# 한국복지패널의 가중치 조정과 향후 과제

*The Weighting Adjustment of Korea Welfare Panel Study*

**손창균** 한국보건사회연구원 부연구위원

패널조사는 횡단면 조사와는 달리 최초 표본이 시간이 지남에 따라 조사 대상 표본으로부터 탈락함으로서 표본의 마모와 그에 따른 대표성 상실의 문제가 발생한다. 그러므로 이러한 표본의 대표성 상실 문제를 적절히 해결하기 위해 가중치 조정을 실시한다. 횡단면 조사에서는 (1) 추출가중치의 조정, (2) 무응답 가중치 조정, (3) 사후층화 가중치 조정과 같이 3단계의 가중치 조정 과정을 수행하지만, 패널 조사의 경우에는 이와 더불어 각 웨이브별 표본의 대표성을 유지하기 위해 종단면 가중치(longitudinal weight) 조정을 함께 고려해야 한다.

이러한 관점에서 본 고에서는 현재 국내·외에서 수행되고 있는 패널조사의 가중치 조정방법을 고찰하고, 한국복지패널(Korea Welfare Panel Study: KOWEPS)의 가중치 산정에 관한 이론적 근거를 마련함과 동시에 향후 과제에 대해 고찰하고자 한다.

## 1. 서론

종단면 조사(Longitudinal survey)의 일종인 패널조사(Panel survey)는 개인 또는 집단의 행태 연구나 사회의 변화가 개인의 행동양식에 미치는 영향 등에 대한 조사를 다년간 수행하는 조사이다. 당연히 횡단면적인 특성과 시계열적인 특성을 모두 가진 데이터로서 주로 사회, 경제, 교육학 등에서 많이 활용되는 자료수집 방법이다. 패널자료는 어느 한 시점에 조사된 횡단면 자료와는 달리 시간의 흐름에 따라 개체들의 동적인 패턴을 연구할 수 있다는 특징을 가진다. 이러한 관점에서 패널 조사의 장점으로는 표본의 크기가 커짐에 따라 자유도가 증가함으로 추정량의 효율이 향상될 수 있으며, 설명변수들 간의 공선성(collinearity) 문제가 감소하며, 추정량의 편향감소 등이 있다. 따라서 개체간의 동적 연관성(dynamic relationship)에 관한 연구가 가능하며, 개체들 간의 이질성(heterogeneity)을 모형화 할 수 있다. 그러나 이러한 장점에도 불구하고 종단면 조사는 최초 표본이 시간이 지남에 따라 조사 대상 표본으로부터 탈락함으로서 발생하는 표본의 마모(attrition)와 그에 따른 대표성 상실의 문제이다. 그러므로 이러한 표본의 대표성 상실 문제를 적절히 해결하기 위해 적용가능한 방법이 가중치 조정 방법이다. 일반적으로 횡단면 조사에서는 (1) 추출가중치의 조정, (2) 무응답 가중치 조정, (3) 사후층화 가중치 조정과 같이 3단계의 가중치 조정 과정을 수행하지만, 패널 조사의 경우 이와 더불어 원 표본의 대표성을 유지하기 위해 종단면 가중치(longitudinal weight)를 함께 고려해야 한다.

이러한 관점에서 본 연구에서는 국내외 패널조사에서 적용하고 있는 가중치 조정 방법에 대해 고찰하고, 2006년의 1차 웨이브를 중심으로 한국복지패널(Korea Welfare Panel Study: KOWEPS)에 대한 향후 분석을 위해 요구되는 가중치 산정에 관한 이론적인 근거를 마련함과 동시에 현재 국내에서 수행되고 있는 다른 여타의 패널조사에 대해 가중치 산정 방법과 비교하고자 한다.

## 2. 해외 패널조사의 가중치 조정 방법

### 1) 미국의 PSID(Panel Study of Income Dynamics)[1]

PSID는 초기에 두개의 독립된 표본으로 구성하였다. 하나는 층화다단계추출에 선정된 생산가능인구에 대한 횡단면 표본이며, 다른 하나는 저소득층에 대한 표본이다. 횡단면 표본은 SRC(Survey Research Center)에 의해 추출되었고, 이 표본은 등확률로 추출된 표본으로 1968년에 2,930가구를 조사완료 하였다. 저소득층에 대한 표본은 PSID가 SEO (Survey of Economic Opportunity)표본으로부터 추출한 1,872가구로 구성되며, 불균등확률 표본이다. 전자를 SRC표본이라 하고, 후자를 SEO표본이라 부른다.

최초 5,000가구로 시작하여 매년 조사를 실시하며, 조사의 초점은 경제상황과 인구학적인 상황, 특별히 소득원과 소득총액, 취업 가구원 구성변동, 주거위치 등에 대한 사항이다. 이와 더불어 사회학적, 심리학적 측도를 포함하고 있다. 1995년 현재 PSID는 약 28년간 생존한 개인들에 대해 약 50,000명 이상의 정보를 수집하였다. 표본은 1968년 이래로 매년 조사된 개인을 포함하며, 1990년에 추가된 남미계통(Hispanic)의 가구주 2,000명의 대표표본을 포함하며, 원 표본 가구원에 의해 형성된 가구원을 포함한다.

#### (1) PSID 표본 추출과 초기 가중치의 계산과정

가중치 계산을 위해 먼저 PSID의 표본추출과정을 이해할 필요가 있다. 1968년 당시 가구표본은 다음 두 가지로 구성된다. (1) 공통적으로 미국에 거주하는 횡단면 표본과 (2) OEO(Office of Economic Opportunity)의 요구에 따라 1967년에 미국 센서스국(Census Bureau)에 의해 조사된 가구들의 부차표본이다.

1969년과 1970년에 표본은 전년도에 계속적

1) McGonagle, K. A. & Schoeni, R.F. (2006). The Panel Study of Income Dynamics: Overview and Summary of Scientific Contributions After Nearly 40 years. Institute for Social Research, University of Michigan.
Heeringa, S.G. & Connor, J.H. (1999). 1997 Panel Study of Income Dynamics Analysis Weights for Sample Families and Individuals. Institute for Social Research, University of Michigan.

으로 조사된 가구에 거주하는 모든 패널 구성원으로 구성된다. 따라서 전년도 웨이브(wave)에 응답하지 않은 구성원에 대해서는 2차년도 조사를 수행하지 않았다. 거처의 횡단면 표본을 SRC(Survey Research Center)의 마스터 프레임(master frame)으로부터 상주인구 전체 추출률로 추출하였다. 1968년의 센서스 표본은 재조사와 같은 형태인데, 왜냐하면 이 가구들에 대해서는 센서스국에서 전년도에 이미 조사가 이루어졌기 때문이다. 8가지의 기본적인 추출률을 가진 확률 표본추출이지만, 센서스국에 의해 조사된 가구들 중 소득이 $2,000+N($1,000) 이하인 가구에 대해서만 조사되었다. 여기서 N은 가구 수를 나타낸다. $2,000+N($1,000)값은 1967년에 사용된 연방의 빈곤선(federal poverty line)의 2배와 거의 일치하는 값이다. 이 값 이상의 소득을 갖는 가구는 제외하였으며, 특히 북동부, 북중부, 서부지역 등 3개 지역에 있는 SMSA(Standard Metropolitan Statistical Areas) 외부 지역의 빈곤가구들은 제외하였다.

(2) 1968년 가중치의 계산

각 표본은 1968년 조사에서는 무응답이었다. 왜냐하면 재면접 표본들이 인구센서스 조사에서 무응답자들이었기 때문이다. 즉, 이들은 센서스국에 의해 조사된 응답자 이름과 주소를 OEO에게 공개하는데 대한 서명을 거부하였다. 또한 OEO로부터 SRC에게 일부 표본 주소를 전달하는데 실패하였다.

1968년 가중치를 결정하기위해 다음과 같은 3가지 확률을 계산하였다. (1) SRC 횡단면 표본에서 대한 확률 (2) 재조사 표본에 대한 확률 (3) 결합된 표본에 대한 확률이다.(횡단면과 재 면접 표본들을 결합하였을 때, 전체 비추정(overall ratio estimation) 방법은 사용하지 않았는데, 이는 모집단 종합에 관한 정보를 가지고 있지 않았기 때문이다.)

다음은 앞에서 언급한 3가지 추출확률에 대해 살펴보기로 한다.

① SRC 횡단면 표본에 대한 확률

횡단면 표본은 표본추출 당시의 전체 미국인에 대한 고정 비율(0.66/10,080)에 따라 추출되었다. 응답률은 지리적 위치, SRC 자체-대표성(self-representing)과 비대표지역, 자체-대표지역(self-representing area)의 중심지역과 기타지역, 비자체-대표지역(nonself-representing area)에서 SMSA와 non-SMSA에 따라 다양하며, 전체적으로 16가지의 응답률을 고려하였다. 횡단면 표본에 대한 면접확률은 "초기 추출률(initial selection rate)× 응답률(response rate)"으로서 (0.66/10,080)×(응답률)과 같다. 예를 들어 뉴욕의 맨하튼의 면접확률은 (0.66/10,080)×(61/100), 또는 1/25,037이 된다.

② 재면접 표본에 대한 확률

센서스국에 의해 원 표본(original sample)을 추출하기 위해 사용된 추출률은 8종이 있다. 357개의 PSU(primary sampling unit)가 두 가지 서로 다른 추출률을 사용하였다. 표본으로 선택된 1차 추출단위(PSU)내에서 이름과 주소를 접수받은 표본가구에 대해 재면접이 실시되었다. 접수율의 차이는 매우 다양하기 때문에 1차 추출단위에 따른 표본가구의 접수율에 대한 조정이 필요하며 이러한 조정 작업은 백인과 유색인종 가구에 대해 수행되었다. 재면접에 대한 무응답 조정은 자체-대표지역과 기타지역에 따라 4개 지역에 대해 수행되었다. 재면접 표본에 대한 확률은 "센서스 표본에 대한 초기추출률×센서스 부차추출률×SRC부차추출률×접수율×응답률" 로 정의된다. 예를 들어 뉴욕, 맨하탄에서 층1에 있는 백인가구의 재면접 확률은 (1/3,158)×(1/1)×(1/1)×(20/100)×(63/100)=1/25,063과 같다.

③ 결합된 표본에 대한 확률

결합된 표본은 다음과 같은 세 부분으로 고려할 수 있다.

㉠ 센서스국으로부터 전달받은 재 면접 표본
㉡ 남부의 SMSA와 non-SMSA로 부터 추출된 횡단면 표본에 있는 빈곤 가구
㉢ 횡단면 표본의 나머지 부분

세부분 중 처음 두 부분 ㉠과 ㉡은 동일한 모집단으로부터 두 개의 독립된 표본을 추출한 것이므로, 어떤 가구는 표본1 또는 표본2 또는 두 부분에서 모두 추출될 수 있다. 따라서 결합된 표본에서 면접확률은 "재 표본에서 면접확률 + 횡단면표본에서 면접확률 - 두 확률의 곱"으로 정의된다.

예를 들어 맨하튼에서의 추출확률은 (1/25037)+(1/25063)-(1/25037×1/25063)=(1/12525)가 된다. 따라서 맨하튼의 가중치는 이 확률의 역수로서 12,525가 된다.

### 2) 캐나다의 SLID(Survey of Labor and Income Dynamics)[2)]

1993년을 기준년도로 시작한 SLID는 개인들에 대한 종단면패널 조사이다. 이 패널의 목적은 개인들의 경제적 풍요(well-being)의 변화와 이러한 변화의 요인들, 특별히 인구학적인 측면과 가구의 특성, 노동시장 활동 등을 주요인자(key factors)로 측정하고자 하였다.

(1) 패널설계

SLID 표본은 2개의 패널이 동시에 존재하며, 각 패널은 6년간 지속된다. 1차 패널은 1993년 1월에 추출되었고, 1992년 12월 31일 현재 캐나다의 10개 지역을 포괄한다. 2차 패널은 1996년 1월에 추출되었으며, 1995년 12월 31일 현재 10개 지역의 모집단을 포괄한다. 두 표본 모두 "인디언 보호지역에 거주하는 자", "군복무자", "6개월 이상 기관이나 시설에 거주하는 자"는 모집단에서 제외하였다. SLID 패널은 매 3년마다 변동하게 되는데, 추가적으로 이러한 규칙에 의한 패널 구성 형태를 도식화하면 [그

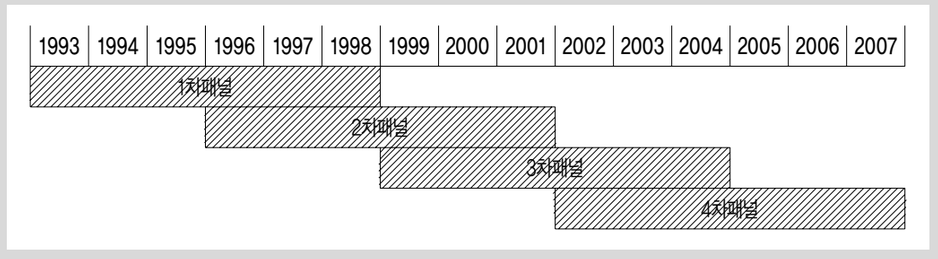
2) Lecvesque, I., & Franklin, S. (2000). Longitudinal and Cross-Sectional Weighting of Survey of Labour and Income Dynamics 1997 Reference year. Statistics of Canada Research Paper series.
Naud, J. F. (2004). Combined-panel longitudinal weighting Survey of Labour and Income Dynamics. Statistics of Canada Research Paper series.
Latouche, M., Dufor, J., and Merkouris, T. (2000). Corss-sectional Weighting: Combining Two or More Panels. Statistics of Canada Research Paper series.

림 1]과 같이 표현된다. 즉, 1차 패널은 1993년부터 1998년까지이고, 2차 패널은 1996년부터 2001년까지이다. 따라서 1996년부터 1998년까지는 2개의 패널이 동시에 존재하게 된다. 1999년부터는 1차 패널이 종료되고 2차와 3차 패널이 동시에 존재하게 된다. 2차 패널의 종료시점은 2001년 12월 31일 이며, 2002년 1월에 4차 패널이 시작된다.

이와 같이 연동 패널(rotation panel)을 사용하게 된 주된 이유를 살펴보면, 조사당시의 횡단면 표본의 대표성을 확보하며, 장기간의 패널유지로 인한 패널 마모효과를 감소시키며, 패널의 응답 부담을 경감시키고자 함이다. 각 개인은 1년에 총 2회의 설문을 수행해야 하는데, 1월에는 노동시장관련 내용을 질문하고, 5월에는 임금과 소득에 관한 설문을 수행한다. 응답 부담을 경감시키기 위해 응답자는 해당 소득관련 설문에 응답하지 않는 대신 5월에 캐나다 통계청(Statistics Canada)이 개인의 소득자료(Revenue Canada tax file)를 사용하도록 허가 한다. 각 패널은 캐나다 노동력조사(Labor Force Survey: LFS)의 1월과 2월에 조사된 표본으로부터 약 15,000가구를 표본으로 추출한다. 이때, LFS는 층화다단계 추출방법을 사용하며, 최종 추출단위는 거처(dwelling)이다. 따라서 표본으로 선정된 거처안의 모든 가구원은LFS 표본이 된다. LFS 표본으로 선정된 가구는 6개월 간 표본에 남아있게 되며, 매월 전체 표본의 1/6씩 새로운 표본으로 대체된다. SLID의 1차 패널은 LFS의 6개의 연동 그룹 중 2개의 그룹(20,000가구)을 먼저 선정하고, 이들의 약 88%인 17,000가구가 SLID 표본으로 동의하였다. 이들 중 2,000가구를 제외한 15,000가구를 1차 패널로 구성하였다. SLID의 목적에 부합하기 위해 관심단위는 개인이다. 왜냐하면, 가구의 특성상 시간의 흐름에 따른 가구의 변동은 매우 적어 종단면 분석을 위한 대상으로 고려하는 것은 적절하지 않기 때문이다. 횡단면적인 분석 목적으로서 가구와 개인은 모두 분석단위로 고려하였다. SLID의 패널 표본으로 선정되면, 패널로 선정된 가구의 모든 구성원은 연령에 무관하게 패널의 종단면 표본으로 남게 된다. 이들은 심지어 이사를 가거나, 사망하거나, 혹은 시설에 입소하거나, 군복무를 한다하더라도 패널지속기간인 총 6년간 종단면 표본으로 고려된다.

SLID는 종단면 개인의 특성뿐만 아니라 가구의 특성과 연관되어 있다. 적어도 한 종단면 개인과 같이 살고 있는 모든 사람은 면접대상이 된다. 그러므로 주어진 해당 년도의 횡단면 표본은 조사기준년도의 12월 31일에 모든 종단면 개인과 그 당시 함께 살았던 모든 사람으로 구성된다. 만일 기준년도 당시 인디언 보호구역, 민간시설, 군대에 6개월 이상 거주한 것을 제외하고, 10개의 지방(province)중 한 지역에 살았던 사람이라면, 주어진 년도의 12월 31일 기준으로 조사대상에 포함된다. 종단면의 범위에서 보면 모든 횡단면 개인은 조사대상에 포함된다. 종단면 표본이 아닌 조사대상자를 동거인(cohabitants)라 한다. 이러한 표본 구성을 바탕으로 각 년도의 횡단면과 종단면 가중치를 고려해야 한다. 종단면가중치(longitudinal weight)는 종단면 표본이 추출되었을 당시의 10개 지방 모집단에 대한 대표성을 확보하기 위해 고려한 가중치이며, 횡단면 가중치(cross-sectional weight)는 주어진 기준년도의 12월 31일의 10개 지방 모집단에 대한 대표성을 확보하기 위한 가중치이다. 1차 패널의 종단면 표본으로부터 구한 추정치는 1992년 12월 31일의 10개 지방의 모집단에 대한 값이며, 2차 종단면 표본으로부터 구한 추정치는 1995년 12월 31일의 10개 지방의 모집단에 대한 값이다. 1997년 기준년도의 횡단면 표본으로부터 구한 추정치는 1997년 12월 31일 지역 모집단에 대한 값이다. 1993년 1월에 39,745명의 개인이 1차 패널의 종단면 표본이었고, 1996년 1월에 43,547명이 2차 패널의 종단면 표본이었다. 1997년 기준의 횡단면 표본은 81,090명의 개인이었고, 이중 70,372명이 종단면 표본이었다.

**그림 1. SLID의 패널 구성 형태**




(2) 패널 가중치 계산

SLID의 가중치는 우선 종단면 가중치와 횡단면 가중치로 구분할 수 있으며, 각각의 가중치를 보다 세분하여 살펴보는 것이 바람직 하지만, 개략적인 가중치 산정 방법만을 소개하기로 한다.

① 종단면 가중치

초기 종단면 가중치는 가구추출확률의 역수와 같고, 모든 종단면 가구원은 동일한 종단면 가중치를 가진다. 1차 패널과 2차 패널의 초기 종단면 가중치는 다음과 같이 계산된다.

1차 패널의 초기 종단면 가중치:

$$w_{int,\,p1}=w_{LFS}\times 3\times 1.19 \qquad (2.1)$$

2차 패널의 초기 종단면 가중치:

$$w_{int,\,p2}=w_{LFS}\times 3 \qquad (2.2)$$

여기서 $w_{LFS}$는 무응답이 조정된 LFS 가중치이며, 숫자 3은 LFS 연동그룹의 추출확률의 역수로서(LFS의 6개 연동그룹 중 2개를 선택함) 2/6의 역수이다. 또한 1.19는 1차 패널에서 초기 면접에서 응답자들의 추출률(=1/0.84)의 역수이다.

다음으로 무응답 조정을 위해 SLID는 개인의 무응답 조정승수(adjustment factors)를 응답자 동질그룹(RHG)의 응답률의 역수로 정의하였다. 적절한 무응답 승수를 계산하기위해 무응답을 정의하고, 이에 필요한 설명변수를 선택한 다음 로지스틱 회귀를 이용하여 적절한 응답자 그룹을 형성하여 무응답 승수를 구성하였다.

이러한 과정으로부터 다음과 같이 무응답이 조정된 가중치를 정의할 수 있다.

$$w_{ADJST}=\begin{cases}0 & \text{, 무응답가구의 개인}\\ w_{int} & \text{, 어린이 또는 횡단면표본에서 제외된 개인}\\ w_{int}/R_{RHG} & \text{, 응답가구의 개인}\end{cases} \quad (2.3)$$

여기서 $w_{ADJST}$는 무응답이 조정된 종단면 가중치이고, $w_{int}$는 초기 종단면 가중치로서 (2.1)과 (2.2)에서 정의되었다. 또한 $R_{RHG}$는 RHG에서 가중된 응답률이다.

SLID는 횡단면적인 소득의 추정치와 분산을 추정치에 영향을 주는 영향력 관찰치에 의한 극단가중치를 조정하였다. 영향력 관찰치에 대한 조정값은 0과 1사이의 값으로 계산되며, 영향력 관찰치에 대한 조정가중치는 다음과 같다.

$$w_{infl}=w_{ADJST}\times\beta_{infl} \quad (2.4)$$

여기서 $w_{infl}$은 영향력 관찰치에 대한 가중치가 조정된 종단면 가중치이고, $w_{ADJST}$는 무응답이 조정된 종단면 가중치, $\beta_{infl}$는 영향력 관찰값에 대한 조정 승수이다.

다음으로 사후층화를 2개의 패널에 대해 독립적으로 수행하였다. 사후층화로부터 계산된 종단면 가중치는 다음과 같은 식에 의해 계산된다.

$$w_{post}=w_{infl}\times t_L/\sum_L w_{infl} \quad (2.5)$$

여기서 $t_L$은 사후층 $L$에 대한 종합이며, 이 값은 캐나다 통계청에서 구한 값이다.

최종적으로 사후층화 가중치에 대해 가구와 개인의 정보 보호를 위해 일정 수준의 잡음(noise)가중치를 결합하여 외부 자료로 공개하고 있다. 따라서 사후층화 조정 가중치에 잡음가중치가 고려된 최종적인 종단면 가중치는 다음과 같다.

$$w_{noise}=\begin{cases}w_{post}\pm(e\times a) & \text{, 가중치가 같은 동일 가구의 개인}\\ w_{post} & \text{, 그 외의 모든 경우}\end{cases} \quad (2.6)$$

여기서 e는 U(0,1)에서 발생시킨 확률잡음(random noise)이고, a는 종단면 잡음(longitudinal noise)이다.

② 횡단면 가중치

SLID의 횡단면 가중치는 특정한 기준년도에서 추정치를 계산하기 위해 2개의 패널이 결합되는 형태의 가중치로 나타난다. 횡단면적 가중치의 목표모집단은 기준년도의 12월 31일 현재 10개 지방 거주하는 자로서 보호지역, 또는 시설 및 군대시설 거주자는 제외된다. 모든 종단면 개인과 그들과 동거하는 자는 횡단면 표본으로 간주한다. 초기 횡단면 가중치는 무응답이 조정된 종단면 가중치로서 각 시점에서 모집단을 대표한다.

횡단면가중치 조정의 첫 번째 단계는 무응답이 조정된 종단면 가중치를 배분 승수(allocation factor)를 이용하여 2개 패널 표본을 결합한다. 이때 패널 분배 승수는 다음과 같이 계산된다.

$$p_1=\frac{n_1}{n_1+n_2\frac{d_1}{d_2}}\ ,\quad p_2=1-p_1 \quad (2.7)$$

여기서 $p_1$과 $p_2$는 각각 1차 패널과 2차 패널의 패널 배분승수이며, $n_1$과 $n_2$는 각각 1차 패널과 2차 패널에서 16세 이상의 응답한 종단면 개인의 수이고, $d_1$과 $d_2$는 각각 조정된 가중치를 적용한 추정치의 분산과 단순임의 표본의 분산의 비로서 각각 1차 패널과 2차 패널의 설계효과를 나타낸다.

결과적으로 1차와 2차 패널을 결합하여 조정된 횡단면 가중치는 다음과 같이 정의된다.

$$w_{com}=\begin{cases}p_1 w_{ADJST} & \text{, 1차 패널의 개인}\\ (1-p_1)w_{ADJST} & \text{, PAF의 적용에서 제외되지 않은 2차 패널의 개인}\\ w_{ADJST} & \text{, PAF의 적용에서 제외된 2차 패널의 개인}\end{cases} \quad (2.8)$$

한편 개인횡단면가중치는 다음과 같다.

$$w_{share}=\begin{cases}w_{com} & \text{, 가구의 모든 개인이 종단면이거나, 초기 동거인이 없는 가구의 종단면 개인.}\\ \sum_h w_{com}\frac{n_{L,h}}{n_{L,h}+n_{IPh}} & \text{, 적어도 1인 이상의 동거인이 있고, 초기에 없었던 동거인이 있는 가구의 동거인.}\end{cases} \quad (2.9)$$

여기서 $n_{L,h}$는 가구 h에 있는 종단면 개인의 수이고, $n_{IPh}$는 가구 h에 있는 초기 동거인의 수이다.

이와 함께 통합 배분된 횡단면가중치는 다음과 같다.

$$w_{share}=\frac{\sum_h w_{com}}{n_{L,h}+n_{IPh}} \quad (2.10)$$

또한 지방 내에서의 이주에 따른 가중치 조정을 수행해야 하는데, 이러한 이주효과를 반영한 조정된 가중치는 다음과 같다.

$$w_{mig}=w_{share}\times a_{mig} \quad (2.11)$$

여기서 $a_{mig}$는 이주효과에 대한 조정 승수로서 가중치의 평균이나, 중위 가중값, 가중치의 4분위 등을 사용할 수 있으며, SLID에서는 95분위수를 사용했다.

종단면 가중치와 마찬가지로 횡단면 가중치의 영향력 관찰치에 대한 가중치 조정은 다음과 같이 이주효과를 조정한 가중치에 영향력 승수를 고려하여 조정할 수 있다.

$$w_{infl}=w_{mig}\times\beta_{infl} \quad (2.12)$$

또한 종단면가중치의 사후층화 조정과 같은 방법으로 식(2.12)의 가중치를 조정하여 사후층화가중치를 구할 수 있다. SLID에서는 지역×성별×연령 그룹으로 분할하여 횡단면 가중치에 대한 사후층화 조정을 수행하였다. 마지막으로 사후층화 조정된 가중치 $w_{post}$에 동일한 가중치를 갖는 동일한 가구의 개인들의 가중치로 $w_{noise}$를 다음과 같이 정의한다.

$$w_{noise}=\begin{cases}w_{post}\pm(e\times a) & \text{, 가중치가 같은 동일 가구의 개인}\\ w_{post} & \text{, 그 외의 모든 경우}\end{cases} \quad (2.13)$$

## 3. 국내 패널조사의 가중치 조정 방법

### 1) 국내 패널조사의 현황

외국 특히 미국, 캐나다 영국 등의 패널조사가 1960년대부터 수행되기 시작하였으며, 국내에서는 1993년부터 1998년까지 5년간 수행된 대우재단의 대우패널이 국내 패널조사의 효시이다. 이를 기반으로 1998년 한국노동연구원의 노동패널이 시작된 후 청년패널(2001), 저소득

층 자활패널(2002), 청소년패널(2003), 복지패널(2004), 연금패널(2005), 고령자패널(2006), 교육고용패널(2005), 장애인패널(2007), 인구패널(예정), 소비자패널(2004), 여성패널(2007) 등으로 다양하고 많은 패널조사들이 수행되고 있거나 혹은 계획되고 있다.

3절에서는 국내에서 수행되고 있는 각종패널들 중 한국가구경제패널과 노동패널의 가중치 조정 방법을 중심으로 살펴보고자 한다.

### 2) 한국가구경제 패널조사(KHPS)

대우경제연구소에서 아시아권에서 처음으로 실시한 패널조사로서 1993년부터 매1년 주기로 수행된 패널조사로서 1998년까지 6년간 지속되었다.

#### (1) 표본추출

제주도를 제외한 전국의 일반가구를 모집단으로 고려하고, 이때, 외국인가구, 특수시설(고아원, 양로원, 기도원 등)의 시설은 조사대상에서 제외하였다. 목표 표본수는 4,500가구로 설정하고, 조사 완료된 4,00가구는 1가구당 2,747가구를 대표하게 된다. 1차 추출단위(PSU)는 시 · 군 · 구로 하고, 2차 추출단위(SSU)는 읍 · 면 · 동, 3차 추출단위(TSU)는 통 · 반 · 리로 하는 3단계집락추출법을 사용하였다. 각 단계에서는 계통추출법을 사용하여 집락을 추출하고, 최종적으로 표본집락에서는 임의추출에 의해 표본가구를 선정하였다. 6대 도시에서는 통 · 반 · 리 당 8가구를 표본으로 선정하였고, 기타 지역의 통 · 반 · 리에서는 7가구를 표본가구로 추출하였다. 패널조사에 사용된 추출틀은 다음과 같이 5가지를 이용하였다.

◯ 전국의 시 · 군 · 구 리스트 (가구수 포함, 1992년 행정구역총감),
◯ 전국의 읍 · 면 · 동 리스트 (1990년 인구센서스, 분할된 동지역은 1992년행정구역총감),
◯ 추출된 읍면동의 통반리 리스트(통반리의 가구수 포함, 1993년 4~5월 조사),
◯ 추출된 통반리별 주소리스트(통반리의 가구수 포함, 1993년 4~5월 조사),
◯ 추출된 주소의 가구주 리스트 (전화번호 포함, 1993년 6~7월 조사),

통계청 조사구를 기반으로 한 추출틀은 인구이동이 심한 경우 센서스 조사시점과 패널조사시점의 차이로 인구수 또는 가구수의 변동이 반영되지 않기 때문에 본 조사의 추출틀을 조사시작 약 1~3개월 전에 작성한 통반리 번지의 가구수 리스트를 추출틀로 이용하였다.

#### (2) 가중치조정 방법

대우패널의 가중치는 1차년도 에만 작성되었고, 가중치의 형태는 각 추출단계별 추출확률의 역수로 계산되었다.

◯ 추출확률:

$$p_{ijkl} = (a \times p_i)(b_i \times p_{ij})(c_{ij} \times p_{ijk})\frac{n_{ijk}}{N_{ijk}} = \left(a \times \frac{N_i}{N}\right)\left(b_i \times \frac{N_{ij}}{N_i}\right)\left(c_{ij} \times \frac{N_{ijk}}{N_{ij}}\right) = (ab_i c_{ij}) \times \frac{N_{ijk}}{N} \quad (3.1)$$

여기서 $a$는 표본 PSU의 수, $b_i$는 $i$번째 표본 PSU 내의 표본 SSU의 수, $c_{ij}$는 $i$번째 PSU, $j$번째 SSU 내의 표본 TSU의 수이다. $n_{ijk}$는 $i$번째 PSU, $j$번째 SSU, $k$번째 표본TSU내의 가구수이며, $N_{ijk}$는 $i$번째 PSU, $j$번째 SSU, $k$번째 TSU내의 총가구수이다. $p_i$는 $i$번째 PSU가 표본으로 추출될 확률, $p_{ij}$는 $i$번째 PSU의 $j$번째 SSU가 추출될 확률, $p_{ijk}$는 $i$번째 PSU, $j$번째 SSU내의 $k$번째 TSU가 추출될 확률이다.

◯ 가중치:

$$W_{ijkl} = constant \times 1/p_{ijkl} \quad (3.2)$$

### 3) 한국 노동 패널조사(KLIPS)[3)]

한국 노동패널은 국내 패널조사 중 대표적인 조사로서 2006년 12월 현재 8차년도 패널조사가 이루어진 상태이다. 노동패널은 노동시장과 관련된 기초 자료를 생산하여 고용정책의 수립과 향후 평가에 이용하고자 1998년에 시작된 조사이다. 비농촌지역에 거주하는 한국의 가구 및 가구원을 대표하는 5,000가구(17,505명)의 개인을 대상으로 매년 개인의 경제활동, 노동시장이동, 소득활동 및 소비, 교육 및 직업훈련, 사회생활 등에 관해 추적 조사하는 종단면 조사이다. 1998년 1차 년도에 조사된 5,000가구의 가구원들은 비농촌지역의 거주인구를 대표한다. 2차년도 조사에서는 1차년도 원표본가구원들에 대한 매년조사가 수행되고, 이사 및 분가한 경우에는 추정조사를 수행한다. 원표본가구에서 출생한 자녀들은 표본가구원으로 추가되며, 원표본가구원 또는 그 자녀가 결혼 등으로 배우자가 있는 경우 그 배우자도 혼인관계가 지속되는 한 조사대상으로 포함한다.

#### (1) 표본추출

1995년 인구센서스의 10% 표본조사구를 모집단으로 한다. 표본추출과정에서 1995년 인구센서스의 10% 표본조사구중에서 5,000가구를 직접 추출하지 않고, 1997년 고용구조특별조사('97고특)의 결과와 상호 비교를 위해 추출된 표본이 '97고특 조사의 표본에 속하도록 하기위해 지역별로 층화한 후 층 내에서는 '97고특 조사의 층화 기준을 적용하였다. 조사구의 추출방법은 계통추출방법을 사용하였고, 제주도를 제외한 시부만을 대상으로 1,000개의 조사구를 선정하였고, 각 조사구에서 '97고특 조사의 조사대상가구중 5~6가구를 단순임의 추출하였다. 계통 추출된 조사구가 '97고특 조사구가 아닌 경우에는 가장 가까운 '97고특 조사구를 표본조사구로 선정하였다. 각 조사구내에서 특정 가구가 추출될 확률은 조사구내의 총가구수, '97고특에서 성공한 가구수, KLIPS에서 추출한 가구수에 따라 결정된다. 약 5개월간의 조사('98.6.2~'98.10.13)로 부터 5,000가구 약 17,505명중 면접에 성공한 가구원은 13,317명이고, 가구수로는 75.3%인 3,773가구가 면접에 성공하였고, 나머지 24.5%인 1,227가구는 대체하였

3) KILPS 1차년도 – 5차년도 USER's GUIDE. 한국노동연구원.
강석훈(1999). KLIPS의 1차웨이브 가중치 부여방안에 대한 연구. 한국노동패널연구.
강석훈(2003). KLIP의 가중치 부여방안 연구. 한국노동패널연구.

다. 이러한 표본대체의 주요한 사유로는 주소불명이 약 1.7% 이사로 인한 추적불가가 6.1%, 이사 후 추적하였으나 응답거절인 경우가 0.6%, 응답을 강력히 거절하는 경우 11.8%(591가구)이고 기타이유가 4.3%로 나타났다.

(2) 가중치 조정방법

1차 년도에는 2단층화집락계통추출법을 사용하여 표본가구를 선정하였다. 1단계에서는 1995년 인구센서스의 10% 표본조사구중에서 도시지역 조사구 19,025개를 먼저 선정하고, 다음으로 이중에서 1,000개의 조사구를 선정하였다. 이 과정에서 최종 표본으로 선정된 조사구는 951개 조사구이다.

1차년도의 조사에서는 1차 패널에 대해서 횡단면 조사이므로 통상적인 횡단면 조사의 가중치를 그대로 적용하게 된다. 일반적인 가중치 조정 단계는 먼저 추출확률을 계산하고, 다음으로 무응답을 조정하고, 마지막으로 사후층화조정을 수행하게 된다. 앞에서 언급한 외국 패널의 경우를 살펴보면 개별 가중치 산정 과정은 다양하지만, 큰 틀에서는 세 단계로 가중치를 조정한다. KLIPS의 경우 개인단위의 사후층화는 통계청의 인구추계자료가 있으나, 세부적인 정보가 없는 관계로 추출확률과 응답률에 의한 가중치 계산만 수행하였다. KLIPS는 SLID와는 다르게 가구 및 개인을 분석단위로 하기 때문에 개인별 가중치와 가구별 가중치를 각각 구해야 한다.

① 1차 웨이브의 가중치 산정 방법

1차 웨이브에서는 가구와 그 가구에 속한 가구원의 추출확률이 동일하기 때문에 개인별 가중치와 가구가중치는 같은 값을 가진다. 추출확률과 응답확률을 모두 고려한 가구가중치는 다음과 같이 계산된다.

○ 서울 및 6대광역시

0.1×(표본조사구수/도시조사구수)×('97고특조사가구수/ED 내 전체가구수 )×(총접촉가구수 /'97고특조사가구수)×(최종조사가구수/총 접촉가구수)

○ 도의 동부

0.1×(해당 도의 동부 표본조사구수/해당도의 동부 조사구수)×('97고특조사가구수/ED 내 전체가구수 )×(총접촉가구수/'97고특조사가구수)×(최종조사가구수/총 접촉가구수)

○ 도의 읍면부

0.1×(해당도의 시에 속한 표본읍면부 조사구수/해당도의 시에 속한 읍면부 조사구수)×('97고특조사가구수/ED 내 전체가구수 )×(총접촉가구수/'97고특조사가구수)×(최종조사가구수/총 접촉가구수)

이때 표본으로 선정된 모든 가구원이 응답하였으므로 동일가구내의 가구원 가중치는 해당 가구가중치와 같게 된다.

② 2차 웨이브이후의 가중치 산정방법

KLIPS에서는 2차 웨이브 이후 Duncan(1995)[4] 의 가중치 부여방법을 적용하고 있다. Duncan의 방법은 우선 미국의 대표적인 패널조사인 PSID에서 적용한 방법으로 논리적으로 일관성이 있기 때문이다. 이는 하나의 가구단위 분석이 현실적으로 어렵다는 점을 고려한 방법이다. 이에 대한 절차를 살펴보면 다음과 같다.

○ 1단계: 초기웨이브에서 가구가중치를 구한다. 이때 추출확률과 조사과정의 응답률을 함께 고려하고, 마지막으로 사후층화조정까지 수행한 가중치를 구한다.

○ 2단계: 초기웨이브에서 구한 가구가중치를 연령이나 응답여부에 관계없이 모든 가구원의 가구원가중치로 사용한다. 이 가중치는 15세 이상의 가구원뿐만 아니라 15세 이하의 모든 가구원에게도 동일하게 적용한다.

○ 3단계: 2차 웨이브 이후부터는 가구원들의 상이한 응답률을 이용하여 가구원들의 가중치를 조정한다. 이때 2차 웨이브에서는 해당가구에 존재하지만, 1차 웨이브에서는 응답하지 않았던 비표본가구원이나 1차 웨이브이후 새로 태어난 자녀의 경우에는 개인차원의 무응답조정과정에 포함시키지 않는다.

○ 4단계: 2차 웨이브에서 산출된 개인가중치의 가구 내 평균을 이용하여 2차 웨이브의 가구가중치를 산출한다. 결혼, 동거 등의 사유로 새롭게 진입한 비표본가구원에게는 0의 가중치를 부여한다.

4) Duncan, G. (1995). A Simple Method for Weighting in Household Panel Survey. Working paper, Northwestern University.

## 4. KOWEPS의 가중치 조정 방법

### 1) KOWEPS의 개요

한국복지패널 조사는 외환위기 이후 빈곤층(또는 working poor)및 차상위계층의 가구형태, 소득수준, 취업상태가 급격히 변화하고 있어, 이들의 규모와 상태변화를 동적으로 파악하여 정책지원을 위한 기초 자료를 생산하고, 소득계층별 경제활동 상태별, 연령별 등 각 인구집단의 생활실태와 복지욕구 등을 역동적으로 파악하고 정책의 효과를 평가함으로써 정책형성과 피드백에 기여하기위한 조사이다. 따라서 도시의 일반 가구를 대상으로 하는 KLIPS와는 지향하는 정책관점이 서로 다르다고 할 수 있으며, 조사대상가구는 일반가구와 저소득층 가구를 각각 50%씩 추출하여 이들에 관한 다양한 복지실태와 욕구 등에 관한 조사를 하고자 한다. 한국복지패널조사는 가구용 설문지와 가구원용 설문지(15세 이상), 아동용 설문지로 구성되어있으며, 가구원용 설문지는 15세 이상 중 · 고등학생을 제외한 경제활동인구 모두에 대해 응답을 받도록 하고 있다. 또한 부가조사로서는 1차 웨이브에서는 아동, 2차는 복지인식, 3차는 장애인을 대상으로 설문을 수행하도록 계획하였다.

### 2) 표본추출과 가중치 조정

(1) 표본 추출

한국복지패널조사는 한국보건사회연구원의

"2006 국민생활실태조사(2006.6.30~10.1)"의 517개 표본 조사구를 바탕으로 7,000가구를 소득기준으로 중위소득 60%이하인 3,500가구와 중위소득 60% 이상인 3,500가구를 각각 표본으로 추출하여 조사를 수행하였다. 이때 저소득층과 일반 가구층을 구분하기 위해 가구의 경상소득의 중앙값을 기준으로 60%이상을 일반가구, 그 이하를 저소득층 가구로 나누어 각각 추출하였다.

〈표 1〉의 지역별 표본가구비율을 살펴보면 전체적으로 일반 및 저소득 가구의 비율은 53.6:46.4 이며 이는 원래의 표본으로부터 저소득층을 과대표집하였기 때문이며, 이 비율은 가중치 조정을 통해 72.6:27.4 으로 재조정하여 추후 가구분포로 이용하였다.

**그림 2. 한국복지패널 표본추출절차**


**표 1. 가구분포현황**

(단위: %)

| 지 역 | 표본 가구 비율* | | | 추계 가구 비율** | | |
|---|---|---|---|---|---|---|
| | 계 | 일반 | 저소득 | 계 | 일반 | 저소득 |
| 전국 | 100.0 | 53.6 | 46.4 | 100.0 | 72.6 | 27.4 |
| 서울 | 18.9 | 23.4 | 13.7 | 21.6 | 23.4 | 16.8 |
| 부산 | 7.5 | 7.3 | 7.7 | 7.7 | 7.4 | 8.5 |
| 대구 | 6.0 | 5.4 | 6.7 | 5.3 | 4.6 | 7.1 |
| 인천 | 6.2 | 6.5 | 5.8 | 5.5 | 5.5 | 5.6 |
| 광주 | 3.5 | 3.2 | 3.7 | 2.6 | 2.5 | 3.1 |
| 대전 | 2.9 | 3.5 | 2.3 | 2.9 | 3.1 | 2.3 |
| 울산 | 2.9 | 3.3 | 2.4 | 2.2 | 2.3 | 2.0 |
| 경기 | 16.0 | 18.6 | 13.0 | 22.3 | 23.2 | 20.0 |
| 강원 | 3.3 | 3.1 | 3.6 | 3.1 | 3.1 | 3.0 |
| 충북 | 3.1 | 3.0 | 3.3 | 3.0 | 2.9 | 3.3 |
| 충남 | 4.5 | 4.3 | 4.9 | 4.1 | 3.9 | 4.9 |
| 전북 | 4.9 | 3.7 | 6.3 | 3.6 | 3.0 | 5.3 |
| 전남 | 5.3 | 2.9 | 8.1 | 3.3 | 2.1 | 6.5 |
| 경북 | 7.0 | 4.3 | 10.0 | 4.8 | 4.0 | 6.7 |
| 경남 | 6.9 | 6.4 | 7.5 | 6.7 | 7.9 | 3.7 |
| 제주 | 1.1 | 1.2 | 1.1 | 1.2 | 1.1 | 1.4 |

주: * 가중치가 적용되지 않은 비율
** 가중치가 적용된 비율

(2) 가중치 조정과정

2006년 국민생활실태조사 본 조사를 바탕으로 한국복지패널 표본을 추출하였기 때문에 기본가중치(base weight)로는 국민생활실태조사 가중치를 사용하였다. 이를 위해 다음과 같은 기호를 정의하자.

$h = 1, 2, \cdots, H$ : $h$층을 나타내는 첨자

$i = 1, 2, \cdots n_h$ : $h$번째 층의 $i$번째 표본 조사구를 나타내는 첨자

$j = 1, 2, \cdots n_{hi}$ : $h$번째 층의 $i$번째 표본조사구내 $j$번째 가구를 나타내는 첨자

$N_h$ : $h$층의 모집단 조사구수

$n_h$ : $h$층의 표본조사구 수

$M_{hi}$ : $h$층의 $i$번째 조사구의 모집단 가구 수

$m_{hi}$ : $h$층의 $i$번째 조사구의 표본 가구 수

$r_{hi}$ : $h$층의 $i$번째 조사구의 응답 가구 수

$w_{hij}$ : $h$층의 $i$번째 조사구의 $j$번째 표본 가구에 대한 가중값

$y_{hij}$ : $h$층의 $i$번째 조사구의 $j$번째 표본 가구에 대한 관측값

각 표본가구가 표본으로 추출될 확률의 역수로서 가구별 확대 승수를 산출하여 이를 기본가중치로 고려한다. 다음으로 각 층내의 조사구내 표본가구들의 무응답으로 무응답조정 가중치를 통합하여 다음과 같은 무응답조정가중치를 산정한다.

$$w_{hij} = \frac{10}{9} \cdot \frac{N_h}{n_h} \cdot \frac{M_{hi}}{m_{hi}} \cdot \frac{m_{hi}}{r_{hi}} = \frac{10}{9} \cdot \frac{N_h}{n_h} \cdot \frac{M_{hi}}{r_{hi}} \quad (4.1)$$

여기서 (10/9)는 인구조사 90% 조사구를 이용하여 이를 100%로 확대하기위한 확대 상수이다.

식(4.1)의 가중치는 기본가중치에 조사구별 무응답을 조정한 가중치로서 소득층별 가중치는 저소득층 가구를 과대표집하였기 때문에 다음과 같은 절차로 재조정하였다.

$$w^*_{hij} = w_{hij} \cdot \frac{M_h}{\widehat{M}_h} \quad (4.2)$$

여기서 $M_h = \sum_i M_{hi}$는 $h$층의 모집단 총 가구수로서 2005년 지역별 가구추계치이며, $\widehat{M}_h = \sum_{i,j} w_{hij}$이다.

(3) 가중치 분석

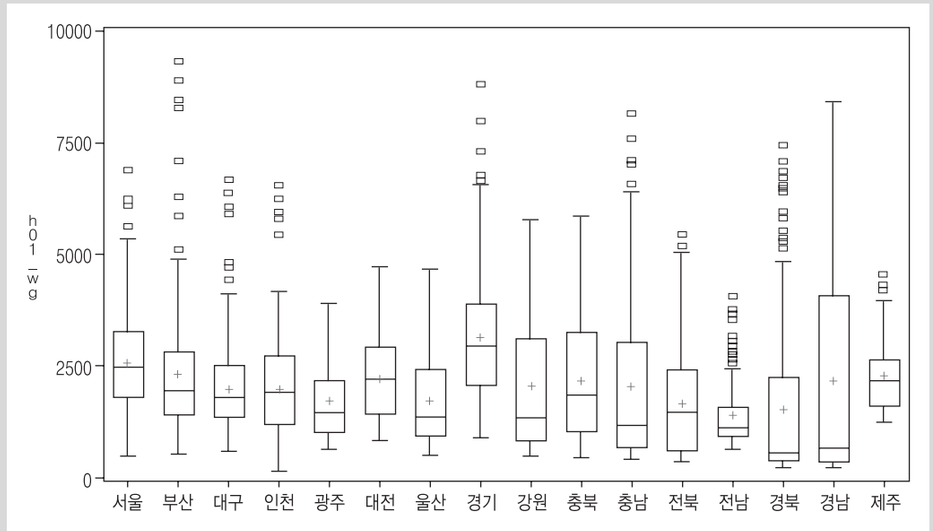
국민기초생활실태조사로부터 조사구당 무응답가구에 대한 가중치를 조정하고,최종적으로 제외조사구에 대한 조정을 마친 후의 가중치에에 대한 기술 통계값을 산출한 결과가 <표 2>와 같다. 평균적으로 1가구가 약 2,246개의 가구를 대표하며, 이들의 변동은 전국적으로 0.72%로 매우 안정적으로 나타났다.

**표 2. KOWEPS 1차 조사 최종가중치의 기술통계값**

| 가중치 변수 | 평균 | 표준오차 | CV(%) |
|---|---|---|---|
| h01_wg | 2246.48 | 16.234243 | 0.72 |

다음의 〈표 3〉은 KOWEPS의 최종 가중치의 지역별 합계인 추계가구수와 2005년 인구센서

**그림 3. 지역별 가중치 분포**




스의 지역별 가구수 통계를 비교한 결과이다. 최종가중치와 센서스자료의 지역별 분포를 살펴보면 모든 지역에 대해 센서스 가구수와 거의 유사함을 알수 있으며, 경기지역의 경우 1.3% 정도 과대 추계된 경향이 있으나, 전체적인 분포에는 영향을 주지 않는 것으로 나타났다.

**표 3. KOWEPS의 지역별 추계가구수와 센서스간의 비교(2005년 기준)**

| 지역 | 추계 가구 수 | 비율 | 센서스 가구 수 | 비율 |
|---|---|---|---|---|
| 전국 | 15,887,128 | 100.0 | 15,887,128 | 100.0 |
| 서울 | 3,433,110 | 21.6 | 3,309,890 | 20.8 |
| 부산 | 1,221,437 | 7.7 | 1,186,378 | 7.5 |
| 대구 | 838,412 | 5.3 | 814,585 | 5.1 |
| 인천 | 872,007 | 5.5 | 823,023 | 5.2 |
| 광주 | 419,680 | 2.6 | 460,090 | 2.9 |
| 대전 | 464,025 | 2.9 | 478,865 | 3.0 |
| 울산 | 351,424 | 2.2 | 339,095 | 2.1 |
| 경기 | 3,541,342 | 22.3 | 3,329,177 | 21.0 |
| 강원 | 487,526 | 3.1 | 520,628 | 3.3 |
| 충북 | 480,457 | 3.0 | 505,203 | 3.2 |
| 충남 | 659,152 | 4.1 | 659,871 | 4.2 |
| 전북 | 578,295 | 3.6 | 619,958 | 3.9 |
| 전남 | 530,542 | 3.3 | 666,319 | 4.2 |
| 경북 | 755,796 | 4.8 | 938,840 | 5.9 |
| 경남 | 1,066,942 | 6.7 | 1,056,007 | 6.6 |
| 제주 | 186,982 | 1.2 | 179,199 | 1.1 |



〈표 3〉 계속

(4) 가중치 효과분석

통상적으로 단순임의 추출에 의한 표본은 표본단위들이 자체가중치를 부여 받기 때문에 별도의 가중치조정과정이 필요 없다. 또한 층화집락 추출과 같은 복합표본추출 설계 하에서 각 층별 규모에 따라 비례배분을 적용하면, 추정과정에서 자체가중의 효과가 있기 때문에 별도의 가중치 조정이 필요 없게 된다. 그러나 KOWEPS 와 같은 복합 표본추출설계를 적용하여 조사가 수행되면 각 층별, 집락별 응답률의 차이와 표본의 차이로 인한 과소추정이 발생하기 때문에 반드시 추정과정에서 가중치 조정을 필요로 한다. 하지만 이와 같이 가중치를 고려하여 추정치를 산정할 경우 추정량의 분산이 확대되는 효과 때문에 추정량의 신뢰구간이 길어져 추정의 정도가 떨어지게 된다. 따라서 추정과정에 적용되는 가중치의 효과를 분석하여 실제로 추정량에 미치는 영향이 어느 정도 인지를 분석할 필요가 있다.

$$L = n \times \frac{\sum_h n_h w_h^2}{(\sum_h n_h w_h)^2} \qquad (4.3)$$

여기서 $n = \sum_h n_h$는 총 표본수, $w_h$는 최종가중치, $n_h$는 층별 표본수를 의미한다.

식(4.3)은 다음과 같이 가중치의 변동계수에 관한 식으로 변환할 수 있다.

$$L = n \times \frac{\sum_j w_j^2}{(\sum_j w_j)^2} = 1 + CV^2(w_j) \qquad (4.4)$$

식(4.4)로부터

$L = 1+(0.0072)^2 = 1.00005184$

이므로 가중치 적용에 따른 분산의 증가분은 거의 없다고 할 수 있다.

## 5. 향후과제

패널조사는 개인 또는 가구의 동적인 현상을 파악하는 중요한 조사 방법으로 현재 국내에서 급속히 확산되어 가고 있다. 패널 조사의 장단점은 이미 많은 연구에서 파악되고 있지만, 다양한 장점에도 불구하고, 시간의 흐름에 따라 표본에서 탈락하는 가구가 증가한다는 사실과 이를 방지하기 위한 패널 유지비용이 일반 횡단면 조사에 비해 상대적으로 월등히 많이 소요된

다는 점이다. 이러한 문제점은 모든 패널조사에서 공통적으로 안고 있는 문제이지만, 가능한 범위 내에서 표본마모율(sample attrition) 축소를 위한 노력이 필요하다.

### 1) 1차 웨이브 이후의 가중치 조정

한국복지패널 조사에서의 가중치 조정과정은 1차 웨이브에서의 가중치를 기본으로 하여 향후 2차 웨이브 이상의 조사에서의 가중치를 조정하고자 한다. 1차 웨이브의 기준년도는 2005년도 12월 31일 기준이며, 이 기준년도는 종단면 가중치를 산정하는 기준이 될 것이다. 2006국민기초생활실태조사에서 재개발, 자연재해, 강력거절 조사구(주로 아파트조사구)에 대해서는 적절한 조사구의 무응답 가중치를 부여함으로서 무응답 조사구 문제를 해결하였다. 또한 2차 웨이브 이상 조사가 진행됨에 따라 가구차원에서는 이주 또는 분가 등으로 인하여 가구의 변동이 발생함으로 이를 위한 추적조사 및 가중치 조정이 반드시 필요하다. 또한 개인 차원에서는 1차 웨이브 당시의 가구원이 아닌 개인이 2차 웨이브에서는 가구에 살고 있는 경우나 새로 태어난 자녀 등에 대해 개인차원의 변동에 대한 가중치의 조정이 필요하다. 이러한 가중치 조정과정은 가능하면 객관적이고, 타당한 방법을 적용하여 추정치의 편향을 가능한 축소시키도록 해야 할 것이다.

### 2) 가중치의 은닉(masking) 작업

통상적으로 동일한 조사구에서의 가구 가중치는 같은 가중치를 가지며, 또한 동일한 가구내에서 개인들의 가중치는 서로 같은 값을 가진다. 따라서 가중치의 특성을 파악하면, 개인정보를 일정 수준까지는 파악할 수 있는 단서를 제공할 수 있다. 이러한 개인 정보 누설을 방지하기위해 SLID의 경우 내부데이터에 대해서는 별도의 은닉을 하지 않지만, 외부로 공표되는 자료에 대해서는 반드시 가중치에 은닉 작업을 추가하는 것으로 나타났다.

그러므로 KOWEPS자료의 특성상 개인정보의 누출을 방지하고자 최종가중치에 일정 수준의 잡음을 부가한 가중치를 대외자료로 공표하고자 한다. 이러한 작업은 이미 SLID에서 진행하고 있으며, 이를 기반으로 KOWEPS의 데이터 환경에 적합한 은닉 작업을 수행할 계획이다.

### 3) 조사방법의 개선

KOWEPS의 현재 조사방법은 종이조사표(Paper and Pencil Interview: PAPI)를 이용한 개별 면접 방식이다. 조사방법의 변경에 따른 조사 자료의 변동성이 예상되지만, 현재 많은 조사기관에서 활용하고 있는 컴퓨터에 의한 조사방식(Computer Assisted Personal Interview: CAPI)으로의 변경을 고려하고자 한다. 이러한 조사방법의 개선을 통한 이점으로는 첫째, 별도의 자료입력 단계가 불필요하며, 둘째, 조사과정에서 프로그램에 의해 자동적으로 에디팅 작업이 수행되기 때문에 별도로 에디팅 작업이 불필요하며, 셋째, 자료 공표기간의 단축으로 이용자 서비스가 증대되는 장점이 있다. 따라서 다양한 테스트를 거쳐 4차 웨이브부터 CAPI를 이용한 조사방법을 적용할 예정이다.

### 4) 표본 탈락률 최소화

KOWEPS의 1차년도 표본유지율은 약 92%로서 국내 유수의 패널조사에 비해 월등이 높은 수준을 유지하였다. 이러한 표본유지율을 지속적으로 관리하기 위해서는 면접자의 조사 참여를 독려할 수 있는 다양한 동기를 부여할 필요가 있다. 현재 KOWEPS에서는 조사 응답자 가구와 응답자 개인들에 대한 조사답례비를 지불하고 있으며, 또한 이사로 인한 패널 탈락률을 최소화 하기위해 별도의 사은품을 제공하고 있다. 이러한 노력과 더불어 패널가구에 대해 다양한 보건복지 관련 정보와 패널자료 관련 책자를 제공하여 패널가구의 자발적인 참여를 유도하고 있다. 향후 보다 개선된 표본관리를 위해 인터넷 망을 이용한 표본관리 프로그램을 개발하고, 패널가구에 대한 인센티브를 확대할 예정이다. 보건복지