강 석 훈 지도교수

석사학위청구논문

패널데이터 가중치 산정에 대한 연구

2010

성신여자대학교 대학원

경제학과

박 지 혜

패널데이터 가중치 산정에 대한 연구

강석훈 지도교수

이 논문을 석사학위논문으로 제출함

2009년 11월

성신여자대학교 대학원

경제학과

박지혜

인 준 서

박지혜의 석사학위 논문으로 인준함.

심사위원 _____인

심사위원 _____인

심사위원 _____인

성신여자대학교 대학원

논문개요

일반적인 횡단면조사에서 가중치는 표본추출확률(selection probability) 계산, 무응답(non-response) 조정, 사후층화(poststratification) 과정을 거쳐 부여된다. 한편, 종단면서베이(longitudinal survey)의 일종인 패널조사에서 1차 웨이브는 횡단면조사에서의 가중치부여방법을 이용할 수 있으나, 2차 웨이브 이후부터는 웨이브 간에 표본탈락(panel attrition)과 비표본가구원(nonsample household)의 진입 등의 문제가 발생하기 때문에 횡단면분석 이외에 종단면분석을 위한 가중치가 부여되어야 한다.

본 논문에서는 먼저 횡단면데이터와 패널데이터의 가중치 부여방안을 위한 이론연구 즉, 패널조사에서의 가중치 부여방법에 대한 선행연구를 실시하였다.

특히 본 논문에서는 구체적인 가중치 산정방법을 알아보기 위해 한국여성정책연구원의 여성가족패널(KLoWF: Korean Longitudinal Survey of Women & Family)의 2차년도 자료를 이용하여 2차 웨이브에서의 종단면기록개인가중치, 종단면응답개인가중치를 도출하였다. 다만 횡단면개인가중치, 횡단면응답가중치, 횡단면가구가중치에 대한 산정방법과 도출부분은 본 논문에서는 자세히 다루고 있지 않아 추후 연구가 필요한 부분이다.

목차

논문개요

1. 서론 ····· 1

2. 가중치의 의미와 부여방법 ····· 2

2.1 가중치의 정의 ····· 2
2.2 일반적인 가중치 부여방안 - 횡단면데이터를 중심으로 ····· 3

3. 패널데이터 가중치 부여방안에 대한 기초 이론 ····· 3

3.1 횡단면 가중치 ····· 4
3.1.1 횡단면가구가중치 ····· 4
3.1.2 횡단면개인가중치 ····· 8
3.2 종단면 가중치 ····· 9

4. 기존 패널데이터에서의 가중치 부여방안 ····· 11

4.1 해외 패널조사에서의 가중치 부여방법 검토 ····· 11
4.1.1 미국 PSID ····· 11
4.1.2 영국 BHPS ····· 14
4.1.3 독일 GSOEP ····· 21
4.1.4 미국 SIPP ····· 23
4.1.5 미국 NLS ····· 26
4.2 한국 패널조사에서의 가중치 부여방법 검토 ····· 30
4.2.1 한국노동패널조사 ····· 30
4.2.2 한국교육고용패널 ····· 35
4.2.3 한국복지패널조사 ····· 43
4.2.4 대졸자 직업이동 경로조사 ····· 47

5. 여성가족패널의 가중치 부여 방안 ····· 50

5.1 1차년도 가중치 ····· 51
5.1.1 가구가중치 ····· 52
5.1.2 가구원가중치 ····· 52
5.2 2차년도 가중치 부여방법 ····· 53
5.2.1 종단면기록개인가중치 ····· 54
5.2.2 종단면응답개인가중치 ····· 60
5.2.3 횡단면가중치 ····· 67

6. 결론 ····· 68

참고문헌
ABSTRACT
부록

표 목차

<표 1> BHPS에서의 가중치의 종류 ····· 15
<표 2> NLS의 그룹별 특성 ····· 27
<표 3> GOMS 기본가중치 ····· 49
<표 4> 2차년도 기록개인 여부 ····· 54
<표 5> 변수의 기초통계량 : 1차 기록되고 2차 기록된 경우 ····· 55
<표 6> 변수의 기초통계량 : 1차 기록되고 2차 기록되지 않은 경우 ····· 56
<표 7> 종단면기록개인 : 로짓분석 결과 ····· 58
<표 8> 종단면기록개인가중치의 기초통계량 ····· 59
<표 9> 종단면기록개인가중치 사용 전과 후의 변수별 기록개인비중 ····· 59
<표 10> 2차년도 응답개인 여부 ····· 60
<표 11> 변수의 기초통계량 : 1차년도 응답하고 2차년도 응답한 경우 ····· 61
<표 12> 변수의 기초통계량 : 1차년도 응답하고 2차년도 미응답한 경우 ····· 62
<표 13> 종단면응답개인 : 로짓분석 결과 ····· 64
<표 14> 종단면응답개인가중치의 기초통계량 ····· 65
<표 15> 종단면응답개인가중치 사용 전과 후의 변수별 응답개인비중 ····· 65
<부표 1> 종단면기록개인가중치 : 프로빗분석 결과
<부표 2> 종단면응답개인가중치 : 프로빗분석 결과

1. 서론

일반적으로 표본조사의 조사결과를 이용하여 분석할 때에는 표본의 선택확률과 응답확률이 상이하기 때문에 가중치를 통해 이러한 차이를 조정한다. 일반적인 횡단면조사에서 가중치는 표본추출확률(selection probability) 계산, 무응답(non-response) 조정, 사후층화(post-stratification) 과정을 거쳐 부여된다. 한편, 종단면서베이(longitudinal survey)의 일종인 패널조사에서 1차 웨이브는 횡단면조사에서의 가중치부여방법을 이용할 수 있으나, 2차 웨이브 이후부터는 웨이브 간에 표본탈락(panel attrition)과 비표본가구원(nonsample household number)의 진입 등의 문제가 발생하기 때문에 횡단면분석 이외에 종단면분석을 위한 가중치가 부여되어야 한다.

본 논문은 가중치의 의미와 가중치 부여방법에 관한 기초 이론을 알아본 후 기존의 국내외 다양한 패널조사에서의 가중치 부여방법에 대해 검토하여 패널데이터 가중치 산정 방법에 대한 이해를 돕고 또한 실증적으로 여성가족패널 데이터를 이용하여 종단면가중치를 구해봄으로써 보다 자세하게 가중치 부여방법을 제시하는데 목적이 있다.

본고의 구성은 다음과 같다. 우선 가중치의 의미와 일반적인 횡단면데이터를 중심으로 한 가중치부여방안에 대해 알아본 후 패널데이터에서의 횡단면, 종단면 가중치의 부여방안에 대한 기초 이론을 살펴본다. 패널데이터의 가중치 부여방안을 알아보기 위해 기존의 국내외 대표적인 패널조사에서의 가중치 부여방법에 대해 알아본다. 또한 보다 구체적인 가중치 부여방법을 알아보기 위해 한국여성정책연구원에서 조사되는 여성가족패널 데이터를 가지고 2차 웨이브에서의 종단면응답개인가중치와 종단면기록개인가중치를 구해본다. 마지막으로 본 연구의 요약과 함께 한계점을 밝히고자 한다.

2. 가중치의 의미와 부여방법

2.1 가중치의 정의

사전적으로 가중치란 평균치(平均値:평균값)를 산출할 때 각 개별치(個別値)에 부여되는 중요도를 말한다. 비중을 서로 달리하는 여러 품목에 대한 하나의 평균치를 산출할 때, 단순한 산술평균만으로는 합리적인 수치를 뽑을 수가 없으므로 비중에 따라 각 개별품목에 알맞은 중요도를 결정하고 이를 적용시켜 평균치를 얻게 된다.

가중치는 서베이 분석에서 다른 것들보다 추출된 요소에 상대적으로 더 큰 중요성을 할당하는 데 사용되며 추출된 요소들이 불균등확률표본에 의해 추출될 때 필요하다. 또한 가중치는 전체 무응답을 위한 조정과 사후층화에 사용된다.

일반적으로 표본조사의 목적은 어떠한 특정 항목에 대하여 모집단의 특성인 총계, 평균 또는 비율 등을 정확하게 추정하는 것이지만 표본추출과정에서 시도별로 또는 주택유형별로 표본을 할당한 후에 각 층별로 계통추출법을 적용하여 최종 추출단위인 가구를 산정하였기 때문에 표본가구들이 동일한 추출확률로 선정되지 않는다. 모평균이나 모비율 추정에서 단순 표본평균이나 표본비율을 사용한다면 추정량은 불편성을 갖지 못하여 편향이 포함된 추정값을 산출하게 되므로 조사대상별로 추출률과 응답률 및 기타외부정보를 이용한 벤치마킹 조정값을 고려하여 가중치를 부여하여 불편추정량(Unbiased Estimator)을 산출하도록 해야 한다.

특히, 표본설계 과정에서 모든 조사 단위들이 동일한 추출률을 갖도록 하였을지라도 조사과정에서 완전한 응답을 얻지 못할 수도 있고 표본설계에 사용한 정보와 조사시점에서 정보간의 차이가 있을 경우에는 이들을 반영하

여 조사시점에서 모집단의 특성을 제대로 추정할 수 있도록 가중치를 산출하여 모수추정에서 사용해야 한다.[1)]

2.2 일반적인 가중치 부여방안 - 횡단면데이터를 중심으로

일반적인 횡단면조사에서 가중치 부여는 기본적으로 3단계과정을 거쳐서 계산된다. 1단계는 표본추출확률을 계산하는 것이다. 대부분의 횡단면조사는 표본추출확률이 상이하므로 이를 조정하기 위해 1단계에서는 표본추출확률을 계산한다. 이 때 표본의 모집단에 대한 부분포함(incomplete coverage)의 문제도 동시에 고려한다. 2단계는 무응답을 조정하는 것이다. 무응답의 경우를 조정하기 위해 사용되면 일반적으로 알려진 변수, 예를 들면 표본이 속한 지역에 대한 자료를 통해 표본이 분할하는 방법, 모든 표본의 알려진 성질을 이용하여 회귀분석(또는 로짓분석)을 실시하는 방안 등이 사용된다. 3단계는 사후층화로 1단계와 2단계를 거친 후 나타난 표본의 성질을 외부적으로 알려진 전체모집단의 성질과 근사시키는 과정이다. 결국 최종적인 가중치는 위에서 구한 각 단계의 확률, 비율 등을 곱하여 산출하게 된다.[2)]

3. 패널데이터 가중치 부여방안에 대한 기초 이론

종단면 조사의 일종인 패널조사(Panel survey)는 개인 또는 집단의 행태연구나 사회의 변화가 개인의 행동양식에 미치는 영향 등에 대한 조사를 다년간 수행하는 조사이다. 패널조사는 횡단면적인 특성과 시계열적인 특성을 모두 가진 데이터로서 주로 사회, 경제, 교육학 등에서 많이 활용되는 자료수집 방법이다. 패널자료는 어느 한 시점에 조사된 횡단면 자료와는 달리

1) 이계오, 고령화연구패널조사 가중치, 한국노동연구원
2) 강석훈(1999), KLIPS 1차 웨이브에서 가중치 부여방법에 관한 연구, 한국노동연구원

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 탈락과 그에 따른 대표성 상실이 나타나는 문제가 있다. 그러므로 이러한 표본의 대표성 상실 문제를 적절히 해결하기 위해 적용가능한 방법이 가중치 조정방법이다.[3)]

이 장에서는 패널데이터에서의 가중치 부여방안을 알아보기 위해 횡단면 가중치에서의 횡단면가구가중치, 횡단면개인가중치와 종단면가중치에 대한 기초 이론을 알아보고자 한다.

3.1 횡단면 가중치[4)]

3.1.1 횡단면가구가중치

Kalton et al(1994)은 1차 웨이브 이후의 횡단면가중치는 결국 비표본가구원의 추출확률과 밀접한 관련이 있으나 실제로는 이 추출확률을 구하지 않고도 불편추정량을 얻을 수 있는 가중치부여방법을 제안하고 있다. 이를 구체적으로 설명하면 다음과 같다.

횡단면가구가중치를 산출하기 위한 방법을 고려하자. 모집단(모집단가구

3) 손창균(2008), 한국복지패널의 가중치 조정과 향후과제, 한국보건사회연구원
4) 강석훈(2003), KLIPS의 가중치 부여방안 연구, 한국노동연구원

수 H)의 변수 Y의 총합, $Y=\sum_{i=1}^{H} Y_i$,을 구하는 추정량으로 다음을 고려할 수 있다.

$$\hat{Y}=\sum_{i=1}^{H} w_i Y_i$$

단, w_i 는 가구 i가 표본에 속해 있을 때 1의 값을 가지고, 아니면 0의 값을 가지는 확률변수이다. 이때 $\hat{Y}$ 가 $Y=\sum_{i=1}^{H} Y_i$ 의 불편추정량이 되기 위해서는 $E(w_i)=1$ 가 만족되어야 한다.

이러한 조건을 만족하는 방법으로 추출확률역수가중치 부여방법(inverse selection probability weighting scheme, ISPWS), 동등가구가중치부여방법(equal household weighting scheme, EHWS), 동등개인가중치부여방법(equal person weighting scheme, EPWS) 등이 있다.

EPWS나 EHWS는 모두 현재 가구원 중에서 원시표본가구에 속한 가구의 추출확률만을 이용하여 가중치를 부여하는 방법으로 불편추정량이며, ISPWS의 경우보다 분산은 증가하나 현실적으로 적용가능하다. 만약 원시표본가구가 유사한 수준의 동등확률로 추출되었다면 가구나 개인차원의 분석에서 EHWS가 이러한 분산증가를 제어하기 위한 준적절한(near optimal) 방법이다. 한편, non-epsem 표본인 경우에는 EHWS나 EPWS는 모두 하위적절한(suboptimal) 방법이다. 그러나 이 경우에도 적절한 가중치 부여방법은 알 수 없는 최초의 추출확률에 의존하기 때문에 EHWS나 EPWS가 모두 사용될 수 있다. 다만 이 경우에도 EHWS는 1차 웨이브에서의 가구 구성에 대한 정보(현재의 표본가구에 속해 있는 원시표본가구원들이 동일한 원시표본가구에서 추출되었는지의 여부, 즉 원시표본가구의 수)를 필요로 하나 EPWS에서는(전체 원시표본가구원수만 필요) 원시표본가구수는 필요

하지 않기 때문에 EPWS가 보다 선호된다.

(1) ISPWS : 가구 H_i 가 표본에 속할 확률은 다음과 같이 표현가능하다.

$$P(H_i) = P(h_j \cup h_k \cup h_l \ldots.) \\ = \sum P(h_j) - \sum\sum P(h_j \cap h_k) + \sum\sum\sum P(h_j \cap h_k \cap h_l) \ldots$$

단, $h_j, h_k, h_l, \ldots$는 적어도 한명의 가구원이 현재의 H_i 가구에 소속되어 있는 경우를 말한다. 이때의 가구가중치는 다음과 같다.

$$w_i = 1/P(H_i)$$

이러한 가중치는 $E(w_i) = P(H_i)\left[1/P(H_i)\right] + \left[1 - P(H_i)\right]0 = 1$이 되어 불편성의 조건을 만족한다. 그러나 ISPWS를 사용하려면 모든 현재 표본가구의 모든 구성원의 원시가구가 최초웨이브에서 추출될 확률과 이들의 결합추출확률을 알아야 하는데, 원시표본가구로 선정된 가구에 대해서는 이 확률이 알려져 있으며, 그렇지 않은 가구들의 경우와 결합확률은 알려져 있지 않기 때문에 현실적으로 사용하기가 어렵다(Rendtel(1992)은 결합확률을 무시하고, 로짓모형들을 이용하여 비표본가구의 추출확률을 계산하는 방법을 제안한 바 있음). 아래의 대안들은 비표본가구의 추출확률을 이용하지 않고 원시표본가구의 추출확률만을 이용하여 가중치를 구하는 방법이다.

(2) EHWS : 다음과 같은 가중치를 고려한다.

$$w_i = \sum_j \alpha_{ij} \cdot w_{ij}^{'}$$

단, $w_{ij}^{'}$은 적어도 하나의 가구구성원이 현재의 표본가구 H_i에 소속되어 있는 가구 h_i가 원시표본가구로 선정된 경우에는 $1/p_j$의 값을 가지고, 그렇지 않은 경우에는 0의 값을 가지는 변수이다. p_j는 가구 h_i의 추출확률, 즉 $p_j = P(h_j)$이며, $\sum_j \alpha_{ij} = 1$이다.

이제 α_{ij}의 값을 가구 H_i에 소속되어 있는 가구원들의 원시표본가구수를 이용하면 가중치는 $w_i = \sum w_{ij}^{'}/C_i$가 된다. 단, C_i는 가구에 속해 있는 가구원들의 원시표본가구수를 의미한다. 이러한 방법이 EHWS이며, Huang(1984)는 multiplicity approach라고 명명하였다.

(3) EPWS : 다음과 같은 가중치를 고려한다.

$$w_i = \sum_j \sum_k \alpha_{ijk} \cdot w_{ijk}^{'}$$

단 $w_{ijk}^{'}$은 가구 h_j의 구성원 k가 원시표본가구원인 경우에는 $1/p_j$의 값을 가지고, 그렇지 않은 경우에는 0의 값을 가지는 변수이다. p_j는 가구 h_j의 추출확률, 즉 $p_j = P(h_j)$이며, $\sum_j \sum_k \alpha_{ijk} = 1$이다.

이제 α_{ijk}의 값을 원시표본으로 추출될 수 있었던 모든 가구원의 수를 이용하면, 가중치는 $w_i = \frac{1}{M_i} \sum M_{ij} w_{ij}^{'}$가 된다. 단, 가구 H_i의 가구원 중에서 1차 웨이브에서 추출된 가구와 동일한 구성원인 모든 경우에 $w_{ij}^{'} = w_{ijk}^{'}$은 상

수이며, M_{ij}는 가구 H_i에 소속되어 있는 가구원 중에서 가구 h_j로부터 온 가구원의 수를 의미하고 $M_i = \sum_j M_{ij}$는 가구 H_i에 소속되어 있는 가구원 중에서 1차 웨이브에서 원시표본으로 추출될 수 있었던 개인의 수를 의미한다. 이 방법이 EPWS이며, Huang(1984), Ernst(1989)에서는 fair share weighting scheme으로 명명되고 있으며 이 방법은 SIPP와 SLID에서 사용하는 방법이다.

3.1.2 횡단면개인가중치

모집단개인의 총합, $Y = \sum_{i=1}^{H} \sum_{k=1}^{N_i} Y_{ik}$ 에 대한 추정량으로 다음의 $\hat{Y}$을 고려할 수 있다.

$$\hat{Y} = \sum_{i=1}^{H} \sum_{k=1}^{N_i} w_{ik} Y_{ik}$$

단, Y_{ik}는 가구 H_i에 소속되어 있는 가구원 k의 변수값이며, H는 전체가구의 수이고, N_i는 가구 H_i의 가구원수이다. 또한 w_{ik}는 가구 H_i에 소속되어 있는 가구원 k가 표본에 포함되어 있지 않으면 0의 값을 가지는 확률변수이다. 이때 $\hat{Y}$가 Y에 대한 불편추정량이 되기 위한 조건은 $E(w_{ik}) = 1$ 이다.

이러한 불편성 조건을 만족하는 가장 단순한 방법으로 원시표본가구원은 추출확률의 역수를 이용하고, 기타의 신규진입자의 가중치는 0으로 하는 경우이다.

w_i를 가구 H_i의 모든 구성원의 가중치라고 하면(동일한 가구내의 모든 가구원의 가중치는 동일), w_i는 다음과 같이 쓸 수 있다.

$$w_i = \sum_j \sum_k \alpha_{ijk} w_{ijk}$$

단 w_{ijk}은 가구 h_j가 원시표본가구이면 원시표본가구의 추출확률의 역수, 즉 $1/p_j$를 갖고, 아니면 0의 값을 가지는 변수이다. 만약 $\sum_j \sum_k \alpha_{ijk} = 1$이면 불편추정량이 된다. α_{ijk}를 모든 원시가구원에 대하여 $1/M_i$를 부여하는 경우가 EPHS방법이며, 이 때 가구 H_i의 모든 가구원(신규진입자 포함)은 $w_i = \sum_j \sum_k w_{ijk}/M_i$의 가중치를 부여받게 된다. 단 $M_i = \sum_j M_{ij}$로서 가구 H_i에 존재하는 원시가구원의 수이다. 만약 원시가구 h_j의 구성원에 대해 $\alpha_{ijk} = 1/C_i M_{ij}$를 부여하는 방법이 EHWS이다. 단 C_i는 현재의 가구 H_i에 구성원이 있는 원시가구 h_j의 수이다.

EPWS, EHWS의 두 가지 방법 모두 불편추정량이다. 이 접근방법은 기본적으로 표본가구원에 대하여 원시표본가구원인가 아닌가에 관계없이 동일한 가중치를 부여하는 방법이다. 전자는 현재 가구에 속한 원시표본가구원의 가중치를 원시웨이브에서 응답가능 했던 가구원으로 나눈 평균가중치를 현재의 모든 가구원에 적용하는 방법이며, 후자는 현재 가구에 속한 원시표본가구원의 가구가중치를 현재 가구원의 원시표본가구의 수로 나눈 평균가중치를 현재의 모든 가구원에 적용하는 방법이다.

3.2 종단면 가중치

종단면서베이의 1차 웨이브는 횡단면서베이에서의 가중치부여방법을 이용할 수 있으나, 2차 웨이브 이후부터는 횡단면분석 이외에 종단면분석이

이루어지기 위해서는 횡단면가구가중치(cross-sectional household weights, CHW), 횡단면개인가중치(cross-sectional individual weights, CIW), 종단면가구가중치(longitudinal household weights, LHW), 종단면개인가중치(longitudinal individual weights, LIW)의 가중치를 고려할 수 있다.

종단면가중치를 고려하기 위해서는 종단면서베이에서 특수하게 발생하는 문제들을 고려하여야 하는데 대표적인 요인으로 표본탈락(panel attrition)과 비표본가구원의 진입을 고려해야 한다.

표본탈락이란 직전 웨이브에서 응답대상이 되었던 주체가 새로운 웨이브에서는 응답거부 등의 요인에 의해 표본에서 탈락한 경우이다. 표본탈락은 가구 전체가 탈락하는 가구표본탈락과 가구는 계속적으로 응답하였지만 가구구성원의 일부가 표본에서 탈락하는 가구원 표본탈락을 구분하여 고려할 수 있다. 이러한 전체 무응답이외에도 개별 항목에 대해서만 응답을 거부하는 항목무응답(item nonreponse)도 발생할 수 있으나 이 경우는 가중치부여 방법 보다는 보정(imputation)방법이 주로 사용된다.

비표본가구원의 문제는 웨이브가 진행됨에 따라 결혼, 이민 등으로 인해 표본가구에 새로운 개인이 포함되거나 탈락하기도 하며, 표본가구원 사이에서 새로운 가구원이 출생하는 경우에 발생되는 문제이다. 특히 비표본가구원의 경우에는 최초 웨이브에서의 가중치가 없기 때문에 이들에 대해 어떠한 가중치를 부여하는 지가 중요한 이슈로 부각하게 된다. 한편, 이러한 웨이브가 진행됨에 따라 비표본가구원의 진입 또는 표본가구원의 탈락에 의해 가구구성이 달라지며, 이에 따라 종단면가중치를 부여할 수 있는 가구의 개념이 모호해 진다. 이와 같은 개념상의 문제로 인하여 기존 패널조사에서는 2차 웨이브 이후부터는 횡단면가구가중치와 횡단면개인가중치, 종단면개인가중치는 산출하나 종단면가구가중치에 대해서는 이를 구하는 경우와 구하지 않는 경우로 나뉘고 있다. PSID(Panel Survey of Income Dynamics), SIPP(Survey of Income Program Participants) 등에서는 종단면가구가중치

를 구하고 있으나, BHPS(British Household Panel Study), GSOEP(German Socio-Economic Panel) 등에서는 종단면가구가중치를 구하지 않고 있다.[5)]

4. 기존 패널데이터에서의 가중치 부여방안

4.1 해외 패널조사에서의 가중치 부여방법 검토

4.1.1 미국 PSID (Panel Study of Income Dynamics)[6)7)]

PSID는 초기에 두 개의 독립된 표본으로 구성하였다. 하나는 층화다단계 추출에 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. 표본은 1968년 이래로 매년 조사된 개인을 포함하며, 원표본 가구원에 의해 형성된 가구원을 포함한다.

(1) 표본추출과정

5) 강석훈(2000), KLIPS 2차 웨이브의 가중치 부여방법에 관한 연구, 한국노동연구원
6) 강석훈(1999), KLIPS 1차 웨이브에서 가중치 부여방법에 관한 연구, 한국노동연구원
7) 손창균(2008), 한국복지패널의 가중치 조정과 향후과제, 한국보건사회연구원

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

1969년과 1970년에 표본은 전년도에 계속적으로 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 확률표본추출이지만, 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) 가중치산출과정

1968년 최초 조사에서 각 표본은 무응답이었다. 왜냐하면 재면접 표본들이 인구센서스 조사에서 무응답자들이었기 때문이다. 즉, 이들은 센서스국에 의해 조사된 응답자 이름과 주소를 OEO에게 공개하는데 대한 서명을 거부

하였다. 또한 OEO로부터 SRC에게 일부 표본 주소를 전달하는데 실패하였다. 1차 웨이브에서는 표본추출확률과 응답률의 곱으로 가중치를 계산하였다. PSID 초기표본에는 일반적인 확률표본인 SRC(Survey Research Center)표본과 빈곤가구를 중심으로 한 SEO(Survey of Economic Opportunity) 표본이 혼재되어 있으며 각각의 추출확률이 상이하다.

1차 웨이브 가중치를 결정하기 위해 다음과 같은 3가지 확률을 계산하였다. ① SRC 횡단면 표본에 대한 확률 ② 재조사 표본에 대한 확률 ③ 결합된 표본에 대한 확률이다(횡단면과 재면접 표본들을 결합하였을 때, 전체 비추정(overall ratio estimation) 방법은 사용하지 않았는데 이는 모집단 총합에 관한 정보를 가지고 있지 않았기 때문이다).

① SRC 횡단면 표본에 대한 확률

횡단면 표본은 표본추출 당시에 전체 미국인에 대한 고정비율(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. 개별 분할 내에서 무응답이 임의적이라고 가정하면, 어떤 가구가 응답할 확률은 최초의 표본추출확률과 응답률의 곱으로 나타난다. 후속연구에 의하면 지역 외에 응답률은 가구주의 연령, 성별, 거주지크기에 따라 달라지는 것으로 나타난다. 횡단면 표본에 대한 면접확률은 “초기 추출률(initial selection rate)×응답률(response rate)″으로서 (0.66/10,080)×(응답률)과 같다.

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×센서스부차추출률×SRC부차추출률(subsampling-rate)×접수율(receiving rate)×응답률"로 정의된다.

③ 결합된 표본에 대한 확률

결합된 표본은 다음과 같은 세 부분으로 고려할 수 있다.

㉠ 센서스국으로부터 전달받은 재면접 표본
㉡ 남부의 SMSA와 non-SMSA로부터 추출된 횡단면 표본에 있는 빈곤가구
㉢ 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 "재표본에서 면접확률+횡단면표본에서 면접확률-두 확률의 곱"으로 정의된다.

4.1.2 영국 BHPS (British Household Panel Study)[8)]

BHPS에는 다양한 가중치가 제공되고 있으며, 가중치의 계산과정에서도 매우 세밀한 부분까지 조정하는 특징을 가지고 있다. 가중치는 일반적으로 응답가구원가중치(respondent individual weights), 기록가구원가중치(enumerated individual weights), 가구가중치(household weight)로 구성되어 있으며, 이는 다시 횡단면분석과 종단면분석 등의 분석목적에 따라 구분된다.

<표 1> BHPS에서의 가중치의 종류

구분		가중치종류	변수명
횡단면분석	1차 웨이브	응답가구원가중치	AXRWGHT
		기록가구원가중치	AXEWGHT
		가구가중치	AHHWGHT
	2차 웨이브 이후	응답가구원가중치	wXRWGHT
		기록가구원가중치	wXEWGHT
		가구가중치	wXHWGHT
종단면분석		응답가구원가중치	wLRWGHT
		기록가구원가중치	wLEWGHT

자료: British Household Panel Survey User Manual

종단면분석에 사용되는 가구가중치는 2차 웨이브부터는 구하지 않고 있으며, 이는 종단면 분석의 관점에서 볼 때 가구의 구성원이 변화하기 때문에 시간에 따라 동일한 가구라고 보기 어렵기 때문이다.

(1) 1차 웨이브

BHPS 1차 웨이브에서는 다음과 같은 4단계를 거쳐 가중치를 구한다. 무

8) 강석훈(1999), KLIPS 1차 웨이브에서 가중치 부여방법에 관한 연구, 한국노동연구원

응답가구원에 대한 별도의 조정을 실시한다는 점과 1차 웨이브의 횡단면가중치가 가구가중치, 기록가구원가중치, 응답가구원가중치의 3가지로 나누어진다는 특징이 있다.

① 상이한 선택확률의 조정을 위한 가중치(디자인가중치)

표본추출과정에서 상이한 추출확률을 조정하기 위해 1단계로 PSU의 추출확률을 구하고 2단계로 1단계에서 PSU로 추출된 상태에서 어떤 지점이 선택될 확률을 구한다. 1단계 및 2단계가 주어진 경우에 선택된 지점에서 어떤 가구가 추출될 확률의 곱으로 추출확률을 계산하고 이에 대한 역수로 가중치를 부여한다. 실제로 3단계에서 어떤 선택된 지점에서 가구가 추출될 확률을 계산하기 위해서는 선택된 지점의 가구수를 알아야 하는데 이것이 불가능한 경우에는 면접을 시도한 모든 가구수를 선택된 지점의 가구수로 한다.

② 가구차원에서의 무응답 조정

가구무응답조정은 지역과 PSU의 특성, 그리고 건물형태 등으로 분할하고 동일분할 내에서는 응답가구와 무응답가구 간에 평균 등에서 차이가 없다는 가정 하에 응답률을 통해 조정된다. 건물 형태가 알려지지 않은 경우에는 지역과 사회경제적 특성을 기준으로 하여 조정되었으며, 건물형태가 알려진 경우에는 이를 추가하여 대략 30-45가구가 되도록 표본을 분할한 후 개별 분할내의 응답가구수가 그 분할 내에서 응답가구 및 무응답가구수와 일치하도록 응답가구에 가중치를 부여한다(가구수를 기준으로 조정).

③ 응답한 가구 내에서의 무응답 가구원에 대한 조정

응답한 가구 내에서의 무응답 가구원에 대한 조정이 필요한 이유는 일부 응답가구에서 응답해야 하는 가구원 중에서 무응답가구원이 있는 경우가 있었기 때문이다. 이 경우에는 응답가구원과 무응답가구원(대용응답자(proxy respondent) 포함)을 대상으로 로짓모형을 설정하였다. 지역, 주거형태, 부유한 정도, 가구 내에서의 응답대상가구원(eligible respondent)의 수, 결혼여부, 고용형태, 연령, 성별, 그리고 이러한 변수들의 상호작용(interaction) 등을 설명변수로 사용하였다. 가중치로 인한 분산의 지나친 확대를 피하기 위해 1.75를 최대로 하여 절단(trimming)하였으며, 가중치는 모든 응답가구원의 응답확률의 역수로 계산하였다. 이 과정에서 모든 응답자의 추정확률(fitted probability)을 구하기 위해 일부변수의 경우 유사자료대체법에 의한 보정(hot-deck imputation)을 하였다.

④ 원시표본수를 일치시키기 위한 가중치의 재조정(rescaling)

재조정은 절삭(truncation), 사후층화, 규모재조정 등의 과정을 거쳐서 이루어진다.

가중치부여에 따른 분산확대 효과를 최소화하고, 1차 웨이브에 근거한 매우 큰 값의 가중치가 나타날 수 있는 잠재적인 가능성의 문제를 피하기 위해, 기록가구원가중치와 응답가구원가중치는 2.5를 넘지 않도록 절삭되었다. 이러한 절삭과정과 관련된 편향과 분산확대의 상반관계는 유효표본수(effective sample size)와 추정치의 분산의 퍼센트 증가를 이용하여 산출되었다.

사후층화는 표본데이터의 한계분포를 모집단의 알려진 한계분포와 같이 조정하는 과정이며, 표본틀의 축소포함(under coverage)의 문제를 해결하고 추정치의 정도와 로버스트성을 높이기 위하여 사용된다. 1991년 영구 센서

스에서 나타난 거주형태(household tenure), 가구크기, 자동차의 수의 한계분포를 표본에서의 가구와 기록가구원의 한계분포와 일치하도록 사후층화 하였다. 기록가구원은 연령과 성별에 의해 다시 사후층화 되었으며, 응답가구원은 위에서 사용한 거주형태, 가구크기, 자동차의 수, 연령, 성별에 따라 사후층화 되었다(기록가구원에 대한 사후층화는 동일한 가구내의 구성원 간에 상이한 가중치를 부여하게 됨에 주목).

절삭과 사후층화과정을 거쳐 최종가중치는 이들의 합이 각각의 그룹 내에서 달성한 표본수가 되도록 규모재조정을 하였으며, 이는 가중된 총 표본수가 가중치를 사용하기 않은 표본수와 일치하도록 하는 과정이다.

최종적으로 가구가중치는 위의 1단계와 2단계의 곱에 4단계(응답가구수에 따른 규모재조정)를 적용하여 산출되었으며, 기록가구원가중치는 1단계와 2단계의 곱에 4단계(기록가구원수에 따른 규모재조정)를 적용하여 산출되었고 응답가구원가중치는 1단계와 2단계 그리고 3단계의 곱에 4단계(응답가구원수에 따른 규모재조정) 조정요인을 곱하여 산출하였다.

(2) 2차 웨이브 이후[9)10)]

① 종단면 응답가중치

현재 웨이브까지 모두 응답한 경우만을 대상으로 사망자나 감옥에 간 경우, 이민을 간 경우 등 응답거부가 아닌 경우에는 모두 응답으로 간주하여 분석한다(이들 계층은 비적격 면접대상(ineligible)으로 분류되며, 모집단상의 변화를 의미하므로 응답거절(refusal)과 구분되어야 한다).

우선, 계속응답자와 응답거절자를 결정하는 요인들을 SPSS CHAID[11)]를

9) Taylor.M.eds(2003), British Household Panel Survey User Manual
10) 강석훈(2009), 여성가족패널 2차 웨이브 이후의 가중치 부여방안 연구, 한국여성정책연구원(내부자료)
11) CHAID에서 나타난 응답률 결정변수들은 이사여부, 연령, 성, 고용상태, 소득총합과 구성, 인종, 조직

이용하여 분석하고 구분된 셀 내에서의 응답비율의 역수를 전년도 사후층화 전 종단면응답개인가중치에 곱하여 현재 웨이브의 1차 무응답조정 후 종단면개인가중치를 산출한다. 이러한 방식은 최종적으로 구분된 셀 내에서의 무응답은 임의적이라는 가정에 기초한다.

이렇게 산정된 1단계 종단면응답개인가중치는 1차 웨이브에서의 개인특성(연령, 성, 주거형태, 자동차 보유대수, 가구크기)에 맞게 사후층화 되었다.

이번 웨이브에 16세가 되어 새롭게 면접을 한 가구원도 종단면개인가중치를 갖는다. 이들에게는 부모의 종단면개인가중치의 최소값 또는 부모가 없는 경우 가구구성원의 종단면개인가중치 중 최소값을 가중치로 부여하였다.

3차 웨이브 이후에는 직전 웨이브 등에서 응답하지 않았던 개인이 다시 응답한 경우가 있었는데 가중치작업에서는 이들을 마치 직전 년도에 응답한 것으로 간주하여 모델링 하였으며, 설명변수는 가장 최근 년도의 관측치를 사용하였다.

② 종단면 기록가구원가중치

종단면 기록가구원가중치는 가구주 특성과 가구에 기반을 둔 가중치를 가진 종단면 응답가구원을 사용함으로써 두 단계를 거쳐 사용된다. 종단면 기록가구원가중치(wLEWGHT)는 탈락이 발생한 경우(종료사건의 발생을 포함)에 조정된다. 무응답가구와 마찬가지로 전화응답자나 대용(proxy)응답자, 아이들에 대한 종단면 가중치는 이 가중치로 제공된다. 표본으로 들어온 새로운 출생자는 그들 부모의 가중치의 평균값으로 주어진다(일반적으로 표본으로 2명의 부모를 가진 아이들이 1명만 표본인 아이들보다 더 높은 가중

및 단체에의 참여정도, 교육 정도와 같은 개인특성과 지역, 거주형태, 자동차의 수, 소비내구재의 보유여부 등이었음

치를 갖도록 하기 위해). 가구는 종단면에서 정의될 수 없기 때문에 종단면 가구가중치는 없다.

③ 횡단면가중치

횡단면가중치는 정의상으로 1차 가중치나 종단면가중치가 없는 신규진입자를 포함하는 것으로 도출된다. 최초 포함과 응답확률을 알지 못하기 때문에 이런 개인들을 포함하기 위한 가정이 필요하다. 'fair shares approach'(Emst 1989, Lavallèe & Hunter 1992, Rendtel 1991)이라 불리는 표준적인 공정배분방식을 사용한다.

기본적으로 이 접근방식은 표본탈락을 조정한 1차 웨이브의 기록개인가중치의 합을 현재 웨이브의 기록개인들이 공유하는 방식이다. 현재 웨이브에서의 모든 가구구성원은 무응답이 조정된 1차 웨이브의 기록개인가중치의 합을 1차 웨이브의 모집단에 있었던 가구구성원의 수로 나눈 값을 부여받는다(즉. 분모의 숫자에는 1차 웨이브의 모집단에 포함되어 있던 신규진입자는 포함하지만 1차년도 이후 신규출생자는 제외된다).

횡단면가중치를 구하는 첫 번째 단계는 1차 웨이브에 존재하던 기록개인가중치를 최신 웨이브까지 오면서 발생한 표본탈락을 조정하여 표본탈락을 조정한 기록개인가중치를 만드는 일이다. 이 과정은 종단면기록개인가중치를 작성하는 경우와 기본적으로 동일한 방법을 사용한다. 응답률은 가구주, 가구특징, 가구원 특징 등을 사용하여 분석하였다. 1차 웨이브의 원시표본가구원의 사후층화 후 가중치에 추정된 응답률을 적용하여 가중치를 계산하였다. 이렇게 계산된 원시표본가구원의 가중치를 공정배분방식을 통해 모든 가구원에게로 적용하여 최종 횡단면기록개인가중치를 구하였다.

이 횡단면기록개인가중치를 바탕으로 1번 웨이브에서 응답대상자였지만 응답거절을 한 경우의 응답률을 적용하여 최종 횡단면응답개인가중치를 구

하였다. 이 가중치는 가구내 무응답을 조정한 결과이다. 이 때 무응답조정은 기본적으로 기록개인과 동일하나 현재 웨이브의 특성만을 이용하여 응답률을 분석하였다.

횡단면가구가중치는 횡단면기록가구원가중치와 같게 설정하였다. 다만, 가구 총수를 다시 재조정하였다.

모든 가중치는 2.5를 기준으로 절삭 하였으며, 가중표본의 수가 원시표본의 수와 같도록 비율조정 하였다.

4.1.3 독일 GSOEP (The German Socio-Economic Panel)[12)]

인구 및 소득 통계의 수립을 목적으로 실시하고 있는 GSOEP는 교육, 고용, 자산 및 정부정책에 관한 개인의 신념 등을 조사하여 패널데이터를 작성하고 있다.

GSOEP는 국립 경제연구기관의 하나인 DIW(Deutsch Institut fűr Wirtschaftsforschung)의 주관 하에 프랑크프루트, 만하임, 그리고 베를린 대학이 공동으로 수행하고 있다. 실사는 독일 내에서 유일한 학술조사 전문기관인 Infratest라는 민간조사기관에 의해 수행되고 있다.[13)]

1차 웨이브의 가중치는 추출확률과 응답확률을 이용하여 산출하였으며 사후층화는 고려하지 않았다. 2차 웨이브 이후의 가중치는 응답확률(또는 표본탈락률)을 이용하여 산출한다.

GSOEP의 표본탈락은 가구차원에서 이루어지는 경우가 대부분이고, 개인차원에서 이루어지는 경우가 매우 적었다. 이에 따라 횡단면분석에 사용하는 횡단면가구가중치와 횡단면개인가중치는 동일하게 된다. 결국 GSOEP의

12) Markus Pannenberg, Rainer Pischner, Ulrich Rendtel, Martin Spiess and Gert G. Wagner(2005), "Sampling and Weighting", Desktop Companion to the German Socio-Economic Panel(SOEP)

13) 강석훈(1997), 유럽의 패널조사 현황과 시사점, 한국노동연구원

횡단면가중치는 가구 단위의 표본탈락만을 조정한 후 가중치를 산출하고 있으며, 비표본가구원의 경우에도 가구의 가중치를 사용하는 것으로 볼 수 있다.[14)]

(1) 1차 웨이브 횡단면 가중치

다른 가중치 도출에 시작점으로 사용되는 값들이기 때문에 1차 웨이브에서의 추출확률은 특히 중요하다. GSOEP의 1차 웨이브 조사설계는 2단계의 추출확률과정을 거친다. 1단계(표본점)는 표본 A에서의 지역(district)과 표본 B에서의 주(counties)를 조사한다. 2단계(1차 웨이브의 가구)는 임의경로절차(random route procedure)를 사용하는 조사지역에서 얻을 수 있고 표본 B에서 사람은 주의 외국인등록으로부터 얻어진다.

(2) 종단면 가중치

t+x차 웨이브에 대한 추정치를 얻기 위해 모집단에서 어떤 하위그룹에 대해 탈락률이 얼마나 큰 지 아는 것이 필요하다. 탈락률의 역수는 가중치를 구하는 요소이다. 탈락률 측정은 교차분류표 또는 로짓 회귀분석에 의해 이루어질 수 있다. t+1차 웨이브에서 탈락한 원인을 결정하기 위해 t=0인 웨이브에서 가구의 특성들이 사용될 수 있다. 추가적으로 t+1차 웨이브에서 현지조사에서의 특성을 이용할 수 있다. 예를 들어 가구의 이동정보나 면접자의 변화에 대한 정보를 이용할 수 있다.

이러한 분석은 각각의 웨이브에서 이루어진다. 종단면가중치는 1차 웨이브(2차 웨이브에서 시작)에서 또 다른 웨이브까지를 조정한다. 다중 웨이브 과정에서 종단면분석을 위한 정확한 가중치를 구하기 위해 종단면 요소들에

14) 강석훈(2000), KLIPS 2차 웨이브의 가중치 부여방법에 관한 연구, 한국노동연구원

서로를 곱하는 것이 필요하다.

(3) 2차 웨이브 이후의 횡단면 가중치

2차 웨이브에는 기존응답자 뿐만 아니라 신규진입자가 존재한다. 문제시되는 해당 년도에 가구의 표본확률이 알려져 있거나 추정될 수 있는 한 신규진입자 존재는 문제가 되지 않는다. 그래서 PSID에서 하는 것처럼 기존가구에 참여하는 개인에는 0의 가중치를 할당하는 것이 불필요하다.

그러나 가구의 표본확률이 추척원칙 뿐만 아니라 패널의 시작에서 가구원들의 표본확률에 의해 유일하게 결정되기 때문에 새로운 진입자를 가진 가구는 그렇지 않은 가구보다 표본으로 선택될 기회가 더 높다(왜냐하면 최소한 그것들이 달성되는 데 두개의 경로가 있기 때문이다).

결론적으로 새로운 진입자를 가진 가구에는 더 낮은 가중치가 할당되어야한다. 만약 가구가중치가 모든 가구 구성원들에게 적용된다면(즉 신규진입자), 이 낮은 가중치는 새로운 진입자에 의해 발생한 경우 가구수의 증가를 보상한다.

4.1.4 미국 SIPP (Survey of Income Program Participants)

SIPP의 주요 목적은 미국 내에서 개인별 가구별 프로그램 참가와 소득에 대한 정보와 소득과 프로그램 참가의 중요한 결정요인에 대한 정보를 정확하고 포괄적으로 제공하는 것이다. SIPP는 이전 연도를 기준으로 현금과 비현금에 대한 상세한 정보를 제공한다. 이 서베이는 또한 세금, 자산, 부채, 정부이전프로그램에 참여에 대한 데이터를 수집한다. SIPP 데이터는 정부가 연방, 주, 지역별 프로그램의 효율성을 평가하는 것을 돕는다.[15]

15) Survey of Income and Program Participation user's guide, Washington.D.C., 2001

(1) 1차 웨이브

1차 웨이브의 횡단면가중치는 표본추출확률, 응답률, 사후층화의 과정을 거쳐서 이루어지며 다음의 4단계를 거쳐 산출된다.

① 기본가중치(BW)

가구추출확률의 역수이며, 무응답이 없고, 표본틀이 모집단을 완전히 포함하는 경우(complete)에는 이 가중치를 이용하면 모집단에 대한 불편추정치를 구할 수 있다.

② 1차년도 가구무응답요인(F_N)

무응답 가구를 조정하기 위하여 사용되었으며, 다음과 같은 변수를 사용하여 무응답조정분할(noninterview adjustment cell)을 구한다.

- 센서스지역 : 북동부/중서부/남부/서부
- 거주지역 : MSA(Metropolitan Statistical Area)/non-MSA
- 도시크기 : non-MSA에서는 Place/not place, MSA에서는 Central City/balance
- 주요가구원(reference person)의 인종 : Black/nonblack
- 거주형태 : 소유자/임대자
- 가구원수 : 1, 2, 3, 4, 5 이상
- Rotation group : 1, 2, 3, 4(1984년 패널에만 적용)

각각의 분할은 최소한 30가구이상이 있어야 하며 무응답조정비율(응답가능한 표본가구수(eligible sample household)/응답표본가구수(interviewed household)이 2.0을 초과하지 않아야 한다. 두 가지 조건중의 하나가 충족되지 않은 경우 두 가지 조건이 모두 충족될 때까지 분할들을 결합한다(결합하는 방법은 Chapman et al(1986). Sin호 and Petroni(1998) 참조).

③ 1단계비율조정(1st stage ratio estimate factor, F_{1s})

PSU의 차이로 인한 표본오차를 줄이기 위하여 사용되며 non-self representing PSU의 표본가구에 대하여 적용된다. PSU는 센서스지역, 거주지역(MSA/other), 중심지여부(central city/balance, MSA의 경우에만 사용), 인종 변수를 사용하여 분할되었다. 각각의 분할에서 조정요인은 1980년도 전체 센서스에 나타난 해당가구를 표본 PSU의 센서스에 나타난 해당가구로 나눈 비율을 사용한다.[16]

④ 2차비율요인(F_{2s})

서베이에서의 과소포함 문제를 부분적으로 조정하여 추정량의 평균제곱오차를 감소시키기 위해, 1차적으로 개인차원의 비율조정을 실시한다. 이 비율조정은 센서스자료와 월별상시인구조사의 결과를 이용하여 사후층화 하는 과정이다.

센서스자료를 이용하는 비율조정요인의 분자는 연령, 인종, 스페인계, 그리고 성별 요인으로 (출생, 사망, 이민 등을 보정한) 센서스자료를 구분하여

16) F_{1s}는 분산감소효과가 적은 것으로 나타난 1996년 이후의 패널부터는 사용하지 않고 있음.

각 분할 내에서 구한 센서스 추계치를 사용한다. 각 분할 내에서의 분모는 기본가중치, 무응답요인 그리고 1차 조정요인을 감안하여 산출된 개인에 대한 추정치를 사용한다.

월별상시인구조사(Monthly Current Population Survey)에서 산출된 인종(black/nonblack)별, 성별, 결혼상태별, 가구주의 가족 내에서의 지위별에 따른 가구원의 수도 모집단통제(population control)변수로 사용한다. 남자의 경우 결혼상태별, 가족 내에서의 지위별 범주는 1차 가구(primary family) 또는 하위가구(subfamily)내에 있는 남자는 1차가구의 남편, 배우자가 없는 남성 가구주, 하위가구의 남편, 그리고 기타 등이며, 1차 가구 또는 하위가구에 속하지 않은 남자는 가구주, 가구주가 아니거나 group quarters에 있는 개인이다. 여성의 경우에도 위의 범주에서 남편이 아내로, 남성이 여성으로 바뀌는 것 외에 본질적으로 동일하다. 남편의 수와 아내의 수를 동일하게 하고, 가구주의 수와 가구의 수를 동일하게 유지하면서 가중치를 사용하는 추정치가 외생적으로 주어지는 총합과 일치하게 하기 위해 반복위수과정(iterative ranking procedure, nelson et al(1985))을 사용한다.

이러한 과정을 거쳐 최종적으로 산출된 1차년도 횡단면개인가중치는 $FW_c = BW \cdot F_N \cdot F_{1s} \cdot F_{2s}$ 이다.

4.1.5 미국 NLS (National Longitudinal Survey)[17]

NLS는 Bureau of Labor Statistics(BLS), 미국 노동부의 후원을 받아 행해지는 조사이다. 남성과 여성의 다양한 그룹에서 노동시장경험을 여러 해에 걸쳐 정보를 수집한다. NLS 표본들은 수십년 동안 조사된 수천 명의 개인들로 구성되어 있다. 초기에는 퇴직, 전업주부의 노동시장 복귀, 학교-직

17) U.S. Department of Labor Bureau of Labor Statistics(2001), "NLS of Young Women User's Guide". National Longitudinal Surveys

장의 전환 등의 특정이슈를 알기위한 코호트였지만 지금은 더 넓은 범위의 주제에 대한 유용한 정보가 제공되고 있다.

<표 2> NLS의 그룹별 특성

Survey Group	연령 코호트	생년코호트	최초표본 크기	최초/최근 조사연도	조사 상태
Older men	45-59세 (66/3/31)	06/4/1-21/3/31	5020	1966/1990	완료
Mature Women	30-44세 (67/3/31)	22/4/1-37/3/31	5083	1967/1999	진행중
Young Men	14-24세 (66/3/31)	41/4/1-52/3/31	5225	1966/1981	완료
Young Women	14-24세 (67/12/31)	43/1/1-53/12/31	5159	1968/1999	진행중
NLSY79	14-21세 (78/12/31)	1957-1964	12686	1979/2000	진행중
NLSY79 Children	출생-14세	-	3	1986/2000	진행중
NLSY79 Young Adults	15세 이상	-	3	1994/2000	진행중
NLSY97	12-16세 (96/12/31)	1980-1984	8984	1997/2001	진행중

자료: NLS of Young Women User's Guide(2001)

(1) 표본추출[18)]

장년여성 표본(Mature Woman Sample)은 첫 조사 시점인 1967년 3월 31일을 기준으로 30~44세인 여성을 조사대상으로 선정하였다. 코호트(cohort)는 미국 센서스국(Census Bureau)이 1964년 초부터 1966년 말까지 수행하였던 매월노동력조사(*Monthly Labor Survey*)의 1,900개 1차 표본틀

18) 강석훈(2006), 여성가족패널 표본설계방안, 한국여성정책연구원(내부자료)

(PSU: primary sample unit)로부터 추출된 다단계 확률표본이다.

PSU는 표준 대도시 지구 (SMSAs: standard metropolitan statistical areas), 카운티(counties), 카운티의 일부(parts of counties), 독립 도시들(independent cities)로 구성된다. 485개 카운티와 독립도시들을 포함하는 총 235 표본지역들은 모든 카운티와 Columbia 특별구를 대표하여 선택되었다. 235개 표본지역(strata)은 사회경제학적 특성에 따라 상대적으로 동질적인 한개 이상의 PSUs로 생성되었으며, 최종적으로 각 PSU에서 가구 단위의 확률표본이 모집단을 대표하기위해 표집 되었다.

장년여성코호트를 위한 조사는 30~44세 여성 5,393명을 대상으로 표본은 약 5,000명의 응답자(非白人 1,500명과 白人 3,500명)로 계획되었다. 여성은 주로 백인 조사구에 있는 白人, 주로 非白人 조사구에 있는 非白人, 주로 非白人 조사구에 있는 白人, 주로 백인 조사구에 있는 非白人 등 4개 형태로 추출되었다.

흑인 응답자에 대한 신뢰성 있는 통계자료를 위해, 총 인구 기대비율의 두 배로 흑인을 초과 표집하였다. 주로 非白人 조사구에 있는 가구들의 표본추출 비율은 주로 백인 조사구에 있는 가구들의 3~4배였고, 1967년 첫 조사 시 5,393명의 계획된 여성 중 94.3%인 5,083명이 조사되었다.

청년여성표본(Young Woman sample)은 첫 조사 시점인 1967년 12월 31일을 기준으로 14~24세인 여성을 대상으로 하였다. 청년여성코호트를 위한 조사는 14~24세 여성 5,533명을 대상으로 표본은 약 5,000명의 응답자(非白人 1,500명과 白人 3,500명)로 계획되었으며, 1968년 첫 조사 시 5,533명의 계획된 여성 중 93.2%인 5,159명이 조사되었다. 그 밖의 사항은 장년여성표본과 동일하다.

(2) 기초년도(base-year) 표본추출 가중치

NLS의 인구 데이터는 다단계 비율(multi-stage ratio) 추정치에 기반을 두고 있다. 먼저, 각 표본 추출의 최종 확률에 상응하는 기초 가중치를 부여한다. 조사된 모든 사람들에 대한 기초년도 가중치는 초기 조사에서 인터뷰되지 않은 사람들뿐만 아니라 흑인의 과대대표를 설명하기 위해 조정된다. 4개의 센서스 지역(Northeast, North Central, South, and West), 도시/농촌 거주, 인종에 기초를 두는 장년여성에 대한 이 조정은 16개 분류로 이루어진다.

비율 가중치 조정의 첫 단계에서, 1960년 센서스 당시, 표본의 PSUs로부터 추정된 인종과 거주 분포 간의 차이 및 4개의 각 주요 지역 총인구가 고려되었다. 1960년 센서스 데이터를 사용하여, 각 지역 인종과 거주로써 추정된 인구총계는 표본의 PSUs을 위한 센서스수치를 적절히 가중하여 계산되었다. 비율은 1960년 센서스에서 보여진 것처럼 그 지역의 실제 총인구와 이들 추정치(표본 PSUs에 기초한) 사이에서 계산되었다.

비율조정의 두 번째 단계에서 표본 비율은 연령, 성, 인종에 의한 독립적인 현행 인구추정치로 조정된다. 이들 추정치는 지속적인 인구 고령화, 사망, 미국과 타국간의 이주를 고려하기 위해 가장 최근의 센서스 데이터를 사용함으로써 제시되었다. 3개 연령그룹에서 인종에 의한 조정이 행해졌다.

(3) 표본추출 가중치 조정

최초 인터뷰 후, 비면접자들로 인해 표본 크기가 축소되는 문제를 해결하기 위해 인터뷰된 개인의 표본추출가중치가 변경된다. 장년여성 코호트는 기초년도 이후 추가된 개인(신규)이 없는 패널이다. 그 결과, 첫 조사 후의 모든 재가중치 부여는 기초년도 인구지표로 조정된다. 이 개정은 두 단계로 이루어진다.

우선, 매년 비면접자 범위 밖의 개인이 센서스국에 의해 정의되고, 비면

접자의 표본으로부터 제거된다. 이 그룹은 수용되거나, 사망하거나, 군 생활 중인, 혹은 외국으로 이주한 사람(더 이상 미국 인구 구성원이 아닌 사람)으로 이루어진다.

조정의 두 번째 단계는 면접의 (존재 가능한) 비대표적인 특성들을 인지한다. 각 조사연도 동안 인터뷰된 사람뿐만 아니라 적합하지만 인터뷰되지 않는 사람들은 1967년에 보고된 인종, 미국에 거주한 기간(9년 이하, 10년 이상)과 교육기간에 기초한 24개 무응답 조정셀로 된다. 각 셀들에서 인터뷰된 사람들의 기초년도 표본추출 가중치는 그 해의 재조사비율(base year 가중치를 사용함)에 상응하는 비율로 증가되었다.

일반적인 가중치는 특정 개인에 대한 큰 가중치로 인해 야기되는 분산의 증가를 최소화 하기위해 계획되었다. 이 과정에 의한 결과는 표본의 특정 하위부분은 동일한 표본추출가중치로 정해졌다는 것이다. CHRR은 이 문제를 피하기 위해 가중치를 조정한다.

4.2 한국 패널조사에서의 가중치 부여방법 검토

4.2.1 한국노동패널조사(Korean Labor & Income Panel Study)

한국노동패널조사는 한국노동연구원에서 조사하는 연구로 비농촌지역에 거주하는 한국의 가구와 가구원을 대표하는 패널표본구성원(5,000 가구에 거주하는 가구원)을 대상으로 1년 1회 경제활동 및 노동시장이동, 소득활동 및 소비, 교육 및 직업훈련, 사회생활 등에 관하여 추적 조사하는 종단면 조사이다.

(1) 1차년도 가중치 부여

KLIPS는 패널조사이지만 1차 조사는 횡단면조사이므로 1차년도의 가중치는 일반적인 횡단면 조사에서의 가중치부여 방법을 사용한다.

① 표본추출확률의 계산

지역별 도시조사구의 추출확률은 지역별 전체조사구 중에서 도시조사구의 비율을 이용하며, 도시조사구 중에서 표본조사구로 추출될 확률은 도시조사구 중에서 표본조사구의 비율을 이용하여 계산한다.

- 지역별 표본조사구 추출확률
  - 서울 및 6대 광역시 : 0.1×(해당 시의 표본조사구수/해당시의 조사구수)
  - 도의 동부 : 0.1×(해당 도의 동부 표본조사구수/해당 도의 동부 조사구수)
  - 도의 읍면수 : 0.1×(해당 도의 읍면부 표본조사구수/해당 도의 읍면 부 조사구수)
- 최종표본가구 추출확률 = 총접촉시도가구/조사구내 97고특[19]응답가구수
- 가구추출확률 = 조사구추출확률 × 최종표본가구 추출확률

② 응답률과 면접확률의 계산

일반적으로 응답률조정을 위해서는 무응답가구에 대한 특성파악이 전제

19) 1997년 고용구조특별조사

되어야 한다. KLIPS 1차년도 조사의 경우 무응답가구의 지역(광역시 또는 기타 도의 동부 및 읍면부) 또는 ED(Enumeration districts: 조사구)를 이용하여 응답률을 계산할 수 있다. KLIPS 1차 웨이브의 응답률은 ED내에서 응답자와 무응답자의 특성이 같다는 가정 하에 ED내에서의 응답률을 면접확률로 이용하였다.

- 응답률 = 최종조사가구수/ED내 총접촉가구수

③ 1차년도 가중치의 계산

추출확률과 응답확률을 모두 고려한 가구가중치는 다음과 같이 계산된다.

- 서울 및 6대 광역시 : 0.1×(표본조사구수/도시조사구수)×(ED내 접촉가구수/ED내 고특조사 가구수)×(최종 성공가구수/ED내 접촉가구수)
- 도의 동부 : 0.1×(표본조사구수/해당 도의 동부 조사구수)×(ED내 접촉가구수/ED내 고특조사가구수)×(최종 성공가구수/ED내 접촉가구수)
- 도의 읍면부 : 0.1×(표본조사구수/해당 도의 읍면부 조사구수)×(ED내 접촉가구수/ED내 고특조사가구수)×(최종 성공가구수/ED내 접촉가구수)

추출된 가구에서는 모든 가구원이 응답하였으므로 동일 가구 내에서는 가구가중치와 가구원가중치는 동일하게 된다. 또한 최종적으로 추출확률과 응답률을 감안하면 특정한 가구의 가중치는 가구추출확률과 가구가 속한 지역의 응답률의 곱의 역수로 나타나게 된다.

(2) 2차년도 이후 가중치 부여

① 종단면 가구 및 개인 가중치 부여 방법론

KLIPS는 미국의 PSID와 동일한 추적원칙을 가지고 있으며 PSID와 동일하게 Duncan(1995)의 가중치 부여방법을 따르고 있다. Duncan(1995)의 가중치 부여방법은 다음과 같다.

㉠ 최초 조사년도에서 가구차원의 가중치를 구한다.

이때 표본추출과정에서 사용된 상이한 추출확률을 감안하여야 하며, 가능한 경우 상이한 응답률로 보정한다. 또한 마지막 단계에서 외부적으로 이용 가능한 모집단의 정보가 있다면 이러한 사항들을 비율조정을 이용하여 적용한다.[20]

㉡ 최초 조사년도에서 작성된 가구가중치를 연령이나 응답여부에 관계없이 모든 가구원의 가중치로 사용한다.

㉢ 2차 조사 이후부터는 가구원들의 상이한 응답률을 이용하여 가중치를 조정한다. 이 단계에서는 가구와 가구원의 정보를 모두 이용한다. 예를 들어 응답여부를 나타내는 로짓모형을 설정하여 모형을 추정한 후 이 계수를 사용하여 모든 가구원들의 응답확률을 추정하며, 이 응답확률의 역수를 최초 개인가중치에 곱하여 2차 조사에서 무응답조정 가중치를 산출한다. 이때 2차 조사에서는 존재하지만 1차 조사에서 응답하지 않았던 비표본가구원이나 1차 조사 이후 새롭게 진입한 가구원의 경우는 개인차원의 무응답조정 과정에 포함하지 않는다.

20) 이 과정은 횡단면조사의 가중치 부여방안과 기본적으로 동일하다.

㉣ 2차 조사에서 산출된 가구원 가중치의 평균을 이용하여 2차 조사의 가구가중치를 산출한다.

㉤ 평균을 구할 때는 원시가구원의 가구원가중치의 합을 전체 가구원수(비표본가구원+표본가구원)로 나누어 계산한다. 원시표본가구원과의 결혼 등의 사유로 새롭게 진입한 비표본가구원의 경우에는 0의 가구원가중치를 부여한다. 새롭게 태어난 가구원의 경우에는 가구가중치를 계산할 때는 분모와 분자 모두에서 제외하며, (이들이 응답대상가구원이 되었을 때는) 이 가구원이 속한 가구의 가구가중치를 부여받게 된다.

이러한 가중치부여과정을 거치게 되면 가구 차원의 무응답조정가중치가 산정되며 최초 조사에 존재했던 가구원의 개인차원의 무응답조정 가중치가 산출된다. 비표본가구원의 경우는 0의 개인가중치를 가지며 새롭게 태어난 가구원의 경우에는 이들이 태어난 조사년도에서의 가구가중치를 개인가중치로 부여받게 된다.

㉥ 3~9차 조사에서도 동일한 방법을 이용한다.

② 횡단면 개인 가중치 부여 방법론

횡단면 개인 가중치는 가구 가중치의 평균을 해당 가구원 전원에게 할당하는 방식으로 부여된다. 이는 가구가 응답할 경우 가구원이 누락되는 경우는 거의 없고, 가구원의 응답이 아니라 가구의 응답이 조사 참여에 결정적이라는 논리에 입각한 것이다.

③ 가중치의 스케일 조정

1998년 가구가중치는 이미 스케일 조정된 것이므로 그대로 사용하였으나, 나머지 연도의 경우에는 통계청의 가구추계자료의 총가구수 증가율을 평균한 값을 KLIPS 1차년도 가중치합계에서부터 매년 곱해주어 산출하였다.

개인가중치의 스케일 조정에는 경제활동인구조사의 비농가 생산가능인구의 평균증가율을 이용하여 1차년도 개인가중치 합계에 평균증가율 값을 매년 곱해주는 방식으로 산출하였다.

4.2.2 한국교육고용패널 (Korean Education & Employment Panel)[21]

한국교육고용패널조사는 우리나라 청소년의 교육경험과 진학, 진로, 직업세계로의 이행 등을 파악하기 위하여 2004년에 시작하여 이후 동일한 표본을 1년 주기로 추적 조사하여 인적자원의 구축과 활용에 대한 패널자료를 구축해 오고 있다.

KEEP 조사는 국가 인적자원개발 정책 수립을 위한 기초자료의 수집, 국가 교육정책 수립의 기초자료 제공, 노동시장 정책 수립의 기초자료 제공 등을 목적으로 이루어지며 조사 자료는 국가 인적자원개발 정책 수립의 기초자료, 관련 기관의 연구자료로는 물론 개인, 학부모 및 교사에 대하여 자기개발 및 평생 직업교육의 방향을 제공하는 등으로 폭넓게 활용되고 있다. KEEP 조사는 기준년도인 2004년에 중학교 3학년생, 일반계 고등학교 3학년생, 전문계 고등학교 3학년생 각각 2,000명씩을 조사대상 패널(학생)로 선정하여, 매년 추적 조사하고 있다.

21) 한국교육고용패널 1차(2004)~3차(2006)년도 조사 사용자지침서, 한국직업능력개발원, 2007

(1) 1차년도 가중치

1차년도의 가중치는 횡단면 조사의 일반적인 가중치 부여 방법을 사용한다.

① 추출확률 : 불균등 추출확률(unequal selection probability) 보정

학생의 추출확률은 학교의 추출확률, 추출된 학교에서의 학급 추출확률, 추출된 학급에서의 학생 추출확률들의 곱으로 계산되며, 학생의 추출확률 보정 가중치는 학생의 추출확률의 역수로 계산된다.

중학교와 일반계 고등학교의 추출확률은 지역별(15개 시, 도) 전체 학교 수와 표본 학교 수에 따라 다르며, 전문계 고등학교의 추출확률은 학교유형별(공업고, 상업고, 기타고) 전체 학교 수와 표본 학교 수에 따라 다르다.

- 중학교, 일반계 고등학교의 학교 추출확률

$$\text{지역별 학교 추출확률} = \frac{\text{지역별 표본 학교 수}}{\text{지역별 전체 학교 수}}$$

- 전문계 고등학교의 학교 추출확률

$$\text{학교유형별 학교 추출확률} = \frac{\text{학교유형별 표본 학교 수}}{\text{학교유형별 전체 학교 수}}$$

추출된 학교에서의 학급 추출확률은 각 학교의 전체 학급 수와 표본 학급 수에 따라 학교별로 다르다.

- 추출된 학교에서의 학급 추출확률

$$\text{추출된 학교에서의 학급 추출확률} = \frac{\text{추출된 학교의 표본 학급 수}}{\text{추출된 학교의 전체 학급 수}}$$

추출된 학급에서의 학생 추출확률은 각 학급의 전체 학생 수와 표본 학생 수에 따라 학급별로 다르다.

- 추출된 학급에서의 학생 추출확률

$$\text{추출된 학급에서의 학생 추출확률} = \frac{\text{추출된 학급의 표본 학생 수}}{\text{추출된 학급의 전체 학생 수}}$$

학생의 최종적인 추출확률은 학교의 추출확률, 추출된 학교에서의 학급 추출확률, 추출된 학급에서의 학생 추출확률을 모두 반영하여 계산된다.

- 학생의 추출확률(중학생, 일반계 고등학생)

  학생의 추출확률 = 지역별 학교 추출확률 × 추출된 학교에서의 학급 추출확률 × 추출된 학급에서의 학생 추출확률

- 학생의 추출확률(전문계 고등학생)

  학생의 추출확률 = 학교유형별 학교 추출확률 × 추출된 학교에서의 학급 추출확률× 추출된 학급에서의 학생 추출확률

학생의 추출확률 보정 가중치는 학생 추출확률의 역수로 계산된다.

- 학생의 추출확률 보정 가중치(BY_w1)

$$\text{학생의 추출확률 보정 가중치}(BY\_w1) = \frac{1}{\text{학생의 추출확률}}$$

② 응답률 : 무응답(non-response) 보정

KEEP 1차년도 조사에서 6,000명의 표본 학생 모두가 응답하여 모든 학생의 응답률은 동일(=1)하다. 따라서 학생의 무응답 보정 가중치는 고려하지 않는다.

- 학생의 무응답 보정 가중치(BY_w2) = 1

③ 가중치 총합 : 사후층화(post-stratification) 보정

학생의 가중치 총합은 학생의 추출확률 보정 가중치와 무응답 보정 가중치로 계산되며, 학생의 사후층화 보정 가중치는 학생의 가중치 총합에 대한 전체 학생 수(2004년 교육 통계연보의 3학년 학생 수)의 비율로 계산된다.

- 학생의 가중치 총합(중학생, 일반계 고등학생)

$$\text{가중치총합} = \sum_{\text{지역별}\cdot\text{성별}} BY\_w1 \times BY\_w2 = \sum_{\text{지역별}\cdot\text{성별}} \text{학생의 추출확률 보정 가중치} \times 1$$

- 학생의 가중치 총합(전문계 고등학생)

$$\text{가중치총합} = \sum_{\text{학교유형별}\cdot\text{성별}} BY\_w1 \times BY\_w2 = \sum_{\text{학교유형별}\cdot\text{성별}} \text{학생의 추출확률 보정 가중치} \times 1$$

학생의 사후층화 보정 가중치는 가중치 총합에 대한 전체 학생 수의 비율로 계산된다.

- 학생의 사후층화 보정 가중치(BY_w3)(중학생, 일반계 고등학생)

$$\text{학생의 사후층화 보정 가중치} = \frac{\text{지역별 전체 학생 수}}{\text{지역별 가중치 총합}}$$

- 학생의 사후층화 보정 가중치(BY_w3)(전문계 고등학생)

$$\text{학생의 사후층화 보정 가중치} = \frac{\text{학교유형별별 전체 학생 수}}{\text{학교유형별 가중치 총합}}$$

④ 1차년도 가중치

학생의 최종적인 가중치(BY_weight)는 학생의 추출확률 보정 가중치(BY_w1), 학생의 무응답 보정 가중치(BY_w2), 학생의 사후층화 보정 가중치(BY_w3)의 곱으로 계산된다.

- 학생의 가중치(BY_weight) $= BY\_w1 \times BY\_w2 \times BY\_w3$

= 학생의 추출확률 보정 가중치 × 1 × 학생의 사후층화 보정 가중치

(2) 2차년도 가중치

① 응답률 : 무응답(non-response) 보정

응답률은 지역, 성별, 학교유형을 기준으로 계산된다.

- 응답률 $= \frac{\text{2차년도 조사성공 표본 수}}{\text{2차년도 유효 표본 수}}$

무응답 보정 가중치는 응답률의 역수로 계산된다.

- 무응답 보정 가중치($F1\_w1$) $= \frac{1}{\text{응답률}}$

$$= \frac{\text{2차년도 유효 표본 수}}{\text{2차년도 조사성공 표본 수}}$$

② 가중치 총합 : 사후층화 보정

가중치 총합은 1차년도 가중치와 2차년도 무응답 보정 가중치로 계산되며, 사후층화 가중치는 가중치 총합에 대한 전체 학생 수의 비율로 계산된다.

- 가중치 총합(일반계 고등학생, 전문계 고등학생)

$$\text{가중치총합} = \sum_{\text{2차년도 학교지역별}\cdot\text{성별}} BY\_weight \times F1\_w1$$

$$= \sum_{\text{2차년도 학교지역별}\cdot\text{성별}} \text{1차년도가중치} \times \text{2차년도 무응답 보정 가중치}$$

- 가중치 총합(대학생)

$$\text{가중치총합} = \sum_{\text{1차년도 학교지역별}\cdot\text{성별}\cdot\text{대학진학자}} BY\_weight \times F1\_w1$$

$$= \sum_{\text{1차년도 학교지역별}\cdot\text{성별}\cdot\text{대학진학자}} \text{1차년도가중치} \times \text{2차년도 무응답 보정 가중치}$$

- 가중치 총합(취업자, 미취업자 등 비진학자)

$$\text{가중치총합} = \sum_{\text{1차년도 학교지역별}\cdot\text{비진학자}} BY\_weight \times F1\_w1$$

$$= \sum_{\text{1차년도 학교지역별}\cdot\text{비진학자}} \text{1차년도가중치} \times \text{2차년도 무응답 보정 가중치}$$

사후층화 보정 가중치는 가중치 총합에 대한 전체 학생 수의 비율로 계산된다.

- 사후층화 보정 가중치($F1\_w2$) = $\frac{\text{지역별 전체 학생 수}}{\text{지역별 가중치 총합}}$

③ 2차년도 가중치

2차년도 최종적인 가중치($F1\_weight$)는 1차년도 가중치(BY_weight), 2차년도 무응답 보정 가중치($F1\_w1$), 2차년도 사후층화 보정 가중치($F1\_w2$)의 곱으로 계산된다.

- 2차년도 가중치($F1\_weight$) = $BY\_weight \times F1\_w1 \times F1\_w2$

= 1차년도 가중치 × 2차년도 무응답 보정 가중치 × 2차년도 사후층화 보정 가중치

(3) 3차년도 가중치

① 응답률 : 무응답 보정

응답률은 지역, 성별, 학교유형을 기준으로 계산된다.

- 응답률 = $\frac{\text{3차년도 조사성공 표본 수}}{\text{3차년도 유효 표본 수}}$

무응답 보정 가중치($F2\_w1$) = $\frac{1}{\text{응답률}}$ = $\frac{\text{3차년도 유효 표본 수}}{\text{3차년도 조사성공 표본 수}}$

② 가중치 총합 : 사후층화 보정

가중치 총합은 1차년도 가중치와 3차년도 무응답 보정 가중치로 계산되며, 사후 층화 가중치는 가중치 총합에 대한 전체 학생 수의 비율로 계산된다.

- 가중치 총합(일반계 고등학생, 전문계 고등학생)

$$\text{가중치총합} = \sum_{\text{3차년도학교지역별}\cdot\text{성별}} BY\_weight \times F2\_w1$$

$$= \sum_{\text{3차년도학교지역별}\cdot\text{성별}} \text{1차년도가중치} \times \text{3차년도 무응답 보정 가중치}$$

- 사후층화 보정 가중치($F2\_w2$)(일반계 고등학생, 전문계 고등학생)

$$\text{사후층화 보정 가중치} = \frac{\text{지역별 전체 학생 수}}{\text{지역별 가중치 총합}}$$

③ 3차년도 가중치

3차년도의 최종적인 가중치($F2\_weight$)는 1차년도 가중치(BY_weight), 3차년도 무응답 보정 가중치($F2\_w1$), 3차년도 사후층화 보정 가중치($F2\_w2$)의 곱으로 계산된다.

- 3차년도 가중치($F2\_weight$)(일반계 고등학생, 전문계 고등학생)

3차년도 가중치($F2\_weight$) = BY_weight × $F2\_w1$ × $F2\_w2$

= 1차년도 가중치 × 3차년도 무응답 보정 가중치 × 3차년도 사후층화 보정 가중치

• 3차년도 가중치($F2\_weight$)(대학생, 비진학자)

3차년도 가중치($F2\_weight$) = BY_weight × $F2\_w1$

= 1차년도 가중치 × 3차년도 무응답 보정 가중치

4.2.3 한국복지패널조사 (Korea Welfare Panel Study)[22)]

한국복지실태조사는 외환위기 이후 빈곤층(또는 working poor) 및 차상위층의 가구형태, 소득수준, 취업상태가 급격히 변화하고 있는 상황에서 이들의 규모와 실태변화를 동태적으로 파악해 정책지원에 기여함과 동시에 정책에 따른 지원효과를 제고하고 소득계층별, 경제활동 상태별, 연령별 등 각 인구집단의 생활실태와 복지욕구 등을 역동적으로 파악하고 정책효과성을 평가함으로써 정책형성과 피드백에 기여하고자 한다.

한국복지실태조사는 국민의 경제 · 사회적 행태변화, 특히 빈곤층 및 차상위층의 규모 · 실태변화에 대한 분석을 통해 현재의 사회복지제도를 효과적으로 개편할 수 있는 방안을 제시하고 정책욕구 및 수요의 체계적 파악을 통해 정책우선지원순위 결정 및 그에 따른 중장기 재정수요 파악을 통한 합리적인 복지지출계획을 수립과 복지의존성이 한 세대에서 다음 세대로 이전되는지를 규명할 수 있는 빈곤의 세대 간 연구 등과 같은 심층 분석자료 제공을 통해 학술연구 발전에 기여하고 장기적인 사회통합 제고에 기여한다. 또한 정부정책 또는 경제적, 인구학적 여건변화에 대한 충격분석 및 각 정책의 효과성 평가를 통한 사회복지정책의 개선방안을 제시할 수 있다.

표본은 2005년도 인구주택총조사 90% 자료로부터 2006년 국민생활실태조사가구 30,000가구를 2단계 층화 집락 추출에 의해 추출하였고 이들 가구 중 소득계층별로 저소득층 가구와 일반가구 각 3,500가구씩을 층화집락계통

22) 한국보건사회연구원(2006),『한국복지패널 1차년도 조사자료 User's Guide』, pp16~24
_____(2007),『한국복지패널 2차년도 조사자료 User's Guide』, pp21~25

추출을 통해 총 7,000가구를 패널가구로 선정하였다. 층화 2중 추출(stratified double sampling)의 형태로 조사완료 가구수 기준으로 7,072가구를 표본으로 최종 추출하였다.

(1) 1차년도 가중치

1차년도 가중치는 2단계의 가중치를 부여한다.

Koweps 표본은 2005년도 인구주택총조사 자료의 90% 모집단으로부터 각 지역별의 크기에 비례하여 표본조사구를 추출하였으므로, 지역별 모집단 조사구와 표본조사구의 비율로서 PSU의 추출확률을 산정할 수 있다. 1단계 가중치는 지역별 조사구의 크기에 따라 확률비례로 부여된 가중치이고, 하나의 조사구에 추출된 가구는 동일한 추출확률을 가지게 됨으로 PSU당 가구가중치로 사용된다. 즉, 2006년 국민생활 실태조사에 적용한 1단계 가구가중치는 다음과 같은 방식으로 계산된다.

- 가구추출확률(selection probability)

$$
\begin{aligned}
p_{ij} &= (a \times p_i)(b_i \times p_{ij})\frac{n_{ij}}{N_{ij}} \\
&= \left(a \times \frac{N_i}{N}\right)\left(b_i \times \frac{N_{ij}}{N_i}\right) \\
&= (ab_i) \times \frac{N_{ij}}{N}
\end{aligned}
$$

여기서 a는 표본 PSU 수, b_i는 i번째 표본 PSU 내의 표본 가구수이다. n_{ij}는 i번째 PSU, j번째 가구수이며, N_{ij}는 i번째 PSU, j번째 표본 SSU내의 총가구수이다. p_i는 i번째 PSU가 표본으로 추출될 확률, p_{ij}는 i번째 PSU, j번째 표본 SSU가 추출될 확률이다.

- 2006년 국민생활실태조사 가구가중치

$$W_{1ij} = constant \times (1/p_{ij})$$

2단계 가중치는 Koweps의 기본 가중치인 W_{1ij}와 복지패널조사를 위해 7,000가구를 추출하는 과정에서의 추출확률과 조사가구에 대한 응답확률을 고려한 가중치이다.

Koweps의 가구가중치를 W_{ph}로 정의하면 일반가구와 저소득가구를 지역별(16개시도)로 구분하고 조사구별 가중치(W^{*}_{1ij}), 소득층별 가중치(W_{st})와 2005년 인구주택총조사 90% 자료를 이용한 사후층화 가중치(W_{pst})를 고려하였다.

- 가구패널가중치

$$H01_{wg} = W^{*}_{1ij} \times W_{st} \times W_{pst}$$

여기서 지역별, 조사구별 사후층화과정으로부터 W_{pst}을 부여하여 전체적인 가중치조정을 수행하였다.

가구원패널가중치는 가구조사에서 응답한 가구의 가구원을 모두 조사함으로서 각 가구원의 응답확률이 가구응답확률과 동일함으로 가구가중치를 해당 가구원에 동일하게 부여하였다. 가구원 가중치에 대한 사후 층화 가중치 조정은 이미 가구 가중치 조정 과정에서 수행하였으므로 원래의 패널 가구가중치를 직접 적용하도록 하였다.

- 가구원패널가중치

$$P01_{WG} = H01_{WG}$$

여기서 $P01_{WG_k}$는 k번째 조사대상가구의 응답 가구원의 가중치를 나타낸다.

패널가구에 속한 4~6학년 초등학교 재학생을 조사한 부가조사로서 “아동조사”에 대한 가중치는 원 패널가구의 가중치를 조사대상 아동에 적용하도록 하였다.

- 아동 패널가중치

$$C01_{WG} = h01_{WG}$$

여기서 $C01_{WG_k}$는 k번째 조사대상가구의 아동 가구원의 가중치를 나타낸다.

(2) 2차년도 가중치 조정

가구가중치는 횡단면과 종단면 가중치를 구분하지 않고 단일가중치를 부여하며, 개인가중치만 횡단면과 종단면 가중치를 부여하였다.

한국복지패널의 가중치 부여 방법은 우선 1차년도에 부여된 개인 가중치를 이용하여 해당 가구의 개인조사표에 응답한 가구원에 대해 가중치 조정을 수행하였다. 1차년도에 부여된 각 개인의 가중치는 해당 가구의 가구가중치와 동일하며, 이때 모든 가구원에 대해 가중치가 부여된 것이 아니라, 가구원 조사표를 작성한 15세 이상 가구원에 대해 가중치가 부여되었기 때문에 개인가중치는 개인조사표에 응답한 모든 가구원에게 부여하도록 하였고 가중치 부여 절차를 자세히 살펴보면 다음과 같다.

먼저 1차년도에 응답한 개인이 2차년도에도 응답한 경우의 개인 가중치는 1차년도 개인가중치를 기준으로 응답확률을 추정하여 그 값의 역수를 곱

한 값으로 가중치를 부여한다. 이때 1차년도에 조사당시에 군입대 또는 유학 등의 사유로 응답하지 않고 2차년도에 새롭게 응답한 원표본가구원의 경우에는 가구의 평균 가중치를 적용하였다.

한편 2차년도에 결혼, 또는 동거 등의 사유로 새롭게 2차년도 조사에 참여한 가구원(신규가구원)인 경우에는 2차년도에 0의 가중치를 부여받게 되는데, 왜냐하면 신규가구원은 1차년도 조사당시 조사대상이 아니었기 때문에 2차년도에 별도의 가중치를 부여할 수 없기 때문이다.

가중치의 평균을 구할 때에는 원표본가구원의 가구가중치의 합을 전체가우원수(비표본가구원+원표본가구원)로 나누어 계산한다.

원표본가구원과 결혼 등의 사유로 신규로 진입한 비표본 가구원(배우자)의 경우에는 0의 가구원 가중치를 부여하고, 원표본 가구에서 새롭게 태어난 가구원의 경우 가구가중치를 계산할 때 가구원수에서 제외하고, 가구원 조사표 조사대상이 아님으로 가중치를 부여하지 않는다. 다음으로 가구가중치는 2차 조사에서 산출된 가구원 가중치의 각 가구당 가중치의 평균을 구하여 2차 조사의 가구가중치를 산출한다.

4.2.4 대졸자 직업이동 경로조사 (Graduates Occupational Mobility Survey:GOMS)[23][24]

대졸자 직업이동 경로조사는 학교에서 노동시장으로의 학교(전공)별 이행현황 분석과 원활한 이행을 지원하기 위한 다각적인 정책수요가 증대하고 대학 졸업자의 경력개발 및 직업(직장) 이동경로를 추적 조사하여 자료를 구축함으로써, 교육 · 노동시장 간 신뢰성 있는 인력수급정보 제공 및 인력

23) 강석훈, 김영원(2007), 2006년 대졸자직업이동 경로조사 가중치 산출방안에 대한 연구, 한국고용정보원

24) 강석훈, 김영원(2008). 대졸자 직업이동경로조사 가중치 산출을 위한 연구, 한국고용정보원

수급불일치 완화를 도모하고 개별 학교 및 전공별로 세부 노동시장 성과를 학부모와 수험생에게 제공하여 대학(전공)의 합리적 선택을 유도하는 것을 목적으로 하고 있다. GOMS는 전문대학 이상 고등교육과정을 이수한 졸업자를 모집단으로 하고, 졸업자의 약 5%를 선정하여 매년 추적조사를 실시한다. 교육시장에서 노동시장으로 이행의 성과분석, 직무불일치 분석, 인력수급모형 등 광범위한 자료 활용을 기대한다. 2006년 첫 조사는 2005년 전문대 이상 대학졸업자를 대상으로 하여 패널표본을 구축하였으며 구축된 표본에 대해서는 최소 8년간 매년 추적조사를 실시할 예정이다.

(1) 1차년도 가중치

대졸자 직업이동 경로조사에서 어떤 표본이 선택될 확률은 다음과 같이 계산할 수 있다.

- Pr (목표모집단에서 조사모집단에 뽑히는 경우(사상A)
  & 조사모집단에서 사용모집단에 뽑히는 경우(사상B)
  & 사용모집단에서 전화접촉되는 경우(사상C)
  & 전화접촉된 표본에서 조사협조로 응답하는 경우(사상D)
  & 조사협조로 응답한 표본에서 실제 표본으로 추출되는 경우(사상E)
  & 실제표본으로 추출된 경우에 조사에 응답할 경우(사상F))

확률A, 확률B, 확률C를 구하기 위한 추가적인 정보를 사용하지 않고 각 150개 셀에서 독립적으로 구한다. 확률D, 확률E, 확률F를 구하기 위해 가능한 많은 정보를 이용하는 것이 좋으나, 현실적으로 존재하는 조사거부자에 대한 정보는 학교급, 지역, 전공, 성, 연령 등의 자료이므로 이러한 정보를 이용하여 응답확률을 계산한다. 각각의 응답확률은 로짓모형을 이용하여 산출한다.

<표 3> GOMS 기본가중치

기본가중치 1	DW1=(확률A)*(확률B)*(확률C)*(확률G) (확률A)*(확률B)*(확률C)=셀별전체전화응답자수/셀별모집단수 (확률G)=전화접촉자 중 최종응답자가 될 확률
기본가중치 2	DW2=(확률A)*(확률B)*(확률C)*(확률D)*(확률E)*(확률F) (확률A)*(확률B)*(확률C)=셀별전체전화응답자수/셀별모집단수(가중치1과 동일) (확률D)=전화접촉자 중 조사협조자가 될 확률 (확률E)=조사협조자수 중 최종 접촉자가 될 확률 (확률F)=최종 접촉자 중 최종응답자가 될 확률

추정된 응답확률을 바탕으로 최종 가중치를 산출하였다. 먼저 DW1과 DW2의 역수를 이용하여 가중치 FW1, FW2를 산정하였다. 최종적으로 모집단의 학교급별(대학, 전문대, 교육대), 성별(남, 여), 권역별(서울권, 경기권, 충청권, 경상권, 전라권), 전공대계열별(인문계, 사회계, 교육계, 공학계, 자연계, 의약계, 예체능계)로 구분된 150개 셀의 모집단수와 일치하도록 조정하여 최종가중치 FW1_AT, FW2_AT를 산출하였다.

(2) 2차년도 가중치

2차년도 가중치는 1차년도의 가중치에 응답률을 곱하여 산출하였다. 사후층화는 1차년도 모집단의 기본적인 분포를 이용한다.

1차년도 사후층화 전 가중치(w1_sf) 응답률에 로짓모형을 이용하여 계산

한 응답률의 역수(1/q_semi)를 곱하여 2차년도 사후층화 전 가중치(w2_sf)를 산정하였다. 2차년도 대졸자 직업이동 경로조사의 경우 새롭게 표본에 진입한 경우는 없고, 표본에서 탈락한 경우만 존재하므로 횡단면 가중치와 종단면가중치가 구분되지 않는다. 만약 3차년도에서 새로운 표본이 추가되거나 1차년도에서 응답하고 2차년도에서 응답하지 않고, 3차년도에서 응답한 경우가 발생하면 횡단면가중치와 종단면가중치의 구분이 필요하게 된다.

사후층화 전 가중치는 표본을 이용하여 산출하는 총계치가 학교급별, 권역별, 전공별, 성별 모집단의 수에 일치하도록 비율조정하여 사후층화 후 최종 2차년도 가중치를 산출하였다.

5. 여성가족패널의 가중치 부여 방안[25)]

여성가족패널조사(KLoWF: Korean Longitudinal Survey of Women & Family)는 2006년부터 여성의 생활세계와 가족의 구조 및 변화실태를 파악하기 위해서 실시하는 전국 규모의 패널조사이다. 이 조사는 여성의 생애주기별 경제활동 지위변화와 가족생활과의 관계를 비롯하여, 가족과 관련한 가치의 변화, 가족관계의 변화, 가족과정(family formation process)과 이벤트의 변화, 가족구조의 변화를 횡단면(cross sectional)뿐만 아니라 종단면적(longitudinal)으로도 추적할 수 있는 방대한 자료 구축을 목적으로 하고 있다. 이러한 자료구축을 통하여 여성가족패널조사는 여성의 경제활동 증가와 이로 인한 일-가족생활 조화(work-family life balance), 일-가족생활 전환실태(work-family life transition) 등을 파악함으로써 여성의 생활세계를 경험적으로 규명하고 우리 사회 가족의 현재와 미래를 전망할 수 있도록 해준다. 또한 여성가족패널조사의 방대한 조사결과는 포괄적인 여성 및 가족정책 수립의 주요 기초 자료로 활용됨으로써 사회구성원들이 부모 된 권리

25) 박수미 외(2008), 2008년 여성가족패널조사 사업보고서 및 제1차 기초분석보고서, 한국여성정책연구원

와 노동자로서의 권리를 조화롭게 영위할 수 있도록 하는데 일조한다.

여성가족패널은 전국에 거주하는 만 19세 이상 64세 이하의 여성 약 10,000명을 대상으로 여성의 경제활동과 가족관계 등에 대한 사항을 매년 추적해 조사하고 있다. 표본기획단계에서는 적격여성가구원 10,000명을 조사하기 위해 예비조사결과에 따라 전국의 약 8,500가구를 추출한다. 통상적인 서베이자료의 경우 모집단분석을 위해 가중치를 사용해야한다. 여성가족패널조사의 경우, 16개 시도별 · 지역별 추출률이 상이하기 때문에 가중치를 사용하여야 할 필요성이 더욱 커진다.

5.1 1차년도 가중치

여성가족패널조사의 1차년도 자료는 횡단면자료이므로 횡단면가중치만을 계산한다.

가중치의 종류로는 가구가중치, 기록가구원가중치, 응답가구원가중치 등 세 가지 종류의 가중치를 고려할 수 있다. 가구가중치는 가구를 분석 단위로 사용할 때 사용할 수 있는 가중치이다. 기록가구원가중치는 가구에 존재하는 적격여성가구원을 단위로 분석할 때 사용할 수 있는 가중치이다. 이때 기록가구원가중치는 적격여성가구원의 응답여부에 관계없이 모두 부여된다. 응답가구원가중치는 가구에서 응답한 적격여성가구원을 단위로 분석할 때 사용하는 가중치이다.

엄격한 확률표본에서는 각 표본이 추출될 확률이 사전적으로 결정되지만 여성가족패널조사의 경우 각 ED내의 구체적인 적격가구 리스트가 없기 때문에 각 표본이 추출될 확률은 최종단계에서 임의로 추출되는 표본의 응답확률과 밀접하게 연관된다. 이에 따라 표본추출확률과 응답률을 분리하여 고려하기 어렵기 때문에 두 가지를 동시에 고려한다.

- 표본가구추출확률=(ED추출확률)*(ED내에서 적격가구가 표본가구로 추출되고 가구용설문서에 응답할 확률)
- 표본기록가구원추출확률=(ED추출확률)*(ED내에서 적격가구가 표본가구로 추출되고 가구용설문서에 응답할 확률)
- 표본응답가구원추출확률=(ED추출확률)*(ED내에서 적격가구가 표본가구로 추출되고 적격가구원이 개인용설문서에 응답할 확률)

위의 세 가지 모든 경우에 첫 번째 부분인 ED추출확률은 각 가구가 속한 지역의 가구수총수의 제곱근에 비례한다.

- ED추출확률∝지역별 가구수제곱근

따라서 세 가지 경우에서 모두 우변의 두 번째 확률을 구하는 방법을 고려해야하지만 이 과정은 매우 복잡하다.

5.1.1 가구가중치

체계적인 응답률 계산이 불가능하기 때문에 최종적으로는 각 ED내에 거주하는 가구의 경우 응답률에 체계적인 차이가 없다는 가정에서 기본가중치는 다음과 같이 구한다.

- 가구가중치 기본가중치(XW1_H)=(ED추출확률)*(적격응답가구/적격가구)

5.1.2 가구원가중치

기록가구원과 응답가구원의 기본가중치는 가구최종가중치를 사용한다.

- 기록가구원 기본가중치(XW1_PE)=(ED추출확률)*(적격응답가구/적격가구)
- 응답가구원 기본가중치(XW1_PR)=(ED추출확률)*(적격응답가구/적격가구)

이제 기록가구원의 최종가중치는 기록가구원기본가중치에 벤치마킹 조정을 실시하여 얻을 수 있다. 기록가구원의 경우에는 16개 시도별 연령별 여성적격가구원수(19~24세, 25~29세, 30~34세, 35~39세, 40~44세, 45~49세, 50~54세, 55~59세, 60~64세)를 벤치마킹변수로 사용하여 조정하였다. 가구최종가중치를 이용하여 벤치마킹조정을 실시한 이후 각 지역별로 평균값의 1/3이하와 3배를 기준으로 절삭하였다. 이렇게 산정된 기록가구원개인가중치는 W1_PE이다.

응답가구원의 경우에는 가구최종가중치를 바탕으로 응답확률을 계산하여 응답률의 역수를 곱한 후 다시 시도별 연령별 여성적격가구원을 벤치마킹변수로 사용하여 조정하였다. 가구최종가중치에 응답확률의 역수를 곱하여 응답개인기본가중치를 산정하였으며, 이 기본응답개인가중치를 여성적격가구원 벤치마킹 변수를 사용하여 조정하였다(벤치마킹변수는 기록개인가구원과 동일하게 사용함). 이렇게 조정된 가중치는 다시 지역별로 평균의 1/3를 최소값으로, 평균의 3배를 최대값으로 절삭한 다음 다시 시도별 적격가구수와 동일하게 평균 비율 조정을 실시하였다.

5.2 2차년도 가중치 부여방법[26)]

26) 강석훈(2009), 여성가족패널 2차 웨이브 이후의 가중치 부여방안 연구. 한국여성정책연구원(내부자료)

여성가족패널자료는 횡단면분석과 종단면분석에 모두 사용될 수 있도록 설계되었으므로 횡단면가중치와 종단면가중치를 각각 작성한다. 분석단위는 가구단위와 개인단위가 있을 수 있으며, 다시 개인의 경우에는 기록개인과 응답개인으로 구분할 수 있다. 현재 한국의 연구자들이 응답개인이외에 기록개인을 이용하여 분석하는 경우는 사례가 많지 않은 편이지만 국제적인 패널조사에서 기록개인에게도 가중치를 부여하는 경우가 많다는 점과 향후 기록개인에 대한 분석수요가 늘어날 수 있다는 점, 기록개인가중치가 2차 웨이브 이후의 횡단면개인가중치를 작성하는 데 기초가중치로 사용된다는 점을 고려하여 응답개인과 기록개인가중치를 모두 계산하고자 한다.

이 절에서는 여성가족패널의 2차 웨이브에서의 종단면 가중치 차원으로 종단면기록개인가중치와 종단면응답개인가중치 부여방법을 보다 자세히 다루고 있으며 횡단면 가중치는 부여방법에 대해서만 설명하고 있다.

5.2.1 종단면기록개인가중치

<표 4>은 여성가족패널의 2차 웨이브 가구데이터에 나타난 기록된 개인 중에서 2차 웨이브에서 탈락한 경우와 기록된 경우의 빈도와 비율을 나타내고 있다. 1차 웨이브의 전체 기록개인은 31,505명이고 그 중 2차 웨이브에도 기록된 경우는 30,127명이며 기록되지 않은 사람은 1,378명으로 나타나 원시표본 대비 95.63%의 높은 기록률과 4.37%의 탈락률을 기록하였다.

<표 4> 2차년도 기록개인 여부

enum	Freq.	Percent	Cum.
0	1,378	4.37	4.37
1	30,127	95.63	100
Total	31,505	100	

종단면기록개인가중치를 구하기 위해 먼저 1차 웨이브에서의 가구최종가중치(h01weight)를 1차 웨이브의 기록개인가중치로 설정한다. 2차 웨이브의 모든 기록개인을 1차 웨이브에 나타났던 기록개인(원시표본가구원)과 1차 웨이브에 나타나지 않았고 새롭게 진입한 비표본가구원으로 구분한다. 사용한 변수의 기초 통계량은 1차 웨이브에서 나타난 모든 기록개인을 2차 웨이브에서 기록개인으로 나타난 경우(enum=1)와 2차 웨이브에서는 기록개인으로 나타나지 않은 경우(enum=0)로 구분하여 각각 <표 5>, <표 6>에 제시되어 있다.

<표 5> 변수의 기초통계량 : 1차 기록되고 2차 기록된 경우[27)]

변수	변수설명	1차 기록되고 2차 기록된 경우			
		평균	표준편차	최소값	최대값
age	연령	32.741	20.130	-9	107
reg_1	서울특별시	0.126	0.332	0	1
reg_2	부산광역시	0.074	0.262	0	1
reg_3	대구광역시	0.054	0.227	0	1
reg_4	인천광역시	0.061	0.238	0	1
reg_5	광주광역시	0.050	0.218	0	1
reg_6	대전광역시	0.049	0.216	0	1
reg_7	울산광역시	0.044	0.204	0	1
reg_8	경기도	0.122	0.327	0	1
reg_9	강원도	0.046	0.209	0	1
reg_10	충청북도	0.048	0.213	0	1
reg_11	충청남도	0.053	0.225	0	1
reg_12	전라북도	0.052	0.222	0	1
reg_13	전라남도	0.059	0.236	0	1

27) age변수의 -9값은 모름/무응답을 의미하는 값으로, 0세와는 다른 의미라고 생각됨. 추후 -9값의 처리에 대해 논의해 볼 필요가 있음. enum=1일 때 age가 -9의 값을 갖는 obs의 수는 2개이고 enum=0일 때 age가 -9의 값을 갖는 obs의 수는 5개에 불과함

reg_14	경상북도	0.065	0.247	0	1
reg_15	경상남도	0.069	0.254	0	1
reg_16	제주도	0.028	0.164	0	1
job_1	일자리유무	0.587	0.492	0	1
wed	결혼유무	0.405	0.491	0	1

<표 6> 변수의 기초통계량 : 1차 기록되고 2차 기록되지 않은 경우

변수	변수설명	1차 기록되고 2차 기록되지 않은 경우			
		평균	표준편차	최소값	최대값
age	연령	27.032	10.700	-9	89
reg_1	서울특별시	0.071	0.257	0	1
reg_2	부산광역시	0.064	0.245	0	1
reg_3	대구광역시	0.041	0.198	0	1
reg_4	인천광역시	0.028	0.164	0	1
reg_5	광주광역시	0.049	0.217	0	1
reg_6	대전광역시	0.069	0.253	0	1
reg_7	울산광역시	0.039	0.194	0	1
reg_8	경기도	0.060	0.237	0	1
reg_9	강원도	0.078	0.269	0	1
reg_10	충청북도	0.084	0.278	0	1
reg_11	충청남도	0.050	0.218	0	1
reg_12	전라북도	0.083	0.276	0	1
reg_13	전라남도	0.128	0.334	0	1
reg_14	경상북도	0.074	0.262	0	1
reg_15	경상남도	0.067	0.250	0	1
reg_16	제주도	0.016	0.125	0	1
job_1	일자리유무	0.159	0.366	0	1
wed	결혼유무	0.434	0.496	0	1

2차 웨이브의 종단면기록개인가중치를 계산하기 위해 1차 웨이브에 이어 2차 웨이브에도 기록된 경우를 1로 하는 로짓분석을 실시하였다. 가구 및 개인특성을 설명변수로 사용하여 계속해서 기록개인으로 지속될 확률을 구

한다. 사용한 설명변수는 연령, 연령의 제곱, 거주지역 더미, 일자리유무, 결혼유무를 사용하였다. <표 7>에 제시되어 있는 추정결과를 보면 대부분의 변수가 유의하게 나왔으나 <표 5>와 <표 6>에서 보여주는 age의 평균값이 2차 웨이브에서의 탈락자의 연령이 더 낮은 것으로 보아 변수간 공선성 문제가 발생한 것으로 보인다.[28)]

이 모형으로부터 도출된 응답확률의 역수를 1차 웨이브의 기록개인가중치에 곱하여 2차 웨이브의 무응답조정 후 사후층화 전 기록개인가중치(pw02_2)를 구하였다.

마지막으로 이용 가능한 1차 웨이브 사후층화정보(여성가족패널조사의 응답대상이 되는 가구에 속하는 모든 개인자료)를 이용하여 사후층화를 실시하여 종단면기록개인가중치를 작성하였다.

이렇게 산정된 종단면기록개인가중치는 enum_pw02이다. <표 8>은 사후층화 전 기록개인가중치와 사후층화 후 최종 기록개인가중치의 기초통계량을 나타내고 있다. 종단면기록개인 응답확률을 구하기 위해서 다양한 모형을 고려해 볼 수 있는데 본고에서는 프로빗 모형을 이용하여 가중치를 계산해보고 결과를 비교해 보았다. 프로빗분석 결과는 부록의 <부표 1>을 참고하면 된다. 두 모형의 응답확률 비교 결과 로짓모형과 프로빗모형 사용에 따른 차이는 거의 없는 것으로 나타나 1차 웨이브에서의 가중치 산정 때와 마찬가지로 일반적인 로짓모형을 이용하여 응답확률을 고려하기로 하였다.

<표 9>는 가중치를 부여하기 전후의 변수별 비중을 제시하고 있다. 단순비중과 가중치비중을 비교해보면 서울의 경우 단순비중은 12.4%이고 가중치비중은 21.38%로서 양자의 차이가 8.98%p에 달하고 있다. 경기도의 경우에도 단순비중은 11.92%이지만 가중치비중은 22.88%로서 양자의 비중은 10.96%p의 차이를 보이고 있다. 이처럼 단순비중과 가중치 비중의 차이가 크

28) age변수만 넣고 로짓분석을 하였을 경우, age의 계수값이 0.015의 값이 나와 연령이 증가할수록 기록률이 높아지는 것으로 나타남

게 나타나는 것은 표본추출과정에서 제곱근 비례방식의 추출방식을 취했기 때문이다.[29] 한편, 다른 지역의 경우에는 모두 가중치 비중이 단순비중보다 2%p 내외로 작았다.

<표 7> 종단면기록개인 : 로짓분석 결과

구분	계수추정치	표준오차	z-값	p-값
age	-0.285[30]	0.011	-25.71	0.000
age2	0.003	0.000	20.06	0.000
reg_2	-0.382	0.157	-2.43	0.015
reg_3	-0.483	0.179	-2.70	0.007
reg_4	0.104	0.201	0.52	0.604
reg_5	-0.679	0.170	-3.99	0.000
reg_6	-0.996	0.157	-6.35	0.000
reg_7	-0.681	0.182	-3.73	0.000
reg_8	-0.157	0.158	-0.99	0.322
reg_9	-1.257	0.155	-8.11	0.000
reg_10	-1.377	0.152	-9.06	0.000
reg_11	-0.863	0.170	-5.07	0.000
reg_12	-1.230	0.152	-8.11	0.000
reg_13	-1.746	0.140	-12.47	0.000
reg_14	-0.983	0.154	-6.40	0.000
reg_15	-0.822	0.157	-5.25	0.000
reg_16	-0.132	0.251	-0.52	0.600
job_1	0.109	0.072	1.51	0.132
wed	3.576	0.111	32.12	0.000
_cons	6.954	0.192	36.21	0.000

Number of obs=31505
LR chi2(19)=2925.04
Prob>chi2=0.0000
Pseudo R2=0.2584

29) 박수미 외(2008), 2008년 여성가족패널조사 사업보고서 및 제1차 기초분석보고서, 한국여성정책연구원
30) age 변수만을 넣고 로짓분석을 한 결과는 (+)인 계수값이 나온다.

<표 8> 종단면기록개인가중치의 기초통계량 (관측치수:30,127)

변수	변수설명	평균	표준편차	최소값	최대값
pw02_2	2차년도기록개인가중치(사후층화전_로짓)	1480.235	822.4664	212.568	5507.792
enum_pw02	2차년도기록개인가중치(사후층화후_로짓)	1270.238	712.7015	178.8783	4925.888
pw02_p2	2차년도기록개인가중치(사후층화전_프로빗)	1480.424	825.1103	212.6674	5597.788
enum_pw02	2차년도기록개인가중치(사후층화후_프로빗)	1270.238	711.9482	180.9156	5000.792

<표 9> 종단면기록개인가중치 사용 전과 후의 변수별 기록개인비중

구분	단순비중	가중치비중	차이
서울특별시	12.4	21.38	-8.98
부산광역시	7.35	7.65	-0.3
대구광역시	5.37	5.39	-0.02
인천광역시	5.91	5.48	0.43
광주광역시	4.99	3.11	1.88
대전광역시	5	3.19	1.81
울산광역시	4.35	2.37	1.98
경기도	11.92	22.88	-10.96
강원도	4.71	2.85	1.86
충청북도	4.94	2.84	2.1
충청남도	5.32	3.65	1.67
전라북도	5.33	3.45	1.88
전라남도	6.21	3.42	2.79
경상북도	6.58	5.11	1.47
경상남도	6.89	6.18	0.71
제주도	2.73	1.05	1.68
일자리없음	59.41	61.21	-1.8

일자리있음	40.59	38.79	1.8
미혼	43.18	44.37	-1.19
기혼	56.82	55.63	1.19

5.2.2 종단면응답개인가중치

종단면응답개인가중치를 구하기 위해 먼저, 1차 웨이브 응답자를 2차 웨이브에서의 응답자(resp=1)와 응답거부자(resp=0)로 구분한다. 이 때 사망자나 해외이주자, 집단시설거주지로 이동한 사람, 비민간거주지역으로 이동한 사람 등은 응답자로 간주한다.

<표 10>은 여성가족패널의 2차 웨이브 개인데이터에 나타난 응답개인 중에서 2차 웨이브에서 탈락한 경우와 응답한 경우의 빈도와 비율을 나타내고 있다. 1차 웨이브의 전체 응답개인은 9,997명이고 그 중 2차 웨이브에도 응답한 경우는 8,364명이고 응답하지 않은 사람은 1,633명으로 나타나 원시표본 대비 83.67%의 응답률과 16.33%의 탈락률이 나타냈다.

1차년도 응답하고 2차년도 응답한 경우와 1차년도 응답하고 2차년도 응답하지 않은 경우의 변수의 기초통계량이 <표 11>와 <표 12>에 제시되어 있다.

<표 10> 2차년도 응답개인 여부

resp	Freq.	Percent	Cum.
0	1,633	16.33	16.33
1	8,364	83.67	100
Total	9,997	100	

<표 11> 변수의 기초통계량 : 1차년도 응답하고 2차년도 응답한 경우

변수	변수설명	1차응답하고 2차응답한 경우			
		평균	표준편차	최소값	최대값
age	연령	42.346	11.539	19	64
job_1	일자리유무	0.445	0.497	0	1
reg_1	서울특별시	0.112	0.315	0	1
reg_2	부산광역시	0.078	0.268	0	1
reg_3	대구광역시	0.057	0.232	0	1
reg_4	인천광역시	0.053	0.224	0	1
reg_5	광주광역시	0.048	0.213	0	1
reg_6	대전광역시	0.054	0.225	0	1
reg_7	울산광역시	0.045	0.207	0	1
reg_8	경기도	0.098	0.298	0	1
reg_9	강원도	0.050	0.218	0	1
reg_10	충청북도	0.052	0.223	0	1
reg_11	충청남도	0.058	0.233	0	1
reg_12	전라북도	0.057	0.232	0	1
reg_13	전라남도	0.065	0.246	0	1
reg_14	경상북도	0.068	0.252	0	1
reg_15	경상남도	0.072	0.259	0	1
reg_16	제주도	0.033	0.179	0	1
wed	결혼유무	0.894	0.308	0	1
ch_birth	출산경험유무	0.857	0.350	0	1
r_form1	모름/무응답/ 무상및기타	0.669	0.471	0	1
r_form2	자가	0.182	0.386	0	1
r_form3	전세	0.108	0.310	0	1
r_form4	보증부월세/ 월세(사글세포함)	0.041	0.199	0	1
eh	적격가구원수	1.234	0.516	1	4
hh[31]	가구원수	3.549	1.202	1	9

<표 12> 변수의 기초통계량 : 1차년도 응답하고 2차년도 미응답한 경우

변수	변수설명	1차응답하고 2차미응답한 경우			
		평균	표준편차	최소값	최대값
age	연령	39.258	11.762	19	64
job_1	일자리유무	0.346	0.476	0	1
reg_1	서울특별시	0.208	0.406	0	1
reg_2	부산광역시	0.077	0.266	0	1
reg_3	대구광역시	0.042	0.200	0	1
reg_4	인천광역시	0.096	0.294	0	1
reg_5	광주광역시	0.037	0.190	0	1
reg_6	대전광역시	0.036	0.185	0	1
reg_7	울산광역시	0.023	0.149	0	1
reg_8	경기도	0.188	0.391	0	1
reg_9	강원도	0.033	0.179	0	1
reg_10	충청북도	0.036	0.187	0	1
reg_11	충청남도	0.029	0.167	0	1
reg_12	전라북도	0.031	0.172	0	1
reg_13	전라남도	0.026	0.158	0	1
reg_14	경상북도	0.051	0.221	0	1
reg_15	경상남도	0.077	0.266	0	1
reg_16	제주도	0.013	0.113	0	1
wed	결혼유무	0.792	0.406	0	1
ch_birth	출산경험유무	0.745	0.436	0	1
r_form1	모름/무응답/무상및기타	0.573	0.495	0	1
r_form2	자가	0.258	0.438	0	1
r_form3	전세	0.135	0.342	0	1
r_form4	보증부월세/월세(사글세 포함)	0.034	0.180	0	1
eh	적격가구원수	1.374	0.613	1	4

31) 가구원수는 동거가구원과 일시적비동거가구원의 합

hh	가구원수	3.519	1.147	1	9

응답확률을 구하기 위해 가구 및 개인특성을 이용하여 응답률을 추정하는 로짓모형을 설정한다. 1차 웨이브에서 응답확률을 계산한 경우에는 적격응답자 중에 실제응답자의 응답확률을 계산하였기 때문에 사용할 수 있는 변수가 제약적이었으나 1차에서 2차로 넘어가면서 사용할 수 있는 변수가 많기 때문에 보다 다양한 응답확률모형을 고려해 볼 수 있다[32].

1차 및 2차 웨이브에서 동시에 응답한 개인을 1로 하는 로짓분석을 하였다. 설명변수에는 개인변수로서 연령, 직업보유여부, 직종, 결혼여부, 가구주와의 관계, 출산경험유무, 가구변수로서 거주지역, 거주형태(자가, 전세, 월세, 기타), 가구원 수 등을 이용하여 응답확률모형을 고려해 보았으나, 모든 응답자에 대하여 가용하지 않았기 때문에 응답확률을 통계적으로 계산하기에는 불가능하였다. 이러한 이유로 본고에서 사용한 응답률 분석 모형은 <표 13>과 같다. 1차년도 응답개인가중치에 이 모형으로부터 도출된 응답확률의 역수를 곱하여 무응답조정 후 사후층화 전 종단면응답개인가중치 pw02_1을 산정하였다.

이 기본응답개인가중치를 16개 시도별, 연령별 여성적격가구원수(19~24세, 25~29세, 30~34세, 35~39세, 40~44세, 45~49세, 50~54세, 55~59세, 60~64세)를 벤치마킹 변수로 사용하여 사후층화를 실시하였다. 이러한 사후층화 과정을 거쳐 구해진 최종기록개인가중치는 pw2이다.

<표 14>는 사후층화 전, 후의 기록개인가중치의 기초통계량을 비교하고 있다. 프로빗분석의 결과가 로짓분석보다 다소 결정계수가 높게 나타났지만 사후층화 후 최종 응답개인가중치의 기초통계량에는 차이가 거의 없게 나타났다.

32) 부록의 <부표 2>에 프로빗모형으로 응답률을 구한 결과가 있음

<표 13> 종단면응답개인 : 로짓분석 결과

구분	계수추정치	표준오차	z-값	p-값
age	-0.026	0.021	-1.21	0.224
age2	0.000	0.000	1.56	0.118
job_1	0.401	0.061	6.60	0.000
reg_2	0.523	0.119	4.39	0.000
reg_3	0.902	0.146	6.15	0.000
reg_4	-0.054	0.115	-0.46	0.642
reg_5	0.641	0.154	4.15	0.000
reg_6	0.970	0.156	6.23	0.000
reg_7	1.188	0.186	6.38	0.000
reg_8	-0.160	0.095	-1.69	0.092
reg_9	0.834	0.161	5.17	0.000
reg_10	0.787	0.156	5.05	0.000
reg_11	1.028	0.168	6.10	0.000
reg_12	0.949	0.166	5.73	0.000
reg_13	1.144	0.176	6.49	0.000
reg_14	0.613	0.137	4.48	0.000
reg_15	0.346	0.121	2.87	0.004
reg_16	1.513	0.239	6.33	0.000
wed	0.322	0.163	1.97	0.049
ch_birth	0.164	0.147	1.12	0.263
r_form1	0.364	0.089	4.08	0.000
r_form2	0.077	0.098	0.78	0.435
r_form4	0.378	0.169	2.23	0.025
eh	-0.332	0.063	-5.23	0.000
hh	0.083	0.031	2.67	0.008
_cons	0.839	0.413	2.03	0.042

Number of obs=9997
LR chi2(25)=581.36
Prob>chi2=0.0000
Pseudo R2=0.0653

<표 14> 종단면응답개인가중치의 기초통계량(관측치수:8,364)

변수	변수설명	평균	표준편차	최소값	최대값
pw02_1	2차년도응답개인가중치 (사후층화전_로짓)	1847.482	1483.332	201.2089	11072.62
pw2	2차년도응답개인가중치 (사후층화후_로짓)	1871.094	1547.767	199.5936	13517.3
pw02_p	2차년도응답개인가중치 (사후층화전_프로빗)	1847.174	1477.403	200.5704	10772.55
pw2_probit	2차년도응답개인가중치 (사후층화후_프로빗)	1871.094	1546.267	199.3233	13517.57

<표 15>는 가중치를 부여하기 전후의 변수별 응답개인비중을 제시하고 있다. 단중비중과 가중치비중을 비교해보면 서울특별시의 경우 단순비중은 12.74%이지만, 가중치 비중은 22.64%로서 양자의 차이가 9.9%p에 달하고 있다. 경기도의 경우에도 단순비중은 11.29%이지만 가중치 비중은 21.87%로서 양자의 비중은 10.58%p의 차이를 보이고 있다. 이처럼 단순비중과 가중치 비중의 차이가 크게 나타나는 것은 종단면기록개인가중치에서와 마찬가지로 표본추출과정에서 제곱근 비례방식의 추출방식을 취했기 때문이다.

<표 15> 종단면응답개인가중치 사용 전과 후의 변수별 응답개인비중

구분	단순비중	가중치비중	차이
서울특별시	12.74	22.64	-9.9
부산광역시	7.78	7.94	-0.16
대구광역시	5.46	5.38	0.08
인천광역시	5.99	5.42	0.57
광주광역시	4.61	3.04	1.57
대전광역시	5.06	3.13	1.93
울산광역시	4.12	2.2	1.92
경기도	11.29	21.87	-10.58

강원도	4.73	2.89	1.84
충청북도	4.96	2.89	2.07
충청남도	5.29	3.6	1.69
전라북도	5.29	3.45	1.84
전라남도	5.84	3.28	2.56
경상북도	6.54	5.09	1.45
경상남도	7.29	6.12	1.17
제주도	2.98	1.05	1.93
일자리없음	57.14	60.47	-3.33
일자리있음	42.86	39.53	3.33
미혼	12.26	19.13	-6.87
기혼	87.74	80.87	6.87
출산경험있음	16.13	23.33	-7.2
출산경험없음	83.87	76.67	7.2
점유형태-모름/무응답/무상및기타	65.29	63.67	1.62
점유형태-자가	19.48	22.7	-3.22
점유형태-전세	11.21	10.26	0.95
점유형태-보증부월세/월세(사글세포함)	4.02	3.37	0.65
적격가구원수1	78.84	71.09	7.75
적격가구원수2	16.9	22.04	-5.14
적격가구원수3	3.99	6.12	-2.13
적격가구원수4	0.27	0.75	-0.48
가구원수1	4.75	4.67	0.08
가구원수2	15.76	13.65	2.11
가구원수3	21.89	22.37	-0.48
가구원수4	40.66	42.23	-1.57
가구원수5	12.96	13.28	-0.32
가구원수6	3.07	2.9	0.17
가구원수7	0.63	0.52	0.11
가구원수8	0.24	0.35	-0.11
가구원수9	0.03	0.03	0
연령(19~24)	6.64	11.9	-5.26
연령(25~29)	8.79	12.01	-3.22

연령(30~34)	14.14	12	2.14
연령(35~39)	16.93	13.57	3.36
연령(40~44)	13.42	12.74	0.68
연령(45~49)	11.49	13.21	-1.72
연령(50~54)	10.53	10.36	0.17
연령(55~59)	9.28	7.8	1.48
연령(60~64)	8.76	6.42	2.34

5.2.3 횡단면가중치

2차년도 데이터에서 횡단면가중치는 횡단면기록개인가중치와 횡단면응답개인가중치, 횡단면가구가중치를 고려할 수 있다.

횡단면기록개인가중치를 구하기 위해서는 앞에서 사용한 종단면기록개인가중치를 작성하는 로짓모형을 이용하여 1차 및 2차 웨이브에 모두 존재한 기록개인들의 무응답조정 후 기록개인가중치를 구한다. 이는 1차 웨이브의 사후층화 후 기록개인가중치에 지속확률의 역수를 곱하여 산출한다. 이렇게 산출된 2차 웨이브 기록개인가중치의 합을 비표본가구원을 포함하여 현재 가구원으로 나누어 모든 가구원의 2차 웨이브 횡단면개인기록가중치로 설정한다. 이렇게 작성된 무응답조정 후 공정배분 방식 적용 후의 횡단면기록개인가중치는 가용한 모집단의 2차 웨이브 정보를 이용하여 사후층화하여 최종 횡단면기록개인가중치를 작성한다.

횡단면응답개인가중치를 구하기 위해서 먼저 2차 웨이브 자료에서 적격응답개인을 응답한 경우와 응답하지 않은 경우로 구분하여 응답률을 추정한다. 주로 2차 웨이브에서의 자료를 이용하며, 로짓모형을 사용하여 계산한다. 2차 웨이브의 횡단면기록개인가중치에 응답확률의 역수를 곱하여 표본가구원과 비표본가구원의 무응답조정 후 횡단면응답개인가중치를 작성한다. 이렇게 작성된 가중치에 가용한 모집단의 2차 웨이브 정보(적격응답대상에

한정)를 이용하여 사후층화하여 최종 횡단면응답개인가중치를 작성한다.

횡단면가구가중치는 횡단면기록개인가중치와 동일하게 설정한 후 알려진 2차 웨이브의 모집단 가구정보를 이용하여 사후층화하여 최종횡단면가구가중치를 작성한다.

6. 결론

본 논문에서는 가중치의 의미와 일반적인 횡단면데이터를 중심으로 한 가중치 부여방안에 대해 알아본 후 패널데이터에서의 횡단면, 종단면 가중치의 부여방안에 대한 기초 이론을 살펴보았다.

또한 기존 패널데이터의 가중치 부여방안을 알아보기 위해 외국의 대표적인 패널조사인 PSID, BHPS, GSOEP, SIPP, NLS와 국내 패널조사인 한국노동패널, 한국교육고용패널, 한국복지패널조사, 대졸자직업이동경로조사에서의 가중치 부여방법에 대해 알아보았다.

개인과 가구를 단위로 하는 패널조사에서는 횡단면분석용 개인, 가구단위 가중치가 필요하며 종단면분석용 개인, 가구단위의 가중치가 필요하고, 개인가중치도 가구조사에서 조사된 모든 개인에게 부여하는 기록개인가중치와 응답한 개인에게만 부여하는 응답개인가중치를 모두 작성하는 경우가 많다. 따라서 본고에서는 보다 구체적인 가중치 산정 방법을 알아보기 위해 여성가족패널 2차년도 자료를 이용하여 종단면기록개인가중치와 종단면응답개인가중치를 산정해 보았다.

먼저 종단면기록개인가중치를 산정하기 위해서는 1차 웨이브에서의 가구최종가중치를 1차 웨이브 기록개인가중치로 설정한 후 2차 웨이브의 모든 기록개인을 1차 웨이브에 나타났던 기록개인과 1차 웨이브에 나타나지 않았고 새롭게 진입한 비표본가구원으로 구분한다. 또한 1차 웨이브에서 나타난 모든 기록개인을 2차 웨이브에서 기록개인으로 나타난 경우와 기록개인으로

나타나지 않은 경우로 구분한다. 다음에 가구 및 개인특성을 설명변수로 사용하는 로짓모형을 통해 기록개인으로 지속될 확률을 구하고, 이 모형으로부터 도출된 응답확률의 역수를 1차 웨이브의 기록개인가중치에 곱하여 2차 웨이브의 무응답조정 후 사후층화전 기록개인가중치를 구한다. 마지막으로 이용 가능한 1차 웨이브 사후층화 정보를 이용하여 사후층화를 실시하여 종단면기록개인가중치를 작성하였다.

종단면응답개인가중치를 산정하기 위해서는 우선 1차 웨이브 응답자를 2차 웨이브에서의 응답자와 응답거부자로 구분한 후에 가구 및 개인특성을 이용하여 응답률을 추정하는 로짓모형을 설정하였다. 이 모형으로부터 도출된 응답확률의 역수를 1차 웨이브 응답개인가중치에 곱하여 무응답조정 후 사후층화 전 응답개인가중치를 구하였다. 마지막으로 기본응답개인가중치를 이용 가능한 1차 웨이브 사후층화 정보를 이용하여 사후층화를 실시하여 최종 종단면응답개인가중치를 작성하였다.

여성가족패널의 2차년도 데이터의 경우에는 종단면기록개인가중치, 종단면응답개인가중치, 횡단면기록개인가중치, 횡단면응답개인가중치, 횡단면가구가중치를 구할 수 있다.

그러나 본 논문에서는 종단면기록개인가중치와 종단면응답개인가중치 산정 방법을 자세히 설명하고 있을 뿐 횡단면기록개인가중치, 횡단면응답개인가중치 등은 다루고 있지 않다. 추후에 횡단면가중치 산정에 대한 자세한 방법을 다룰 필요가 있을 것이다.

참고문헌

강석훈 · 김영원(2007), 「2006년 대졸자직업이동 경로조사 가중치 산출방안에 대한 연구」, 한국고용정보원

_____(2008), 「대졸자 직업이동경로조사 가중치 산출을 위한 연구」, 한국고용정보원

강석훈(1997), 「유럽의 패널조사 현황과 시사점」, 워킹페이퍼 시리즈, 한국노동연구원

_____(1999), 「KLIPS 1차 웨이브에서 가중치 부여방법에 관한 연구」, 『제1회 노동패널학술대회 자료집』한국노동연구원

_____(2000), 「KLIPS 2차 웨이브에서 가중치 부여방법에 관한 연구」, 『제2회 노동패널학술대회 자료집』한국노동연구원

_____(2003), 「KLIPS의 가중치 부여방안 연구」, 『한국노동패널연구』 Working Paper Series 2003-04, 한국노동연구원

_____(2006), 「여성가족패널 표본설계방안」, 한국여성정책연구원

_____(2009), 「여성가족패널 2차 웨이브 이후의 가중치 부여방안 연구」, 한국여성정책연구원

박수미 외(2008), 『2008년 여성가족패널조사 사업보고서 및 제1차 기초분석 보고서』, 한국여성정책연구원

손창균(2008), 「한국복지패널의 가중치 조정과 향후과제」, 보건복지포럼 통권 제145호

이계오(2007),「고령화연구패널조사 가중치」,한국노동연구원 (http://www.klosa.re.kr) (2009.07.03)

한국보건사회연구원(2006),『한국복지패널 1차년도 조사자료 User's Guide』

_____(2007),『한국복지패널 2차년도 조사자료 User's Guide』

한국직업능력개발원(2007), 『한국교육고용패널 1차(2004)~3차(2006)년도 조사 사용자지침서』(http://keep.krivet.re.kr) (2009.07.03)

Markus Pannenberg, Rainer Pischner, Ulrich Rendtel, Martin Spiess and Gert G. Wagner(2005), "Sampling and Weighting", *Desktop Companion to the German Socio-Economic Panel(SOEP)*

U.S. Department of Labor Bureau of Labor Statistics(2001), "NLS of Young Women User's Guide". *National Longitudinal Surveys*

U.S. Census Bureau(2001), *Survey of Income and Program Participation Users' Guide*, third edition Washington, DC: U.S. Census Bureau

Taylor, M. eds(2003), "British Household Panel Survey User Manual Volume A : Introduction, Technical Report and Appendices" , University of Essex

ABSTRACT

The study for calculating the weighting of the panel data

Park, Jee Hye
Department of Economics
Graduate School of
Sungshin Women's University

The weighting of the cross-section analysis is obtained by the process of three stages: calculation of selection probability, adjustment of non-response and post-stratification. This weighting can be used for the first wave of the panel survey, however, the weighting has to be newly calculated and provided for the second wave of the longitudinal section analysis in order to avoid some problems such as the panel attrition among the waves and the entry of non-sample household member.

This paper deals with the concept of weighting and how to figure it out . Furthermore, it examines the methods that are used for the panel survey to understand what kind of methods have been used for the previous panel data in and around the country.

This paper also examines empirically and thoroughly how to compute longitudinal weighting with the second year data of Korean Longitudinal Survey of Women & Family which is provided by Korean Women's Development institute.

부록

<부표 1> 종단면기록개인응답률 : 프로빗분석 결과

구분	계수추정치	표준오차	z-값	p-값
age	-0.117	0.004	-26.97	0.000
age2	0.001	0.000	21.59	0.000
reg_2	-0.102	0.073	-1.39	0.163
reg_3	-0.187	0.081	-2.30	0.021
reg_4	0.065	0.088	0.74	0.461
reg_5	-0.314	0.078	-4.05	0.000
reg_6	-0.393	0.074	-5.27	0.000
reg_7	-0.186	0.088	-2.11	0.035
reg_8	-0.018	0.071	-0.26	0.798
reg_9	-0.508	0.075	-6.82	0.000
reg_10	-0.552	0.073	-7.55	0.000
reg_11	-0.297	0.080	-3.70	0.000
reg_12	-0.482	0.072	-6.67	0.000
reg_13	-0.715	0.066	-10.78	0.000
reg_14	-0.359	0.072	-4.96	0.000
reg_15	-0.259	0.074	-3.50	0.000
reg_16	-0.021	0.115	-0.18	0.854
job_1	-0.030	0.036	-0.85	0.395
wed	1.776	0.050	35.17	0.000
_cons	3.236	0.078	41.40	0.000

Number of obs=31505
LR chi(19)=2766.10
Prob>chi2=0.0000
Pseudo R2=0.2444

<부표 2> 종단면응답개인응답률 : 프로빗분석 결과

구분	계수추정치	표준오차	z-값	p-값
age	-0.016	0.012	-1.37	0.170
age2	0.000	0.000	1.75	0.080
job_1	0.227	0.033	6.82	0.000
reg_2	0.300	0.067	4.48	0.000
reg_3	0.509	0.080	6.39	0.000
reg_4	-0.027	0.068	-0.39	0.694
reg_5	0.365	0.084	4.33	0.000
reg_6	0.544	0.084	6.50	0.000
reg_7	0.667	0.097	6.87	0.000
reg_8	-0.089	0.056	-1.59	0.113
reg_9	0.467	0.087	5.39	0.000
reg_10	0.447	0.084	5.29	0.000
reg_11	0.575	0.088	6.52	0.000
reg_12	0.527	0.087	6.05	0.000
reg_13	0.633	0.090	7.01	0.000
reg_14	0.355	0.075	4.71	0.000
reg_15	0.203	0.068	2.97	0.003
reg_16	0.814	0.119	6.84	0.000
wed	0.184	0.093	1.97	0.049
ch_birth	0.099	0.083	1.20	0.231
r_form1	0.199	0.050	3.96	0.000
r_form2	0.036	0.056	0.64	0.520
r_form4	0.196	0.092	2.13	0.034
eh	-0.191	0.036	-5.30	0.000
hh	0.045	0.017	2.60	0.009
_cons	0.563	0.233	2.42	0.016

Number of obs=9997
LR chi(25)=588.89
Prob>chi2=0.0000
Pseudo R2=0.0662