

## ■ 원표본(2005년) 가중치 산출과정

### ○ 기본가중치

- (의 미) 기본 가중치는 표본 설계에서 사용된 표본 추출 확률을 바탕으로 하되 50세 이상 가구원을 가지는 가구 수의 분포를 이용하여 50세 이상의 가구에 대해 대표성을 가지도록 결정
- (필요성) 표본 설계 당시에 추출되었던 표본 가구들이 접촉 불능, 부적격, 거절 등의 현실적인 이유로 최종 표본으로 선정되지 못하고 현장에서 유사 가구로 대체되는 경우가 종종 있으므로 이러한 현실들을 반영하여 기본 가중치가 현재의 목표 모집단에 대표성을 최대한 가지도록 결정
- (방 법) 모집단에서 추출단위가 표본에 포함될 확률(=표본추출확률)의 역수

$$\text{기본가중치} = \frac{1}{\text{조사구 추출 확률}} \times \frac{1}{\text{조사구내 표본 추출 확률}}$$

$$\text{조사구내 표본 추출 확률} = \frac{\text{표본가구수}}{\text{조사구내 유효(50세 이상)가구수}}$$

### ○ 무응답가중치 보정

- (의 미) 2차년도 이후 탈락(attrition)이나 단위 무응답(unit nonresponse)으로 인한 편향을 제거
- (필요성) 표본 대표성 유지 및 추정 값의 정확성 제고
- (방 법) 무응답층을 이용한 무응답 가중치 조정
  - 모집단이  $G$ 개의 무응답층으로 나누어진다고 할 때

$$U = \bigcup_{g=1}^G U_g, \quad S = \bigcup_{g=1}^G S_g$$

- 층 내의 모든 원소들의 응답률은

$$\begin{aligned} P[R_i=1 | X_i] &= P[R_i=1 | i \in S_g] \\ &= \frac{g \text{ 번째 층의 총 응답 가구수}}{g \text{ 번째 층의 총 가구수}} \end{aligned}$$

- 무응답층을 이용한 가중치 조정 추정량은 다음과 같이 표현

$$\hat{\Theta}_r = \sum_{g=1}^G \left( \sum_{i \in S_g} w_i \right) \cdot \frac{\sum_{i \in S_g} w_i R_i Y_i}{\sum_{i \in S_g} w_i R_i}$$

- 각 무응답층 내에서 무응답률이 일정하다고 할 때 이 무응답층을 이용한 가중치 조정 추정량은 다음과 같은 통계적 성질이 알려짐

①  $\hat{\Theta}_r$ 은 각 무응답층에서 비조정(ratio adjustment)을 한 개별비 추정(separate

ratio estimator)의 일종으로 비추정량을 사용함으로써 나타나는 편향(ratio bias)이 존재하지만 무응답층내의 표본크기가 크거나 가중치들이 동질적이면 무시할 수 있을 정도로 작아짐. 따라서  $\hat{\Theta}_R$ 는 근사적 비편향 추정량이 됨.

② 분산의 형태는 다음과 같이 표현

$$V[\hat{\Theta}_R] \doteq V[\hat{\Theta}_n] + E\left[\sum_{g=1}^G n_g^2 \left(\frac{1}{r_g} - \frac{1}{n_g}\right) \hat{\sigma}_{weg}^2\right]$$

분산의 형태를 보면 등식 오른쪽 첫 번째 항은 무응답과 관계없는 항으로 점추정치의 표본 설계로 인한 분산이며 두 번째 항은 무응답으로 인해 증가되는 분산 값으로 일반적으로 층의 개수가 많아질수록 그리고 층 내 Y 변수의 분산이 적을수록 적어진다는 것을 알 수 있음

- (무응답층 설정) Dufour 등 (2001)의 연구결과에 따라 나무모형(tree model)이 로지스틱회귀모형(logistic regression)보다 더 효율이 좋다는 것이 밝혀졌으므로 나무모형을 사용. 나무모형 중 응답률의 동질성을 분류하는 기법으로는 CHAID (Chi-Square Automatic Interaction Detection) 알고리즘이 이용. CHAID 알고리즘은 Kass (1980)가 제안한 방법으로 주어진 층 분류 옵션 중에서 가능한 모든 층의 분류 중에서 응답률의 동질성 검정 통계량인 카이제곱 통계량의 값을 최대로 해 주는 층 분류를 찾아서 순차적으로 분류해 나가는 층 분류 기법임

#### ○ 횡단 가중치와 종단 가중치

- 아래 표는 패널조사에서 가능한 응답자의 형태를 보여줌

패턴	응답상태	조사시점		
		1차조사	2차조사	3차조사
1	완전응답	○	○	○
2	감소	○	○	X
3		○	X	X
4	간헐적 무응답	○	X	○

주: 응답(○), 무응답(X)

- 각 조사시점의 횡단면 분석을 고려하면, 2차년도 조사의 횡단면 분석은 위 표의 패턴 1과 2에 관련된 조사 대상자들이 대상임. 이 때 패턴 3과 4에 속하는 조사대상자는 무응답자로 간주하여 가중치를 설정. 또한 3차년도 조사의 횡단면 가중치는 패턴 1과 4에 속하는 조사 대상자들을 대상
- 종단 가중치를 고려하면, 3차년도의 종단면 분석은 위 표의 패턴 1에 속하는 표본이 대상임. 종단 가중치는 1차, 2차, 3차년도 모두 응답한 개인을 대상으로 산출하고 종단 가중치는 1차년도 조사의 가중치, 2차년도 조사의 횡단가중치, 그리고 1차년도 조사와 2차년도 조사에서 모두 응답했으나 3차년도 조사에서는 무응답한 조사 대상자의 정보를 이용하여 구한 가중치의 곱으로 결정

### ○ 사후가중치 보정

- (의미) 모집단의 분포를 반영하기 위해서는 기본 가중치에 보정(adjustment)
- (필요성) 2005년 인구 주택 총 조사 전수 조사 자료를 분석하여 각 지역별로 모집단 분포와 표본 자료의 여러 속성별 최종 가중합이 동일해지도록 가중치를 결정. 즉, 벤치마킹 변수에 대해서 전수조사에서 얻어진 알려진 값과 표본에서 추정되는 추정값을 일치시켜 줌으로써 (또는 아주 근접하게 해 줌으로써) 통계의 일관성(consistency)도 유지하고 다른 항목의 추정에도 효율을 높이고자 함
- (방법) 2005년 인구 주택 총 조사 전수 조사 자료를 분석하여 각 지역별로 모집단 분포와 표본 자료의 여러 속성별 최종 가중합이 동일해지도록 가중치를 결정
  - 가구 자료의 경우 각 지역 내에서 동부 읍면부별 분포, 가구주 연령대별 분포, 성별 분포, 거처 종류별 분포, 점유 형태별 분포, 그리고 가구원수별 분포 사용
  - 개인 자료의 경우 보정에 사용되는 변수는 각 지역 내에서 동부/읍면부별 분포, 개인 연령대별 분포, 성별 분포가 사용
  - 가중치 절삭(weight trimming) : 각 지역별로 가중치 이상점을 평균 가중치보다 1/3 보다 작거나 3배 이상이 되는 경우로 규정하여 가중치 절삭을 한 후에는 다시 가중치 평균을 계산하여 절삭전과 절삭후의 평균의 비를 곱해줌으로써 지역별 가중치 합이 같아지도록 함

## ■ 통합 표본 가중치 산출과정

### ○ 추가 표본 구축 현황

- 표본 추가는 모집단 신규 유입과 패널 마모를 보완하도록 표본을 설계함. 따라서 추가 표본과 원표본(05표본)을 결합한 통합표본을 구축하였으며 두 표본 자료의 표본 설계 추출 확률을 반영하고 횡단면적 대표성을 극대화하는 가중치 작업을 수행함

### ○ 특징

- 통합 표본 가중치는 2010년 인구주택 총 조사 자료를 바탕으로 추출된 추가 표본과 2005년 인구주택 총 조사 자료를 바탕으로 추출된 원자료 표본을 합치고자 할 때 사용하게 될 통합 가중치를 통계학적 방법을 사용하여 개발
- 최종 가중치는 초기 표본 추출 확률과 반비례하도록 결정하되 이를 사용하여 얻어지는 연령대/성별/지역 등 주요 변수에 대한 통계치가 모집단 총계의 분포와 일치해 지도록 조정

### ○ 방법

#### - 모집단과 표본

- 2005년 처음 표본 설계 당시의 모집단  $U_1$ 이 있고 이 모집단에서 표본  $s_1$ 을 추출. 현재 시점에서 신규 유입된 모집단을  $U_2$ 라고 하면 최종 모집단은  $U = U_1 \cup U_2$ 으로 표현. 표본 추가 단계에서  $U_2$ 를 2013년 기준으로 만 50세부터 58세 사이의 가구주가 있는 가구와 그 가구원으로 정하였고 이로부터 표본  $s_2$ 를 추출

- 이상적으로는  $U_1$ 과  $U_2$ 는 서로 교집합이 없는 배반 집합(exclusive set)으로 정의하는 것이 표본 추가와 가중치 작업이 수월하게 되지만 실제로는  $S_1$ 중에서도  $U_2$ 의 정의에 포함되는 가구나 가구원이 존재하게 되므로 이를 가중치 작업에 반영하여 제대로 처리
- 기존 표본  $S_1$ 중에서  $U_2$ 의 정의에 포함되는 가구나 가구원 표본은  $S_2$ 에 포함. 즉,  $U_2$ 의 정의 기준에 따라  $S_1 = S_1' \cup S_{1O}$ 으로 나누고 (여기서  $S_{1O}$ 은 새로운 기준에 의해  $S_2$ 에 포함되는 표본 집합)에 대한 가중치 처리를 하고 그 후  $S_2' = S_2 \cup S_{1O}$ 에 대한 가중치 처리
- 기존 표본  $S_1'$ 에 대한 가중치 설정
  - 기존 표본  $S_1'$ 은 패널 마모 등으로 인해 그 부분 집합인  $S_{R1}$ 이 남게 되고 이에 대한 가중치는  $\sum_{i \in S_{R1}} w_i(x_{i1}, \dots, x_{ip}) = \sum_{i \in S_1'} d_i(x_{i1}, \dots, x_{ip})$ 을 만족하도록 결정.
  - 여기서  $d_i$ 는 무응답이 없을 경우 사용했었을 기본 가중치이고  $(x_{i1}, \dots, x_{ip})$ 는 가중치 작업에 사용되는 변수로써 지역, 연령대, 성별, 가구주 학력, 거주 형태 등의 변수가 됨
- 추가 표본  $S_2' = S_2 \cup S_{1O}$ 에 대한 가중치 설정
  - 먼저  $S_2$ 에서 표본 설계 시 얻어지는 기본 가중치(표본 추출 확률의 역수)가 있고 예도 기존 표본에서 사용되는 가중치가 있으므로 이를 바탕으로 통합하여 최종 가중치를 결정. 모집단  $U_2$ 에 대한 기본 정보를 2010년 센서스 자료를 통해서 얻을 수 있으므로 이를 바탕으로 하여 가중치 작업을 실시. 즉, 기본 가중치로부터 calibration 작업을 해야 하는데 (Deville and Sarndal, 1992),
  - 이 때 사용되는 제한 조건은  $\sum_{i \in S_2'} w_i(x_{i1}, \dots, x_{ip}) = \sum_{i \in U_2} (x_{i1}, \dots, x_{ip})$  으로 표현