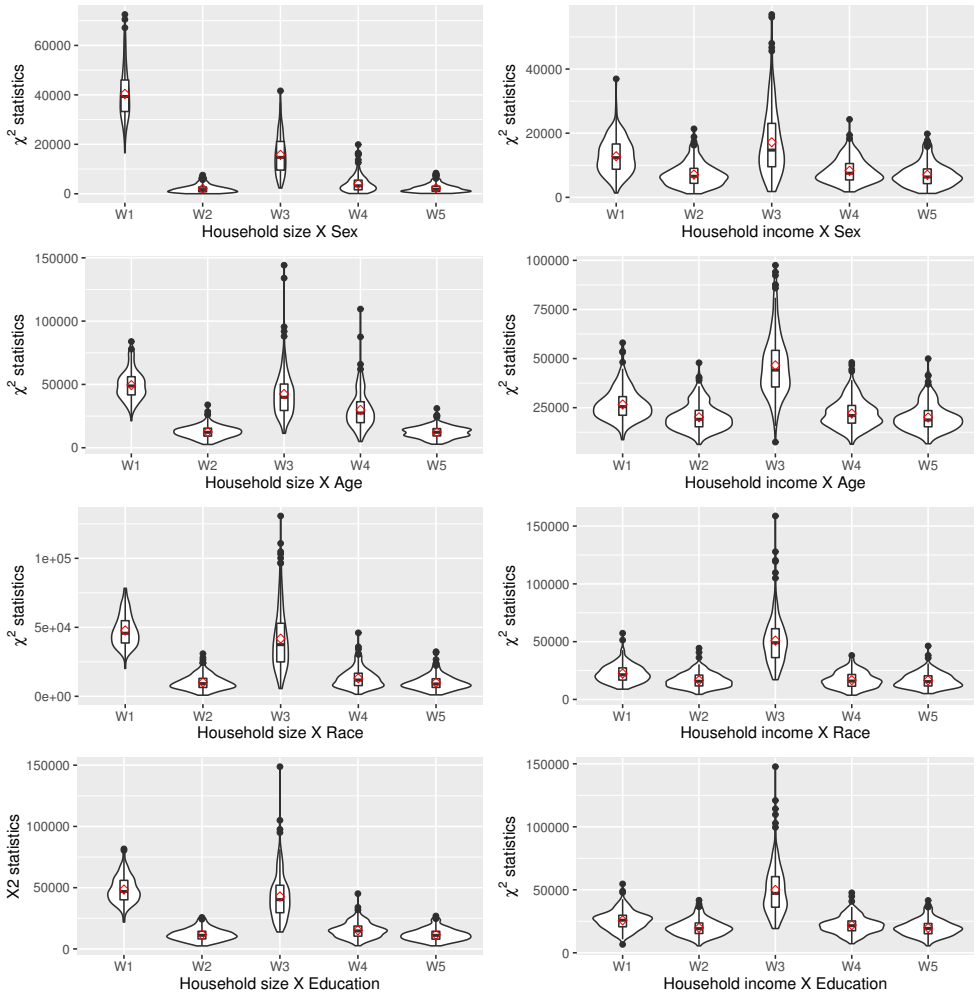

Figure 1: *Viloin plots of goodness-of-fit statistics by calibration weighting method.*

Table 6: Relative bias and root mean squared error of total estimates by calibration weighting method

Variable	Category	Statistics	Weighting method				
			W1	W2	W3	W4	W5
Household	Linguistically isolated	RB(%)	0.10	−0.02	0.01	0.10	−0.03
		RRMSE(%)	0.49	0.50	0.69	0.49	0.50
Individual	Married	RB(%)	6.06	−0.08	−1.52	−0.57	−0.06
		RRMSE(%)	6.30	1.41	2.76	1.56	1.42
	Employed	RB(%)	0.10	−0.03	−0.39	−0.41	−0.05
		RRMSE(%)	1.42	1.43	1.97	1.66	1.44
	Unemployed	RB(%)	−1.41	−0.57	1.41	0.88	−0.43
		RRMSE(%)	8.39	8.41	11.64	9.56	8.35
	Not in labor force	RB(%)	0.04	0.14	0.41	0.64	0.16
		RRMSE(%)	2.00	2.03	2.87	2.39	2.03

중 복합보조정보를 개체화 한 후 단일단계의 칼리브레이션 조정을 통해 가중치를 산출하되 집락가중치가 집락 내 개체가중치들의 평균이 되도록 하는 W2 방법과 개체과 집락 수준 각각의 보조정보를 이용한 수준별 칼리브레이션 조정을 상호 반복적으로 수행하되 역시 집락가중치가 집락 내 개체가중치들의 평균이 되도록 하는 W5 방법이 가중치의 변동량을 크게 늘리지 않고도 수준간 다변량적 특성을 잘 반영할 수 있음을 확인할 수 있었다. W2와 W5는 집락과 개체의 상호간 보조정보 총합추정의 적합도에 다른 방법들에 비해 매우 양호한 것으로 나타났다. 더불어 W2와 W5는 칼리브레이션 조정에 포함되지 않는 조사정보의 총합추정에 대한 상대편향 및 상대 평균 제곱근 오차가 작게 나타났다. W2와 W5는 가중치 변동성은 물론 적합도와 상대편향 및 상대 평균 제곱근 오차에서 서로 매우 유사하게 나타나고 있는데, 집락 및 개체간 응답성향 구성에 따른 수준별 특성 및 무응답의 조정력 등에 대한 상세한 비교연구는 흥미로운 추후 과제가 될 것이다.

References

Deville JC and Särndal CE (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376–382.

Estevao VM and Särndal CE (2006). Survey estimates by calibration on complex auxiliary information, *International Statistical Review*, **74**, 127–147.

Kim S (2019). Iterative calibration approach to integrated weighting for household surveys, *Unpublished Master thesis*, Pukyong National University, Busan, Korea.

Kolenikov S and Hammer H (2015). Simultaneous ranking of survey weights at multiple levels, *Survey Methods: Insights fromm the Field, Special issue: 'Weighting: Practical Issues and "How to" Approach'*, Retrieved from: http:// surveyinsights.org/?p-5099. DOI:10.13094/SMIF-2015-00010.

Lemaitre G and Dufour J (1987). An integrated method for weighting persons and families, *Survey Methodology*, **13**, 199–207.

Park I and Lee H (2002). A revisit of design effects under unequal probability sampling, *The Survey Statistician*, **46**, 23–26.

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

Valliant R, Dever JA, and Kreuter F (2013). *Practical Tools for Designing and Weighting Survey Samples*, Springer, New York.

Received May 18, 2021; Revised May 19, 2021; Accepted May 20, 2021

통합 칼리브레이션 가중치 산출 비교연구

박인호[1,a], 김수진[b]

[a]부경대학교 통계학과 [b]국민건강보험공단 건강보험연구원

요 약

이단추출은 개체와 집락 단수준별 모집단 특성을 함께 추정할 수 있게 해준다. 단위수준별 보조정보가 함께 주어질 때, 단위수준별 정보 및 가중치 구성을 통합적으로 고려한 칼리브레이션 가중치를 산출한다면 단위수준별 특성은 물론 수준간의 다변량적 특성도 적절히 반영할 수 있을 것이다. 본 연구는 Estevao와 Särndal (2006)과 Kim (2019)이 고려한 통합 칼리브레이션 가중치 산출 방법에 대해 살펴보았다. 간단한 모의실험을 통해 기존의 통합 칼리브레이션 가중치 산출방법의 효율성을 비교하였다. 이 중 복합보조정보를 개체화한 후 단일단계의 칼리브리이션 조정으로 가중치를 산출하되 집락가중치가 집락 내 개체가중치 평균이 되도록 정의하는 방법과 단위수준별 보조정보를 이용한 수준별 칼리브레이션 조정을 상호 반복적으로 수행하되 집락가중치가 집락 내 개체가중도치 평균이 되도록 하는 방법이 조정전 가중치의 변동량을 크게 늘리지 않고도 수준간 다변량적 특성을 잘 반영할 수 있음을 확인할 수 있었다. 집락과 개체의 상호간 보조정보에 대한 총합추정의 적합도 측면에서 매우 양호하였고, 칼리브레이션 조정에 포함되지 않는 조사특성들의 총합추정에 대한 상대편향 및 상대 평균 제곱근 오차가 작게 나타났다.

주요용어: 이단추출, 통합가중치, 칼리브레이션 추정, 복합보조정보, 미국사회조사

이 논문은 제 2저자의 부경대학교 석사학위논문의 일부를 발췌하여 수정함

[1]교신저자: (48513) 부산광역시 남구 용소로 45, 부경대학교 통계학과. E-mail: ipark@pknu.ac.kr