

## 농가경제조사의 무응답 대체군 형성 방안

이기재<sup>1)</sup>, 김규성<sup>2)</sup>, 김진<sup>3)</sup>

### 요 약

본 연구에서는 표본 농가의 교체나 무응답으로 인한 데이터의 손실을 최소화하기 위하여 핫덱방법을 적용할 때 필요한 무응답 대체군 형성 방안을 제안하였다. 농가경제조사의 무응답 현황과 특성을 살펴보고, 대체군 형성 방안들을 비교할 수 있는 측도를 제안하였다. 제안된 비교 측도를 이용해서 대체군 형성 방안들을 비교하였다.

주요용어 : 무응답 대체군, 의사결정나무모형, 가중핫덱

### 1. 서론

농가경제조사는 우리나라의 농가경제 및 농업경영의 실태를 정확히 파악하기 위한 통계조사로 농촌경제의 동향과 구조변화를 파악함으로써 농업정책 수립과 농업경영의 개선을 위한 기초자료로 제공되고 있다(한국통계학회, 2002).

농가경제와 관련된 주요지표 항목들을 매월 집계한 후 12개월간의 조사결과를 집계하여 연간소득, 수입·지출 항목과 농가자산, 농가부채 등을 공표하게 되는데, 이 중 한 달 이상의 결측치를 포함하는 농가는 연간추정에서 제외된다. 농가경제조사는 표본가구가 부적격이나 조사불응으로 유고가 발생하면 조사구 내에서 같은 영농형태를 갖고 농업소득이 유사한 가구로 즉시 교체(substitution)한 후 계속조사하고 있지만, 이들 교체 가구에서 조사된 결과는 당해연도 추정과정에는 이용하고 있지 않다. 아울러 교체가 원활히 이뤄지지 않거나, 표본대상가구의 단기적인 출타, 자연재해 등으로 인해 1~2개월의 결측치가 발생할 가능성이 늘 존재한다(김진, 2004).

농가경제조사는 패널조사의 일종으로 볼 수 있다. Lepkowski(1989)는 패널조사에서 무응답 유형을 구분하고, 무응답의 처리에 대해서 다루었다. 본 논문은 농가경제조사에서 무응답이나 교체로 인한 데이터 손실을 막기 위해서 조사과정에서 발생한 가구 무응답을 대체할 수 있는 대체군의 구성방안을 살펴보는 것이다. 이를 통해서 많은 비용과 시간이 소요되어 조사되고 있는 농가경제조사의 데이터 손실을 줄여 추정의 효율을 높이는 것을 목적으로 한다.

본 연구는 농가경제조사의 무응답 대체와 분산 추정 연구 결과의 일부이다. 무응답 대체군을 형성하는 것은 궁극적으로 분산추정 방법과 관련이 있다. 이 연구에서는 대체군을 구성하고, 대체군 내에서 무응답의 대체는 가중핫덱대체법을 사용하는 것을 원칙으로 하였다. 가중핫덱대체는 응답값 중 하나를 선정할 때 가중값을 두어 선정하는 방법으로 복합표본설계에 의해서 표본이 추출되어 대체군 내의 응답값이 서로 다른 추출확률을 가질 때 유용하다. 가중핫덱대체를 이용하면 대체 후에 조사변수의 분포가 그대로 유지될 수 있다는 장점이 있고, 조사에서 다루어지는 변수가 대단히 많기 때문에 무응답 대체방법으로 가중핫덱대체 방법이 적합하다고 할 수 있다. 본 논문에서는 농가경제조사를 위한 무응답 현황과 대체군 형성 방안을 비교하여 농가경제조사에 알맞은 무응답 대체군 형성에 주안점을 두고자 한다.

- 1) 한국방송통신대학교 정보통계학과 부교수, 서울시 종로구 동숭동 169번지
- 2) 서울시립대학교 통계학과 부교수, 서울시 동대문구 전농동 90번지
- 3) 통계청 조사관리과 사무관, 대전광역시 서구 둔산동 920번지 정부대전청사 3동

## 2. 농가경제조사의 무응답 및 표본교체 현황

### (1) 농가경제조사 표본설계

농가경제조사 표본설계는 각 도를 부차모집단으로 하여 각 도에서 층화 2단계락추출법을 사용하였는데, 1차 추출단위는 농업조사구나 부락이고, 2차 추출단위는 농가가 사용되었다. 현행 표본설계의 층화는 영농형태를 기준으로 하였고, 층 내에서 각 표본 농가의 가중치는 동일하다. 결과적으로 각 도별로 10개 층을 구성하였고, 전체 320개의 조사구를 추출하여 각 조사구에서 10 농가씩 총 3,200 농가가 표본으로 선정하였다

### (2) 농가경제조사의 무응답 유형

2003년도 농가경제조사에서는 전체 3,200가구 중에서 3,043(95.1%)의 농가에서는 교체가 일어나지 않았으며, 148 농가(4.6%)에서 1회 이상 교체가 발생하였다. 또한 1달 이상 무응답이 발생한 농가는 모두 27 농가이다. 대부분의 무응답은 표본 농가를 교체하는 과정에서 표본에서 이탈하는 가구와 새로이 표본에 포함되는 가구를 섭외하는 과정에서 발생하고 있다. 표본 농가 중 교체가 1회 이상 발생한 경우는 148 농가이다. 148 농가 중 표본 교체가 2회 이상 발생한 가구는 모두 9가구이며, 2회 7가구, 3회 2가구이다.

<표 1> 농가경제조사의 교체 및 무응답 유형

| 교체회수  | 완전응답  | 무응답 |    | 전체    |
|-------|-------|-----|----|-------|
|       |       | 중간  | 연결 |       |
| 무교체   | 3,043 |     |    | 3,043 |
| 1회 교체 | 116   |     | 23 | 139   |
| 2회 교체 | 4     |     | 3  | 7     |
| 3회 교체 | 1     |     | 1  | 2     |
| 기타    | 9     |     |    | 9     |
|       | 3,173 |     | 27 | 3,200 |

무응답의 지속 기간을 보면, 26농가에 대해서 1달이며, 2개월이 1가구로 나타났다. 즉, 표본 농가의 이탈이 발생하면, 거의 대부분 1달 이내에 새로운 농가로 교체가 되어 확정되고 있다. <표 2>는 2003년 1월부터 12월 사이에 발생한 표본 농가 교체 승인 내역을 교체 사유별로 정리한 것이다.

<표 2> 표본 농가의 교체 사유

| 교체사유 | 표본 농가 수    | 교체 승인 내역   |
|------|------------|------------|
| 비농가  | 44 (28.0%) | 62 (27.9%) |
| 전출   | 25 (15.9%) | 32 (14.4%) |
| 단독가구 | 45 (28.7%) | 60 (27.0%) |
| 조사불능 | 23 (14.6%) | 33 (14.9%) |
| 조사불응 | 15 ( 9.6%) | 30 (16.5%) |
| 장기출타 | 5 ( 3.2%)  | 5 ( 2.3%)  |
| 합계   | 157 (100%) | 222 (100%) |

### (3) 농가경제조사의 무응답 특징

통상적으로 무응답을 처리하는 방법은 가중치를 조정하는 방법과 대체를 하는 방법이 있다.

단위 무응답이 발생하면 가중치 조정을 하며, 항목 무응답이 발생하면 대체를 하는 것이 일반적이다. 농가경제조사의 무응답은 기존의 분류에 의하면 모두가 단위무응답에 속한다. 따라서 통상적인 접근방식으로 보면 가중치 조정을 하는 것이 타당해 보인다. 그러나 농가경제조사의 속성을 보면 가중치 조정보다는 대체를 하는 것이 더 합리적인데 그 이유는 다음과 같다.

첫째, 일계부 조사는 12개월 계속되는 패널조사이기 때문에 12개월 중 한 두 달에서 발생한 무응답은 시계열 과정에서 보면 항목 무응답으로 간주할 수 있다.

둘째, 농가경제조사에는 수십여 개의 항목이 있고, 소분류 항목을 합해서 중분류 항목을 만들고, 중분류 항목을 집계하여 대분류 항목을 만든다. 또한 일부 항목은 일계부 데이터와 원부 데이터가 결합되어 집계된다. 이러한 이유로 현행 농가경제조사에서 표본 농가의 교체나 무응답이 발생한 경우에는 이들 농가의 전체 데이터를 추정과정에서 제외하여 분석하고 있다.

무응답은 속성에 따라 (i) 무시 가능한 무응답과 (ii) 무시하기 어려운 무응답으로 구분하는데, 농가경제조사에서 발생하는 무응답의 주요 원인은 농가의 비농가화, 단독농가화, 진출, 조사불능 등은 무시가능한 무응답으로 분류할 수 있다. 반면에 자연재해 등으로 인해 발생한 무응답은 무시하기 어려운 무응답으로 분류할 수 있다. 무시가능한 무응답인 경우는 가중치 조정이나 무응답 대체를 통하여 처리가 가능지만, 무시하기 어려운 무응답은 가중치조정이나 대체와는 다른 조치가 필요하다.

### 3. 무응답 처리를 위한 대체군 형성

#### (1) 대체군 형성을 위한 기초 분석

2003년도 농가경제조사에 대해서 전국의 영농형태 및 전·겸업별 표본크기 현황을 분석한 결과에 의하면 영농형태와 전·겸업 구분에 따라서 대단한 편차가 있음을 알 수 있다. 대부분의 표본 농가는 논벼(전업), 논벼(1종겸업), 과수(전업), 채소(전업), 축산(전업), 2종겸업 등으로 구분됨을 알 수 있다. 나머지 영농형태는 표본 농가의 수가 상대적으로 대단히 작다. 한편, 각도의 영농형태 및 전·겸업별 표본크기 현황을 분석한 결과에 의하면 각 도별로 충분한 표본 농가가 있는 논벼, 채소, 2종 겸업농 등에 대해서는 각 도별로 이들 농가들을 각각 묶어서 대체군으로 구성할 수 있는 정도의 표본크기인 것으로 나타났다.

<표 3> 농가소득과 농업소득에 유의한 영향을 미치는 설명변수

| 구분     | 농가소득 기준                              | 농업수입 기준                  |
|--------|--------------------------------------|--------------------------|
| 1월 조사  | 지역‡, 농업종사자수‡, 경영주 연령‡                | 지역‡, 농업종사자수‡             |
| 2월 조사  | 농업종사자수‡                              | 농업종사자수‡, 경지면적‡           |
| 3월 조사  | -                                    | 영농형태‡                    |
| 4월 조사  | 영농형태‡, 종사자수‡, 연령‡, 경지면적‡             | 영농형태‡                    |
| 5월 조사  | 농업종사자수‡                              | 지역‡, 영농형태‡               |
| 6월 조사  | 영농형태‡, 종사자수‡, 연령‡                    | 지역‡, 영농형태‡, 종사자수‡, 연령‡   |
| 7월 조사  | 연령‡                                  | 연령‡                      |
| 8월 조사  | 지역‡, 영농형태‡, 연령‡                      | 지역‡, 영농형태‡               |
| 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‡, 경지면적‡                           | 지역‡                      |
| 연단위    | 지역‡, 영농형태‡, 전·겸업‡, 경지면적‡, 종사자수‡, 연령‡ | 영농형태‡, 경지면적‡, 종사자수‡, 연령‡ |

주) †, ‡ : 각각 유의수준 5%, 1% 이하에서 통계적으로 유의한 것을 표시함

## 농가경제조사의 무응답 대체군 형성 방안

영농형태별, 전·겸업별 가구당 소득 평균 현황을 분석한 결과에 의하면 1종겸업 농가는 전업농가와 비교할 때 겸업소득, 사업의 소득에서 크게 차이가 나타나고 있다. 그러나 농가소득이나 농업소득과 비교하면 그 차이는 상대적으로 적게 나타나고 있다. 한편, 전업농과 1종겸업 농가는 농업소득에서 거의 차이가 나타나고 있지 않다. 다만 화훼농인 경우에는 농업소득에서 차이가 발생하고 있고, 전작 농가의 경우에 농업소득이 다른 영농형태에 비해서 적게 나타나고 있다. 겸업소득과 사업의 소득을 고려한다면 전·겸업 구분도 필요한 변수로 볼 수 있다.

앞선 <표 3>은 농가소득, 농업소득 변수를 각각 종속변수로 하고, 영농형태, 전·겸업, 지역(도), 경지면적, 농업종사자 수, 경영주 연령 등을 독립변수로 하는 회귀분석을 실시하여 얻은 결과이다. 전체적으로 영농형태, 전·겸업 구분, 지역, 경지면적, 농업종사자 수, 경영주의 연령 등이 소득을 예측하는 데 중요한 설명변수인 것으로 나타났다.

### (2) 대체군 형성 방법

무응답 대체를 할 때에 대체군(Imputation class)을 형성하는 것이 필수적이다. 대체군을 형성하는 이유는 무응답으로 인한 편향을 대체를 통하여 줄이기 위함이다. 대체군을 만드는 방법은 보조정보를 이용하여 (i) 사후 증화를 하거나 (ii) 주요 관심변수를 기준으로 비슷한 가구들을 묶어 대체군을 형성하는 방법이 있다. 다른 방법으로 로지스틱회귀분석을 통하여 응답률을 추정한 후, 비슷한 응답률을 갖는 가구를 대체군으로 묶는 방법 등이 있다. 그런데 2003년도 농가경제조사 결과를 보면 무응답 또는 표본교체 비율이 4.6%~8%로 전체적으로 작게 나타나고 있어서, 응답률이 비슷하도록 대체군을 묶는 방법은 현실적으로 적용하기 어렵다.

농가경제조사는 일계부를 통해서 월 단위로 데이터가 정리되고 있지만 최종적인 통계는 연 단위로 집계하여 발표하고 있다. 본 연구에서 농가경제조사의 무응답 대체군은 월 단위 데이터와 연단위 데이터를 종합적으로 분석하여 설정하였다.

핫덱대체를 적용하기 위해서 무응답 대체군을 설정할 때 가장 기본원칙은 주요 관심변수들이 동질적이라도 구분하여 대체군을 구성하는 것이다. 이 과정은 증화나 사후증화의 과정과 유사하다고 할 수 있다. 농가경제조사 주요 관심변수는 소득이나 수입과 같은 연속형 변수이다. 본 연구에서는 2003년도 농가경제조사 자료에 대해서 SAS Enterprise Miner의 의사결정나무모형(Decision Tree Model)을 이용하여 설명변수 중에서 대체군을 결정하는 방법을 사용하였다. 각 가구의 농가소득과 농업소득을 목표변수(target variable)로 농가의 특성을 나타내는 나머지 변수들을 설명변수(input variable)로 하였다. 본 연구에서 적용된 무응답 대체군 설정 과정은 다음과 같다.

- 1단계 : 관심변수(target variable)로 사용할 변수 선택  
농가경제조사에서 주요 관심변수는 농가소득, 농업소득, 겸업소득, 사업의 소득 등이다.
- 2단계 : 관심변수에 영향을 주는 주요 설명변수 선택  
설명변수로 영농형태, 전·겸업, 지역, 경지면적, 가구원 수, 연령 등을 고려하였다.
- 3단계 : 대체군 설정 과정에서 사용될 설명변수의 우선순위 결정  
회귀분석 결과와 영농형태, 전·겸업, 지역별 표본 농가 분포를 확인하여 영농형태와 전·겸업 구분을 우선 고려하고, 지역, 경지면적, 종사자 수, 연령 등을 부차적으로 고려하였다.
- 4단계 : 의사결정 나무모형을 이용하여 대체군 구성  
본 연구에서는 영농형태, 전·겸업, 지역 구분을 고려해서 일차적으로 대체군을 구성하고, 이들 대체군 중에서 표본이 충분히 큰 경우에 대해서 경지면적을 설명변수로 하는 의사결정 나무모형을 적용하여 대체군을 세분하는 작업을 진행하였다.
- 5단계 : 대체군 구성의 타당성 확인을 위한 검토 작업

### (3) 대체군 형성 방안

무응답 대체군을 만드는 과정은 층화와 유사해서 대체군을 세분하면 대체군 내의 변수값은 유사하게 되지만 적용할 때는 현실적인 문제가 발생한다. 이 조사에서 농가경제 관련 변수에 영향을 주는 변수인 영농형태, 전·겸업 구분, 지역, 경지면적, 종사자 수, 연령 등의 변수들을 모두 무응답 대체군을 만들 때 고려할 수는 없다. 이들 변수를 모두 고려해서는 지나치게 무응답 대체군이 세분화되어 현실적으로 적용할 수 없다. 본 연구에서는 다음의 세 가지 방안을 고려하여 이를 비교해서 최종적인 무응답 대체군 형성 방안을 제시하고자 한다.

방안 1은 앞선 분석 결과에 기초해서 다음의 원칙에 따라서 무응답 대체군을 구성하였다.

첫째, 무응답 대체군 구성을 위해서 우선 표본 농가의 영농형태를 고려한다. 전국적으로 영농형태별 표본 농가 수가 적은 '화훼'와 '기타'는 각각 독립적인 무응답 대체군으로 결정한다.

둘째, 영농형태 내의 전·겸업별 표본 농가수가 50가구 미만인 경우에는 영농형태 구분에 따라서 대체군을 결정한다. 이에 따라 과수(1종겸업), 특용작물(전업), 전작(1종겸업), 전작(전업), 전작(2종겸업), 축산(1종겸업) 등은 독립적인 무응답 대체군으로 결정된다.

셋째, 각 도별로 충분한 표본 농가가 있는 논벼(전업), 채소(전업), 2종 겸업농 등에 대해서는 각 도별로 이들 농가들을 각각 묶어서 무응답 대체군으로 구성한다.

넷째, 논벼(1종겸업), 과수(전업), 채소(1종겸업), 축산(전업)에 대해서는 표본 농가 수가 적어서 지역별로 대체군을 형성할 수 없다. 본 연구에서는 이들 구분 내에서 지역을 우선적으로 고려하고, 농가소득을 목표변수로 하여 의사결정 나무모형을 적용하여 대체군을 구성하였다. 방안 1의 무응답 대체군은 모두 44개이다.

방안 2는 방안 1을 기초하되 각 지역별 변동이 적은 논벼(전업), 논벼(1종겸업), 2종겸업 등의 영농형태에 대해서는 지역 구분을 두지 않고, 경지면적을 고려해서 대체군을 형성하는 방안이다. 나머지 영농형태에 대해서는 방안 1과 같게 구성한다. 또한 전작, 기타 영농형태는 논보다는 발농사가 중심이며, 이러한 영농형태는 변동이 심하다는 점을 고려하여 영농형태를 병합해서 대체군을 구성하였다. 방안 2에 의해서 구성되는 무응답 대체군은 모두 34개이다.

방안 3은 방안 2를 기초로 하여 경지면적에 따라서 구분된 무응답 대체군 중에서 대체군 내의 표본크기가 200 농가 이상인 대규모 대체군을 유사한 성격의 지역으로 세분하는 방안이다. 위의 과정을 거쳐서 최종적으로 방안 4에 의해서 구성되는 무응답 대체군은 모두 39개이다.

#### (4) 무응답 대체군 형성 방안의 비교

무응답 대체의 효율성을 고려한다면 가급적 대체군을 세분하는 것이 바람직하지만, 농가의 영농형태는 매년 변동될 수 있기 때문에 무응답 대체군을 세분하게 되면 어느 해에는 대체군에 속한 표본수가 대단히 적거나 없는 경우가 발생할 수 있다. 따라서 무응답 대체군을 형성할 때는 무응답 대체의 효율성 측면과 안정성 측면을 함께 고려해야 한다.

앞서 제시한 세 가지의 무응답 대체군 작성 방안을 비교하자. 먼저, 무응답 대체군을 층으로 간주하고, 무응답의 발생 빈도는 대체군의 크기에 비례한다고 가정할 수 있다. 이와 같은 경우에 대체군을 형성하여 무응답을 대체하는 경우와 대체군을 형성하지 않고 전체 조사 데이터에 대하여 핫덱방법을 적용하는 경우를 비교해 보자.

대체군 형성과정이 효과적이라면 대체군 내의 조사값들이 동질적이어서 대체군 내 분산( $S_h^2$ )이 작을 것이다. 각 변수에 대해서 대체군 형성 효과는 비례배정을 하는 경우에 설계효과를 계산하여 그 값을 작을수록 효과적이라고 할 수 있다.

① 전체를 하나의 대체군으로 간주하는 경우의 무응답 대체에 따른 변동 크기

$$V_{SRS}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}, \quad \text{여기서 } n : \text{무응답 농가수}, N : \text{전체 표본 농가수}$$

② 무응답 대체군을 구성하는 경우의 무응답 대체에 따른 변동 크기

$$V_{prop}(\bar{y}_{st}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{h=1}^H \frac{N_h}{N} S_h^2$$

여기서,  $n$  : 무응답 농가수,  $N$  : 전체 표본수,  $N_h$  : 각 대체군에서 표본크기

③ 무응답 대체군 형성의 효과

$$\text{무응답 대체군 형성 효과} = 1 - \frac{\sum_{h=1}^H \frac{N_h}{N} S_h^2}{S^2}$$

무응답 대체군 형성 효과는 백분율로 나타내어 그 값이 커질수록 무응답 대체군이 효과적으로 구성되었다고 해석한다. 연간 데이터를 이용해서 분석할 때 방안 3이 가장 효과적인 것으로 나타났는데, 무응답 대체로 인한 변동을 농가소득 변수에 대해서는 약 16%, 농업소득 변수는 약 24%, 겸업소득 변수는 10%, 사업의 소득 변수는 약 37%, 농가부채 변수는 약 13% 정도 낮출 수 있다(<표 4> 참고). 이러한 결과를 종합해 볼 때 농가경제조사를 위한 무응답 대체군 형성방법은 경지면적 변수를 대체군 형성에 사용하는 방안 3이 적합할 것으로 판단된다.

<표 4> 각 방안별 주요변수에 대한 무응답 대체군 형성 효과 비교

| 구분  | 농가소득  | 농업소득  | 겸업소득 | 사업의 소득 | 농가부채  |
|-----|-------|-------|------|--------|-------|
| 방안1 | 10.5% | 17.5% | 8.7% | 37.9%  | 6.1%  |
| 방안2 | 15.9% | 24.0% | 9.5% | 36.9%  | 12.4% |
| 방안3 | 16.3% | 24.2% | 9.4% | 37.1%  | 12.5% |

#### 4. 결론 및 추후 연구과제

본 논문에서는 농가경제조사를 위한 무응답 현황과 대체군 형성 방안을 비교하여 농가경제 조사에 알맞은 무응답 대체군 형성 방법에 대해서 살펴보았다. 무응답 대체군은 영농형태, 전·겸업, 지역 구분을 고려해서 대체군을 구성하고, 표본크기가 큰 경우에 대해서 경지면적을 설명 변수로 하는 의사결정 나무모형을 적용하여 대체군을 세분하였다.

무응답 대체군을 형성하는 것은 궁극적으로 분산추정 방법과 관련이 있다. 무응답 대체군을 형성하여 얻는 추정의 효율 증가를 파악하기 위해서는 핫택대체 후의 분산추정과 관련이 있다. 따라서 제안된 무응답 대체방법의 타당성을 확인하기 위하여 핫택대체 후의 추정량의 분산과 편향 정도를 파악하기 위한 모의실험을 통한 연구가 계속 진행될 예정이다.

#### 참고문헌

- [1] 김규성 (2000), “무응답 대체 방법과 대체 효과”, 「조사연구」 1(2), 1-19
- [2] 김진 (2004) “농가경제조사의 대체법 비교”, 「통계연구」 출간예정
- [3] 조사통계연구회 (2000), “무응답오차”, 자유아카데미
- [4] 한국통계학회 (2002), “농가경제조사 표본설계 최종보고서”, 미간행보고서.
- [5] J.N.K.Rao and J. Shao (1992), “Jackknife variance estimation with survey data under hot deck imputation”, *Biometrika*, 79(4), 811-822
- [6] J.M. Lepkowski(1989), “Treatment of Wave Nonresponse in Panel Surveys”, in *Panel Surveys*(D. Kasprzyk, G. Duncan, G. Kalton, M.P. Singh, eds.), John Wiley & sons. New York, pp. 348-374