응용통계연구 제 18권 2호, 2005년, pp. 311-328

농어가경제조사에서 가중핫덱 무응답 대체법의 활용

김규성[1] 이기재[2] 김진[3]

요 약

본 논문은 농어가경제조사에서 발생하는 무응답을 처리하는 방법에 관한 것이다. 농어가경제조사는 모두 층화다단표집을 한 후 가중평균으로 모평균을 추정하므로 이에 적합한 대체법으로 가중핫덱 대체법을 고려하여 가중핫덱 대체 절차와 모평균 추정법, 그리고 대응되는 분산추정법을 고찰하였다. 그리고 모의실험을 통하여 가중핫덱 대체가 두 조사에 적용될 수 있음을 보였고 수정된 잭나이프 분산추정법을 사용하면 추정치의 신뢰도도 효과적으로 나타낼 수 있음을 보였다. 또한 두 조사에 적용할 수 있는 대체군 형성 절차를 제시하고, 예로써 각각 4가지 방안을 비교, 분석하였다. 그리고 그 중 가장 효율적인 방안을 결과로써 제시하였다.

주요용어: 대체군, 의사결정 나무모형, 잭나이프 분산추정

1. 서론

통계청에서 실시하는 농어가경제조사는 우리나라 농어가경제 및 농어업 경영실태를 파악하기 위한 조사로 전국의 농어가를 대상으로 하는 표본조사이다. 농가경제조사는 1953년부터, 어가경제조사는 1963년부터 실시해온 계속조사로써 농어가에서 발생하는 수입, 지출, 그리고 재산과 부채의 변동 사항을 조사한다. 이를 위하여 일계부 조사에서는 수입, 지출 상황을 매일 기록하여 월 단위로 집계하며, 원부 조사에서는 연중의 재산 변동상황을 연초와 연말에 조사하여 그 차이를 집계하고 있다. 연간 통계는 일계부 집계와 원부 집계가 결합되어 작성된다(예를 들면 김규성 1998; 통계청 2003a; 통계청 2003b).

표본농어가는 표본설계를 통하여 전국의 농어가를 대표할 수 있도록 선정되며 자연교체가 이루어질 때까지 표본농어가로서의 임무를 맡게 된다. 그러나 표본농어가의 사정에 의하여, 혹은 표본농어가로서의 임무에 대한 피로감 때문에 표본에서 탈퇴를 하는 경우가 수시로 발생한다. 즉, 표본설계 시에는 농어가였으나 전업 등으로 일반 가구로 전환되는 경우, 이사 등으로 타 지역으로 이동하는 경우, 혹은 응답의 피로감으로 인하여 응답을 거절하는 경우가 발생하게 된다. 이러한 경우 조사의 연속성 및 조사 결과의 신뢰도 유지를 위하여 표본에서 탈퇴하는 농어가는 다른 농어가로 교체된다.

1) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 통계학과, 부교수.
E-mail: kskim@uos.ac.kr
2) (110-791) 서울특별시 종로구 동숭동 169, 한국방송통신대학교 정보통계학과, 교수.
E-mail: kjlee@knou.ac.kr
3) (302-701) 대전광역시 서구 둔산동 920, 정부대전청사, 통계청 조사관리과, 사무관.
E-mail: jink@nso.go.kr

표본조사에서 모집단 및 표본의 상황 변동으로 인한 표본교체 및 무응답 발생은 불가피한 것으로 보인다. 원천적으로 표본교체 및 무응답 발생을 방지할 수는 없기 때문에 표본교체 및 무응답 발생을 최소한으로 하는 것이 최선일 것이다. 그러나 일단 이루어진 교체 및 무응답에 대해서는 사후적인 통계 처리를 하는 것이 일반적인 추세이고, 이에 대한 여러 가지 기법들이 최근에 많이 개발되고 있다. 본 연구는 농어가경제조사에서 발생하는 무응답을 처리하는 방법에 관한 것이다. 무응답이 발생한 상황에서 무응답 발생 이유를 검토하고 적절한 무응답 처리를 하여 월평균 혹은 연평균 추정치의 신뢰도를 유지하는 방법을 찾고자한다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 농어가경제조사의 표본설계와 무응답 특성을 알아보고 3절에서는 농어가경제조사에서 활용하기에 적합한 가중핫덱 대체법과 이와 연관된 잭나이프 분산추정법을 살펴본다. 또한 모의실험을 통하여 본 연구에서 사용하는 대체법이 모평균 추정 및 분산 추정에 적용할 수 있는지 살펴본다. 4절에서는 무응답 대체의 효율성을 더 높이기 위하여 무응답 대체군을 형성하는 방안을 제시하고, 마지막으로 5절에서는 본 연구의 요약과 결론을 덧붙인다.

2. 농어가경제조사

2.1. 표본설계

2002년에 재설계된 농가경제조사 표본설계는 각 도를 부차모집단으로 하고 각 도에서 층화2단집락표집을 하고 있다. 이때 1차 추출단위로는 농업조사구나 부락이 선정되었으며 2차 추출단위로는 농가가 사용되었다. 또한 층내에서는 동일한 추출확률을 사용하여 동일층 내에서의 가중치는 동일하다. 결과적으로 각 도별로 10개의 층이 만들어지고 320개의 조사구(12개 주산지 포함)가 선정되었으며 각 조사구에서 10 농가씩 총 3,200 농가가 표본으로 선정되었다(한국통계학회 2002a). 어가경제조사 표본설계도 2002년에 재설계되었으며, 전국을 대상으로 층화2단집락표집과 층내에서 동일한 추출확률이 사용되었다. 최종적인 층의 수는 20개이며 276조사구에서 1,175 어가가 표본으로 선정되었다(한국통계학회 2002b).

두 표본설계가 제안하는 모평균 추정량은 가중치를 이용한 가중평균 형태이다.

$$\bar{y}_w = \frac{\sum_{hik \in s} w_{hik} y_{hik}}{\sum_{hik \in s} w_{hik}} \tag{2.1}$$

여기서 w_{hik}와 y_{hik}는 h번째 층의 i번째 조사구의 k번째 표본가구에 대한 가중값과 관심변수 값이며, s는 표본을 나타낸다. 두 표본설계 모두 모평균 추정량의 분산추정은 선형화 방법에 의한 설계기반 분산추정법을 사용하는데, 유한모집단 수정을 위하여 각 층에서 구한 표본추출률 $f_h = n_h/N_h$를 추가한 추정공식을 사용하고 있다.

$$v(\bar{y}_w) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi\cdot} - \bar{e}_{h\cdot\cdot})^2 \tag{2.2}$$

여기서 $e_{hi\cdot} = \sum_k w_{hik}(y_{hik} - \bar{y}_w)/\sum_{hik\in s} w_{hik}$, $\bar{e}_{h\cdot\cdot} = \sum_{i=1}^{n_h} e_{hi\cdot}/n_h$, 그리고 n_h는 h층의 조사구수이다. 이 공식은 통계패키지 SAS에 구현되어 있으므로 SAS를 이용하여 손쉽게 사용할 수 있는 장점이 있다.

2.2. 표본농어가의 무응답 특성

2003년의 3,200개 표본농가 중 무응답이 발생한 농가는 27농가였다. 농가경제조사에서 발생한 무응답 패턴의 첫 번째 특징은 무응답이 표본을 교체하는 과정에서 주로 발생한다는 것이다. 2003년의 경우는 모두 기존의 표본농가가 표본에서 이탈하여 새로운 농가를 표본으로 섭외하는 과정에서 발생하였다. 두 번째 특징은 무응답 지속 기간이 길지 않다는 점이다. 27농가 중 26농가는 무응답 기간이 1달이며, 1농가만이 2달로 나타났다. 표본농가의 이탈이 발생하면 거의 대부분이 1달 이내에 새로운 농가로 교체가 되어 확정됨을 보여준다.

전체 1,175 표본어가 중 무응답이 발생한 어가는 19어가였다. 이 중 1어가에서만 계속되는 조사 중간에 무응답이 발생했을 뿐, 나머지 18어가에서는 모두 표본어가를 교체하는 과정에서 무응답이 발생하였다. 그리고 무응답이 지속되는 개월 수는 14어가는 1개월, 3어가는 2개월, 2어가는 4개월이었다. 표본어가의 이탈이 발생하면 대부분 1달 이내에 새로운 어가로 교체가 되지만 25% 정도는 2달 이상이 걸렸다.

농어가경제조사의 무응답 패턴은 모두 단위무응답(unit nonresponse) 형태이다. 따라서 통상적인 접근방식에 의하면 가중치 조정을 하는 것이 타당해 보인다. 그러나 농어가경제조사의 내용을 들여다보면 가중치 조정보다는 대체를 하는 것이 더 합리적인데 그 이유는 다음과 같다. (i) 일계부 조사는 12개월 계속되기 때문에 12개월 중 한 두 달에서 발생한 무응답은 시계열 과정에서 항목무응답으로 간주될 수 있다. (ii) 더 근본적인 이유는 추정과정에 있다. 농어가경제조사에는 수 십여 개의 항목이 있고, 소분류 항목이 더해져서 중분류 항목을 만들고, 중분류 항목이 더해져서 대분류 항목을 만든다. 또한 일부 항목은 일계부 데이터와 원부 데이터가 결합되어 집계가 된다. 따라서 월별로 무응답 횟수가 다르므로, 월별로 가중치 조정을 하게 되면 12개의 가중치 조정 계산 알고리즘이 있어야 하고, 연간 추정에는 또 하나의 계산 알고리즘이 있어야 한다. 현실적으로 사용하기 어려운 방법이다. 이러한 이유로 농어가경제조사에서는 가중치 조정보다는 무응답 대체가 더 합리적인 방법이라고 할 수 있다.

무응답 처리를 위해서는 무응답 속성을 파악하여 무응답 패턴이 무시가능한 형태인지, 혹은 무시하면 안되는지를 살펴보아야 한다. 농어가경제조사에서 주로 발생하는 무응답의 원인은 농어가의 비농어가화, 단독농어가화, 전출 등으로 나타났다. 따라서 농어가경제조사에서 발생하는 무응답의 패턴은 무시가능한 무응답으로 간주될 수 있다.

3. 가중핫덱 대체 및 분산추정

무응답 대체방법으로 보조정보가 없는 경우는 핫덱 대체가 널리 쓰이며 활용가능한 보조정보가 있는 경우는 비대체나 회귀대체가 유용하게 쓰인다. 통상적인 문헌에 나오는 대

체 방법은 대부분 단순임의표본을 전제로 했을 경우에 개발된 방법들이다. 그런데 농어가 경제조사의 경우는 서로 다른 추출확률로 인하여 표본 가구들이 서로 다른 가중치를 가지고 있으므로 가중치를 고려하는 대체방법을 선택해야 한다. 가중치를 고려하는 핫덱 대체 방법으로 가중핫덱 대체를 생각할 수 있다.

3.1. 가중핫덱대체

층화다단표본에 무응답이 발생한 경우를 고찰하자. 편의상 응답 메카니즘은 전 조사단위에서 무응답 확률이 동일한 경우를 고려하고 대체군은 1개인 경우를 고려하자. 표본 중에서 응답을 한 집단을 s_r로 표시하고 무응답이 발생한 집단을 s_m으로 표시하자. 그리고 y^*_{hij}를 대체된 데이터라고 하자. 그러면 대체 후 모평균 추정량은 다음과 같이 표현된다.

$$\bar{y}_I = \frac{\sum_{(hik)\in s_r} w_{hik}\, y_{hik} + \sum_{(hik)\in s_m} w_{hik}\, y^*_{hik}}{\sum_{hik\in s} w_{hik}} \tag{3.1}$$

위의 추정량의 성질은 대체 방법에 따라 다르다. 보조 변수가 없는 경우 쉽게 사용할 수 있는 대체 방법으로는 평균대체, 핫덱 대체 등을 고려할 수 있는데 평균대체는 사용하기는 편리하지만 대체후 표본분포에 심한 왜곡이 발생하여 바람직하지 않은 것으로 알려져 있고, 핫덱 대체는 대체후 표본분포를 유지하는 장점이 있지만 가중치가 있는 경우는 핫덱 대체 후 추정량이 편향이 되는 단점이 있다(Rao & Shao, 1992, p.816). 따라서 비편향성을 유지하면서 핫덱대체를 하려면 가중치를 이용하는 가중핫덱 대체를 해야 한다. 이 방법은 층화다단추출인 경우에 Rao & Shao (1992)에 의해서 제안 되었는데 그 절차는 다음과 같다.

- 절차1 : 응답군 s_r에서 무응답을 대체할 단위, (gjl)를 가중치 $w_{gjl}/\sum_{(hik)\in s_r} w_{hik}$에 비례하여 복원으로 뽑는다.
- 절차2 : 선정된 단위를 이용하여 무응답을 대체한다. $y^*_{hik} = y_{gjl}$.

가중 핫덱 방법으로 무응답을 대체한 후 만든 대체후 추정량 $\bar{y}_I$는 모평균에 대한 근사 비편향 추정량이 된다.

3.2. 잭나이프 분산추정

무응답 대체를 할 때에 많이 나타나는 오류는 모수 추정보다는 분산추정에서이다. 무응답을 대체하면 대체로 인한 변동이 추가되어 대체후 추정량 $\bar{y}_I$의 분산이 무응답이 없는 추정량의 분산보다 크기 때문에, 대체로 인한 분산의 증가분을 추정해야 모평균 추정치의 분산을 비편향 추정할 수 있다(예를 들면 Sarndal 1992).

가중핫덱으로 무응답을 대체한 후, 수정된 잭나이프 분산추정법을 알아보기 위하여 다음의 기호를 도입하자. g층에서 j번째 조사구를 제외한 통계량을 다음과 같이 나타내자.

$$\widehat{S}(gj) = \sum_{(hik)\in s_r,\ h\neq g} w_{hik} y_{hik} + \frac{n_g}{n_g - 1} \sum_{(gik)\in s_r,\ i\neq j} w_{gik} y_{gik} \tag{3.2}$$

$$\hat{T}(gj) = \sum_{(hik)\in s_r,\ h\neq g} w_{hik} + \frac{n_g}{n_g - 1} \sum_{(gik)\in s_r,\ i\neq j} w_{gik} \tag{3.3}$$

그리고 다음을 정의하자. $(hik) \in s_m$인 (hik)에 대하여,

$$z_{hik}^{*(gj)} = \begin{cases} y_{hik}^{*} + \dfrac{\hat{S}(gj)}{\hat{T}(gj)} - \dfrac{\hat{S}}{\hat{T}}, & (hi) \neq (gj) \\ y_{hik}^{*}, & (hi) = (gj) \end{cases} \tag{3.4}$$

위의 식을 이용하여 (gj)번째 조사구를 제외한 후 만든 잭나이프 반복 추정량은 다음과 같다.

$$\hat{Y}_I^a(gj) = \hat{S}(gj) + \sum_{(hik)\in s_m,\ h\neq g} w_{hik} z_{hik}^{*(gj)} + \frac{n_g}{n_g - 1} \sum_{(hik)\in s_m,\ i\neq j} w_{gik} z_{gik}^{*(gj)} \tag{3.5}$$

마지막으로 다음과 같은 수정된 잭나이프 분산추정량을 얻는다.

$$v_J(\hat{Y}_I) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}^a(gj) - \hat{Y}_I)^2 \tag{3.6}$$

위 추정량은 대체후 분산추정량 $Var(\hat{Y}_I)$의 근사 일치추정량이다 (Rao & Shao, 1992).

위에 식 (3.6)은 모총계 추정량의 분산추정량이다. 따라서 모평균 추정량의 분산 추정량을 구하기 위해서는 약간의 변형이 필요하다. 모집단 크기를 M으로 나타내자. 모평균을 추정하기 위해서는 모집단 크기 M을 추정하는 과정이 병행되어야 한다. 모집단 크기 M의 추정은 모총계 추정과 동일한 방법을 이용하되 $y_{hik} = 1$인 성질을 이용한다. 앞에서와 동일한 방법으로 $\hat{S}_M(gj)$, $\hat{T}_M(gj)$ 그리고 $z_{hik}^{*(gj)}$를 계산한 후에 (gj)번째 조사구를 제외한 후 만든 잭나이프 반복 추정량을 다음과 같이 얻는다.

$$\hat{M}_I^a(gj) = \sum_{(hik)\in s,\ h\neq g} w_{hik} + \frac{n_g}{n_g - 1} \sum_{(hik)\in s,\ i\neq j} w_{gik} \tag{3.7}$$

그러면 모평균에 대한 잭나이프 반복 추정량은

$$\bar{y}_I^a(gj) = \hat{Y}_I^a(gj) / \hat{M}_I^a(gj) \tag{3.8}$$

이며 수정된 잭나이프 분산추정량을 다음과 같다.

$$v_J(\bar{y}_I) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\bar{y}^a(gj) - \bar{y}_I)^2 \tag{3.9}$$

3.3. 모의실험

농어가경제조사의 주요 변수에 대하여 대체후 모평균 추정량과 잭나이프 분산추정량의 효율성을 알아보기 위하여 모의실험을 실시하였다. 모의실험에 이용된 데이터는 2003년 1월 농어가경제조사 표본데이터이며 계산의 편의를 위하여 단위를 10,000원으로 변환하였

다. 무응답 가구는 2, 4, 6, 8, 10개인 경우를 고려하였고 응답 메커니즘은 등확률 메커니즘을 고려하였다. 표본데이터로부터 등확률 메커니즘을 이용하여 무응답 가구를 생성한 후 표본에서 제외하여 응답 표본을 만들고 각각의 무응답 가구에 대해서는 가중핫덱 대체를 실시하여 완비데이터(complete data)를 생성하였다. 그리고 이러한 시행을 독립적으로 500번 반복하여 500개의 완비데이터를 생성하였다.

대체후 추정량의 성질을 알아보기 위하여 500개의 대체후 추정치를 구한 후 평균을 구하였다.

$$AVE = \frac{1}{K}\sum_{k=1}^{K} \bar{y}_I^{(k)}, \quad K = 500$$

여기서 $\bar{y}_I^{(k)}$는 k번째 반복에서 구한 추정치로 식 (3.1)의 형태를 갖는다. 분산추정량으로는 통상적인 잭나이프 분산추정량(naive Jackknife variance estimator), v_{naive}과 식 (3.9)의 수정된 잭나이프 분산추정량, v_J을 구하였다. v_{naive}은 대체된 무응답을 조사값으로 간주하고 식 (2.2)의 공식으로 계산한 값이다. 두 분산추정량의 효율은 대체후 추정량 $\bar{y}_I$의 분산과 비교하여 평가할 수 있다. 이를 위하여 500개의 분산추정치를 구하고 대체후 추정량의 분산과 비교하여 상대편향과 상대안정성을 계산하였다.

- 상대편향 : $$RB(\%) = \sum_{k=1}^{K} \frac{(v_k - V(\bar{y}_I))/K}{V(\bar{y}_I)} \times 100$$
- 상대안정성 : $$RS(\%) = \sum_{k=1}^{K} \frac{\sqrt{(v_k - V(\bar{y}_I))^2/K}}{V(\bar{y}_I)} \times 100$$

여기서 v_k는 k번째 표본으로 구한 분산추정치이며, $V(\bar{y}_I)$는 대체후 추정량 $\bar{y}_I$의 분산이다. 상대편향이 음수이면 분산추정치가 분산을 과소평가하고 있다는 의미이며 양수이면 과대평가하고 있다는 의미이다. 또한 상대안정성은 작을수록 분산추정량이 분산에 가깝다는 뜻이 된다.

표 3.1 - 표 3.2는 농가경제조사의 주요변수인 농가소득과 농업소득에 대한 모의실험 결과이고, 표 3.3 - 표 3.4는 어가경제조사의 어가소득과 어업소득에 대한 결과이다.

표 3.1: 잭나이프 분산추정량의 상대편향과 상대안정성 : 농가소득 (단위 : 만원)

무응답 가구수	AVE	평균		$V(\bar{y}_I)$	상대편향(%)		상대안정성(%)	
		v_{naive}	v_J		RB_{naive}	RB_J	RB_{naive}	RB_J
2	190.76	119.37	119.67	119.43	-0.0444	0.2070	4.6845	5.7970
4	190.75	119.17	119.62	119.48	-0.2583	0.1156	6.5151	7.2131
6	190.72	119.06	119.66	119.56	-0.4172	0.0838	9.5855	9.5552
8	190.77	119.13	119.89	119.74	-0.5120	0.1198	14.0814	12.0143
10	190.75	118.87	119.77	119.74	-0.7280	0.0261	15.7698	13.0780

표 3.2: 잭나이프 분산추정량의 상대편향과 상대안정성 : 농업소득 (단위 : 만원)

무응답 가구수	AVE	평균		$V(\bar{y}_I)$	상대편향(%)		상대안정성(%)	
		v_{naive}	v_J		RB_{naive}	RB_J	RB_{naive}	RB_J
2	66.84	89.23	89.45	89.39	-0.1795	0.0742	4.9277	5.5075
4	66.87	89.15	89.49	89.47	-0.3644	0.0167	8.9158	7.7803
6	66.82	89.06	89.52	89.53	-0.5274	-0.0193	11.9367	9.8396
8	66.83	88.99	89.56	89.66	-0.7482	-0.1096	15.8874	12.1535
10	66.79	89.06	89.75	89.82	-0.8389	-0.0758	22.6150	16.3371

표 3.3: 잭나이프 분산추정량의 상대편향과 상대안정성 : 어가소득 (단위 : 만원)

무응답 가구수	AVE	평균		$V(\bar{y}_I)$	상대편향(%)		상대안정성(%)	
		v_{naive}	v_J		RB_{naive}	RB_J	RB_{naive}	RB_J
2	193.56	179.32	179.94	179.61	-0.1629	0.1823	23.3670	18.3966
4	193.52	178.41	179.62	179.50	-0.6122	0.0665	19.1682	13.9881
6	193.54	178.44	180.24	179.94	-0.8330	0.1679	26.5608	20.0207
8	193.51	178.43	180.85	181.10	-1.4743	-0.1403	46.7094	28.4109
10	193.48	177.96	180.99	181.19	-1.7845	0.1119	48.5796	31.7441

표 3.4: 잭나이프 분산추정량의 상대편향과 상대안정성 : 어업소득 (단위 : 만원)

무응답 가구수	AVE	평균		$V(\bar{y}_I)$	상대편향(%)		상대안정성(%)	
		v_{naive}	v_J		RB_{naive}	RB_J	RB_{naive}	RB_J
2	105.24	138.84	139.32	139.37	-0.3744	-0.0329	11.7437	9.2362
4	105.23	138.83	139.79	139.98	-0.8215	-0.1386	30.1480	20.9044
6	105.19	138.55	139.99	140.01	-1.0467	-0.0130	30.5357	21.9496
8	105.19	138.64	140.55	140.63	-1.4161	-0.0556	49.0301	32.2641
10	105.17	138.78	141.16	141.02	-1.5870	0.1011	61.0757	39.5155

모의실험을 통하여 다음과 같은 사항을 알 수 있다. 첫째, 가중핫덱 대체 후 계산한 가중평균은 모평균에 근사적으로 비편향이다. 농어가 표본 모두 무응답수가 늘어도 추정치의 평균인 AVE는 크게 바뀌지 않는다. 즉, 모평균에 근사적으로 비편향이라고 할 수 있다.

둘째, 통상적인 잭나이프 분산추정량(v_{naive})은 무응답 가구수가 커지면 음의 방향으로 상대편향이 증가하지만 수정된 잭나이프 분산추정량(v_J)은 상대편향이 증가하는 경향이 있다고 보기 어렵다. 수정된 잭나이프 분산추정량은 근사적으로 비편향이라는 경험적인 증거이며, 통상적인 잭나이프 분산추정량은 무응답의 수가 많아질수록 분산을 과소평가하는 경향이 증가한다는 경험적인 증거이다. 셋째, 무응답 가구수가 증가하면 두 분산추정량 모두 상대안정성의 수치가 커진다. 그러나 수정된 잭나이프 분산추정량의 수치가 통상적인 잭나이프 분산추정량의 수치보다 대체로 작게 나타난다. 이와 같은 사실을 종합하면 가중핫덱 대체방법은 농어가경제조사의 무응답 대체에 적합하며, 수정된 잭나이프 분산추정량은 대체후 추정량의 신뢰도를 효과적으로 표현한다는 사실을 알 수 있다.

4. 무응답 대체군 형성

4.1. 무응답 대체군 형성

앞 절의 결과는 무응답 대체군이 하나인 경우이다. 그런데 농가경제 표본수는 3,200개, 어가경제 표본수는 1,175개 이므로 표본전체를 하나의 대체군으로 간주하기에는 다소 큰 규모로 보인다. 따라서 표본을 몇 개의 대체군으로 나누어 무응답 처리에 사용하면 추정의 효율을 더 높일 수 있을 것으로 기대된다. 이 절에서는 농어가경제조사에서 여러개의 무응답 대체군을 만드는 과정을 소개하기로 한다.

무응답 대체군은 무응답으로 인한 무응답 편향의 효과를 줄이고, 대체후 추정량의 효율을 높일 수 있도록 만들어져야 한다. 이를 위해서는 응답확률이 비슷하고 또한 관심변수의 값이 비슷한 단위들로 대체군을 구성하고, 대체군 간에는 응답확률은 물론 관심변수의 값들도 차이가 나도록 하는 것이 좋다(예를 들면 Oh & Scheuren 1983).

대체군을 만드는 방법은 보조정보를 이용하여 (i) 주요변수들이 대체군 내에서 동질적이도록 만들거나, (ii) 대체군 내의 응답률이 비슷하도록 대체군을 만드는 방법을 생각할 수 있다. 혹은 두 방법을 절충할 수도 있다.

단위무응답의 처리를 위한 무응답층 구성과 관련된 기존 연구로는 Rizzo 외 2인(1996), Dufour 외 4인(2001), 김재광 외 2인(2004) 등이 있다. Rizzo 외 2인(1996)은 SLID(Survey of Labour and Income Dynamics) 자료에 분석을 바탕으로 무응답층 구성을 위한 방법론의 선택보다는 무응답층 구성에 사용되는 보조변수들의 선택이 더 중요할 것이라고 주장하였다. Dufour 외 4인(2001)은 단위무응답의 처리를 위한 무응답층 구성방법으로 로지스틱 회귀모형 방법과 의사결정 나무 모형 등을 이용한 방법 등을 제시하고 모의실험을 통해서 두 방법을 비교하였다. 김재광 외 2인(2004)는 가계조사에서 발생하는 단위무응답을 처리하기 위하여 CHAID(Chi-Square Automatic Interaction Detection) 알고리즘을 이용하여 무응답층을 구성하였다.

본 연구에서 대체군을 만드는 중요한 목적은 단위무응답을 처리하기 위한 것이라기보다는 특정 월에 무응답인 농어가의 조사자료를 대체군 내의 조사결과로 대체함으로써 완비데이터를 얻는 데 있다. 농어가경제조사에서 무응답 대체를 위한 대체군 형성을 위해서는 대체군 형성에 이용될 유용한 보조변수를 확보하는 일이 선행되어야 한다. 농가경제조

사의 경우 활용 가능한 보조변수로 지역, 영농형태, 전·겸업 구분, 논·밭 면적, 농업경영주 나이, 가구원수, 농업종사자수 등이 있고, 어가경제조사에는 지역, 어업형태, 전·겸업 구분, 어장면적, 어선톤수, 어업경영주 나이, 가구원수 등이 있다. 농어가경제조사의 경우 활용 가능한 보조변수가 대부분 범주형 변수이고 논·밭 면적 및 어장면적, 그리고 어선톤수 등은 범주화 할 수 있는 연속형 변수이므로 이들 변수만 적절하게 범주화 하면 첫 번째 방법은 사용 가능하다. 그러나 각 변수가 적게는 3개, 많게는 10개 이상의 수준을 갖고 있으므로 전체 변수의 수준 조합을 고려하기에는 그 수가 너무 많다. 따라서 전체 수준 조합의 수를 줄여서 적절한 수의 대체군을 만드는 일이 현실적으로는 중요하다. 두 번째 방법을 사용하기 위해서는 각 농어가의 무응답 확률을 추정해야 하는데, 농어가경제조사의 경우 무응답률이 높지 않아서 무응답률이 비슷하도록 묶는 대체군 형성 방법은 현실적으로 적용하기 어렵다. 따라서 위의 두 방법 중 첫 번째 방법만을 고려하기로 한다.

본 연구에서 사용한 무응답 대체군 구성 단계는 다음과 같다.

- 1단계 : 관심변수 선택

농어가경제조사에는 수십 개가 넘는 관심변수가 있다. 모든 변수를 대체군 형성에 활용하기는 현실적으로 어려우므로 가중 중요한 변수인 농(어)가소득, 농(어)업소득을 관심변수로 한다. 그리고 부차적으로 겸업소득과 사업외소득도 고려한다.

- 2단계 : 설명변수 선택

활용 가능한 설명변수 중 관심변수에 유의한 영향을 주는 주요 설명변수를 선택한다. 회귀분석을 통하여 수준별로 관심변수가 차이를 보이는 설명변수를 찾는다.

- 3단계 : 설명변수의 우선순위 결정

앞에서 찾은 설명변수에 우선순위를 부여한다. 우선순위 결정에는 회귀분석의 결과 뿐 아니라 설명변수간의 결합분포, 관심변수의 수준별 평균 등을 동시에 고려한다.

- 4단계 : 대체군 구성

앞에서 선정한 설명변수와 설명변수의 우선순위를 고려하여 대체군을 구성한다. 대체군 구성은 SAS Enterprise Miner에서 제공되는 의사결정 나무모형(decision tree model)을 이용한다. 의사결정나무모형을 직접적으로 적용해서 대체군을 구성하면 대체군 내의 표본이 너무 적게 배정되는 등 실제적으로 활용할 수 없는 경우가 발생한다. 이를 방지하기 위하여 주요 변수로 일차적으로 대체군을 형성하고, 이들 대체군 중에서 표본이 충분히 큰 경우에 대해서 부차적인 설명변수로 의사결정 나무모형을 적용하여 대체군을 세분하는 작업을 진행한다.

- 5단계 : 대체군 구성의 타당성 검토

주요 관심변수에 대하여 대체군 구성이 타당한지를 검토한다. 농어가경제조사에서는 무응답률이 높지 않고, 각 대체군 내의 무응답률이 거의 비슷하기 때문에 대체군 구성의 타당성은 대체군 형성 전·후의 분산비를 비교하여 검토해 볼 수 있다. 대체군 형성효과를 대체군 구성전 표본평균의 분산과 대체군 형성후 표본평균의 분산의 비로 표현하자. 즉,

$$\text{대체군 형성 효과} = 1 - \frac{\sum_{h=1}^{H} W_h S_h^2}{S^2} \tag{4.1}$$

여기서 W_h는 h군의 비율이며, S_h^2, S^2은 h군의 표본분산 및 전체 표본분산이다. 값이 클수록 대체군 형성효과가 있음을 의미한다.

대체군 형성의 타당성을 확인하는 또 다른 방법으로는 통상적인 방법으로 구한 분산 추정량으로 각 방안에 대해서 가중핫덱을 반영하여 구한 분산 추정량을 나누어 주고 이 값을 1에서 뺀 값을 이용하는 것이다. 이 값은 대체군 형성 후 핫덱대체로 인한 분산의 감소율을 의미하므로 이 값이 클수록 대체군 형성효과가 크다고 할 수 있을 것이다.

4.2. 농가경제조사를 위한 대체군

농가소득, 농업소득, 겸업소득, 사업외소득을 관심변수로 하여 회귀분석을 실시한 결과 영농형태, 전·겸업 구분, 지역, 경지면적, 농업종사자 수, 경영주의 연령 등이 중요한 설명변수로 나타났다. 표 4.1에 월별로 농가소득과 농업소득에 유의한 영향을 미치는 설명변수가 정리되어 있다.

표 4.1: 월별 농가소득과 농업소득에 유의한 영향을 주는 설명변수

구분	농가소득	농업소득
1월 조사	지역‡, 농업종사자수†, 경영주 연령†	지역‡, 영농형태*, 전·겸업*, 농업종사자수†
2월 조사	농업종사자수‡	농업종사자수†, 경지면적†
3월 조사	–	영농형태†
4월 조사	영농형태‡, 종사자수‡, 연령‡, 경지면적‡	지역*, 영농형태‡, 전·겸업*, 연령‡, 경지면적*
5월 조사	농업종사자수†	지역‡, 영농형태†
6월 조사	영농형태‡, 전·겸업*, 종사자수†, 연령‡	지역‡, 영농형태‡, 종사자수†, 연령†
7월 조사	연령‡	연령‡
8월 조사	지역‡, 영농형태‡, 종사자수*, 연령†	지역‡, 영농형태‡, 연령*
9월 조사	지역‡, 영농형태‡, 경지면적‡	지역‡, 영농형태‡, 경지면적‡
10월 조사	지역‡, 영농형태‡, 종사자수‡, 연령†, 경지면적‡	지역‡, 영농형태‡, 종사자수‡, 경지면적‡
11월 조사	영농형태‡, 종사자수†, 연령†, 경지면적‡	영농형태‡, 종사자수†, 경지면적‡
12월 조사	지역‡, 영농형태*, 경지면적‡	지역‡, 경지면적*
연단위 데이터	지역†, 영농형태‡, 전·겸업†, 종사자수‡, 연령‡, 경지면적‡	영농형태‡, 종사자수‡, 연령‡, 경지면적‡

주) *: 유의수준 10% 이하에서 통계적으로 유의한 것을 표시함.
†: 유의수준 5% 이하에서 통계적으로 유의한 것을 표시함.
‡: 유의수준 1% 이하에서 통계적으로 유의한 것을 표시함.

위의 주요 설명변수를 중심으로 대체군을 만들 때 다음의 두 가지 사항을 고려해야 한다. 첫째는 대체군의 수이다. 대체군의 수가 많으면 무응답 편향을 줄일 수 있는 장점이 있

는 반면, 표본이 외부 환경 변화에 민감하게 반응하여 이로 인한 불확실성이 증가하는 단점이 있다. 반면, 대체군의 수가 너무 적으면 대체군 형성의 효과가 줄어드는 단점이 있다. 따라서 무응답 편향의 감소와 분산의 증가를 고려하여 대체군의 수를 정해야 한다. 두 번째는 대체군의 크기이다. 대체군이 너무 작거나 혹은 너무 크면 대체군의 효율성 및 안정성에 부정적인 효과를 준다. 따라서 대체군 크기의 범위를 정할 필요가 있다.

본 논문에서는 다음의 네 가지 대체군 구성 방안을 고려하였다.

- 방안1 : 영농형태, 전·겸업, 지역 구분을 고려해서 일차적으로 대체군을 구성하고, 이들 대체군 중 규모가 큰 대체군에 대해서 경지면적 변수에 의사결정나무모형을 적용하여 세분한다.
- 방안2 : 방안1과 유사하게 영농형태, 전·겸업, 지역 구분을 고려해서 일차적으로 대체군을 구성하되, 지역별 변동이 적은 논벼, 2종겸업 등의 영농형태에 대해서는 지역 구분을 하지 않고, 규모가 큰 대체군에 대해서 경지면적 변수에 의사결정나무모형을 적용하여 대체군을 세분한다.
- 방안3 : 방안2와 같이 일차적인 대체군을 구성하되, 특작, 전작, 기타 영농형태에 대해서는 이들을 병합한다. 그리고 규모가 큰 대체군은 경지면적 변수에 의사결정나무모형을 적용하여 대체군을 세분한다. 방안3은 영농형태 중 특작, 전작, 기타는 주로 밭작물로서 매년 변동되는 폭이 적다는 특징을 고려한 것이다.
- 방안4 : 방안3을 기초로 하되, 논벼(전업) 농가 중에 경지면적 12,000평 미만의 표본농가로 구성된 대체군과 2종 겸업농가를 위한 대체군 중에서 대체군 내의 표본수가 200농가 이상인 대규모 대체군을 유사한 특성을 지닌 지역으로 묶어서 세분한다.

위의 4가지 방법에 의하여 대체군을 형성한 결과, 방안1에서는 44개의 대체군, 방안2에서는 34개의 대체군, 방안3에서는 33개의 대체군, 방안4에서는 39개의 대체군이 만들어졌다.

4가지 대체군 구성 방안을 비교해 보기 위하여 식 (4.1)에서 정의한 무응답 대체군 형성 효과를 비교해 보았다 (표 4.2).

표 4.2: 농가경제조사에서 4가지 방안의 무응답 대체군 형성 효과 비교

구분	대체군 수	농가소득	농업소득	겸업소득	사업외 소득	농가부채
방안1	44	10.5%	17.5%	8.7%	37.9%	6.1%
방안2	34	15.8%	23.8%	9.4%	36.9%	12.4%
방안3	33	15.9%	24.0%	9.5%	36.9%	12.4%
방안4	39	16.3%	24.2%	9.4%	37.1%	12.5%

2003년도의 연간 데이터를 이용해서 분석할 때 방안4가 가장 효과적인 것으로 나타났다. 대체군의 수는 방안1이 가장 많지만 방안4보다 효과가 미미한 것은 대체군 구분이 관심변수에 큰 영향을 주지 못했기 때문이다. 4가지 방안에서 공통적으로 대체군 구분이 농업소득과 사업외 소득에 효과가 큰 것으로 나타난 반면 겸업소득과 농가부채는 대체군 효

과가 상대적으로 작게 나타나는데, 이는 대체군 구분에 사용한 주요 설명변수들이 농업소득과 사업외 소득에 밀접한 연관성이 있기 때문인 것으로 풀이된다. 만일 겸업소득 및 농가부채에 대체군 구성효과를 더 반영되려면 겸업소득 및 농가부채와 연관성이 높은 보조변수를 추가로 확보해야 할 것이다.

4.3. 어가경제조사를 위한 대체군

어가소득, 어업소득을 관심변수로 하여 회귀분석을 한 결과 어업형태, 전·겸업 구분, 지역, 어장면적, 어선톤수 등이 소득을 예측하는 데 중요한 설명변수인 것으로 나타났다. 아래의 표 4.3에 월별 주요 설명변수가 나타나 있다.

표 4.3: 월별 어가소득과 어업소득에 유의한 영향을 주는 설명변수

구분	어가소득	어업소득
1월 조사	어선톤수‡	어업형태*, 전·겸업‡, 어선톤수‡
2월 조사	어업형태*, 어장면적‡	어업형태‡, 전·겸업‡, 어장면적*
3월 조사	어업형태*, 어선톤수*	어업형태‡, 전·겸업‡
4월 조사	어선톤수‡, 어장면적*	전·겸업‡, 어선톤수‡, 어장면적*
5월 조사	어업형태*, 어선톤수*	전·겸업‡, 어선톤수*
6월 조사	-	전·겸업‡
7월 조사	지역‡, 어업형태*	지역‡, 전·겸업‡
8월 조사	지역*, 어업형태‡	지역†, 전·겸업‡
9월 조사	어업형태†	어업형태†
10월 조사	지역†, 어선톤수‡, 어장면적†	지역†, 전·겸업‡, 어선톤수‡, 어장면적†
11월 조사	어업형태*, 어선톤수‡, 어장면적‡	전·겸업‡, 어선톤수‡, 어장면적†
12월 조사	어선톤수‡	어업형태*, 전·겸업†, 어선톤수‡
연단위 데이터	어업형태‡, 어선톤수‡, 어장면적†	전·겸업‡, 어선톤수‡, 어장면적†

주) *: 유의수준 10% 이하에서 통계적으로 유의한 것을 표시함.
†: 유의수준 5% 이하에서 통계적으로 유의한 것을 표시함.
‡: 유의수준 1% 이하에서 통계적으로 유의한 것을 표시함.

어가경제조사를 위한 무응답 대체군은 농가경제조사에서와 유사한 방법으로 만든다. 본 논문에서는 4가지 방안을 고려하였다.

• 방안1 : 우선적으로 어업형태와 전·겸업 구분을 고려하여 무응답 대체군을 형성한다. 지역은 표본설계 상의 지역층(몇 개의 도를 묶어서 구성되었음)을 고려하여 유사한 특성을 나타내는 지역끼리 묶어서 무응답 대체군을 구성한다.

• 방안2 : 방안 1과 마찬가지로 어업형태와 전·겸업 구분을 우선적으로 고려하여 무응답 대체군을 형성하되, 지역은 행정구역상의 각 시·도 구분을 고려하여 유사한 특성을 나타내는 지역끼리 묶어서 무응답 대체군을 구성한다.
• 방안3 : 방안 1과 마찬가지로 어업형태와 전·겸업 구분을 우선적으로 고려하여 무응답 대체군을 형성하되, 동력선 사용어가와 양식업 어가에 대해서 선박톤수와 양식장 면적을 대체군 형성에 이용하고, 이후에 지역을 대체군 세분 과정에서 이용한다.
• 방안4 : 방안 3과 마찬가지로 어업형태와 전·겸업 구분을 우선적으로 고려하고, 지역을 다음으로 고려한다. 이후에 선박톤수와 양식장 면적을 무응답 대체군 세분하는 과정에서 이용한다.

4가지 방안으로 무응답 대체군을 형성한 결과, 방안1은 12개 대체군, 방안2는 20개 대체군, 방안3은 22개 대체군, 방안4는 23개의 대체군이 만들어졌다. 그리고 4가지 대체군 구성 방안에 대한 무응답 대체 효과를 계산한 결과가 표 4.4에 주어져 있다.

표 4.4: 어가경제조사에서 4가지 방안의 무응답 대체군 형성 효과 비교

구분	대체군 수	어가소득	어업소득	겸업소득	사업외 소득
방안1	12	1.6%	16.4%	9.3%	17.0%
방안2	20	2.3%	18.1%	14.4%	19.8%
방안3	22	6.1%	20.9%	14.9%	21.7%
방안4	23	3.7%	19.0%	14.6%	20.5%

무응답 대체군 형성 효과를 비교하면 방안3이 가장 효율적인 것으로 나타났다. 방안3과 방안4가 상대적으로 효과적인 이유는 설명변수로 선박톤수 및 어장면적을 활용했기 때문이다. 또한 방안2가 방안1 보다 효과적으로 나타났는데, 그 의미는 표본설계 상의 지역 구분 대신에 본 연구에서 제시한 지역 구분이 더 효과적이라는 것이다. 그리고 4가지 방안 모두 어업소득, 겸업소득과 사업외 소득 변수에 대해서는 대체군 형성 효과가 큰 반면, 어가소득에 대해서는 대체군 효과가 그리 크지 않은 것으로 나타났다. 어가소득은 어업소득, 겸업소득 그리고 사업외 소득의 합으로 계산되므로 각각의 소득은 대체군에 의해서 효과적으로 구분되지만 세 소득의 합인 어가소득은 대체군 내에 혼재한다는 의미로 해석된다.

앞서 제시한 무응답 대체군 형성 방안 중 가장 높은 효율을 보인 농가경제조사의 방안 4와 어가경제조사의 방안 3의 무응답 대체군 형성 결과를 부록1, 2에 수록하였다.

5. 결론

본 논문에서는 농어가경제조사에서 발생하는 무응답의 처리 방법에 관하여 알아보았다. 가중치가 있는 경우에 활용 가능한 가중핫덱 대체법을 고찰하고, 모의실험을 통하여 적용 가능성을 알아보았다. 모의실험을 통하여 얻은 경험적인 결과는 다음과 같은 점을 시사한다. 첫째, 가중 핫덱 대체법을 농어가경제조사에 적용 가능하다. 대체후 가중평균은 모평

균에 관하여 근사적 비편향성이 있음이 발견된다. 둘째, Rao & Shao가 제안한 수정된 잭나이프 분산추정법은 대체후 가중평균의 신뢰도를 나타내는데 적절하다. 근사적으로 비편향성이 있으며 통상적인 잭나이프 분산추정량에 비하여 상대적으로 안정적이다.

본 연구에서 실시한 모의실험과 유사한 결과로는 Kovar & Chen (1994)의 결과가 있다. 이들은 모의실험을 통하여 평균대체, 핫덱대체, 비대체, 최근방대체 등 4가지 대체법에 대하여 무응답률 5%와 30%에서 수정된 잭나이프 분산추정량이 효율적임을 보이고 있다. 본 연구의 결과와 유사한 결과이긴 하지만, 본 연구는 우리나라에서 실제 실시되고 있는 농어가경제조사에 적용했다는 점에서 실용적인 의의를 찾을 수 있다. 가중핫덱 대체법이 아닌 다른 방법으로 미국이나 캐나다에서 주로 사용하고 있는 최근방 대체법도 고려해 볼 수 있는 유력한 방법이다. 농어가경제조사에서 최근방 대체법에 대한 활용성 여부는 향 후 연구과제로 남겨둔다.

무응답률이 높아지면 대체군 형성의 필요성은 증가할 것이다. 농어가경제조사의 무응답률은 그리 높지 않기 때문에 대체군을 형성하지 않고도 좋은 모의실험 결과를 얻을 수 있었다. 그러나 대체군을 이용하면 더욱 높은 효율을 얻을 수 있을 것이며, 무응답률이 높아지면 대체군의 효과는 더 크게 나타날 것으로 예상된다. 또한 대체군 내의 무응답률도 다르게 나타날 것이므로 각 대체군에서 서로 다른 무응답률을 고려한 대체를 실시해야 할 것이다.

우리나라 농어가경제조사는 조사의 복잡성에 비하여 교체율 및 무응답률이 높은 편은 아니다. 조사에 대한 경험이 축적되어 조사 체계가 안정되고, 또한 현지 조사원이 적극적으로 조사에 임하기 때문일 것이다. 그러나 교체 및 무응답을 원천적으로 방지할 수는 없으므로 1차적으로는 교체 및 무응답률을 줄이는 노력을 해야 하며 2차적으로는 사후적인 무응답을 처리 방법을 활용해야 한다. 본 연구의 결과가 이러한 2차적인 무응답 처리에 효과적으로 활용되어 우리나라 농어가경제조사의 수준을 한층 높이는데 일조하기를 기대한다.

참고문헌

김규성(1998). 농가경제조사의 현황과 개선방향. <응용통계연구>, 11권, 1호, 29-39.

김재광, 한근식, 윤연옥(2004). 가계조사 무응답 처리기법 연구. <통계연구>, 9권, 1호, 79-102.

통계청(2003). <2002년 농가경제통계>.

통계청(2003). <2002년 어가경제통계>.

한국통계학회 (2002). <농가경제조사 표본설계>.

한국통계학회 (2002). <어가경제조사 표본설계>.

Dufour, J., Gagnon, F., Morin, Y., Renaud, M., Sarndal, C.E. (2001). A better understanding of weight transformation through a measure of change, *Survey Methodology*, **27**, 97-108.

Kovar, J.G. and Chen, E. (1994). Jackknife variance estimation of imputed survey data, *Survey Methodology*, **20**, 45-52.

Oh, H.L. and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse, *Incomplete data in sample surveys*, Chap 13, 143-184.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation under inputation for missing survey data, *Biometrika*, **79**, 811-822.

Rizzo, L., Kalton, G. and Brick, M. (1996). A comparison of some weighting adjustment methods for panel nonresponse, *Survey Methodology*, **22**, 43-53.

Sarndal, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used, *Survey Methodology*, **18**, 241-252.

[2005년 2월 접수, 2005년 4월 채택]

부록 1: 농가경제조사를 위한 무응답 대체군 구분 : 방안4

대체군	대체군 크기	영농 형태	전·겸업	지역 및 경지면적
1	127	논벼	전업	경지 5,000평 미만 & 경기, 강원, 충남
2	79	논벼	전업	경지 5,000평 미만 & 충북, 경북, 제주
3	201	논벼	전업	경지 5,000평 미만 & 전북, 전남, 경남
4	74	논벼	전업	경지 5,000평 이상 - 12,000평 미만 & 경기, 강원, 충남
5	109	논벼	전업	경지 5,000평 이상 - 12,000평 미만 & 충북, 전북, 경북, 제주
6	52	논벼	전업	경지 5,000평 이상 - 12,000평 미만 & 전남, 경남
7	51	논벼	전업	경지 12,000평 이상
8	108	논벼	1종겸업	경지 5,000평 미만
9	104	논벼	1종겸업	경지 5,000평 이상 - 12,000평 미만
10	52	논벼	1종겸업	경지 12,000평 이상
11	89	과수	전업	경기, 강원, 충북, 충남, 전북, 전남, 경남
12	55	과수	전업	경북
13	31	과수	전업	제주
14	54	과수	1종겸업	전국
15	29	채소	전업	경기
16	48	채소	전업	강원
17	61	채소	전업	충북
18	57	채소	전업	충남
19	21	채소	전업	전북
20	62	채소	전업	전남
21	91	채소	전업	경북
22	81	채소	전업	경남
23	25	채소	전업	제주
24	74	채소	1종겸업	경기, 강원, 충남, 경남, 제주
25	39	채소	1종겸업	충북, 전북, 전남, 경북
26	55	특작	전업	전국
27	27	특작	1종겸업	전국
28	53	화훼	전업, 1종겸업	전국
29	50	전작,기타	전업	전국
30	29	전작,기타	1종겸업	전국
31	111	축산	전업	경기, 강원, 충북, 충남, 전북, 제주
32	90	축산	전업	전남, 경북, 경남
33	61	축산	1종겸업	전국
34	194	2종겸업		경지 3,500평 미만 & 경기, 충남, 제주
35	99	2종겸업		경지 3,500평 이상 & 경기, 충남, 제주
36	137	2종겸업		경지 5,000평 미만 & 강원, 충북
37	125	2종겸업		경지 5,000평 미만 & 전북, 전남
38	134	2종겸업		경지 5,000평 미만 & 경북, 경남
39	46	2종겸업		경지 5,000평 이상 & 강원, 충북, 전북, 전남, 경북, 경남

부록 2: 어가경제조사를 위한 무응답 대체군 구분 : 방안3

대체군	대체군 크기	어업형태	전·겸업	지역, 선박톤수 및 양식장 면적
1	31	비어선 어가	전업, 1종겸업	전체
2	66	비어선 어가	2종겸업	전북, 전남, 제주
3	58	비어선 어가	2종겸업	전북, 전남, 제주 제외 나머지
4	111	무동력선, 동력선 사용 어가	전업어가	선박 톤수 < 5.0
5	63	무동력선, 동력선 사용 어가	전업어가	선박 톤수 ≥ 5.0
6	37	무동력선, 동력선 사용 어가	1종겸업	경기, 인천, 충남
7	52	무동력선, 동력선 사용 어가	1종겸업	강원, 경북 & 선박 톤수 < 2.5
8	36	무동력선, 동력선 사용 어가	1종겸업	강원, 경북 & 선박 톤수 ≤ 2.5
9	34	무동력선, 동력선 사용 어가	1종겸업	전북, 전남, 제주 & 선박 톤수 < 2.5
10	68	무동력선, 동력선 사용 어가	1종겸업	전북, 전남, 제주 & 선박 톤수 ≥ 2.5
11	50	무동력선, 동력선 사용 어가	1종겸업	부산, 울산, 경남 & 선박 톤수 < 2.5
12	41	무동력선, 동력선 사용 어가	1종겸업	부산, 울산, 경남 & 선박 톤수 ≥ 2.5
13	28	무동력선, 동력선 사용 어가	2종겸업	경기, 인천, 충남
14	45	무동력선, 동력선 사용 어가	2종겸업	강원, 경북
15	62	무동력선, 동력선 사용 어가	2종겸업	전북, 전남, 제주
16	60	무동력선, 동력선 사용 어가	2종겸업	부산, 울산, 경남
17	46	양식업	전업	전북, 전남, 제주
18	27	양식업	전업	전북, 전남, 제주를 제외한 나머지
19	71	양식업	1종겸업	전북, 전남, 제주
20	48	양식업	1종겸업	전북, 전남, 제주를 제외한 나머지
21	46	양식업	2종겸업	양식장 면적 ≤ 0.8
22	47	양식업	2종겸업	양식장 면적 > 0.8

Weighted Hot-Deck Imputation in Farm and Fishery Household Economy Surveys

Kyu-Seong Kim[1] Kee-Jae Lee[2] Jin Kim[3]

ABSTRACT

This paper deals with a treatment of nonresponse in farm and fishery household economy surveys in Korea. Since the samples in two surveys were selected by stratified multi-stage sampling and weighted sample means has been used to estimate the population means, we choose a weighted hot-deck imputation method as an appropriate method for two surveys. We investigate the procedure of the weighted hot-deck as well as an adjusted jackknife method for variance estimation. Through an empirical study we found that the method worked very well in both mean and variance estimation in two surveys. In addition, we presented a procedure of forming imputation class and formed four imputation classes for each survey and then compared them with analysis. As a result, we presented two most efficient imputation classes for two surveys.

Keywords: Decision tree model, Imputation class, Jackknife variance estimation.

1) Associate Professor, Department of Statistics, University of Seoul, 90 Jeonnong-Dong Dongdaemun-Gu, Seoul, 130-743, Korea
E-mail : kskim@uos.ac.kr
2) Professor, Department of Information Statistics, Korea National Open University, 169 Dongsung-Dong Chongro-Gu, Seoul, 110-791, Korea
E-mail : kjlee@knou.ac.kr
3) Field Management & Sampling Devision, Korea National Statistical Office, Government Complex Daejeon, 139 Seonsaro(920 Dunsan 2-dong), Seo-Gu, Daejeon, 302-701, Korea
E-mail : jink@nso.go.kr