Journal of the Korean Data & Information Science Society
2014, **25**(3), 523–531

http://dx.doi.org/10.7465/jkdi.2014.25.3.523
한국데이터정보과학회지

설계효과모형 적용에 관한 연구[†]

박인호[1]

[1]부경대학교 통계학과

접수 2014년 3월 5일, 수정 2014년 4월 4일, 게재확정 2014년 4월 18일

요 약

설계효과는 새로운 표본설계를 계획하거나 기존 표본조사에 적용된 설계요소의 효율성을 평가하는데 널리 사용된다. 본 연구에서는 Gabler 등 (2006)이 제시한 설계효과모형을 층화이단집락추출의 표본설계로 이루어진 2013 식품소비행태조사에 응용하여 적용하였다. 조사결과를 통해 표본설계모형의 유용성과 적절성에 대해 논의하였다.

주요용어: 2013 식품소비행태조사, 가중치, 집락내상관계수, 집락추출, 층화추출.

1. 머리말

설계효과 (design effect)는 복잡한 형태를 갖는 표본설계 (complex sampling design) 하에서 관심추정량이 갖는 분산을 동일한 크기의 가상적인 단순확률표본이 주는 분산에 비교하여 정의된다 (Kish, 1965). 예를 들어, 관심변수 y의 평균추정치의 설계효과는 다음과 같이 정의된다.

$$Deff\,(p, \bar{y}_p) = V_p(\bar{y}_p)/V_{srs}(\bar{y})$$

여기서 $\bar{y}_p = \Sigma_{k\in s} w_k y_k / \Sigma_{k\in s} w_k$와 $V_p(\bar{y}_p)$는 각각 $p = p(s)$으로 표기하는 복잡표본설계에서 결정되는 표본가중치 w_k를 이용한 (가중)평균과 분산을 나타내며, $\bar{y} = \Sigma_{k\in s} y_k/n$와 $V_{srs}(\bar{y})$는 크기 n의 단순확률표본 (simple random sample)에서 정의되는 표본평균과 분산을 뜻한다. 따라서 설계효과란 기존 표본설계에 적용되었거나 새로운 표본설계에 적용할 설계요소들이 단순확률추출 (simple random sampling, srs)과 비교하여 추정량 분산의 크기를 결정하는 영향력, 즉 효율성을 평가하는데 널리 사용된다. 이때 단순확률추출에는 복원과 비복원의 방식이 모두 고려될 수 있지만 유한모집단 수정계수를 고려하지 않아도 되는 복원추출을 가정하고자 하는데, 이는 통계학에서 흔히 접하는 표본확률구조인 *i.i.d.* (independent and identically distributed) 상황에 해당된다.

설계효과는 새로운 표본설계의 표본크기를 결정함에 있어서 매우 중요한 역할을 하기도 한다. 이는 복잡표본설계로부터 얻게 되는 추정량의 분산과 같은 크기의 분산을 가질 수 있는 단순확률표본의 크기, 즉 유효표본크기인 $n_{eff} = n/Deff(p, \bar{y}_p)$으로 계산된다. 만약 단순확률추출에서 n개의 표본크기가 필요하다면, 복잡표본설계에서는 $n \times Deff$개의 표본이 필요하다고 할 수 있을 것이다.

설계효과의 개념을 가장 유용하게 사용되기 위해서는 모집단 구조나 표본설계요소 등이 갖는 개별적 영향을 가능한 명확히 함수형태, 즉 설계효과모형으로 표현해야 한다. 이러한 경우, 설계효과모형을 통해 새롭게 계획되는 표본설계의 효율성을 향상시킬 수 있게 된다 (Rust와 Broene, 2010). Kish

[†] 이 논문은 부경대학교 자율창의학술연구비 (2013)에 의하여 연구되었음.
[1] (608-737) 부산광역시 남구 용소로 45, 부경대학교 통계학과, 조교수. E-mail: ipark@pknu.ac.kr

(1987)는 집락추출과 불균등 가중치의 설계요소를 갖는 복잡표본설계에 대한 가중표본평균의 설계효과 모형을 경험과 직관에 근거하여 다음과 같이 제시하였다.

$$mdeff_p(\bar{y}_p) = \left\{1 + \rho_{yp}(\bar{b} - 1)\right\}\left(1 + cv_w^2\right) \tag{1.1}$$

여기서 ρ_{yp}는 관심변수 y의 집락내상관계수 (intracluster correlation coefficient)이고, $\bar{b}$는 평균집락표본크기이며, cv_w^2은 표본가중치의 상대분산을 나타낸다. Kish의 설계효과모형 (1.1)은 가중표본평균의 분산이 집락 내 개체간 동질성과 평균표본수 $\bar{b}$가 클수록 집락효과 (clustering effect)가 크고, 표본가중치의 변동성에 클 때 가중치효과 (weighting effect)가 커져서 복잡표본설계의 적용이 추정량의 정도에 좋지 않은 영향을 주는 모집단 구조와 설계요소의 함수적 관계를 잘 표현해주고 있다.

Gabler 등 (1999)은 불균등추출확률을 갖는 이단집락추출방식의 표본설계 하에서 1요인 분산분석모형을 가정하여 설계효과모형 (1.1)에 대한 모형에 근거한 타당성을 제시하였다. Gabler 등 (2006)은 서로 다른 종류의 집락들에 의한 차별적 집락효과를 반영할 수 있는 분산분석모형을 통해 좀 더 일반화된 설계효과모형을 제시하였는데 2절에서 이에 대해 논의한다.

Lee (2012)는 기존 연구들을 종합하여 층화, 집락, 가중치의 설계요소별 설계효과모형에 대한 기본적 논의를 정리하였다. 또한 Gabler 등 (2006)이 제시한 설계효과모형을 층화이단집락추출과 연계하여 표현하였고, 더불어 설계효과모형이 적합한 상황들을 관련된 가정들과 함께 요약하여 정리하여 주고 있다.

본 연구에서는 2013 식품소비행태조사에 적용된 설계효과모형에 대해 살펴보고 조사자료를 토대로 모형적용의 결과에 대해 논의하고자 한다. 2013 식품소비행태조사는 우리나라 가구조사에서 흔히 채택하는 층화이단집락추출에 의해 수행된 조사로 Gabler 등 (2006)이 제시한 설계효과모형을 응용하기에 매우 적합한 표본설계의 형태인 것으로 판단된다. 2013 식품소비행태조사의 표본설계는 유럽사회조사의 표본설계 가이드와 통계청 사회조사의 결과를 참고하여 이루어졌다 (Park 등, 2013). 2절에서는 먼저 Gabler 등 (2006)이 제시한 기본 모형과 가정들에 대해 간단히 살펴보고 이에 대해 Lee (2012)이 고려한 층화이단집락추출방식에 대한 적용에 대해 논의한다. 3절에서는 2013 식품소비행태조사의 표본조사 개요와 더불어 표본설계에 적용된 설계효과모형을 구체적으로 살펴본다. 또한 조사결과를 토대로 추정된 주요변수에 대한 설계효과와 설계효과모형의 적절성에 대한 논의를 포함한다. 4절에서는 본 논문에 대한 간단한 논의를 정리한다. Gabler 등 (2006)의 모형은 우리나라 가구조사에 적용한 층화이단집락추출방식의 표본설계를 계획할 때 유용한 것으로 판단된다. 다만 유럽사회조사와는 다소 상의한 설계효과를 경향을 보여주고 있다.

2. 설계효과모형

과학적 조사연구를 위해 채택하는 대부분의 표본설계는 단순확률추출을 따르지 않고 다양한 설계요소를 채택하게 된다. 예를 들면, 모집단을 서로 겹치지 않는 층으로 구분한 뒤 층별로 표본을 뽑는 층화추출을 통해 통계적 효율성은 물론 조사 관리의 효율성을 높이게 된다. 표본개체를 여러 단계에 걸쳐 뽑는 다단계추출을 고려함에 따라 표본개체들은 자연스럽게 집락화되며, 이를 통해 조사방문을 위한 지역이 축소될 수 있어 조사비용의 절감과 조사효율을 높일 수 있게 된다. 특정한 부분모집단으로부터 더 많은 표본을 확보하기 위해 모집단 개체들에 대한 불균등한 비율로 표본추출을 적용하게 됨에 따라 개체별로 상이한 표본가중치가 부여된다. 따라서 이러한 설계요소들은 표본추정치의 정확성이나 정도에 영향을 미치게 된다. 특히 설계요소들이 주는 추정량의 분산에 미치는 영향력은 단순확률추출과 비교하여 설계효과의 형태로 평가된다.

Gabler 등 (2006)은 분석영역 (domain)별로 근본적인 차별성을 갖는 표본설계를 고려하는 복잡표본설계에 적합한 설계효과모형을 제시하였다. 이는 국제비교조사인 유럽사회조사 (European social surveys; ESS)에서 참여국에 대해 비교 가능한 국가별 표본설계 가이드를 제시하기 위함이다 (Lynn 등,

2007). 유럽사회조사의 참여국 별로 사용가능한 표본틀의 특성은 물론 국가별 지역분석의 목적이 다르기 때문에 분석영역별로 불균등한 층화를 고려하는 표본설계의 유형 차이가 존재한다 (Lynn과 Gabler, 2005).

총 N개의 집락으로 구성된 모집단은 H개의 분석영역으로 나뉘고 개별 표본개체에 대해 관측값 y_{hik}과 표본가중치 w_{hik}가 함께 주어진다면, 모평균 추정량으로 가중표본평균으로 고려할 수 있다.

$$\bar{y}_p = \left(\Sigma_{h=1}^{H}\Sigma_{i=1}^{n_h}\Sigma_{k=1}^{b_{hi}} w_{hik}\right)^{-1} \Sigma_{h=1}^{H}\Sigma_{i=1}^{n_h}\Sigma_{k=1}^{b_{hi}} w_{hik} y_{hik}$$

여기서 (h_{ij})는 분석영역 h의 i번째 표본집락내 k번째 표본단위를 나타내며, n_h는 h번째 분석영역내 표본집락수를, b_{hi}는 h번째 분석영역의 i번째 표본집락내 표본개체수를 나타낸다.

만약 (i) 분석영역별 조사변수 y의 분산이 같고 (모든 분석영역 h에 대해 $\sigma_{yh}^2 = \sigma_y^2$이고), (ii) 동일 집락내에서만 0이 아닌 급내상관계수 ($cor(y_{hik}, y_{hik'}) = \rho_{yh}(k \neq k')$)를 갖고, (iii) 분석영역내 조사변수 y_{hik}와 표본가중치 w_{hik}간에 상관관계가 없다 ($cov(w_{hik}, y_{hik}) = 0$)는 가정 (Lee, 2012) 하에서, Gabler 등 (2006)은 다음의 설계효과모형을 도출하였다.

$$mdeff(\bar{y}_p, p) = \sum_{h=1}^{H} \left(\frac{\widehat{M}_h}{\widehat{M}}\right)^2 \left(\frac{b}{b_h}\right) mdeff_h \tag{2.1}$$

여기서 $\widehat{M}_h = \Sigma_{i=1}^{n_h}\Sigma_{k=1}^{b_{hi}} w_{hik}$와 $\widehat{M} = \Sigma_{h=1}^{H}\widehat{M}_h$은 각각 h번째 분석영역과 전체 모집단의 크기추정량이고, $b_h = \Sigma_{i=1}^{n_h} b_{hi}$와 $b = \Sigma_{h=1}^{H} b_h$는 각각 h번째 분석영역과 전체 표본크기를 나타낸다. 또한 $mdeff_h$는 h번째 분석영역의 설계효과로 Kish (1987)의 설계효과모형 (1.1)과 매우 유사한 형태로 다음과 같이 정의된다.

$$mdeff_h = mdeff_{wh} \times mdeff_{ch} \tag{2.2}$$

여기서 식 (2.2)의 오른편에 있는 두 요소는 분석영역 h의 표본가중치와 집락추출로 기인하는 설계효과로 각각 아래와 같이 정의된다.

$$mdeff_{wh} = 1 + cv_{wh}^2 \tag{2.3}$$

$$mdeff_{ch} = 1 + \rho_{yh}(b_h^* - 1) \tag{2.4}$$

또한 $cv_{wh}^2 = \Sigma_{i=1}^{n_h}\Sigma_{k=1}^{b_{hi}}(w_{hik} - \bar{w}_h)^2/(b_h\bar{w}_h)^2$는 분석영역 h내 표본가중치의 상대분산 (relative variance)을 나타내고, $\bar{w}_h = \Sigma_{i=1}^{n_h}\Sigma_{k=1}^{b_{hi}} w_{hik}/b_h$는 분석영역 h내의 평균가중치이다. 또한 ρ_{yh}는 분석영역 h내 집락내상관계수이며, b_h^*는 일종의 분석영역 h의 가중집락표본크기의 개념으로 다음과 같이 정의된다.

$$b_h^* = (\Sigma_{i=1}^{n_h}\Sigma_{k=1}^{b_{hi}} w_{hik}^2)^{-1}\Sigma_{i=1}^{n_h}(\Sigma_{k=1}^{b_{hi}} w_{hik})^2 \tag{2.5}$$

설계효과모형 (2.1)에 의하면 전체 표본이 갖는 설계효과는 분석영역별 설계효과에 분석영역 h의 인자 $\delta_{sh} = (\widehat{M}_h/\widehat{M})^2(b_h/b)^{-1}$가 곱해져 더해지는 형태로 나타난다. 이때 분석영역별 설계효과는 집락효과와 가중치효과의 곱 형태가 된다. 분석영역 h의 인자 δ_{sh}는 해당영역의 상대적 크기 ($\widehat{W}_h = \widehat{M}_h/\widehat{M}$)와 표본할당 ($a_h = b_h/b$)의 함수로 표현되는데, 이는 더해져 1이 되지 않으므로 설계효과모형 (2.1)은 영역별 설계효과의 볼록결합 (convex combination)이 아님을 알 수 있다. 만약 분석영역 h의 설계효과모형 $mdeff_h$가 분석영역에 관계없이 1의 값을 갖는다면 식 (2.1)은 다음과 같이 간단한 형태로 표현된다.

$$mdeff_s = \Sigma_{h=1}^{H}(\widehat{M}_h/\widehat{M})^2/(b_h/b)$$

분석영역을 층으로 고려한다면 위의 식은 비례할당과 비교한 불균등할당이 갖는 층효과 (stratification effect)에 대한 설계효과모형이 된다 (Lee, 2012, Kalton 등, 2005). 또한 분석영역은 Lee (2012)가 지적하였듯이 층으로도 고려할 수 있다.

h번째 분석영역의 집락내상관계수 ρ_{yh}에 대한 추정은 분산분석법 (analysis of variance)을 통해 추정할 수 있지만 연속형 변수에 대해서만 가능하게 된다. 따라서 ρ_{yh} 추정은 Kish (1965, 162쪽)가 제시한 합성추정 (synthetic estimation)을 고려할 수 있다 (Heeringa 등, 2010, 30쪽). Kish 방식 하에서 Kalton 등 (2005, 115쪽)이 제안한 합성추정치를 이용하면 식 (2.2)에 의해 다음과 같이 집락내상관계수의 추정치를 구할 수 있다.

$$\widehat{\rho}_{yh} = (b_h^* - 1)^{-1}[(\widehat{Deff}_h / mdeff_{wh}) - 1] \tag{2.6}$$

여기서 $\widehat{Deff}_h$는 표본조사자료를 바탕으로 분석영역 h의 설계효과 추정량을 뜻한다. 식 (2.6)으로 추정되는 집락내상관계수의 추정값은 동질성비율 (rate of homogeneity; roh)이라고 불리는데 설계효과모형에 대한 기본가정에 대한 만족을 근거로 하기 때문에 주의가 필요하다 (Park과 Lee, 2004).

유럽사회조사의 표본설계지침에 따르면 참가국들의 표본설계에서 유사조사를 통해 ρ_{yh}값을 알 수 없는 경우라면 $\rho_{yh} = 0.02$의 값을 쓸 것을 권유하고 있다. 제시된 급내상관계수는 일부 국가의 태도관련 문항에서 추정한 값들의 근사적 평균값에 해당한다. 우리나라 통계청의 사회조사의 경우는 이보다 큰 $\rho_{yh} = 0.096$인 것으로 알려져 있다.

3. 적용사례: 2013 식품소비행태조사

3.1. 설계효과 예측 및 결과 비교

식품소비행태조사는 우리나라 소비자의 식품소비행태에 대한 현황파악과 트렌드분석을 위해 한국농촌경제연구원이 주관하여 일년주기의 반복조사 형태로 2013년에 처음 실시되었다. 식품소비행태조사의 조사대상은 전국의 1인 및 혈연가구의 식품 주구입자와 성인·청소년 구성원이며, 전국과 수도권, 충청권, 호남권, 대경권, 동남권, 강원권의 6개 권역을 주요 분석영역으로 고려하고 있다. 단, 제주도는 크기제약으로 인해 호남권에 포함시켰다. 통계청의 2010년 인구주택총조사에서 파악한 조사구 명부를 표본틀로 사용하였고, 16개 특·광역시도를 표본층으로 하였다. 층별로 조사구와 가구를 차례로 뽑는 이단집락추출을 적용하였는데 1단계에서는 추출 전에 특성에 따라 조사구를 나열하는 내재적 층화와 더불어 가구 수에 비례한 계통확률로 표본조사구를 추출하였다. 조사구내 표본가구는 단순확률에 따라 추출하였다.

예산상의 제약으로 인해 약 3,000 가구만이 조사 가능함에 따라 전국과 권역별 추정량의 정도수준을 동시에 제고할 수 있도록 절충할당을 고려하였다. 서울이나 경기도와 같이 상대적으로 규모가 큰 지역은 세부지역별 추정도 가능할 수 있도록 권역 수준보다는 표본층 수준의 절충을 고려하였다. 먼저, 규모가 가장 작은 강원도는 적절한 정도수준을 갖도록 총 216가구를 할당한 뒤, 나머지 2,784가구에 대해서는 표본층별로 제곱근할당 (square-root allocation)을 실시하였다. 여기서 제곱근할당은 멱할당 (power allocation)의 특수한 형태로 상세한 기술은 Bankier (1988)이나 Kim과 Kwak (2013) 등에 설명되어 있다. 표본조사구 수는 조사구 당 5개의 응답가구 를 목표로 총 602개의 조사구를 추출할 것을 고려하였다.표본할당에 따라 예상되는 정도수준의 평가를 위해 설계효과모형 (2.1)을 사용하였고, 권역수준의 설계효과를 평가하기 위해서 다음의 식을 고려하였다.

$$mdeff(\bar{y}_p^d, p) = \frac{\sum_{h \in H_d}(W_h^2/a_h) mdeff_h}{\sum_{h \in H_d}(W_h^2/a_h)} \tag{3.1}$$

여기서 $\bar{y}_p^d$는 권역 d의 표본추정량을 정의하며, H_d는 권역 d를 구성하는 표본층을 나타내며, $W_h = M_h/M$와 $a_h = b_h/b$는 각각 표본층의 상대적 크기와 (목표응답) 표본할당비 (allocation rate)를 나타낸다. 다시 말해 권역수준의 설계효과를 표본층별 설계효과에 대한 가중평균으로 추정할 수 있다. 식 (2.1)과 달리 식 (3.1)에서는 표본설계 단계에서 파악된 표본층 크기를 추정량에 대신하여 사용하였다.

유사한 조사 및 관련정보의 부재로 인해 설계효과평가를 위해 다음의 기본 설정을 가정하였다. 무응답조정 등과 같은 가중치 조정에 따른 불균등한 가중치 효과를 반영하기 위해 층별 가중치의 상대분산을 $cv_{wh}^2 = 20\%$으로 가정하였고, 층내 조사구 집락내상관계수는 통계청 사회조사의 추정값인 $\rho_h = 0.096$을 가정하였다. 설계효과모형 (2.1)과 (3.1)에 의해 예측한 전국과 권역수준의 설계효과는 각각 1.9460과 1.6608이었다. 단순확률추출에서의 모비율의 추정치 분산이 $p(1-p)/b$이므로 상대표준오차는 $cv(\hat{p}) = b^{-1} \times mdeff \times (1-p)/p$이다. 따라서 $p = 0.5$에 대한 예측상대표준오차는 전국이 2.5%이고 권역별로는 수도권 4.2%, 충청권과 호남권은 6.0%, 대경권 6.6%, 동남권 5.5%, 강원권 8.8%으로 나타났다. 상세한 설계내역 및 예측설계효과에 대한 논의는 Park 등 (2013)을 참고할 수 있다.

Table 3.1은 2013 식품소비행태조사에 조사한 대표적인 몇 가지 문항들에 대해 특성여부와 동향 및 전망 (혹은 물가)지수를 산출한 후, 이에 대한 추정치, 상대표준오차, 설계효과 추정치를 각각 나열하고 있다. 특성여부로는 가구내 취사, 인터넷 식품구입, 친환경식품구매 등을 포함하며 전망 (혹은 물가)지수로는 식품소비 지출, 체감 장바구니 물가수준, 식품별 지수들을 포함한다. 전망지수는 5점의 리커트 척도, 즉, 매우 감소, 약간 감소, 변화 없음, 약간 증가, 매우 증가를 -1, -0.5, 0, 0.5, 1.0의 값으로 부여한 후 100을 기준으로 환산한 값을 사용하였다 (예, OECD, 2003).

Table 3.1 Estimate, coefficient of variation and design effect by region

estimate (%) coefficient of variation (%) design effect	Region						
	Capital	Chung- chung	Honam	Daekyung	Dongnam	Gangwon	Total
Cooking in household (SQ7)*	89.13%	89.05%	89.03%	87.03%	93.92%	91.07%	89.75%
	2.00%	3.54%	2.43%	3.41%	2.31%	2.89%	1.19%
	2.8394	4.6759	2.6837	2.9536	4.6045	1.7159	3.7469
Online purchase (A4)	21.43%	10.41%	13.09%	6.61%	10.39%	18.94%	15.84%
	9.08%	17.65%	17.40%	26.19%	16.85%	19.30%	6.66%
	1.9482	1.6652	2.5547	1.7782	1.8374	1.7485	2.5152
Eco-food purchase (A9)	46.64%	23.74%	36.07%	22.89%	30.44%	41.43%	37.64%
	5.85%	12.07%	9.16%	15.75%	14.75%	14.75%	4.16%
	2.5954	2.0869	2.6532	2.6959	3.4261	3.0921	3.1583
Eating functional foods (A13)	53.06%	44.48%	46.68%	40.99%	45.71%	42.20%	48.56%
	4.59%	7.54%	6.41%	9.70%	9.16%	11.81%	3.07%
	2.0593	2.0949	2.0132	2.3943	3.9416	2.0457	2.6815
Grocery purchasing frequency (A1)	87.22%	70.42%	69.91%	70.86%	82.55%	67.54%	80.24%
	1.76%	4.22%	4.34%	5.66%	2.76%	8.54%	1.32%
	1.8274	1.9508	2.4525	2.8505	2.0119	3.0481	2.1403
Rice purchasing frequency (B1)	69.11%	53.64%	53.33%	70.74%	76.89%	67.76%	66.95%
	3.19%	6.37%	5.95%	5.95%	3.35%	6.61%	1.98%
	1.9795	2.1591	2.2650	2.5429	2.0820	1.8467	2.4056
Vegetable purchasing frequency (C1)	81.54%	57.16%	67.07%	64.82%	81.48%	59.98%	74.70%
	2.77%	6.61%	4.78%	6.77%	3.24%	9.63%	1.86%
	2.9388	2.6835	2.6083	3.0917	2.5724	2.7967	3.0732
Fruit purchasing frequency (C5)	78.02%	68.43%	54.45%	68.10%	78.16%	68.43%	72.89%
	2.41%	5.06%	6.52%	5.43%	3.86%	8.29%	1.71%
	2.4066	2.9692	2.5325	2.4028	2.9301	4.0311	2.7087
Food expense change for this year (A16)	122.80	119.85	119.33	124.92	126.21	119.67	122.74
	1.53%	1.75%	2.33%	2.35%	2.28%	2.16%	0.92%
	2.0309	1.5304	2.4007	2.5042	3.4022	1.2096	2.6390
Basket price level for this year (A17)	131.02	124.03	127.14	126.45	126.50	127.54	128.49
	1.47%	1.81%	1.94%	2.25%	1.35%	3.11%	0.84%
	1.9064	2.6607	2.4900	2.0296	3.0666	2.1133	2.5942
Prediction of overall food consumption (A18)	120.64	113.79	128.25	128.12	123.58	118.72	121.98
	1.23%	1.81%	1.68%	1.77%	2.00%	3.05%	0.76%
	1.5688	2.4841	2.3072	1.9601	4.3791	2.5036	2.3737
Prediction of rice consumption (A18-1)	106.82	102.36	108.18	108.24	111.95	103.34	107.35
	1.31%	1.40%	1.86%	1.79%	2.21%	2.38%	0.80%
	1.7471	1.7806	2.0860	1.6811	5.1333	1.6462	2.5538
Prediction of meat consumption (A18-1)	108.90	102.69	111.99	111.40	116.77	104.12	109.95
	1.47%	1.85%	1.70%	2.05%	2.33%	2.42%	0.89%
	1.9251	2.1689	1.9635	1,8742	4.4817	1.4199	2.7181
Prediction of vegetable consumption (A18-1)	118.00	113.04	117.86	115.92	122.25	111.33	117.68
	1.21%	1.59%	1.76%	2.00%	2.08%	2.17%	0.76%
	1.7246	2.0381	2.3050	1.8307	4.5571	1.6844	2.4959
Prediction of fruit consumption (A18-1)	118.52	114.74	122.54	118.36	123.29	109.97	119.04
	1.38%	1.55%	1.54%	2.21%	2.20%	2.41%	0.83%
	2.0173	1.7764	1.9136	2.1146	4.1864	1.5397	2.7013
Prediction of dine-out spending (A18-1)	97.30	98.39	102.14	106.32	107.86	97.36	100.63
	1.94%	2.41%	2.97%	2.91%	2.63%	3.54%	1.15%
	1.8697	2.9066	3.3199	2.5417	4.7125	1.7911	2.8124

Note: * Question for main purchaser (Lee *et al.*, 2013)

특성여부 변수들에 대해서는 표본설계에서 예측한 설계효과에 비해 대부분 실측값이 높게 나타났다. 특성값이 $p = 0.5$와 거의 유사한 기능성식품 복용여부에 대한 상대표준오차의 예측값에 비해 추정치는 약간 높음을 확인할 수 있다. 그 외 0.5보다 작은 특성치를 갖는 인터넷 구매, 친환경식품구매는 예측값에 비해 높은 값을 갖는 반면, 0.5보다 큰 추정치를 갖는 가구내 취사, 식료품구매빈도 등은 낮은 값을 갖는다. 이러한 경향은 모비율 추정치의 상대표준편차 $cv(\widehat{p})$가 정의상 추정치의 단조감소 (monotone decreasing) 형태인 $[(1-\widehat{p})/\widehat{p}]^{0.5}$에 비례하기 때문인데, 특히 0에 가까운 모비율 추정치의 상대표준편차는 매우 크게 나타나게 된다

반면, 소비지출 및 예측 등과 같은 지수형 변수들은 소수의 경우를 제외하고는 설계효과 추정치의 값들이 예측치에 비해 높게 나타났다. 하지만, 상대표준오차의 경우는 0.76∼3.72%의 범위로 매우 안정적임을 볼 수 있다.

3.2. 설계효과모형 평가

본 절에서는 설계효과모형 (2.1)과 관련하여 2013 식품소비행태조사의 결과를 근거로 개별 설계요소 및 설계효과결과에 대해 논하고자 한다. Table 3.2는 표본층별 모집단 크기, (응답) 조사구 수, 가구 수와 비율, 가중치효과를 정리하고 있다. 2013 식품소비행태조사에서는 표본대체를 지양하고자 예측표본기법을 통해 차별적인 표본크기조정이 적용하였다. 조사구와 가구 응답률은 전국기준으로 각각 84.8%와 44.5%를 나타냈고, 총 응답조사구와 응답가구는 각각 515개와 3,018가구이었다. 상세한 조사수행의 평가는 Park (2014)를 참고할 수 있다.

무응답에 따른 실제 조사구당 응답가구 수는 목표값 5개보다는 조금 많은 5.86개였다. 따라서 불균등한 추출률과 무응답과 래이킹 비 등의 가중치 조정으로 인한 불균등 가중치의 사용은 전국기준으로 1.5016의 분산증가를 보여주고 있고 표본설계에서 가정한 1.2000보다 약 25.1%나 큰 것으로 나타났다. 층별로는 서울, 인천, 제주 등은 가정치 1.2000에 매우 근접한 값을 보였고, 울산과 경남은 약 2.0000의 값을 보였다.

표본층별 집락효과모형에 영향을 주는 식 (2.5)의 계수 b_h^*는 경남, 강원, 광주, 제주를 제외한 모든 층에서 5보다 작게 나타났다. 전국기준의 평균크기는 $\bar{b} = 4.61$로 조사구당 목표평균응답가구 5보다 약간 작은 영향력을 보여주고 있다.

Table 3.2는 표본틀을 기준으로 한 표본층의 상대규모 W_h는 물론이고 표본가중치를 사용한 상대규모 추정치 $\widehat{W}_h$를 나타내어 주고 있다. 대부분의 표본층은 전체의 10% 미만을 보이고 있지만 서울과 경기는 각각 20%을 넘는 높은 비중임을 알 수 있다. 더불어 층별 절충 (응답가구) 할당률 a_h을 함께 고려한 표본층별 기여율 $\delta_{sh} = \widehat{W}_h^2/a_h$은 서울과 경기의 경우에는 W_h에 비해 커진 반면, 다른 층은 오히려 감소하였음을 알 수 있다. 따라서 절충할당으로 인해 비례할당에 비해 규모가 큰 표본층의 설계효과가 전체의 설계효과에 더욱 많은 영향을 미칠 수 있음을 보여준다.

Table 3.3은 집락내상관계수에 대한 추정치를 표본층별로 정리하여 주고 있다. 추정치는 식 (2.6)으로 계산하였다. 특성여부에 대한 변수들은 표본설계에서 가정한 통계청 사회조사의 계수값 0.096보다는 전반적으로 높은 값을 보여주고 있다. 광주, 부산, 울산 지역을 제외하고는 모두 평균값을 기준으로 0.164∼0.362의 값을 갖는다. 지수형 변수들에 대해서 특·광역시와 경기도는 0.035∼1.55의 상대적으로 낮은 평균값을 갖는 반면, 제주도를 제외한 기타 도지역들은 0.181∼0.374의 비교적 높은 평균값을 나타내었다.

Table 3.2 Various features of sample design and weighting effect by stratum

Stratum	Sample design features and weighting effect								
	W_h	$\widetilde{W}_h$	n_h	b_h	$\bar{b}_h$	a_h	b_h^*	$\widetilde{W}_h^2/a_h$	$mdeff_{wh}$
Seoul	0.204	0.196	62	343	5.53	0.114	4.74	0.3379	1.2600
Incheon	0.053	0.054	34	169	4.97	0.056	4.37	0.0520	1.2137
Gyunggi	0.223	0.226	56	356	6.36	0.118	4.82	0.4348	1.3929
Daejeon	0.030	0.031	28	141	5.04	0.047	4.35	0.0208	1.2646
Chungbuk	0.032	0.033	29	150	5.17	0.050	3.76	0.0214	1.5077
Chungnam	0.043	0.045	34	170	5.00	0.056	3.88	0.0355	1.4619
Gwangju	0.031	0.030	24	143	5.96	0.047	5.02	0.0189	1.3398
Jeonbuk	0.038	0.038	30	168	5.60	0.056	3.85	0.0255	1.6191
Jeonnam	0.037	0.039	29	160	5.52	0.053	4.03	0.0281	1.4965
Jeju	0.011	0.011	15	90	6.00	0.030	5.03	0.0039	1.2717
Daegu	0.050	0.049	30	157	5.23	0.052	3.78	0.0469	1.5728
Gyungbuk	0.058	0.058	40	210	5.25	0.070	3.69	0.0479	1.5845
Busan	0.071	0.070	28	216	7.71	0.072	7.03	0.0694	1.4672
Ulsan	0.022	0.022	17	133	7.82	0.044	4.94	0.0106	1.9966
Gyungnam	0.066	0.066	25	210	8.40	0.070	5.38	0.0634	2.0344
Gwangwon	0.032	0.032	34	202	5.94	0.067	5.15	0.0157	1.5428
Total	1.000	1.000	515	3,018	5.86	1.000	4.61*	1.2327	1.5016*

Note: W_h and $\widetilde{W}_h$ denote, respectively, the actual and estimated relative sizes of stratum; n_h denote a number of responding enumeration district; b_h and b denote, respectively, numbers of responding households within domain h and for the entire sample; $a_h = b_h/b$ represents the average number of responding households within stratum; b_h^* is the quantity defined in (2.5); and $mdeff_{wh}$ is the weighting effect.

Table 3.3 Estimated intracluster correlation by stratum

Stratum	Categorical variable				Index variable			
	mean	1st quartile	median	3rd quartile	mean	1st quartile	median	3rd quartile
Seoul	0.165	0.082	0.188	0.203	0.098	0.080	0.105	0.134
Incheon	0.222	0.066	0.202	0.277	0.078	0.038	0.071	0.166
Gyunggi	0.164	0.072	0.165	0.218	0.059	0.046	0.065	0.101
Daejeon	0.212	0.129	0.180	0.257	0.090	0.022	0.103	0.141
Chungbuk	0.234	0.132	0.182	0.277	0.261	0.203	0.242	0.356
Chungnam	0.296	0.097	0.211	0.375	0.215	0.200	0.234	0.238
Gwangju	0.043	0.017	0.021	0.077	0.150	0.049	0.139	0.177
Jeonju	0.313	0.220	0.324	0.411	0.239	0.170	0.214	0.318
Jeonnam	0.175	0.122	0.191	0.216	0.199	0.115	0.131	0.198
Jeju	0.362	0.313	0.437	0.498	0.035	-0.029	0.064	0.081
Daegu	0.326	0.261	0.313	0.365	0.096	0.032	0.050	0.130
Gyungbuk	0.236	0.183	0.231	0.265	0.181	0.137	0.146	0.194
Busan	0.047	-0.021	0.011	0.065	0.155	0.111	0.123	0.195
Ulsan	0.059	0.006	0.039	0.073	0.144	0.077	0.137	0.213
Gyungnam	0.172	0.062	0.136	0.199	0.374	0.298	0.388	0.435
Gwangwon	0.181	0.071	0.215	0.237	0.031	0.000	0.022	0.039

Table 3.1에 나열된 조사 자료로부터 추정한 층별 가중치의 상대분산 cv_{wh}^2와 집락내상관계수 ρ_{yh}의 값들이 표본설계시 가정하였던 값들, 즉 $cv_{wh}^2 \equiv 20\%$와 $\rho_{yh} = 0.096$과는 매우 차이가 있음을 확인할 수 있었다. 따라서 향후 동일한 구조의 표본설계를 고려한다면 Table 3.2와 Table 3.3의 결과를 참고하여 좀 더 현실적인 값과 유사한 설계효과를 예측할 수 있을 것이다. 더불어 b_h^*는 설계효과에 대한 예측값과 비교할 때 큰 차이를 보여주고 있지 않다.

4. 논의

설계효과는 단순확률추출과 비교하여 층화, 집락, 가중치 등의 복잡한 설계요소들이 결합적으로 영향을 주는 효율성을 분산변동으로 평가하는 표준화된 측도이다. 설계효과의 개념과 개별적 설계요소에 대

한 설계효과모형을 통해 조사비용 및 오류의 측면에서 최적이 될 수 있는 표본설계를 선택할 수 있게 된다.

Gabler 등 (2006)의 설계효과모형 (2.1)은 상이한 집락들로 이루어진 분석영역과 불균등 가중치로 구성되는 복합설계의 효율성을 평가할 수 있는 측도이다. Lee (2012)의 지적과 같이 이는 층화설계에 대해서도 적용될 수 있다. 다만 2013 식품소비행태조사의 결과 (Table 3.1)에서 확인할 수 있듯이 설계효과모형 (2.1)이 가정하는 층분산 σ_{yh}^2의 동일성이 실제 자료에서는 만족하지 않을 가능성이 있다. 따라서 층분산이 불균등한 상황에서 설계효과모형 (2.1)이 얼마나 강건할 수 있는지와 이러한 상황에 보다 적합한 설계효과모형에 대한 연구가 필요하다고 할 수 있다. 더불어 Gabler 등 (2006)의 식 (2.1)은 모형에 근거 (model-based)한 설계효과에 대한 모형이므로 설계에 근거 (design-based)한 설계효과 추정량이 얼마나 유사한 지에 대해서도 추가적인 연구가 필요할 것이다.

Lee (2012)는 불균등가중치가 갖는 평균추정량 $\bar{y}_p$의 분산에 대해 갖는 가중치효과모형 (2.3)이 가중치와 측정변수간의 상관관계가 존재하지 않는 한 부정확한 크기측도 (measure of size; MOS) 무응답조정, 사후층화 등과 같은 가중치 조정들의 영향을 반영할 수 있다고 하였다. 표본가중치는 불균등한 추출확률, 일련의 가중치 조정 등을 통해 개별 응답개체가 대변하는 모집단의 개체수를 나타내므로 일종의 대표성 개념의 확률로 고려할 수 있을 것으로 보인다. 하지만 Gabler 등 (2006, 120쪽)은 무응답조정에 대한 적절성이 추가로 검증되어야 할 것으로 언급하고 있어 이에 대한 추가적인 고찰도 필요할 것으로 판단된다.

References

Bankier, M. D. (1988). Power allocation: Determining sample sizes for subnational areas. *The American Statistician*, **42**, 174-177.

Gabler, S., Hader, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, **25**, 105-106.

Gabler, S., Hader, S. and Lynn, P. (2006). Design effects for multiple design samples. *Survey Methodology*, **32**, 115-120.

Heeringa, S. G., West, B. T. and Berglund, P. A. (2010). *Applied survey data analysis*, Chapman & Hall/CRC, Boca Raton.

Kalton, G., Brick, M. and Le, T. (2005). Estimating components of design effect for use in sample design. In *The Household Sample Surveys in Developing and Transition Countries*, United Nations, New York.

Kim, H. and Kwak, H. (2013). A sampling design for e-learning industry status survey on the business demand sector. *Journal of the Korean Data & Information Science Society*, **24**, 701-712.

Kish, L. (1965). *Survey sampling*, John Wiley & Son, New York.

Kish, L. (1987). Weighting in Deft2. *The Survey Statistician*, June, 1987.

Lee, H. (2012). How should one find out the contributions to the design effect (variance) made by each of the design components (stratification, clustering, weighting) of a complex sample design? *Survey Statistician*, **66**, 16-20.

Lee, K. I., Hwang, Y. J., Kim, D. W., Ban, H. J. and Park, I. (2013). 2013 *consumer behavior survey for food*, Korea Rural Economic Institute, Republic of Korea.

Lynn, P. and Gabler, S. (2005). Approximations to b^* in the prediction of design effects due to clustering. *Survey Methodology*, **31**, 101-104.

Lynn, P., Gabler, S., Hader, S. and Laaksonen, S. (2007). Methods for achieving equivalence of samples in cross-national surveys. *Journal of Official Statistics*, **27**, 107-124.

OECD. (2003). *Business tendency surveys*, A handbook, Paris.

Park, I. (2014). Performance analysis of household survey : 2013 consumer behavior survey for food. Under revision.

Park, I., Byun, J. S. and Im, C. S. (2013). *Sampling design and weighting for the consumer behavior survey for food*, The Korean Statistical Society, Republic of Korea.

Park, I. and Lee, H. (2004). Design effects for the weighted mean and total estimator under complex survey sampling. *Survey Methodology*, **30**, 183-193.

Rust, K. and Broene, P. (2010). Design effects for totals in multi-stage samples. *Proceedings of the American Statistical Association, Section of Survey Research Methods*, 2174-2181.

Journal of the Korean Data & Information Science Society
2014, **25**(3), 523–531

http://dx.doi.org/10.7465/jkdi.2014.25.3.523
한국데이터정보과학회지

A study on design effect models for complex sample survey[†]

Inho Park[1]

[1]Department of Statistics, Pukyong National University

Received 5 March 2014, revised 4 April 2014, accepted 18 April 2014

Abstract

Design effect is often used in designing and planning sample surveys and/or in evaluating the efficiency of complex design features of the surveys. In this study, we applied Gabler *et al*. (2006)'s design effect model to 2013 Consumer behavior survey for food that was carried out by stratified two-stage sampling. Usability and adequacy of the design model to a real survey data are discussed and evaluated.

Keywords: 2013 Consumer behavior survey for food, clustering, intraclass correlation coefficient, stratification, weighting.

[†] This research was supported by a Research Grant of Pukyong National University (2013 Year).
[1] Assistant Professor, Department of Statistics, Pukyong National University, Busan 608-737, Korea. E-mail: ipark@pknu.ac.kr