

# 재정패널 가중치

2012. 02.

## 1. 1차년도(2008년) 가중치

복합표본조사(complex sample survey) 자료의 분석을 위한 가중치는 일반적으로 표본설계 가중치, 무응답에 대한 조정 그리고 모집단 정보를 이용한 사후층화 또는 레이킹(raking)에 의한 조정의 세 단계를 거쳐 산출된다. 1차년도 재정패널 자료를 위해서는 표본설계 가중치, 무응답 조정, 그리고 레이킹 조정을 고려하여 가중치를 작성하였다. 모집단의 특성치인 모수에 대한 비편향 추정량(unbiased estimator)을 산출하기 위해서는 표본설계 가중치를 기반으로 적절하게 보정된 가중치를 적용해야 한다. 재정패널의 구축을 위하여 사용된 표본설계 및 모집단 정보를 통해 작성된 가중치를 이용하여 모수의 불편 추정량(근사 불편 추정량)을 구하는 것이 가중치 산출의 목적이다.

### 1.1 표본설계

표본추출을 위한 표집틀(sampling frame)은 2005년 인구주택총조사 90% 자료와 2006~2008년 신규 아파트 자료를 병합하여 구성되었다. 이는 2005년 이후 발생한 가구 모집단의 변화를 표집틀에 반영하기 위함이다. 인구주택총조사 자료와 신규 아파트 자료의 구분 변수를 1차 층으로 정의하고 각 층으로부터 독립적인 표본을 추출함으로써 재정패널 표본을 구성하였다.

각 표집틀로부터 가구 추출을 위해서는 각 표집틀을 층화한 후 각 층내에서 조사구를 추출하고 추출된 조사구내에서 다시 가구를 추출하는 층화 이단계 추출법이 사용되었다. 90% 인구주택총조사 자료의 층화를 위해서는 지역(제주도 제외 15개 광역시도), 읍면/동부, 일반/아파트가 사용되었으며 층내의 조사구 정렬을 위해서는 시군구와 조사구내의 가구 연건평의 최빈값을 사용하였다. 신축아파트 자료의 경우, 근접 가구들을 묶어서 크기가 50~100인 일종의 조사구를 구성하였으며 구성된 조사구의 층화를 위해서는 지역(제주도 제외 15개 광역시도)을 사용하였고 정렬변수로는 시군구와 대표평수를 사용하였다.

각 표집틀에서 정의된 층의 조사구 크기에 비례하여 표본 조사구를 추출하였다. 실제로 빈곤 지역과 부유지역을 구분하고 이들 지역으로부터 충분한 수의 패널을 확보하고자 해당 지역의 할당 조사구수를 늘리는 방안을 채택하였다. 추출된 조사구로부터 평균적으로 5가구씩 추출하였다.

## 1.2 표본설계 가중치 및 무응답 보정 가중치 산출

1차년도 원패널 5,014개 가구와 2차년도에 추가된 620개 대체가구를 합하여, 총 5,634개 가구에 표본설계 가중치를 부여하였다. 620개 대체가구는 2차년도에 조사 실패한 618개 원패널 가구를 대체하기 위하여 추가된 가구이다(실패한 가구에서 두 가구가 추가로 조사되어 차이 발생). 따라서 1차년도 가중치 산출에서는 620개 대체가구를 무응답 가구로 처리하고 원패널 5,014개 가구에 무응답 보정 가중치를 부여하였다.

조사된  $h$ 번째 층의  $i$ 번째 조사구의  $j$ 번째 가구의 표본설계 가중치는 다음과 같다.

$$w_{hij} = \frac{N_h}{n_h} \times \frac{N_{hi}}{n_{hi}}.$$

위 식에서  $N_h$ 는  $h$ 층의 모집단 조사구수,  $n_h$ 는  $h$ 층의 표본 조사구수,  $N_{hi}$ 는  $h$ 층 조사구  $i$ 의 모집단 가구수, 그리고  $n_{hi}$ 는  $h$ 층 조사구  $i$ 의 표본 가구수를 나타낸다.

1차년도 무응답 보정 가중치는 표본설계 가중치에 응답률의 역수를 곱하여 계산되어 진다. 응답률은  $r_{hi}/n_{hi}$ 로 계산되어 지며,  $r_{hi}$ 는  $h$ 번째 층의  $i$ 번째 조사구의 응답 가구수이다. 조사된  $h$ 번째 층의  $i$ 번째 조사구의  $j$ 번째 가구의 무응답 보정 가중치는 다음과 같다.

$$w_{hij}^{na} = \frac{N_h}{n_h} \times \frac{N_{hi}}{n_{hi}} \times \left( \frac{r_{hi}}{n_{hi}} \right)^{-1} = \frac{N_h}{n_h} \times \frac{N_{hi}}{r_{hi}}.$$

## 1.3 레이킹(raking) 가중치 보정

산출된 무응답 보정 가구 가중치( $w_{hij}^{na}$ )는 다시 통계청에서 제공하는 2008년 가구 및 인구 추계 통계를 바탕으로 레이킹을 이용하여 보정되었다. 레이킹을 위해서 사용된 가구 변수는 각 지역별(제주도 제외 15개 시도) 가구주의 성(남/여), 연령(39 미만/40~49/50~59/60 이상) 그리고 가구원 수(4인 미만/4인 이상)이다. 레이킹을 이용하여 보정된 가중치는

$$w_{hij}^{rak} = w_{hij}^{na} \exp\left(\underline{x}'_{hij} \underline{\lambda}\right)$$

이다. 여기서  $\underline{x}_{hij}$ 는 레이킹을 위해 사용된 변수들을 지시변수들로 표현한 벡터이며  $\underline{\lambda}$ 는  $\sum w_{hij}^{rak} x_{hij} = t_x$ 의 해이며  $t_x$ 는 통계청에서 제공하는 2008년 각 레이킹 변수들의 추계 값을 나타낸다. 레이킹을 통한 가중치의 보정 결과 2008년 추계 자료와의 벤치마킹 과정에서 지나치게 가중치가 크게 나타나는 관측치가 존재하여 가중치의 최댓값을 9,000으로 제한하였다. 가중치가 9,000을 초과하는 극단 관측치들의 경우 가중치를 9,000으로 조정하고, 레이킹 보정 가중치에서 9,000을 제외한 나머지 값을 극단 관측치가 속한 벤치마킹범주(지역\*가구주 성\*가구주 연령\*가구

원 수) 내의 관측치들에게 동일하게 배분하였다.

## 2. 2차년도(2009년) 가중치

재정패널 2차년도 조사결과 분석을 위해서는 2개년도(2008년과 2009년)의 종단면 분석을 위한 가중치와 2차년도 횡단면 분석을 위한 가중치가 각각 산출되었다.

### 2.1 2차년도 종단면 가중치 산출

종단면 분석을 위한 종단면 가중치는 1차년도와 2차년도 모두 응답한 가구에 부여되며, 1차년도의 무응답 보정 가중치( $w_{hij}^{na}$ )에 2차년도 무응답으로 인한 실제 표본크기의 감소를 반영하기 위한 무응답 보정과 2009년 추계 자료를 이용한 레이킹 보정의 2단계에 걸쳐서 계산되어 진다.

패널 탈락으로 인한 무응답 보정을 위해서는 2009년 종단면 응답여부 변수와 2008년 변수들의 관계를 로지스틱 회귀모형을 통하여 설명하고 이를 바탕으로 2009년 응답가구의 응답확률을 예측하였다. 2009년 종단면 응답여부 변수는 2008년과 2009년 모두 응답하면 1, 그렇지 않으면 0을 갖는다. 2008년에 작성된 각 가구의 무응답 보정 가중치( $w_{hij}^{na}$ )를  $w_{2008}$ 이라 표기하고 로지스틱 회귀분석을 통해 예측된 예측 응답확률을  $\hat{p}_{2009}$ 라 할 때 무응답 보정을 통하여 얻어지는 가중치는

$$w_{2009} = w_{2008} \times \hat{p}_{2009}^{-1}$$

로 표현된다. 실제  $\hat{p}_{2009}$ 의 예측을 위하여 적합된 로지스틱 회귀 모형은

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \text{가구주성별}\beta_1 + \text{가구원수}\beta_2 \\ + \text{가구주연령}\beta_3 + \text{지역}\beta_4 + \text{가구연간소득총액}\beta_5$$

이다.

무응답 보정을 통하여 얻어진 가중치는 2009년 인구 및 가구 추계 자료를 이용하여 보정하였다. 2009년 추계 자료의 벤치마킹을 위해서는 1차년도에 사용된 레이킹 기법이 사용되었다. 2009년 레이킹을 위하여 지역과 가구주의 성, 연령, 그리고 가구원수의 주변분포를 사용하였다. 레이킹 이후 계산되어진 2차년도 종단면 가중치가 지나치게 크게 나타나는 경우 1차년도와 동일한 방법을 통해 레이킹 이후 가중치의 최댓값을 12,500으로 제한하였다.

## 2.2 2차년도 횡단면 가중치 산출

2차년도 횡단면 가중치는 2차년도 조사에 응답한 가구에 부여되며, 기존가구(원패널 가구, 대체가구)와 분가가구(2008년 6월 이전 분가)에 따른 가중치의 산출과 이후 2009년 추계 자료 이용한 레이킹 보정의 2단계에 걸쳐 이루어 졌다.

각 가구형태별 1단계 가중치 산출을 살펴보면, 먼저 2차년도에 조사된 기존가구(원패널 가구, 대체가구)의 경우, 1차년도의 표본설계 가중치( $w_{hij}$ )를 사용하고 패널 탈락으로 인한 무응답 보정을 통하여 1단계 가중치가 산출되었다. 무응답 보정 방법은 종단면의 무응답 보정 방법과 동일하다. 로지스틱 회귀모형의 종속변수로는 2009년 횡단면 응답여부 변수를 사용했으며, 2009년 횡단면 응답여부 변수는 2009년에 응답하면 1, 그렇지 않으면 0을 갖는다. 설명변수로는 원패널 가구의 경우 종단면과 동일하게 2008년도 지역, 가구주 성, 가구주 연령, 가구원 수, 그리고 가구연간소득총액을 사용했다. 2008년 자료가 없는 대체가구는 대응되는 원패널 가구(2차년도 조사실패가구)의 2008년도 5개 변수(지역, 가구주 성, 가구주 연령, 가구원 수, 가구연간소득총액)를 그대로 사용했다.

분가가구의 경우, 분가사유에 따라 가중치를 다르게 부여하였다. 분가사유가 결혼인 경우, 원가구의 당해년도 무응답이 보정된 1단계 가중치의 1/2을 부여하였고 기타 사유의 분가가구 경우, 원가구의 당해년도 무응답이 보정된 1단계 가중치를 그대로 부여하였다. 또한 분가사유가 결혼인 35가구 중 응답하지 않은 12가구에 부여된 가중치를 조사 성공한 23가구에 부여하였다. 이를 위하여 분가사유가 결혼인 35가구를 원가구의 2008년 연간총소득액 4,000만원을 기준으로 두 그룹으로 나누고 각 그룹내에서 2차년도 무응답 가구의 가중치를 응답가구에 동일하게 배분하였다.

기존가구와 분가가구의 형태별로 1단계에서 부여된 가중치는 2009년 인구 및 가구 추계 자료를 이용한 레이킹을 통해 보정되었다. 레이킹을 위해 사용된 변수와 모집단 분포는 종단면 레이킹 보정시 사용된 것과 동일하다.

## 3. 3차년도(2010년) 가중치

재정패널 3차년도 조사결과 분석을 위해서는 3개년도(2008년~2010년)의 종단면 분석을 위한 가중치와 1, 2차년도와 마찬가지로 3차년도 횡단면 자료 분석을 위한 가중치가 각각 산출되었다.

### 3.1 3차년도 종단면 가중치 산출

3개년도 종단면 분석을 위한 종단면 가중치는 1차년도, 2차년도 그리고 3차년도 모두 응답한

가구에 부여되며 가중치 산출에는 1차년도의 무응답 보정 가중치( $w_{hij}^{na}$ )를 사용하였다. 3차년도 무응답으로 인한 실제 표본크기 감소를 반영하기 위한 보정은 2차년도 종단면의 무응답 보정 방법을 사용하였다. 로지스틱 회귀모형은 2010년 종단면 응답여부 변수를 종속변수로 사용했으며, 2010년 종단면 응답여부 변수는 2008년, 2009년 그리고 2010년 3개년도 모두 응답하면 1, 그렇지 않으면 0을 갖는다. 설명변수는 2차년도와 동일하게 2008년 지역, 가구주 성, 가구주 연령, 가구원 수, 그리고 가구연간소득총액을 사용했다.

무응답 보정을 통하여 얻어진 가중치는 2010년 인구 및 가구 추계 자료를 이용한 레이킹을 통해 보정되었다. 레이킹을 위해 사용된 변수와 모집단 분포는 2차년도 종단면 레이킹 보정시 사용된 것과 동일하다. 레이킹 이후 계산되어진 3개년도 종단면 가중치가 지나치게 크게 나타나는 경우 1차년도와 동일한 방법을 통해 레이킹 이후 가중치의 최댓값을 13,300으로 제한하였다.

### 3.2 3차년도 횡단면 가중치 산출

3차년도 횡단면 가중치는 3차년도 조사에 응답한 가구에 부여되며, 기존가구(원패널 가구, 대체가구), 기존분가(2008년 6월 이전 분가) 그리고 신규분가(2008년 7월 ~ 2009년 6월 이전)에 따른 가중치의 산출과 이후 2010년 추계 자료를 이용한 레이킹 보정의 2단계에 걸쳐 이루어 졌다.

기존가구의 1단계 가중치 산출은 패널 탈락으로 인한 무응답 보정으로 2차년도 횡단면 무응답 보정에서 기존가구의 무응답 보정 방법과 동일하다. 다만, 2010년 횡단면 응답여부 변수를 로지스틱 회귀모형의 종속변수로 사용했다. 2010년 횡단면 응답여부 변수는 2010년에 응답하면 1, 그렇지 않으면 0을 갖는다. 설명변수는 2차년도와 동일하다. 기존분가의 경우, 3차년도에 모든 가구가 응답하여 2차년도에 부여한 횡단면 1단계 가중치를 그대로 사용하였다. 신규분가는 2차년도 분가가구의 가중치 부여방법과 동일한 방법을 통해 1단계 가중치를 부여하였다.

각 가구형태별로 1단계에서 부여된 가중치는 2010년 인구 및 가구 추계 자료를 이용한 레이킹을 통해 보정되었다. 레이킹을 위해 사용된 변수와 모집단 분포는 종단면 레이킹 보정시 사용된 것과 동일하다. 레이킹 이후 계산되어진 3차년도 횡단면 가중치가 지나치게 크게 나타나는 경우 1차년도와 동일한 방법을 통해 레이킹 이후 가중치의 최댓값을 11,700으로 제한하였다.

## 4. 4차년도(2011년) 가중치

재정패널 4차년도 조사결과 분석을 위해서는 4개년도(2008년~2011년)의 종단면 분석을 위한 가중치와 4차년도 횡단면 자료 분석을 위한 가중치가 각각 산출되었다.

### 4.1 4차년도 종단면 가중치 산출

4개년도 종단면 분석을 위한 종단면 가중치는 4개년도(2008년~2011년) 모두 응답한 가구에 부여되며, 패널 탈락으로 인한 실제 표본크기 감소를 반영하기 위한 보정은 3차년도 종단면 무응답 보정 방법과 동일하다.

무응답 보정을 통하여 얻어진 가중치는 2011년 인구 및 가구 추계 자료를 이용한 레이킹을 통해 보정되었다. 레이킹을 위해 사용된 변수와 모집단 분포는 2차년도 종단면 레이킹 보정시 사용된 것과 동일하다. 레이킹 이후 계산되어진 4개년도 종단면 가중치가 지나치게 크게 나타나는 경우 1차년도와 동일한 방법을 통해 레이킹 이후 가중치의 최댓값을 15,000으로 제한하였다.

## 4.2 4차년도 횡단면 가중치 산출

4차년도 횡단면 가중치는 4차년도 조사에 응답한 가구에 부여되며 기존가구(원패널 가구, 대체가구), 기존분가(2009년 6월 이전 분가)와 신규분가(2009년 7월 ~ 2010년 6월 이전)에 따른 가중치의 산출과 이후 2011년 추계 자료를 이용한 레이킹 보정의 2단계에 걸쳐 이루어 졌다.

기존가구와 기존분가의 1단계 가중치 산출은 패널 탈락으로 인한 무응답 보정으로 2차년도 횡단면 무응답 보정에서 기존가구의 무응답 보정방법과 동일하다. 기존가구는 1차년도의 표본설계 가중치( $w_{hij}$ )를 사용하고, 기존분가는 원가구의 1차년도 표본설계 가중치를 사용하였다. 예측 응답확률은 기존가구와 기존분가를 합하여 로지스틱 회귀분석을 통해 계산된다. 2011년 횡단면 응답여부 변수를 로지스틱 회귀모형의 종속변수로 사용했다. 2011년 횡단면 응답여부 변수는 2011년에 응답하면 1, 그렇지 않으면 0을 갖는다. 기존가구의 경우 설명변수는 2차년도와 동일하게 2008년도 지역, 가구주 성, 가구주 연령, 가구원 수, 그리고 가구연간소득총액을 사용하였고, 기존분가는 관측치별 조사된 최근 자료의 지역, 가구주 성, 가구주 연령, 가구원 수, 그리고 가구연간소득총액을 사용했다. 신규분가는 2차년도 분가가구의 가중치 부여방법과 동일한 방법을 통해 1단계 가중치를 부여하였다.

각 가구형태별로 1단계에서 부여된 가중치는 2010년 인구 및 가구 추계 자료를 이용한 레이킹을 통해 보정되었다. 레이킹을 위해 사용된 변수와 모집단 분포는 종단면 레이킹 보정시 사용된 것과 동일하다. 레이킹 이후 계산되어진 4차년도 횡단면 가중치가 지나치게 크게 나타나는 경우 1차년도와 동일한 방법을 통해 레이킹 이후 가중치의 최댓값을 14,000으로 제한하였다.