조사연구(2018. 11) 19권 4호: 119−136
DOI http://dx.doi.org/10.20997/SR.19.4.4

연구논문

가구 내 1인 조사를 위한 통합가중치 연구: 농어업인복지실태조사 사례를 중심으로

Integrated Weighting for One-person-per Household Surveys

김수진[a] · 박진우[b] · 박인호[c]
Sujin Kim · Jin Woo Park · Inho Park

가구조사의 분석대상으로 가구와 가구원이 모두 고려된다면 단위수준별로 표본가중치가 산출되어야 하며 가능한 단위간 구성적 측면도 반영하는 것이 바람직하다. 이때 구성적 측면의 반영은 주로 관련된 보조정보를 이용한 통합가중방식을 통해 가능하게 된다. 가구당 한 명만을 추출하는 가구조사에서는 가구와 응답자라는 상이한 추출 및 분석단위 간의 개념적·물리적 차이로 인해 가중치 산출은 물론 자료해석이 다소 모호해질 수 있다. 본 연구는 농어업인복지실태조사의 사례를 중심으로 개인화 통합가중치의 산출방법에 대해 살펴보며 추가적으로 가구당 1인조사임에도 두 종류의 가중치를 제공함에 따라 발생할 수 있는 혼선을 피할 수 있는 대리응답을 가정한 통합가중치 산출방법에 대해서도 논의한다. 개인화 통합가중치와 대리응답 통합가중치는 비통합의 개별적 가중치에 비해 변동 증가가 많이 크지 않지만 보조정보 기준의 적합성은 물론 정도수준의 변동도 많지 않음을 확인하였다.

주제어 : 칼리브레이션 조정, 대리응답, 다단추출, 적합도, 상대포함오차, 농어업인복지실태조사.

---

* 본 논문은 농촌진흥청 국립농업과학원 농업과학기술연구개발사업(세부과제번호 : PJ00998903)의 지원에 의해 이루어졌음.

a) 부경대학교 통계학과 대학원생.

b) 수원대학교 데이터과학부 교수.

c) 교신저자(corresponding author): 부경대학교 자연과학대학 통계학과 부교수 박인호.
E-mail: ipark@pknu.ac.kr.

Many household surveys involve sampling and analysis of both household and persons within household. Thus, it is desirable to ensure their hierarchical relationship between these two levels in developing their sample weights. An integrated weighting procedure can achieve this goal utilizing the auxiliary variables on the household and person levels. When only one person per household is sampled, there may exist an ambiguity in either implementing integrated weighting or analyzing complex survey data due to conceptual and/or physical differences in the role of the sampled person within the household, house and/or personal representatives per se. In this article, we compare several weighting approaches including two single-step calibration methods, one with assigning equal shares value for persons and the other with assuming proxy-response. The comparisons are carried out using data from a Survey on the Rural Well-being by Rural Development Administration. We find that both methods aforementioned tend to have better goodness of fits in terms of benchmarking totals with not inducing much variation on the integrated weights when compared with non-integrated calibration.

Key words: calibration adjustment, proxy response, multistage sampling, goodness-of-fit, relative coverage error, survey on the rural well-being

Ⅰ. 서론

가구조사의 분석대상으로 가구와 가구원이 모두 고려된다면 단위수준별로 표본가중치가 산출되어야 하며, 가능한 단위 간 구성적 측면도 함께 반영하는 것이 바람직할 것이다. 예를 들면, 아파트에 거주하는 2인 가구는 아파트 거주자 2명으로

구성되며, 가구소득은 가구원 소득의 합계와 같다. 단위간 구성적 측면을 반영한 가중치를 통합가중치(integrated weight)라고 하는데 이와 관련된 연구는 주로 칼리브레이션 조정(calibration adjustment)의 측면을 통해 제시된다(예, Estevao & Särndal 2006, Steel & Clark 2007).

통합가중치는 가구 내 다수의 대상을 추출하는 조사는 물론 한 명만을 추출하는 조사에서도 적용될 수 있다(Kolenikov & Hammer 2015). 가구 내 조사대상자로 가구주 혹은 배우자 중 오직 한 명만을 가구를 대표하는 자로 선택한다면 가구라는 추출단위(sampling unit)와 개인이라는 분석단위(analysis unit) 간의 개념적 차이로 인해 가중치 산출은 물론 자료해석이 다소 모호해질 수 있다. 예를 들어, 농진청이 주관하는 농어업인복지실태조사는 가구면접조사로 진행되며 가구를 대표하는 만 19세 이상의 가구주 혹은 배우자 중에 한 명을 조사대상 가구에서 자율적으로 선택(self-selection)하도록 한다. 농어업인복지실태조사의 기존 가중치 산출은 가구별로 (i) 가구추출확률의 역수로 정의되는 설계가중치(design weight)와 (ii) 단위무응답 조정 및 모집단 보조정보특성을 반영하는 래이킹-비 조정(raking-ratio adjustment)의 곱의 형태로 이루어진다. 모집단의 보조정보로는 응답자의 연령과 가구의 주택유형이라는 상이한 단위수준의 특성이 동시에 고려되고 있다.

이와 같이 가중치 산출과정에서 가구와 개인특성이 혼재되어 반영된다면 이로 인해 통계량의 추정대상이 전체 가구인지 아니면 전체 가구대표자(가구주와 배우자)인지 다소 모호해질 수 있게 된다. 예로, 가구 내 조사대상자에게 연간 가구소득을 묻는다면 가구를 대표하는 응답이 되며 지역생활에 대한 만족도를 묻는다면 (설문문항에 제시되는 문구에 따라서는) 가구를 대표할 수도 있고 응답자 개인의 의견이 될 수도 있다. 따라서 적절한 자료분석이 되기 위해서는 가구와 개인의 단위수준별 가중치가 산출되어야 하고 자료이용자는 조사항목에 따라 가구와 개인의 가중치를 선택해야 한다.

본 연구는 가구 내 대표자 한 사람만을 선택하는 가구조사에서 고려할 수 있는 통합가중치의 방법들에 대해 살펴본다. 더불어 가구별 상이한 수준의 가중치를 제시함에 따른 이용자의 가중치 선택의 혼선과 자료해석의 모호성을 극복해 줄 수 있는 대리응답(proxy response) 가정에 기초한 가구(화) 통합가중치의 산출방식을 살펴본다. 2장은 가구 내 1인 조사의 표본설계를 논의하고 조사자료에 적용할

수 있는 개별 가중치와 개인화 통합가중치의 산출방법에 대해 다룬다. 더불어 가구대표자로서의 배우자를 대리응답으로 고려하는 대안적 접근인 가구화 통합가중치 산출을 소개한다. 3장은 2017년 농어업인복지실태조사에 대해 앞서 논의한 다양한 방식의 가중치를 산출하고 보조변수에 대한 가중총합을 비교하고 자료분석의 정도수준에 대해 논의한다. 4장은 연구결과에 대한 결론 및 논의를 포함한다.

Ⅱ. 가구 내 표본설계 및 통합가중치 산출

1. 표본설계와 표본가중치

가구추출확률이 $\pi_{j=}\Pr(j\in s_c)$이면 가구 설계가중치는 다음과 같이 정의된다.

$$w_j^d = \pi_j^{-1} \tag{1}$$

여기서 $j=1,...,n$는 가구 지시자이고 s_c는 크기 n개인 가구표본을 나타낸다. 표본가구 내 M_j명의 가구 대표자가 있고 이 중에서 한 명, $m_j \equiv 1$을 조사대상자로 추출한다면 개인 설계가중치는 다음과 같이 정의된다.

$$w^{d_k} = \begin{cases} 2w_j^d & \text{배우자가 있는 가구}(M_j=2) \\ w_j^d & \text{배우자가 없는 가구}\,(M_j=1) \end{cases} \tag{2}$$

여기서 $k=1,...,M_j(\leq 2)$는 개인 지시자이고 $U_j=\{k=1,...,M_j\}$와 $s_j=\{k=1,...,m_j\}$는 가구 내 가구 대표자 전체와 표본을 나타낸다.

표본가중치는 개별응답개체가 대표하는 모집단 내 개체수의 추정량으로 주로 (i) 표본추출확률, (ii) 단위무응답 보정, (iii) 칼리브레이션 조정 등의 세 가지 요소들을 반영한다. 특히, 칼리브레이션 조정(calibration adjustment)은 (설계 혹은 조정 전) 가중치를 조정하여 보조정보 x의 모집단 총합과 표본가중합이 서로 일치하도록 하는 것을 일컫는다. 칼리브레이션 조정은 무응답(non-response)에 따른 편향(bias)

을 줄임은 물론 표본추출틀의 결함에 따른 포함률 오차(coverage error)의 보정에도 매우 효과적인 것으로 알려져 있다. 만약 다수의 보조정보에 대한 모집단 결합분포는 알려져 있지 않지만 개별 보조정보의 주변분포만 주어진다면 개별 보조정보에 대해서 반복적으로 칼리브레이션 조정을 수행하는 래이킹-비 조정(혹은 반복비례적합) 방법을 고려할 수 있다(Deville & Särndal 1992). 다음 절에서는 마지막 요소인 칼리브레이션을 통한 몇 가지의 최종 가중치 산출방법에 대해 논의한다.

2. 개별적 칼리브레이션 가중치

본 절에서는 개체수준별로 적용하는 칼리브레이션 가중치에 대해 살펴본다. Estevao & Särndal (2006)의 표기방식에 따라 가구와 개인 수준의 보조정보(예로, 주택유형과 응답자 연령대)을 각각 $x_{(c)j}$과 $x_{(u)k}$라 하고 가구표본과 개인표본은 $s_{(c)}$와 $s_{(u)}$라 정의한다. 여기서 칼리브레이션은 무응답에 대한 조정도 반영하므로 $s_{(c)}$와 $s_{(u)}$는 무응답을 제외한 응답가구와 응답가구대표를 나타낸다. 이때 괄호 속 첨자 c와 u는 집락(cluster)과 개체(unit)의 단위수준을 나타내며 괄호밖 첨자 j와 k는 가구와 개인의 지시자를 나타낸다. 단위수준별 보조정보의 모총합이 $X_{(c)}$와 $X_{(u)}$으로 주어진다고 가정하자. 래이킹-비 조정은 설계가중치와 다음의 거리측도를 최소화하되

$$D(w^d, w) = \sum_{i \in s} [w_i \log(\frac{w_i}{w_i^d}) - w_i + w_i^d] \tag{3}$$

보조정보의 표본가중합이 모총합과 같아지도록 반복적으로 적합시켜 다음과 같은 단위수준별 래이킹-비 조정의 칼리브레이션 가중치를 얻게 된다.

$$w_j^S = w_j^D \ \exp(\lambda_{(c)}^{S\prime} x_{(c)j})$$
$$w_k^S = w_k^D \ \exp(\lambda_{(u)}^{S\prime} x_{(u)k})$$

여기서 $\lambda_{(c)}^S$와 $\lambda_{(u)}^S$는 다음의 단위수준별 칼리브레이션 조건을 만족시키는 해(solution) 벡터이다.

$$\sum_{j \in S_{(c)}} w_j^S x_{(c)j} = X_{(c)}$$
$$\sum_{j \in S_{(u)}} w_k^S x_{(u)k} = X_{(u)}$$

식(3)의 거리측도는 승법방식(multiplicative method)이라고도 불린다. 칼리브레이션 가중치 산출을 위해 승법방식 이외에 다양한 거리측도들이 있다. 예를 들면, 선형방식 $D(w^d, w) = \Sigma_s (w_i - w_i^d)^2 / 2w_i^d$이 사용되면, 일반화선형추정(generalized regression estimation)에 의한 가중치 조정이 된다. 앞서의 래이킹-비 조정은 가중치 절삭(weight truncation) 과정을 삽입한다면, 비 조정의 성능의 결과를 크게 헤치지 않고 좀 더 안정적인 통합가중치를 얻을 수 있다. 혹은 가중치 조정의 상한과 하한을 정하여 가중치를 조정할 수 있는 로짓방식(logit method)나 절사선형방식(truncated linear LU method) 등도 있다. 혹은 일반화 래이킹-비 추정과 다양한 거리함수에 대한 설명은 Deville et al.(1993)을 참고할 수 있다.

개별적 개인 칼리브레이션 가중치는 가구 칼리브레이션 가중치 w_j^S에 가구 내 개인의 조건부 설계가중치 $w_{k|j}^D = M_j$를 곱한 값을 개인 칼리브레이션 가중치의 기초가중치 w_k^D를 대신할 수도 있을 것이다.

개별적 칼리브레이션 가중치는 산출이 간편하다는 장점이 있다. 하지만 결과적으로 개인수준 보조정보를 통한 칼리브레이션 보정으로 얻을 수 있는 단위간 구조적 정보를 반영하지 못하여 가구가중치의 정확도 개선이 이루어지지 못하는 단점을 갖는다(Kolenikov & Hammer 2015).

3. 개인화 통합 칼리브레이션 가중치

본 절에서는 가구와 개인간 구조적 특성을 반영하는 통합가중치 산출방법 중 가구수준 보조정보를 개인수준으로 통합하는 Kolenikov & Hammer(2015)의 개인화 통합 칼리브레이션 가중치에 대해 살펴본다. 첫째, 가구수준 특성 $x_{(c)j}$을 개인화시켜 개인수준으로 다음과 같이 정의한다.

$$x_{(c)k} = x_{(c)j} \quad (4)$$

여기서 첨자 $(c)k$는 가구수준의 특성으로 개인화된 특성을 정의한 것이다. 둘째, 다음과 같이 통합한 개인화 보조정보

$$x_{(cu)k} = (x'_{(c)k}, x'_{(u)k})'$$

와 해당 통합 보조특성의 모총합 $X^I_{(cu)}$ 를 생성한다. 여기서 첨자$(cu)k$는 가구와 개인의 개인화 통합을 나타내며, $X^I_{(cu)} = \sum_{k \in U_{(u)}} x_{(cu)k}$은 개인화 통합 보조정보의 모총합이고 $U_{(u)}$는 개인모집단을 나타낸다. 셋째, 개인 설계가중치 w^d_k로부터 개인화 통합 보조정보를 이용한 래이킹-비 조정을 통해 개인화 개인 통합가중치 $w^I_k = w^d_k \exp(\lambda^I_{(cu)}{}' x_{(cu)k})$을 산출한다. 여기서 $\lambda^I_{(cu)}$는 다음의 칼리브레이션 조건을 만족시키는 해 벡터가 된다.

$$\sum_{k \in s_{(u)}} w^I_k x_{(cu)k} = X^I_{(cu)} \tag{5}$$

넷째, 식 (2)의 단위수준간 가중치 구성관계를 이용하여 개인화 가구 통합가중치는 다음과 같이 산출한다.

$$w^I_j = w^I_k / M_j$$

즉 배우자가 없는 가구의 경우에는 $w^i_j = w^i_k$이고 배우자가 있는 가구는 $w^i_j = w^i_k/2$이 된다.

물론 가구와 개인간 구성특성을 가구수준으로 통합하는 가구화 통합가중 방식(Estevao & Särndal 2006; Kolenikov & Hammer 2015)을 고려할 수도 있다. 주의할 점은 개인화 통합가중이나 가구화 통합가중 방식 모두 통합의 기준이 되는 수준이 아닌 다른 수준에서의 칼리브레이션 조건을 만족하지 못할 수도 있다는 비효율성을 갖는다(Kolenikov & Hammer 2015).

4. 대리응답에 의한 가구화 통합 칼리브레이션 가중치

대리응답(proxy interview)이란 표본조사의 대상 개체를 대신하여 다른 개체가 응답하는 간접응답(indirect interview)을 지칭한다(Thomsen & Villund 2011).

본 절에서는 가구주를 대신하여 배우자가 응답하는 경우에 이를 가구주의 대리응답으로 간주하여 추출단위로서의 가구와 응답자 개인이 갖는 단위수준별 (모집단) 대표성 문제를 가구수준으로 단순화시켜는 가구화 통합 칼리브레이션 가중치에 대해 살펴본다. 대리응답의 가정하에서는 식(1)의 설계가중치를 사용하여야 한다.

대리응답 가구화 통합가중치 산출은 다음의 단계로 산출할 수 있다. 첫째, 응답자에 상관없이 가구주 특성 $x_{(u)k}$을 다음과 같이 가구수준 특성으로 간주한다.

$$x_{(u)j} = x_{(u)k} \tag{6}$$

여기서 첨자 $(u)j$는 가구주 특성을 가구수준로 나타냄을 뜻한다. 둘째, 다음과 같은 가구화 통합보조정보

$$x_{(cu)j} = \left(x'_{(c)j}\,,\, x'_{(u)j}\right)'$$

와 모총합 X_{cu}^{P}를 생성한다. 여기서 첨자 $(cu)j$는 가구와 개인(가구주)의 가구수준 통합을 나타내며, $X_{(cu)}^{P} = \sum_{j \in U_{(c)}} x_{(cu)j}$은 가구화 통합 보조정보의 모총합이고 $U_{(c)}$는 가구 모집단을 나타낸다. 셋째, 가구 설계가중치 w_j^D로부터 가구화 통합보조정보를 이용한 래이킹-비 조정을 통해 가구화 가구 통합가중치 $w_j^P = w_j^D \exp(\lambda_{cu}^{P\prime} x_{(cu)j})$을 산출한다. 여기서 λ_{cu}^{P}는 다음의 칼리브레이션 조건을 만족시키는 해 벡터가 된다.

$$\sum_{j \in s_{(c)}} w_j^P \; x_{(cu)j} = X_{(cu)}^{P}$$

대리응답으로 간주하게 되면 가구수준 분석만을 고려하는 것이고, 가구주와 배우자간 개인특성을 칼리브레이션을 통한 가중치 산출에 반영할 수 없는 반면 2.3절과는 달리 가구화 개인 통합가중치의 산출은 필요 없게 된다. 또한 식(6)에서 보듯이 가구주 특성이 가구 특성이므로 동일수준 보조정보의 모총합만을 확보하면 된다. 따라서 식(4)와 식(5)와 같이 모집단 내 가구 및 가구 내 구성원수를 알 수 있는 특정한 경우에만 사용이 가능한 2.3절의 개인화 통합 칼리브레이션 가중치에 비해 실현화 가능성이 크다는 장점을 갖는다.

Ⅲ. 사례연구

1. 농어업인복지실태조사

농어업인복지실태조사는 농업진흥청이 주관하며 농어촌 특성에 맞는 복지증진 및 지역 개발시책의 수립과 시행 등에 필요한 기초자료 제공을 목적으로 한다. 농어촌(읍·면) 지역의 가구들을 대상으로 5년 주기로 조사하며 매년 특정 부문별로 조사하여 공표함으로써 농어촌의 복지실태 상황 등에 관한 변화추이를 파악한다. 주기 첫 해에는 읍·면지역은 물론 동지역을 포함하여 비교하는 도·농 종합조사로 진행한다. 표본추출은 동·읍·면 지역구분과 동 지역 내 서울특별시, 광역시, 기타 지역과 읍·면 지역 내 9개 도지역으로 층화한 뒤 층별로 읍·면·동, 조사구, 가구의 순으로 선택하며, 마지막으로 가구 내 가구를 대표하는 19세 이상 가구주 혹은 배우자 중 한 명을 조사대상 가구에서 자율적으로 선택하도록 한다.

농어업인복지실태조사의 가중치 산출은 2013년 조사에서 가구 내 응답자의 연령대만을 보조변수로 고려하여 가구 설계가중치를 조정하여 얻은 최종가중치를 사용한 표본가중합이 성인기준 모총합과 일치하도록 하는 사후층화(post-stratification)를 적용하였고, 2016년 이후 조사에서 주택유형을 추가로 포함하여 래이킹-비(raking-ratio) 조정방식에 의해 가구 설계가중치를 조정하여 얻는 최종가중치를 사용한 표본가중합이 성인기준 모총합과 일치하도록 하였다. 단, 성인과 가구라는 상이한 기준을 사용함에 따른 불일치를 조정하고자 성인 총수에 주택유형별 비율을 곱하여 성인기준 주택유형 총계를 산출하였다. 가중치 산출에 가구와 개인 특성이 다소 혼재하고 있음을 알 수 있다(농촌진흥청 2013).

더불어 조사문항은 “귀 댁의”로 문의하는 가구문항과 “귀하께서는”으로 문의하는 개인문항이 함께 포함된다. 가구문항으로는 연구 가구소득, 장애가구원 존재, 국민기초생활보장 수급 정도 등이 있고 개인문항으로는 생활여건 만족도, 행복요인, 복지여건 등이 포함된다. 따라서 기존 가중치로는 수준별 문항에 대한 대표성을 적절히 반영하지 못할 수도 있게 된다.

2. 가중치 비교

본 연구를 위한 보조정보는 통계청의 마이크로 데이터 통합서비스(https://mdis.kostat.kr)에서 제공하는 인구주택총조사 2015년 2% 표본자료를 이용하였다. 모집단은 농어업인복지실태조사의 조사대상자인 읍·면의 일반 조사구와 아파트 조사구 내에 거주하는 만 19세 이상의 가구대표자인 가구주와 배우자이다. 따라서 가구는 해당 가구대표가 거주하는 가구를 대상으로 하여 2% 표본추출을 반영하는 가구가중치를 이용하여 산출하였다. 2015년 인구주택총조사 기준으로 (추정하여) 파악한 전체 가구수는 3,524,897개이고 가구대표자 수는 5,639,799명이며 성인 수는 6,893,716명이다.

〈표 1〉 가구 및 개인 가중치 방법

가중치 방법	가중치		보조정보	
	가구	가구 대표(성인)	가구	가구대표(성인)
WD	w_j^D	w_k^D		
WS	w_j^S	w_k^S	주택유형	대표자 연령
WI	w_j^I	w_k^I	주택유형	대표자 연령
WP	w_j^P	-	주택유형, 가구주 연령	
WSG	w_j^{SG}	w_k^{SG}	주택유형, 가구주 성별	대표자 연령, 대표자 성별
WIG	w_j^{IG}	w_k^{IG}	주택유형, 가구주 성별	대표자 연령, 대표자 성별
WPG	w_j^{PG}	-	주택유형, 가구주 연령, 가구주 성별	
WO	-	(w_l^O)		(성인 연령) (주택유형 비율×성인 총수)

가중치 비교 및 자료분석은 2017년 농어업인복지실태조사에 응답한 총 3,995개 가구에 대해서 수행하였다. <표 1>은 가중치 산출방법에 따른 수준별 가중치 구성과 표기는 물론 칼리브레이션 조정에 고려된 보조정보를 비교하고 있다. 설계가중치 WD는 식(1)과 식(2)에 의해 산출되었다. 농어업복지실태조사에서 사용한 기존가중

치 WO는 3.1절에서 기술한 대로 산출되었는데 가중총합은 우리나라 전체 만 19세 이상 성인수와 같아지도록 칼리브레이션 조정이 수행되어 궁극적으로는 성인수준의 가중치로 이해할 수 있다. 개별적 칼리브레이션 가중치 WS와 개인화 통합가중치 WI는 각각 농어업복지실태조사에 고려한 주택유형과 가구대표자 연령대의 서로 다른 수준의 특성을 고려한 칼리브레이션을 적용하여 산출하였다. 반면 대리응답 가정의 가구화 가중치 WP는 주택유형과 가구주 연령대의 동일한 가구수준 특성을 고려한 칼리브레이션을 적용하여 산출하였다. 마지막으로 가구주 성별 특성을 추가하여 개별적 가중치 WSG, 개인화 통합가중치 WIG, 대리응답 가구화 가중치 WPG를 각각 칼리브레이션을 적용하여 산출하였다.

〈표 2〉 가중치 비교

가중치방법		평균	CV	총합	최소	Q_1	Q_2	Q_3	최대	L_w	$D(d,u$
가구	WD	821.1	0.45	3280166	185.0	531.0	710.0	1024.0	1580.0	1.20	-
	WS	882.3	0.58	3524897	172.0	553.4	749.1	1125.8	4337.3	1.33	0.18
	WI	830.4	0.97	3317612	60.8	339.5	555.4	1077.9	7147.3	1.94	0.83
	WP	882.3	0.93	3524897	80.9	388.3	639.0	1112.6	7492.6	1.86	0.80
	WSG	882.3	1.03	3524897	118.7	288.6	497.7	1243.3	8443.0	2.05	1.02
	WIG	871.8	1.13	3482955	38.1	315.2	556.1	1066.4	8867.9	2.27	1.10
	WPG	882.3	1.41	3524897	39.2	207.6	526.4	1049.8	16113.1	2.98	1.69
	모집단			3524897							
가구대표	WD	1343.7	0.57	5368140	185.0	684.0	1136.0	1894.0	3160.0	1.32	-
	WS	1411.7	0.98	5639799	132.8	531.6	882.7	1978.5	8790.3	1.96	0.97
	WI	1411.7	1.06	5639799	60.8	444.7	803.4	1987.0	12864.6	2.13	1.35
	WSG	1411.7	1.04	5639799	89.7	407.8	867.0	1922.2	10458.9	2.08	1.16
	WIG	1411.7	1.13	5639799	38.1	475.0	789.5	1818.1	16503.2	2.28	1.70
	모집단			5639799							
성인	WO	1725.6	1.41	6893716	91.8	504.0	972.9	1968.3	34157.6	3.00	4.72
	모집단			6893716							

<표 2>는 가중치 방법별 가구가중치, 가구대표가중치, 성인가중치의 분포들을

비교하고 있다. 먼저, 가구가중치 중 WD, WI, WIG를 제외한 모든 가중치의 평균은 882.3로 같았다. WD는 낮은 평균값을 갖는데 이는 무응답으로 인해 탈락한 가구들에 대한 무응답 조정이 수행되지 않았기 때문이다. 또한 WI와 WIG의 평균도 830.4와 871.8로 다른 가중치의 평균보다 작거나 큰데, 그 이유는 2장 3절에서 논의한 바와 같이 가구특성을 개인화시켜 통합함에 따라 통합가구가중치에 비효율성, 즉 불일치가 생겼음을 보여준다. 이에 반해 가구대표가중치는 WD를 제외한 모든 가중치의 평균이 동일하게 1,411.7의 값을 갖는다. 가구대표가중치 WD는 앞서의 가구가중치와 동일하게 무응답 조정이 반영되지 못함에 따른 결과이다. 기존 농어업인복지실태조사의 성인 가중치 WO의 평균값은 1,725.6으로 이는 전체 성인인구를 대상으로 하기 때문임을 확인할 수 있다.

가중치 변동을 나타내는 CV는 WS에 비해 WI(WIG)와 WP(WPG)이 다소 높게 나타나고 있다. 이는 WI(WIG)는 가구와 개인이라는 서로 다른 수준의 통합적 구조를 반영시켰기 때문이고 WP(WPG)는 추가적 가구수준 특성을 반영시켰기 때문이다. 최솟값, Q1, Q2, Q3, 최댓값의 기술통계의 퍼짐은 물론 범위와 불균등가중치로 인한 분산증가분 $L_w = 1 + cv_w^2$와 식(3)의 조정 전후간 거리측도 $D(d, w)$ 값도 동일한 양상을 보여준다. 더불어 주택유형과 대표자 혹은 가구주 연령대만을 고려한 WS, WI, WP에 비해 가구주와 대표자 성별을 추가로 고려한 WSG, WIG, WPG의 변동이 좀 더 크게 나타났다. 특히 WPG의 변동이 가장 크게 나타났는데 이는 단일수준을 기준으로 가장 많은 보조변수를 고려하였기 때문인 것으로 판단된다.

<표 3>은 가중치방법에 따른 주택유형별 총합추정치를 비교하고 있다. WD를 제외한 다른 6개 가중치 모두 보조변수로 주택유형을 고려하였다. 주택유형별 가중총합은 WI와 WIG를 제외하고 상대포함오차는 없거나 낮은 값을 보인다. WI와 WIG의 가중총합은 전체적으로 -5.9%와 -1.2%의 낮은 상대포함오차를 보인 반면 세부 주택유형 중 기타분류가 -15.9%와 -9.5%로 다소 높은 음의 상대포함오차를 보인다. 이는 개인화를 통한 통합가중치 산출이 주는 가구수준 통합가중치의 비효율성임을 알 수 있다. 주의할 점은 보조변수인 주택유형을 표본층별로 고려했기 때문에 표본수가 상대적으로 작은 특정층에서 두 가지 주택유형을 하나의 범주로 묶어서 조정함에 따라 칼리브레이션 조정에 의한 표본가중합이 모총합과 약간의 차이를 갖게 되었다.

〈표 3〉 **가구 가중총합 비교: 주택유형** (단위: 만 가구, %)

가중치 방법	총합				상대포함오차			
	단독주택	아파트	기타	전체	단독주택	아파트	기타	전체
WD	200.9	102.9	24.2	328.0	-2.5%	-2.7%	-40.4%	-6.9%
WS	206.4	105.5	40.7	352.5	0.1%	-0.3%	0.0%	0.0%
WI	188.2	109.4	34.2	331.8	-8.7%	3.5%	-15.9%	-5.9%
WP	206.4	105.5	40.7	352.5	0.1%	-0.3%	0.0%	0.0%
WSG	206.4	105.5	40.7	352.5	0.1%	-0.3%	0.0%	0.0%
WIG	197.4	114.1	36.8	348.3	-4.2%	7.9%	-9.5%	-1.2%
WPG	206.4	105.5	40.7	352.5	0.1%	-0.3%	0.0%	0.0%
모총합	206.1	105.7	40.7	352.5				

〈표 4〉 **가구 가중총합 비교: 가구주 연령대** (단위: 만 가구, %)

가중치 방법	총합					상대포함오차				
	2-30대	40대	50대	60대	70+	2-30대	40대	50대	60대	70+
WD	20.6	36.2	45.1	68.4	157.6	-63.1%	-43.6%	-42.0%	5.6%	75.7%
WS	22.6	41.2	48.7	76.2	163.8	-59.5%	-35.8%	-37.5%	17.6%	82.5%
WI	41.1	68.9	73.0	63.5	85.3	-26.6%	7.4%	-6.2%	-2.1%	-5.0%
WP	55.9	64.2	77.9	64.8	89.7	0.0%	0.0%	0.0%	0.0%	0.0%
WSG	22.7	41.6	49.5	77.0	161.8	-59.5%	-35.2%	-36.5%	18.8%	80.3%
WIG	47.0	70.3	77.1	66.4	87.4	-15.9%	9.6%	-1.0%	2.5%	-2.6%
WPG	55.9	64.2	77.9	64.8	89.7	0.0%	0.0%	0.0%	0.0%	0.0%
모총합	55.9	64.2	77.9	64.8	89.7					

<표 4>는 가중치방법에 따른 가구주 연령대의 총합추정치를 비교해주고 있다. 가구수준 특성인 가구주 연령대가 반영된 WP와 WPG를 이용한 가중총합은 상대포함오차가 없지만 반영하지 않은 WS, WSG, WI, WIG는 작지 않은 상대포함오차를

보이고 있다. 가구와 대표자를 따로 고려한 WS와 WSG에 비해 통합방식으로 산출한 WI와 WIG이 다소 작은 상대포함오차를 갖는다. 특히 WI보다 많은 보조정보를 사용한 WIG의 상대포함오차가 작음을 알 수 있다. 이는 가중치 산출에서 미세하나마 통합적 가중치 산출의 효과를 나타낸다고 할 수 있다.

<표 5>는 가구대표 연령대를 반영한 가중치들의 총합추정치를 비교하고 있다. 예상할 수 있듯이 가구 설계가중치 WD와 기존 성인가중치 WO를 제외한 모든 가중치의 표본가중합은 모총합과 일치하는 것을 확인할 수 있다.

〈표 5〉 개인 가중총합 비교: 가구대표(성인) 연령대 (단위: 만 명, %)

가중치 방법	총합					상대포함오차				
	2-30대	40대	50대	60대	70+	2-30대	40대	50대	60대	70+
WD	51.2	61.7	84.1	120.6	219.2	-50.3%	-44.9%	-36.8%	18.9%	91.3%
WS	103.1	111.9	133.0	101.4	114.6	0.0%	0.0%	0.0%	0.0%	0.0%
WI	103.1	111.9	133.0	101.4	114.6	0.0%	0.0%	0.0%	0.0%	0.0%
WSG	103.1	111.9	133.0	101.4	114.6	0.0%	0.0%	0.0%	0.0%	0.0%
WIG	103.1	111.9	133.0	101.4	114.6	0.0%	0.0%	0.0%	0.0%	0.0%
WO	(186.9)	(122.1)	(139.8)	(108.7)	(131.9)	(2.7%)	(-3.1%)	(-1.3%)	(1.8%)	(-0.8%)
모총합	103.1	111.9	133.0	101.4	114.6					
	(182.0)	(126.0)	(141.6)	(106.8)	(133.0)					

<표 6>은 가중치 방법에 따른 가구주, 가구대표, 성인의 성별 총합추정치를 비교해주고 있다. 2015년 인구주택총조사 2% 표본자료로 추정한 남성과 여성이 가구주인 모집단 가구수는 각각 2,592,806가구와 932,091가구이다. 또한 남성과 여성의 가구대표자 모집단 총수는 각각 2,733,033명과 2,906,766명이고 성인기준 남성과 여성의 모집단 총수는 3,386,528명과 3,507,178명이다.

〈표 6〉 **가중총합 비교: 가구, 가구대표, 성인 성별** (단위: 만 가구, 만 명)

가중치 방법	가구총합		상대포함오차		가구대표(성인) 총합		상대포함오차	
	남성	여성	남성	여성	남성	여성	남성	여성
WD	234.6	93.4	-9.5%	0.2%	234.6	302.2	-14.2%	4.0%
WS	253.0	99.5	-2.4%	6.8%	225.0	339.0	-17.7%	16.6%
WI	264.7	67.1	2.1%	-28.1%	230.9	333.1	-15.5%	14.6%
WP	279.4	73.1	7.8%	-21.6%	-	-	-	-
WSG	259.3	93.2	0.0%	0.0%	273.3	290.7	0.0%	0.0%
WIG	254.2	94.1	-2.0%	-2.0%	273.3	290.7	0.0%	0.0%
WPG	259.3	93.2	0.0%	0.0%	-	-	-	-
WO					(275.9)	(413.5)	-18.5%	17.9%
모총합	259.3	93.2			273.3	290.7		
					(338.7)	(350.7)		

〈표 7〉 **보조특성별 적합도 비교**

X^2적합도	가구 및 가구주 특성			가구태표(성인) 특성	
가중방법	주택유형	연령	성별	연령	성별
WD	68542.0	997920.2	23528.2	1657384.5	59311.1
WS	10.4	1021025.6	5803.8	0.0	165544.1
WI	27150.9	48514.6	74570.5	0.0	127552.0
WP	10.4	0.0	59116.1	-	-
WSG	10.4	982833.7	0.0	0.0	0.0
WIG	14030.9	21222.9	1066.8	0.0	0.0
WPG	0.0	0.0	0.0	-	-
WO	-	-	-	(3254.7)	(228900.4)

$$X_{w=}^{2}\sum_{g=1}^{G}\frac{(\widehat{X}_{g}(w)-X_{g})^{2}}{X_{g}}$$

먼저, 대리응답 통합가중치 WP는 보조특성 총합추정의 적합도 측면에서 매우 양호하며 가구주 성별과 같은 추가적인 보조정보를 활용한다면 WPG와 같이 개선되는 것을 확인할 수 있다. 물론 대리응답 가정은 가구수준의 분석만 가능하게 되며 가구주 성별 등의 추가적인 보조정보를 이용함에 따른 제약으로 인해 가중치 변동량은 분산증가측도 기준으로 $L_w(\mathrm{WP})=2.09$에서 $L_w(\mathrm{WPG})=2.15$로 증가하는 단점을 갖게 된다(<표 2> 참조). 수준별 분석을 가능하게 하는 가중치 방법에 있어서는 WS(WSG)보다는 WI(WIG)이 전반적으로 나은 적합도를 보여준다. 더불어 수준별 성별 특성을 추가로 고려한다면 WSG와 WIG에서 보듯 그렇지 않은 가중치에 비해 나은 적합도를 보여준다. 주의할 점은 다른 칼리브레이션 가중치들은 인총 2% 표본정보로 모총합을 산출한 반면에, 농촌진흥정의 기존가중치 WO는 통계청 통계포털의 모집단 자료를 이용하였다. 이에 성인특성인 연령과 성별 총합에 대한 적합도 값은 비교적 큰 값을 보인다.

〈표 8〉 농어촌생활 및 기초생활기반 종합만족도

만족도	농어촌생활		표준오차		기초생활기반		표준오차	
가중방법	가구주	가구대표 (성인)	가구주	가구대표 (성인)	가구주	가구대표 (성인)	가구주	가구대표 (성인)
WD	56.3	56.7	0.46	0.48	53.4	53.4	0.84	0.84
WS	56.2	56.3	0.51	0.58	53.4	53.0	0.82	0.98
WI	55.6	56.0	0.54	0.58	52.1	52.2	0.88	0.91
WP	55.6		0.53		51.8		0.86	
WSG	56.2	56.3	0.51	0.55	53.2	52.9	0.82	0.96
WIG	55.5	56.0	0.52	0.54	52.0	52.1	0.89	0.91
WPG	55.5		0.53		51.7		0.86	
WO		(55.4)		(0.58)		(51.3)		(0.90)

<표 8>은 가중치 방법별로 추정한 농어촌생활 종합만족도와 기초생활만족도 및 해당 표준오차를 정리하고 있다. 만족도 추정값은 WS(WSG)에 비해 WI(WIG)와 WP(WPG)에서 매우 미세하게 작은 값을 가지며 큰 차이를 보이지는 않는다.

Ⅳ. 논의

가구 내 한 명만을 조사하는 가구조사에서 조사문항에 따라 응답자는 가구를 대표하기도 하며 개인을 나타낼 수도 있다. 분석단위별 자료분석을 수행하기 위해서는 가구와 개인의 가중치 산출이 필요하며, 가능한 단위간 구성적 측면을 칼리브레이션을 통해 반영한 통합가중치의 산출이 바람직하다. 본 논문에서는 흔히 고려되는 개별적 칼리브레이션 가중치와 가구 보조정보의 개인화를 통한 통합 칼리브레이션 가중치를 비교하여 살펴보았다. 단위수준별 보조정보를 함께 고려하는 통합가중치는 개별적으로 고려하는 가중치에 비해 변동성이 증가할 수 있지만 이로 인한 추정량의 분산증가는 매우 미미하다. 또한 적절한 보조정보 선택을 통해 해당 특성은 물론, 이와 상관관계가 높은 조사변수에 대해서는 적합도가 매우 우수할 수 있음을 확인하였다.

대리응답을 가정하면 가구가중치만 제공할 수 있고 이에 따라 문항에 따른 가중치 선택을 달리해야 하는 혼선을 방지할 수 있다. 더불어 적절한 가구 보조정보를 고려한 칼리브레이션을 수행하면 해당 특성의 적합도는 매우 향상될 수 있게 된다. 하지만, 대리응답을 가정하면 표본개체와 대리응답개체간 응답 차이로 인해 가구 내 대리응답자의 조사참여가 무작위(random)적이지 않을 수 있다. 전자는 측정오차(measurement error)에 해당하고 후자는 선택효과(selection effect)로 해석된다(Thomsen & Villund 2011). 측정오차에 대한 평가는 용이하지 않은 반면, 선택효과에 대한 평가는 대리응답의 성향분석을 통해 가능하다. Thomsen & Villund (2011)는 대리응답의 두 가지 메카니즘을 다음과 같이 정의하고 있다. 대리응답의 확률이 조사특성 y에는 좌우되지 않지만 보조변수 x와 서로 무관하다면 이는 완전임의대리응답(Proxy Completely At Random, PCAR)이고, 이는 마치 대리응답이 전체

표본가구로부터 임의로 확률추출된 것으로 자료분석에서 대리응답의 영향력이 무시될 수 있는 완전히 임의적인 상황을 뜻하며 대리응답자료에 의한 분석이 표본전체를 대표하는데 문제는 없다. 반면 대리응답의 확률이 조사특성 y에는 좌우되지 않지만 보조변수 x의 일부 혹은 전부에 좌우된다면 이는 임의대리응답(Proxy At Random, PAR)이라 하고, 대리응답은 보조변수 x에 의존하는 대리응답성향모형을 통해 충분히 설명할 수 있다. 임의대리응답의 상황에서는 보조변수의 적절한 선택을 통한 칼리브레이션 조정에 의해 대리응답에 의한 편향도 축소시킬 수 있을 것으로 기대된다.

참고문헌

농촌진흥청. 2013. “2013 농어업인 복지실태조사 표본설계 최종보고서.”

Deville, J.C. and C-E. Särndal. 1992. “Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association* 87: 376-382.

Estevao, V.M. and C-E. Särndal. 2006. “Survey Estimates by Calibration on Complex Auxiliary Information.” *Internatinal Statistical Review* 74: 127-147.

Kolenikov, S. and H. Hammer. 2015. “Simultaneous Raking of Survey Weights at Multiple Levels.” *Survey Methods: Insights from the Field, Special Issue: `Weighting: Practical Issues and ‘How to’ Approach*'. Retrieved from http://surveyinsights.org/?p=5099.

Steel, D.G. and R.G. Clark. 2007. “Person-level and Household-level Regression Estimation in Household Surveys.” *Survey Methodology* 33: 51-60.

Thomsen, I. and O. Villund. 2011. “Using Register Data to Evaluate the Effects of Proxy Interviews in the Norwegian Labour Force Survey.” *Journal of Official Statistics* 27: 87-98.

<접수 2018/10/19, 수정 2018/11/21, 게재확정 2018/11/28>