한국노동패널 표본의 대표성과 가중치 보정 방법

김영원*, 김재광**, 이기재***, 조유미***

한국노동패널(KLIPS)과 같은 패널조사의 경우 표본설계시에 선정된 표본이 대표성을 갖고 있더라도 장기간에 걸친 연속조사 과정에서 무응답 등에 의해 표본 마모(attrition)가 발생하게 되고, 모집단 자체에 변동이 발생하기 때문에 꾸준히 표본의 대표성을 검토해 볼 필요가 있다. 만약 표본이 현재의 모집단을 설명하는 데 문제가 있다면, 특정 보조정보를 기준으로 표본의 대표성을 보완하기 위해 갈퀴조정(raking) 또는 가중치 보정(calibration) 기법 등을 적용하는 것이 필요하다. 본 연구에서는 KLIPS 자료를 대상으로 표본 마모를 발생시키는 대표적인 원인인 무응답 발생 패턴을 점검해 본다. 또한 성, 연령, 교육수준이란 측면에서 통계청의 주민등록 및 2000년 인구주택총조사 자료와 비교해 봄으로써 현행 KILPS 표본의 대표성을 검토해 보고, 나타난 차이를 가중치 보정기법을 통해 보완하는 방안을 제시한다. 아울러 가중치 보정효과를 보기 위해 KLIIPS 가중치를 사용하는 경우와 제시된 보정 가중치를 사용하는 경우 근로소득 등과 같은 주요 변수에 대한 추정결과에서 발생하는 차이를 비교해 본다.

Ⅰ. 서 론

패널조사는 일정한 시간적 간격을 두고 동일한 조사단위에 대하여 조사를 계속 하는 것으로 횡단면 자료(cross-sectional data)가 줄 수 있는 여러 가지 정보 이외에도 시간에 따라 달라지는 특성의 변화를 분석하는 데 특히 효율적이다. 하지만 실제 조사를 위해 표본을 계속 관리하면서 조사를 실행하고, 적절히 분석하는 데는 현실적으로 많은 어려움이 있다. 특히 패널조사 표본은 표본 설계 당시의 모집단을 대표하기 때문에 시간의 흐름에 따라 변화해 가는 모집단의 변동을 반영하는데 미흡한 측면이 있다. 이런 문제를 해결하고 적절한 추정결과를 산출하기 위기 위해서는 가중치 조정 문제를 중요하게 다룰 필요가 있다.

패널조사에서는 조사가 진행됨에 따라 무응답(또는 표본 이탈)이 발생하여 표본의 대표성이 크게

* 숙명여자대학교 통계학과 교수
** 연세대학교 응용통계학과 교수
*** 한국방송통신대학교 정보통계학과 교수
**** 숙명여자대학교 통계학과 대학원생

왜곡될 수 있으며, 이를 해결하기 위해 적절한 가중치 조정이 필요하게 된다. 한국패널조사(Korean Labor Income in Panel Survey; KLIPS)의 경우, 통계청의 경제활동인구조사와 대우패널을 비교 대상으로 하여 KLIPS 표본의 대표성 및 표본 이탈자의 특성에 관한 분석이 김대일 등(2000)에 의하여 이루어진 이후, 모집단 변동에 따른 KLIPS 표본의 대표성을 검증하기 위한 본격적인 연구가 수행되지 않고 있다. 또한 대표성 검토 결과 특정 인구통계학적인 측면에서 표본의 대표성에 문제가 있다는 것이 파악되면 이런 문제점을 보완할 수 있는 적절한 가중치 보완 기법이 적용되는 것이 필요하다.

이런 관점에서 본 연구에서는 다음과 같은 두 가지 측면을 연구대상으로 한다. 첫째, 대표성 유지에 많은 영향을 주게 되는 무응답과 관련하여, 현행 KLIPS 표본의 무응답 패턴을 정리해 봄으로써 무응답이 표본의 대표성에 미치는 영향을 검토해 본다. 둘째, 현재 제공되는 KLIPS 가중치를 적용하는 경우 표본이 현재의 모집단을 얼마나 잘 설명할 수 있는지 평가해 보고, 성, 연령, 교육수준 등의 인구통계학적 관점에서 발생하는 차이를 조정하기 위한 가중치 보정(calibration) 방법을 검토해 본다.

이를 위해 통계청의 주민등록자료 및 2000년 인구주택총조사 자료를 기준으로 지역(광역시·도), 성, 연령, 교육수준 측면에서 현행 KLIPS 가중치의 적절성을 평가해 본다. 또한 외부에서 얻어진 보조적인 벤치마크 통계와 표본 추정결과가 일치하도록 조정해 주는 가중치 보정기법을 캐나다 통계청(Statistics Canada)에서 개발하여 사용하고 있는 SAS Macro인 CALJACK을 사용하여 구현해 본다(Bernier and Lavallee(1994) 참조). 새로운 보정 가중치(calibration weight)는 일반적인 횡단면 가중치 부여방법으로 계산된 원래의 가중치(original weight)와 가장 차이가 적게 나면서 주어진 벤치마크 통계와 표본 추정치가 일치하도록 산출된 것이다. 산출된 보정 가중치의 효과를 점검해 보기 위해 근로소득 등 주요 변수들을 대상으로 이렇게 얻어진 보정 가중치를 적용하는 경우와 KLIPS 가중치를 적용하는 경우 추정결과에 있어서 어떤 차이가 발생하는지 살펴본다.

본 연구는 다음과 같이 구성되어 있다. 제Ⅱ장에서는 2차년도부터 6차년도까지의 KLIPS 자료에서 발생하는 무응답 패턴을 정리해 본다. 특히 무응답자 특성 파악을 위해 성별, 연령, 학력 등에 따른 범주별 무응답률을 살펴보기로 한다. 제Ⅲ장에서는 통계청의 주민등록 및 인구주택총조사 자료를 기준으로 성, 연령, 교육수준 등의 측면에서 KLIPS 가중치의 적절성을 검토해 보고, 이런 차이를 보완해 줄 수 있는 보정 가중치를 산출한다. 아울러 산출된 가중치를 적용하는 경우 발생하는 추정상의 효과를 분석한다.

Ⅱ. 한국노동패널 표본 현황 및 무응답 패턴

1. 한국노동패널 개요

한국노동패널조사(KLIPS)는 한국노동연구원이 실시하는 노동관련 가구패널조사로써 우리나라를

대표하는 5,000가구를 추출하여 1998년부터 1차 조사가 시작되어 2003년 6차년도 조사가 완료되었으며, 현재는 7차년도 조사가 진행 중에 있다. 주로 표본가구 및 개인들의 경제활동 및 노동시장 이동, 고용 및 실업, 소득 및 임금, 근로조건 등 전반에 걸친 주제들을 다루고 있다. 패널조사는 매년 동일한 가구를 반복, 추적 조사하므로 표본 이탈을 최소화하는 것이 성공적인 조사를 위한 핵심 과제이다. KLIPS 원표본 가구 유지율은 2차 년도부터 6차 년도까지 88%, 81%, 77%, 76%, 77%로 하락하는 추세이나 외국의 대표적인 패널조사들에 비해 상대적으로 높은 유지율을 보이고 있다.

KLIPS 자료는 개략적으로 다음과 같은 과정을 거쳐 생산된다. 예를 들어 6차년도(2003년) 조사의 경우 2002년 1월부터 2004년 6월까지 총 2년 반이 소요되었다. 2002년 1월부터 3월까지 조사 설계, 4월부터 9월까지 현장실사, 2000년 10월부터 2004년 6월까지 데이터를 가공 및 클리닝 하는 과정을 거친다. 데이터 분석은 연중 계속된다. 현장실사에서 사용하는 설문 방식은 면접 타계식인데, 면접원과 응답자가 직접 대면하여 질문하고 이에 대한 응답을 면접원이 기입하는 방식이다. 마지막으로 2004년 7월이나 8월에 자료를 발간하게 된다. 자료는 가구용, 개인용, 신규용으로 구성되나 본 연구에서는 가구용 데이터만을 사용한다.

KLIPS의 추적조사 원칙은 1차년도 원가구 및 원가구원을 대상으로 한다. 원가구란 1차년도 조사 당시 표본이 되었던 5,000가구이며 원가구원이란 1차년도 조사 당시 원가구에 소속되었던 모든 가구원 17,505명이다. 반면 신규가구란 원가구원이 혼인과 취업 등으로 분가하여 새롭게 만든 가구이고 비원가구원이란 출생과 원가구와의 혼인 등으로 1차년도 이후에 새롭게 표본조사 대상으로 진입한 가구원을 일컫는다(한국노동연구원, 2004).

본 연구에서는 KLIPS 자료 중 2003년 6차년도 가구자료를 주로 사용하고 있으며, 연도별 무응답 패턴에 대한 분석의 경우 1차년도부터 6차년도까지의 자료를 기준으로 한다. 여기서 사용하는 KLIPS 가중치는 2003년도 기준 KLIPS 가구자료 가중치를 말한다.

<표 1> 인구주택총조사 및 KLIPS 표본의 광역시·도별 가구수 분포 현황

광역시·도	총조사 가구수	구성비율 (%)	단순 표본가구수	구성비율 (%)	KLIPS 가중치 표본가구수	구성비율 (%)
서울	3,085,936	21.80	1,051	22.91	1,105	24.10
부산	1,120,186	7.91	426	9.29	399	8.69
대구	759,351	5.36	289	6.30	341	7.43
대전	747,297	5.28	149	3.25	141	3.07
인천	408,527	2.89	269	5.86	261	5.69
광주	413,758	2.92	153	3.34	162	3.53
울산	306,714	2.17	121	2.64	94	2.05
경기	2,668,886	18.86	893	19.47	906	19.75
강원	487,420	3.44	103	2.25	135	2.95
충북	461,463	3.26	99	2.16	92	2.00
충남	589,144	4.16	140	3.05	123	2.68
전북	601,965	4.25	197	4.29	193	4.21
전남	664,287	4.69	142	3.10	129	2.81
경북	887,917	6.27	260	5.67	239	5.20
경남	951,393	6.72	295	6.43	268	5.84
총계	14,154,244	100.00	4,587	100.00	4,587	100.00

우선 KIIPS 현재 표본가구 지역별 현황을 살펴보기 위해, 6차년도 표본 가구의 광역시・도별 분포와 2000년 인구주택총조사 결과를 비교해 보면 <표 1>과 같다. <표 1>은 가중치 없이 단순 계산한 표본 가구수와 KLIPS 가중치를 적용한 표본 가구수를 보여주고 있다. 전체 표본 중 서울특별시와 경기도가 22.91%와 19.47%를 차지하고 있다. 대체적으로 서울특별시와 광역시의 경우 총조사에 비해 표본 가구비율이 높은 것으로 나타났고, 도지역에서는 반대 현상이 나타나고 있다. 한편 총조사와 가장 차이가 많이 나는 지역은 대전광역시와 인천광역시로 총조사와 2%P 이상 차이가 나고 있다.

2. 가구 및 가구원 무응답 패턴

패널조사를 위해 선정한 최초의 표본이 대표성을 가진다고 하더라도 이후의 계속된 조사에서 표본의 무응답이 클 경우 확보된 대표성을 상실할 수 있다. 무응답에 따른 조정 필요성을 검토하고, 구체적인 무응답 처리 방안을 모색하기 위해서는 조사가 진행됨에 따라 발생하는 무응답 패턴을 파악하는 것이 필요하다. 이를 위해 무응답 가구 및 무응답 가구내의 가구원을 기준으로 무응답 패턴을 정리해 본다. 가구원의 무응답패턴은 KLIPS의 개인(가구원)자료를 기준으로도 수행될 수 있지만, 여기서는 연구의 일관성을 고려해 가구자료를 기준으로 분석하고 있다는 점에 유의하기 바란다. 한국노동패널은 기본적으로 가구를 추출단위로 하는 조사이므로 가구차원에서 발생하는 무응답이 중요할 수 있다. 가구차원에서 무응답이 발생하면 해당 가구 모든 가구원에 대해 무응답이 발생하게 된다. 한편 가구차원에서 무응답이 발생하지 않는 경우에도 해당 가구의 개별 가구원에 대해서는 무응답이 발생할 수 있기 때문에 동일한 가구자료를 기초로 하더라도 가구대상 무응답과 가구원대상 무응답 패턴에 대한 분석은 서로 보완적인 정보를 제공할 수 있다.

<표 2>에는 전체 6개 년도에 대한 가구의 웨이브 무응답(wave nonresponse) 패턴이 정리되어 있다. <표 2>에서 “1”과 “0”은 각 조사 웨이브(wave)에서 표본 가구의 “응답”과 “무응답”을 나타낸다. KLIPS의 경우 연도별로 신규 표본이 추가로 발생하기 때문에 무응답 패턴을 각 연도별로 구분하여 정리한 것이다. 1차년도의 경우 처음 조사에 응한 5,000가구를 기준으로 한 것이고(즉 1차년도에는 무응답 가구가 없음), 2차년도부터는 신규 표본 가구까지 포함한 것이다. 1차년도부터 6차년도까지 모든 연도의 조사에 응답한 가구는 3,087가구로 61.74%를 차지한다. 2차부터 6차까지 모두 응답한 가구는 60.65%, 3차부터 6차까지 모두 응답한 가구는 56.85%, 4차부터 6차까지 모두 응답한 가구는 62.61%, 5차부터 6차까지 모두 응답한 가구는 68.14%인 것으로 나타났다. 따라서 이들 완전 응답패턴을 보이는 가구의 구성비율이 다른 응답패턴의 구성비율에 비해 높지만 이들 완전 응답률이 만족할 수준으로 높다고 볼 수는 없을 것 같다. 한편, 완전 응답패턴 다음으로 비중이 높은 것은 각 경우마다 1번이나 2번의 무응답이 있는 패턴들이다. 예를 들어, 1차년도부터 6차년도까지 1번의 무응답이 있는 패턴의 비율을 합하면 11.62%, 2번의 무응답이 있는 패턴들의 비율을 합하면 7.84%이다. 특히 한번 무응답이 발생하면 그 이후 조사에 전혀 응답을 하지 않는 100000, 110000, 111000, 111100, 111110과 같은 무응답 패턴이 상대적으로 높은 비율을 보이고 있다는 점에 유의할 필요가 있다.

<표 2> 가구 웨이브 무응답 패턴 (0:무응답가구, 1:응답가구)

1차년도 부터			2차년도 부터			3차년도 부터			4차년도 부터			5차년도 부터		
응답 패턴	도수	(%)	응답 패턴	도수	(%)	응답 패턴	도수	(%)	응답 패턴	도수	(%)	응답 패턴	도수	(%)
100000	273	5.46	00000	273	5.24	0000	485	8.31	000	719	12.32	00	926	15.86
100001	67	1.34	00001	83	1.59	0001	395	6.77	001	488	8.36	01	614	10.52
100010	9	0.18	00010	12	0.23	0010	31	0.53	010	60	1.03	10	320	5.48
100011	38	0.76	00011	43	0.83	0011	210	3.60	011	323	5.53	11	3,978	68.14
100100	15	0.30	00100	20	0.38	0100	50	0.86	100	207	3.55			
100101	5	0.10	00101	7	0.13	0101	23	0.39	101	126	2.16			
100110	12	0.24	00110	15	0.29	0110	42	0.72	110	260	4.45			
100111	54	1.08	00111	69	1.32	0111	336	5.76	111	3,655	62.61			
101000	30	0.60	01000	34	0.65	1000	234	4.01						
101001	5	0.10	01001	6	0.12	1001	93	1.59						
101010	2	0.04	01010	2	0.04	1010	29	0.50						
101011	3	0.06	01011	4	0.08	1011	113	1.94						
101100	9	0.18	01100	11	0.21	1100	157	2.69						
101101	7	0.14	01101	9	0.17	1101	103	1.76						
101110	15	0.30	01110	15	0.29	1110	218	3.73						
101111	78	1.56	01111	99	1.90	1111	3,319	56.85						
110000	200	4.00	10000	212	4.07									
110001	73	1.46	10001	74	1.42									
110010	4	0.08	10010	4	0.08									
110011	40	0.80	10011	43	0.83									
110100	16	0.32	10100	16	0.31									
110101	8	0.16	10101	12	0.23									
110110	15	0.30	10110	17	0.33									
110111	127	2.54	10111	137	2.63									
111000	182	3.64	11000	187	3.59									
111001	82	1.64	11001	84	1.61									
111010	26	0.52	11010	26	0.50									
111011	100	2.00	11011	105	2.02									
111100	142	2.84	11100	145	2.78									
111101	88	1.76	11101	90	1.73									
111110	188	3.76	11110	196	3.76									
111111	3,087	61.74	11111	3,160	60.65									
총계	5,000	100.0	총계	5,210	100.0	총계	5,838	100.0	총계	5,838	100.0	총계	5,838	100.0

KLIPS는 매년 동일한 가구를 반복, 추적 조사하여 무응답을 최소화한다. 따라서 한 가구가 어떤 연도에 무응답 하더라도 표본에서 완전히 이탈되는 것이 아니라, 다음 연도에 다시 응답할 수도 있다. 또한 1차년도 조사 당시 원가구에 속해있던 가구원(원가구원)이 혼인, 취업 등으로 분가하여 새롭게 가구를 구성하거나, 비원가구원이 출생이나 원가구원과의 혼인 등으로 새롭게 진입하기도 한다.

가구 무응답률과 가구원 무응답률을 각 연도별 신규진입 가구(원)를 제외하고 연도별로 계산한 결과는 <표 3>과 같다. 여기서 무응답 가구(원)는 전년도에 응답하였으나, 올해에는 응답하지 않은 가

구(원)를 나타낸다. 즉, 6개 연도 동안 발생할 수 있는 모든 웨이브 무응답 패턴을 고려하지 않고, 단순히 연속된 2개 연도를 기준으로 전년도에 응답하였으나 올해는 응답하지 않은 경우만 무응답 가구(원)에 포함한 것이다. 예를 들면, 2000년의 가구 무응답률은 515/4508=0.1142이고, 2000년의 가구원 무응답률은 1793/12537=0.1430이다. KLIPS 가구자료에는 응답 가구원수라는 변수가 따로 존재하지 않는다. 따라서 표본에서 응답 가구원수, 무응답 가구원수는 KLIPS에서 제공하는 가구자료에서 가구원수를 파악하여 산출된 결과이기 때문에 가구원 자료를 직접 활용해 산출된 결과와 비교했을 때 약간 차이가 있을 수 있다는 점에 유의하기 바란다.

<표 3> 표본 가구(원)의 무응답률

	1998년 (1차)	1999년 (2차)	2000년 (3차)	2001년 (4차)	2002년 (5차)	2003년 (6차)
응답 가구수	5.000	4.508	4.266	4.248	4.298	4.592
무응답 가구수		622	515	469	333	320
가구 무응답률(%)		12.44	11.42	10.99	7.84	7.45
응답 가구원수	13,107	12,537	12,186	12,678	13,264	14,961
무응답 가구원수		2,104	1,793	1,527	1,123	1,010
가구원 무응답률(%)		16.05	14.30	12.53	8.86	7.61

<표 3>을 보면, 조사가 거듭될수록 가구와 가구원 모두에 있어서 무응답 비율이 감소한다는 것을 확인할 수 있다. 가구의 경우, 2차 조사에서 12.44%였던 무응답률이 3차 조사에서 11.42%, 4차 조사에서 10.99%, 5차 조사에서 7.84%, 6차 조사에서는 7.45%로 낮아졌다. 특히 4차 조사에 비해 5차 조사에서 무응답률 하락폭이 3.15%P(=10.99%-7.84%)로서 3차 조사의 1.02%P, 4차 조사의 0.43%P, 6차 조사의 0.39%P보다 상당히 크게 나타나고 있다. 가구원의 경우에도 비슷한 양상을 보여 주고 있다.

이와 같이 조사가 여러 차례 진행되어 표본 패널이 안정화되기 전까지 표본의 무응답률이 매우 높다는 사실은, 패널 조사를 위해 1차 년도에 대표성 있는 표본을 추출하였다고 해도 해가 지남에 따라 표본의 대표성이 크게 훼손될 가능성이 있음을 시사한다. 따라서 표본의 무응답을 줄여 대표성을 유지함으로써 패널조사가 성공적으로 정착되도록 하기 위해서는 패널조사 수행 초기에 해당하는 2차 또는 3차 조사에서 무응답률을 줄이기 위한 노력을 집중해야 함을 알 수 있다.

김대일 등(2000년)은 경제활동 인구조사」,「임금구조기본 통계조사」및「영세규모 사업체 근로실태조사」와 비교분석하여 노동패널조사 초기 표본의 대표성을 검토한 연구결과를 제시하고 있다. 그들의 연구에 의하면, 노동패널 표본은 경제활동 분포 측면에서 차이를 보이고 있으며, 이런 양상이 성・연령・학력별로 차이를 보이고 있어, 대표성 보완을 위해 가중치를 적극 활용해야 한다는 점을 지적하고 있다. 노동패널과 같은 연속조사에서는 모집단 변동 및 누적적으로 발생하는 표본 무응답 때문에 초기 표본이 갖고 있던 대표성 문제가 점차 더 악화될 소지가 많다. 따라서 현재 KLIPS

표본의 대표성을 반드시 검토해 볼 필요가 있다.

3. 가구(원) 특성에 따른 무응답 발생 현황

무응답이 발생하는 경우, 특히 무응답 가구(원)들과 응답 가구(원)들이 인구통계학적 특성 등에 따라 크게 차이가 없다면 무응답에 의한 대표성 훼손 문제는 심각하지 않을 수도 있다. 이런 관점에서 표본에서 어떤 특성을 갖는 가구(원)들에서 무응답이 많이 발생하게 되는지 살펴보기로 한다.

연도별로 무응답 가구의 모든 가구원들에 대한 특성별 무응답률을 정리해 보면 <표 4>와 같다. 여기서 무응답률은 해당 연도 무응답 가구원수를 총 조사대상 가구원수로 나눈 것으로, 2003년에 무응답한 남성의 퍼센트는 (2003년 무응답한 남성의 수/2003년 남성 표본 수)*100으로, 2003년 성별 총계의 퍼센트는 (2003 무응답 남성 수+2003년 무응답 여성 수)/(2003년 남성표본 수+2003년 여성표본 수)*100 등으로 계산한 것이다.

<표 4>를 보면 조사가 진행될수록 남녀 모두의 무응답률이 하락하였다. 남녀별 무응답률 차이는 1차년도부터 0.3%p, 0.4%p, 0.2%p, 0.5%p, 0.3%p으로 하락과 상승을 반복하였고, 성별에 따라 큰 차이를 보이고 있지는 않다. 무응답률을 연령별로 보면 대체로 연령이 높을수록 무응답률이 낮은 특징이 나타난다. 특히 1차년도와 2차년도의 경우 55세 미만의 연령 범주와 그 이상의 연령 범주 간에 뚜렷한 차이를 보였는데, 55세 미만의 범주에서는 10%가 넘는 높은 무응답률을, 55세 이상의 범주에서는 10% 미만의 낮은 무응답률을 보이고 있다.

한편 가구주와의 관계를 기준으로 무응답률을 살펴보면, 가구주, 배우자, 자녀의 무응답률이 거의 동일하게 나타나고 있으며, 그 외의 가구원들은 특별한 패턴을 보이지 않음을 알 수 있다. 여기서 가구주, 배우자. 자녀 3개 범주의 비율이 거의 같이 변화하는 것은 가구주와 동일한 가구를 형성하는 가족 단위인 가구주, 배우자, 자녀의 경우 가구 무응답으로 인해 동일 가구 내의 모든 가구원에 대해 무응답이 동시에 발생하기 때문이다.

또한 교육수준별 무응답 상황을 살펴보면, 대졸이상의 계층에서는 2000년까지 무응답률이 증가하고 그 이후 매우 가파른 하락폭을 보이며 감소하고 있다. 이러한 현상은 학력 간 임금격차와도 밀접한 관계를 가지는 것으로 보이는데, 김대일 등(2000년)의 연구에서는 이런 현상을 임금이 높을수록 조사 응답에 대한 기회비용이 커지므로 초기 무응답률이 높다고 해석하고 있다.

<표 4> 가구원 특성에 따른 무응답자수 및 가구원 무응답률

		1999년		2000년		2001년		2002년		2003년	
		도수	(%)	도수	(%)	도수	(%)	도수	(%)	도수	(%)
성별	남	1,068	13.5	916	12.4	775	10.7	546	7.6	498	6.6
	여	1,036	13.2	877	12.0	752	10.5	577	8.1	512	6.9
	총계	2,104	13.3	1,793	12.2	1,527	10.6	1,123	7.8	1,010	6.8
연령별	14세 이하	481	14.7	372	12.8	316	11.1	226	8.2	211	7.6
	15~24세	387	14.0	330	13.0	260	10.7	210	9.0	157	6.9
	25~34세	354	13.1	377	15.1	310	12.8	211	8.5	208	7.7
	35~44세	438	16.3	298	12.1	270	11.3	178	7.6	169	7.0
	45~54세	248	13.4	233	13.0	179	9.6	163	8.8	118	5.9
	55~64세	118	8.6	113	8.6	110	8.6	77	5.9	84	6.0
	65세 이상	78	6.9	70	6.1	82	7.0	58	4.6	63	4.4
	총계	2,104	13.3	1,793	12.2	1,527	10.6	1,123	7.8	1,010	6.8
가구주와의 관계	가구주	622	13.8	514	12.1	468	11.0	333	7.8	316	6.9
	배우자	473	13.2	411	12.3	340	10.2	250	7.4	223	6.4
	자녀와 배우자	900	13.4	770	12.5	624	10.4	486	8.3	417	6.9
	부모	39	8.0	36	7.9	39	9.3	35	8.6	24	5.7
	조부모	2	7.1	2	9.5	0	0.0	0	0.0	2	25.0
	형제자매와 배우자	23	15.6	21	17.1	30	33.0	3	3.1	7	7.4
	손자손녀와 배우자	36	12.6	37	14.5	19	7.6	16	6.7	18	7.5
	기타 친인척	6	30.0	2	7.1	5	20.0	0	0.0	2	6.3
	비인척관계 동거인	3	50.0	0	0.0	2	66.7	0	0.0	1	20.0
	총계	2,104	13.3	1,793	12.2	1,527	10.6	1,123	7.9	1,010	6.8
학력	미취학	209	14.1	189	15.4	139	11.1	114	9.7	107	8.7
	무학	57	7.0	46	6.1	53	7.6	36	5.1	38	5.6
	초등학교	330	11.8	233	8.6	231	8.9	163	6.4	157	6.1
	중학교	265	12.2	202	9.9	183	9.0	152	7.7	135	6.8
	고등학교	705	14.0	606	12.9	520	11.8	339	7.8	315	7.1
	전문대학	130	13.6	124	13.1	124	12.4	85	7.7	82	6.7
	대학교	369	16.4	354	17.1	241	11.1	204	9.3	150	6.0
	대학원 석사	28	13.5	33	17.5	27	13.9	24	11.4	22	9.0
	대학원 박사	10	40.0	6	26.1	6	23.1	4	14.3	2	4.8
	총계	2,103	13.3	1,793	12.2	1,524	10.6	1,121	7.8	1,008	6.7

한편 무응답 가구의 성향을 파악하기 위해 가구를 대표하는 가구주를 중심으로 가구주 특성에 따른 가구 무응답률을 정리해 보면, <표 5>와 같다. 2000년을 제외하고 여성이 가구주인 가구의 무응답률이 더 높았으며, 가구주의 연령이 15~24세에 속하는 가구의 무응답률이 상대적으로 높았고, 25~34세, 35~44세의 가구주 순으로 높은 가구 무응답률을 보이고 있다. 학력별로는 대졸이상의 학력을 가진 가구주 가구의 무응답률이 4차년도까지 가장 높았다. 가구주의 근로소득 유무 관점에서 보면 근로소득이 없는 경우 무응답률이 더 높았다. 따라서 이러한 범주에 속하는 가구에 대한 무응답률을 낮추기 위한 노력과 학력 범주별 무응답률 차이 때문에 발생할 수 있는 표본의 대표성 저하 문제를 관리하는 것이 필요하다는 것을 알 수 있다.

<표 5> 무응답가구의 가구주 특성에 따른 가구 무응답률

		1999년		2000년		2001년		2002년		2003년	
		도수	(%)	도수	(%)	도수	(%)	도수	(%)	도수	(%)
성별	남	532	13.8	441	12.1	388	10.8	276	7.7	256	6.8
	여	93	14.2	74	11.8	80	12.5	57	8.5	60	7.4
	총계	625	13.9	515	12.1	468	11.0	333	7.8	316	6.9
연령	14세 이하	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	15-24세	15	34.9	11	28.9	8	19.0	11	34.4	7	17.9
	25-34세	122	15.9	127	18.5	107	15.9	66	9.5	72	9.9
	35-44세	237	17.7	143	11.9	145	12.8	94	8.6	89	7.9
	45-54세	141	13.7	133	13.3	93	8.9	89	8.7	60	5.5
	55-64세	67	8.6	70	9.3	66	8.9	48	6.3	56	6.8
	65세 이상	40	7.3	30	5.2	49	8.0	25	3.7	32	4.2
	총계	622	13.6	514	11.4	468	11.0	333	7.8	316	7.4
학력	미취학	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	무학	23	8.0	11	3.9	28	10.6	11	4.1	15	6.4
	초등학교	66	9.1	50	7.2	50	7.2	46	6.9	48	8.0
	중학교	73	10.6	57	8.7	52	7.5	48	7.2	41	6.5
	고등학교	252	15.0	196	12.3	187	12.2	120	7.7	120	7.8
	전문대학	37	15.0	39	16.8	35	14.2	23	8.4	29	10.4
	대학교	141	19.1	137	20.7	94	13.6	70	9.9	48	6.4
	대학원 석사	23	18.9	20	17.9	18	15.9	11	9.2	14	10.9
	대학원 박사	6	33.3	4	26.7	3	20.0	4		1	4.3
	총계	621	13.8	514	12.1	467	11.0	333	7.8	316	7.5
근로소득	유	527	13.6	443	12.0	395	10.9	288	7.7	265	6.7
	무	94	14.8	71	12.3	73	11.9	45	8.2	51	8.1
	총계	621	13.8	514	12.1	468	11.0	333	7.8	316	6.9

이러한 무응답으로 발생되는 대표성 저하 문제는 무응답 편향(nonresponse bias)의 직접적인 원인이 되는데, 이를 보완하기 위해서는 가중치 조정 또는 표본 추가 등의 작업이 필요하게 된다. 가중치 조정은 보통 두 단계로 적용되는데, 첫 번째 단계는 무응답 성향이나 사회경제적 특성이 비슷한 사람끼리 무응답층을 만들어 그 층내에서는 같은 비보정(ratio adjustment)을 사용하는 방법이고, 두 번째 단계는 모집단의 인구통계학적 정보를 이용하여 표본 추정치가 신뢰할 수 있는 통계와 일치하도록 조정해 주는 가중치 보정(calibration) 단계가 된다. 본 연구에서는 첫 번째 단계에 대해서는 다루지 않고 두 번째 단계인 보정 문제만을 고려하기로 한다(김재광 등(2004) 참조). 이에 대해서 다음 장에서 다루기로 한다.

Ⅲ. 가중치 보정(calibration) 기법의 적용

1. KLIPS 가중치 산출 개요

현재 한국노동연구원에서는 KLIPS 자료의 분석을 위한 가구가중치와 개인(가구원)가중치를 제공하고 있다. 패널조사에서도 표본설계시점에 해당하는 1차 웨이브의 경우에는 일반적인 횡단면조사에서의 추출확률을 기반으로 한 가중치 부여 방법이 그대로 적용될 수 있다. KLIPS는 패널조사이지만 1차 웨이브 자체만으로는 횡단면조사이므로 1차 웨이브의 가중치는 일반적인 횡단면조사에서의 가중치 부여 방법을 사용할 수 있다.

일반적인 통계조사에서 가중치 부여방법은 기본적으로 다음과 같은 3단계 과정을 거쳐서 계산된다. 1단계는 불균등한 추출확률(selection probability)을 반영한 가중치의 산출이다. 대부분의 횡단면조사에서는 불균등 확률추출법이 적용되기 때문에 표본단위에 따라 추출확률이 다르고, 이를 조정하기 위해 추출확률의 역수에 해당하는 설계 가중치(design weight)를 산출한다. 2단계에서는 무응답에 따른 가중치 조정을 한다. 이를 위해 사용될 수 있는 방법으로는 알려진 특성이나 지역 등의 변수를 이용하여 표본을 분할하여 가중치를 조절하거나, 표본의 알려진 성질을 이용하여 회귀분석(또는 로짓분석)을 실시하여 무응답률을 산출하여 이를 가중치에 반영하는 것이다. 3단계에서는 1단계와 2단계를 거친 후 얻어진 가중치를 적용한 표본 추정결과가 외부적으로 알려진 전체 모집단의 특성(벤치마크 보조 정보)과 일치하도록 조정하는 것이다. 이런 과정을 수행하기 위해 흔히 사후층화(post-stratification), 갈퀴조정(raking) 또는 가중치 보정(calibration) 등의 방법들이 사용된다. 최종적인 가중치는 이렇게 3단계의 가중치 조정과정을 거쳐 산출되게 된다.

KLIPS의 가중치 부여 방법은 강석훈(2003)과 김기헌(2003)의 연구를 중심으로 정리하면 다음과 같다. 우선 추출확률과 응답확률을 모두 고려한 가구가중치는 다음과 같이 계산된다. 서울 및 5대 광역시의 경우 0.1*(표본조사구수/도시조사구수)*(97고특조사가구수/ED내 전체 가구수)*(총접촉 가구수/97고특조사가구수)*(최종조사가구수/총접촉 가구수), 도의 동부의 경우 0.1*(해당 도의 표본조사구수/해당 도의 동부 조사구수)*(97고특조사가구수/ED내 전체 가구수)*(총접촉 가구수/97고특조사가구수)*(최종조사가구수/총접촉 가구수) 이고, 도의 읍면부의 경우 0.1*(해당 도의 시에 속한 표본읍면부 조사구수/해당 도의 시에 속한 읍면부 조사구수)*(97고특조사가구수/ED내 전체 가구수)*(총접촉 가구수/97고특조사가구수)*(최종조사가구수/총접촉 가구수) 가 된다. 여기서 97고특은 97년의 고용구조특별조사를, ED는 조사구를 뜻한다. 최종적으로 추출확률과 응답률을 감안하면 특정한 가구의 가중치는 가구추출확률과 가구가 속한 지역의 응답률의 곱의 역수로 나타난다.

계산결과 KLIPS 1차 웨이브에서 1가구는 전국적으로 평균 2,255가구를 대표하며, 최소값은 1,513가구, 최대값은 4,515가구로 나타났다. 가중치의 합이 11,276,900가구로 전체 모집단인 11,100,320 가구와 차이가 있으므로 scale조정을 하였다.

2차 웨이브 이후부터는 Duncan(1995)의 방법론을 이용한다. Duncan의 방법론은 사용하는 가정이 가장 약하며, 계산과정이 비교적 간단하여 쉽게 이해할 수 있다. 그 세부내용은 다음과 같다.

첫째, 최초 웨이브에서 가구차원의 가중치를 구한다. 둘째, 최초 웨이브의 가구가중치를 연령이나 응답여부에 관계없이 모든 가구원의 개인가중치로 사용한다. 셋째, 2차 웨이브 이후부터는 가구원들의 상이한 응답률을 이용하여 개인가중치를 조정한다. 넷째, 2차 웨이브에서 산출된 개인가중치의 가구 내 평균을 이용하여 2차 웨이브의 가구가중치를 산출한다. 다섯째, 3차~6차 웨이브에서도 동일한 방법을 사용한다.

따라서 KLIPS에서 제공하고 있는 가중치는 일반적인 가중치 부여 과정 중 1단계에 해당하는 설계가중치와 2단계에 해당하는 무응답 가중치가 반영된 것이다. 하지만 앞에서 언급한 표본의 추정결과가 외부적으로 알려진 전체 모집단의 특성(벤치마크 보조 정보)과 일치하도록 조정하는 사후층화 또는 가중치 보정(calibration)을 적용하는 3번째 단계의 가중치 조정은 고려하고 있지 않다. 본 연구에서는 우선 KLIPS의 가구 가중치를 적용하는 경우, 주요 특성에 있어서 현재의 우리나라 전체 모집단을 적절하게 설명할 수 있는지 검토하고, 다시 말해 성, 연령, 교육수준 등과 같은 벤치마크 보조 정보를 활용한 가중치 보정의 필요성을 점검하고, 추가적으로 가중치 보정(calibration)을 하는 경우 가구 소득 등 주요 변수의 추정결과에 있어서 어떤 차이가 발생하게 되는지 분석해 보기로 한다.

2. KLIPS 가중치와 보정(calibration) 가중치 비교

현재 한국노동연구원에서는 KLIPS 자료의 분석을 위해 가구 및 가구원 가중치를 제공하고 있다. 이 가중치는 앞에서 살펴본 것과 같이 표본설계 초기의 추출확률과 연도별 무응답률을 반영하여 산출된 가중치이다. 하지만 이런 가중치가 얼마나 적합한 추정결과를 제공해 줄 수 있는지 검토해 볼 필요가 있다. 특히 만약 사후층화에 해당하는 보정(calibration) 기법을 적용하면 추정결과에 어떤 차이가 발생하는지 검토해 보기로 한다.

이를 위해 본 연구에서는 2003년(6차년도) KLIPS 가구자료를 기준으로 KLIPS에서 제공하고 있는 가중치와 보정(calibration)기법을 통해 산출한 가중치를 비교한다. 여기서 보정(calibration) 가중치는 성과 연령 또는 교육수준을 나타내는 변수들을 대상으로 2002년 통계청 주민등록인구 또는 2000년 인구주택총조사를 통해 얻은 통계를 벤치마크로 사용하여 보정(calibration) 기법을 적용해 산출한 가중치이다. 본 연구에서는 캐나다 통계청에서 개발한 CALJACK 프로그램을 활용하여 통계청의 자료를 통해 얻을 수 있는 정보를 주변분포 벤치마크 통계로 설정하는 경우에 얻어지는 보정가중치를 사용하기로 한다. 여기서는 표본에서 추정된 전국 성별-연령대별 구성비가 통계청 주민등록인구에서 나타난 구성비와 일치하도록 한 전국 성-연령 보정(calibration) 가중치, 유사한 방법으로 성-연령 구성비가 15개 특·광역시와 도별로 일치하도록 한 후 이를 종합한 시도 성-연령 보정가중치, 또한 2000년 인구주택총조사 자료에서 얻어진 교육수준별 인구수를 기준으로 한 보정 가중치 등 세 가지 경우를 고려하고 있다. 또한 이런 과정을 통해 얻어진 가중치의 특성 및 소득관련 주요변수에 대한 추정결과를 비교해 본다.

참고로 CALJACK에서는 보정(calibration) 과정에서 사용되는 거리함수를 옵션으로 선택할 수 있게 되어 있다. 사용 가능한 거리함수 중 본 연구에서는 선형방법(linear method)을 적용하고 있기 때문에 여기서 얻어지는 추정결과는 결과적으로 조건으로 주어진 벤치마크 보조정보를 사용한 일반화 회귀추정(generalized regression estimation; GREG)에 해당한다. 최선의 가중치 보정방법을 구현하는 것보다는 가중치 보정을 통해 얻어지는 추정상의 변화를 간단히 살펴보는 것을 목적으로 하기 때문에 여기서는 다른 거리함수를 사용하는 경우 발생하게 되는 반복 수행과정을 피할 수 있는 가장 간단한 형태의 보정방법을 사용한 것이다. 이와 관련된 자세한 내용은 Deville, Sarndal and Sautory(1993)과 Deville and Sarndal(1992)를 참고하기 바란다.

먼저, KLIPS에서 제공하고 있는 KLIPS 가중치(original weight)와 보정(calibration) 가중치를 적용하여 산출한 성별-연령대별 구성비 추정결과는 <표 6>과 같다. <표 6>에서 통계청 주민등록인구의 성별-연령별 구성비를 주변(marginal) 조건으로 주고 CALJACK을 이용하여 산출한 전국 성-연령 기준 보정(calibration) 가중치를 이용하여 구한 성별-연령별 구성비율 추정결과를 보면 의도했던 것과 같이 전국 성-연령 보정 가중치의 경우 주민등록인구의 성-연령 구성비율과 매우 유사해짐을 확인할 수 있다. 여기서 15세 이하의 연령 범주의 경우 경제활동과 거의 무관하다는 점을 고려하여 편의상 보정(calibration)과정에서 제외하였기 때문에 구성비 추정결과에 있어서 약간의 차이를 보이고 있다.

<표 6> 통계청 주민등록인구와 보정 가중치에 의한 성-연령 구성비율 추정결과

인구 및 구성비	표본 현황		통계청 주민등록인구		KLIPS 가중치		전국 성-연령 보정 가중치		시도 성-연령 보정 가중치	
	(명)	(%)	(명)	(%)	(명)	(%)	(명)	(%)	(명)	(%)
남	7,522	50.31	24,200,192	50.18	20,296,039	50.54	24,190,108	50.17	23,926,321	50.18
여	7,430	49.69	24,029,756	49.82	19,862,888	49.46	24,021,376	49.83	23,752,796	49.82
총계	14,952	100.00	48,229,948	100.00	40,158,927	100.00	48,211,484	100.00	47,679,117	100.00
0～4세	797	5.33	2,927,044	6.07	1,712,008	4.26	2,409,369	5.00	2,385,380	5.00
5～9세	961	6.43	3,504,981	7.27	2,822,165	7.03	3,746,508	7.77	3,717,322	7.80
10～14세	1,009	6.75	3,332,883	6.91	2,907,534	7.24	3,607,665	7.48	3,541,325	7.43
15～19세	979	6.55	3,303,134	6.85	2,758,327	6.87	3,303,134	6.85	3,266,614	6.85
20～24세	1,295	8.66	4,048,734	8.39	3,612,520	9.00	4,046,527	8.39	4,003,132	8.40
25～29세	1,324	8.86	4,098,137	8.50	3,320,650	8.27	4,098,137	8.50	4,053,989	8.50
30～34세	1,363	9.12	4,640,520	9.62	3,324,798	8.28	4,632,995	9.61	4,587,832	9.62
35～39세	1,174	7.85	4,207,110	8.72	3,288,773	8.19	4,204,862	8.72	4,158,944	8.72
40～44세	1,239	8.29	4,445,814	9.22	3,537,599	8.81	4,445,814	9.22	4,400,138	9.23
45～49세	1,120	7.49	3,450,959	7.16	3,143,045	7.83	3,450,959	7.16	3,415,730	7.16
50～54세	871	5.83	2,518,841	5.22	2,408,443	6.00	2,518,841	5.22	2,492,424	5.23
55～59세	719	4.81	2,026,553	4.20	1,956,981	4.87	2,026,553	4.20	2,002,409	4.20
60～64세	677	4.53	2,012,612	4.17	1,782,662	4.44	2,012,612	4.17	1,989,304	4.17
65～69세	563	3.77	1,506,190	3.12	1,426,899	3.55	1,501,071	3.11	1,486,908	3.12
70～74세	407	2.72	1,001,689	2.08	1,010,916	2.52	1,001,689	2.08	989,507	2.08
75～79세	234	1.57	631,409	1.31	595,949	1.48	631,409	1.31	624,083	1.31
80～84세	134	0.90	359,925	0.75	343,577	0.86	359,925	0.75	354,774	0.74
85+	86	0.58	213,413	0.44	206,080	0.51	213,413	0.44	209,302	0.44
총 계	14,952	100.00	48,229,948	100.00	40,158,926	100.00	48,211,484	100.00	47,679,117	100.00

한편 <표 7>은 2000년 인구주택총조사에서 얻어진 교육수준별 인구수를 벤치마크 통계로 하여 표본에서 구한 추정결과가 벤치마크 통계와 일치하도록 하는 보정 가중치를 산출하고, 이를 이용하여 범주별 구성비를 추정한 결과를 보여 주고 있다. 여기서도 마찬가지로 보정 과정을 통해 표본 추정결과가 벤치마크 통계와 일치하고 있음을 확인할 수 있다.

<표 7> 총조사 교육수준별 구성비와 보정 가중치에 의한 교육수준별 구성비 추정결과

인구 및 구성비	표본현황		2000년 인구주택총조사		KLIPS 가중치		교육수준 보정 가중치	
	(명)	(%)	(명)	(%)	(명)	(%)	(명)	(%)
초등학교	2,577	19.79	4,023,228	14.77	6,963,818	19.70	4,023,228	14.77
중학교	1,971	15.14	3,693,314	13.56	5,331,922	15.08	3,693,314	13.56
고등학교	4,465	34.29	12,210,058	44.84	11,881,720	33.61	12,210,058	44.84
대학	1,231	9.45	2,613,695	9.60	3,332,646	9.43	2,613,695	9.60
대학교	2,492	19.14	4,187,405	15.38	7,019,128	19.85	4,187,405	15.38
대학원석사	244	1.87	408,805	1.50	700,717	1.98	408,805	1.50
대학원박사	42	0.32	96,045	0.35	124,275	0.35	96,045	0.35
총계	13,022	100.00	27,232,550	100.00	35,354,225	100.00	27,232,550	100.00

KLIPS 가중치와 제시된 세 가지 보정 가중치의 분포를 파악하기 위해 각 가중치의 평균과 표준편차를 구하면 <표 8>과 같고, 상자그림을 그린 결과는 <그림 1>과 같다. 표준편차는 KLIPS 가중치가 가장 작고, 시도 성-연령 보정 가중치의 경우 가장 크다. 즉, 시도 성-연령 보정 가중치가 평균을 중심으로 가장 넓게 분포해 있고 가중치의 변동이 가장 심하다. 특히 시도 성-연령 보정 가중치 중에는 일부 음수값도 존재한다. 이는 GREG을 산출하는 거리함수를 사용했기 때문에 발생한 현상으로, 비록 극히 적은 수이지만 가중치가 음수가 나오는 경우를 배제하기 위해서는 Deville and Sarndal(1992) 등이 제시한 것과 같이 상한과 하한을 설정할 수 있는 다른 형태의 거리함수를 사용하는 것이 필요하다. 또한 시도 성-연령 보정의 경우 보정을 위한 층이 너무 세분화됨으로써 보정(calibration) 과정에서 제한조건이 너무 많이 부여되어 생기는 현상으로 추측된다.

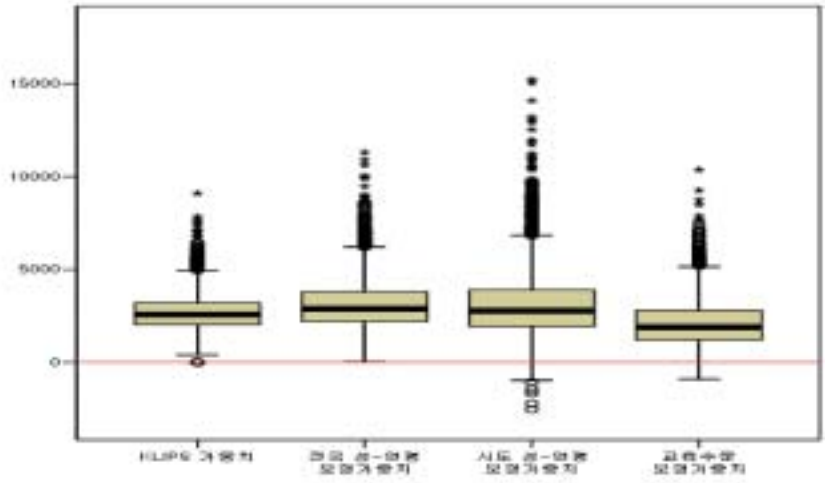
<표 8> 보정 가중치별 평균과 표준편차

	KLIPS 가중치	전국 성-연령 보정 가중치	시도 성-연령 보정 가중치	교육수준 보정 가중치
평균	2655.2	3104.6	3096.2	2062.3
표준편차	1000.6	1300.3	1703.1	1241.7

한편, 균등 가중치 대신 불균등 가중치를 사용하게 되면 편향을 줄일 수 있는 반면 일반적으로 분산이 증가하게 되는데, 가중치의 변동에 따른 추정량의 상대적인 분산 증가분은 일반적으로 $1+CV_W^2$로 설명된다. 여기서 CV_W는 사용된 가중치의 변동계수(Coefficient Variation)를 나타내며, 이런 관점에서 시도 성-연령 보정 가중치와 교육수준 보정 가중치는 비록 추정량의 편향은 줄일 수 있을지 모르지만 분산을 상대적으로 많이 증가시킬 수 있기 때문에 추정의 효율성 측면에서 사용

에 유의해야 할 필요가 있다. 실제 KLIPS 가중치, 전국 성-연령 보정 가중치, 시도 성-연령 보정 가중치, 및 교육수준 보정 가중치의 적용에 따른 분산증가분은 각각 1.1419, 1.1753, 1.3026, 1.3625인 것으로 나타났다.

<그림 1> KLIPS 가중치 및 보정(calibration) 가중치들의 분포 비교



3. 가중치 보정에 따른 주요변수 추정결과 비교

보정(calibration) 가중치를 적용하는 경우와 KLIPS 가중치를 적용하는 경우 주요 변수들에 대한 추정결과에 있어 얼마나 차이가 있는지 살펴봄으로써 보정(calibration)의 적용 필요성을 점검해 볼 수 있을 것이다.

노동패널 가구자료에서 주요 변수라고 볼 수 있는 가구당 연평균 근로소득, 연평균 총소득, 자녀 1인당 월평균 교육 및 보육비, 월평균 저축액, 평균 금융자산 총액, 평균 부채 잔액 등에 대한 전국단의 추정결과를 단순평균, KLIPS 가중치, 전국 성-연령 보정 가중치, 시도 성-연령 보정 가중치, 교육수준 보정 가중치를 적용하여 산출한 결과는 <표 9>와 같다.

<표 9>를 보면 어떤 보정방법에 따른 가중치를 적용하는지에 따라 변동이 심하지는 않지만 추정결과에 있어서 약간의 차이를 보이고 있다. 이 결과를 토대로 여기서 보정방법에 따라 추정결과에 어떤 영향을 미치는지를 일관성 있게 설명할 수는 없을 것 같다. 따라서 어떤 보정 기법이 적절한 것이지 또한 이런 보정 기법의 도입이 필요한 것인지 등에 대해서는 관련 전문가들에 의해 보다 심층적으로 분석될 필요가 있다. 각 보정방법에 따라 산출된 가중치를 적용하여 구한 각 변수들에 대한 시도별 추정결과는 <부표 1>부터 <부표 6>에 수록되어 있다. 참고로 부록에 있는 시도별 추정결과에 있어서 일부 시도의 경우 추정값이 크게 나타나는 경우들이 발생하고 있다. 이는 시도별 표본크기가 충분히 크지 않은 상태에서 일부 지역의 경우 다른 값들에 비해 상대적으로 특이하게 큰 값(outlier)들이 포함되어 있기 때문에 나타나는 현상으로 파악되었다. 예를 들어, 연평균 근로소득에

있어서 울산광역시와 같은 경우가 이에 해당한다.

<표 9> 가중치 보정에 따른 주요변수 추정결과 비교

(단위 : 만원)

	단순평균	KLIPS 가중치	전국 성-연령보정 가중치	시도 성-연령보정 가중치	교육수준 보정가중치
연평균 근로소득	2,633.2421	2,678.8306	2,716.8274	2,683.8964	2,603.0388
연평균 총소득	2,716.5251	2,808.7094	2,866.3416	2,816.1427	2,792.2566
월평균 교육 및 보육비	25.0098	25.8635	25.4388	24.8326	25.9152
월평균 저축액	61.9654	62.0869	62.7837	63.3868	59.9839
평균 금융자산 총액	2,536.2840	2,632.4914	2,527.8393	2,477.2017	2,547.7518
평균 부채 잔액	5,053.7604	5,105.5840	5,021.1335	4,983.1778	4,907.1885

Ⅳ. 결론 및 향후과제

본 연구에서는 KLIPS 표본의 대표성을 검토해 보기 위해 먼저 표본의 대표성을 저해하는 가장 큰 요인으로 지적되고 있는 무응답 패턴을 검토해 보았다. 연차별 무응답의 변동 추이를 살펴보았을 때, 대체적으로 표본설계 후 2~3년 동안에 많은 표본 마모가 발생하고 있으며 일단 어느 정도 표본 패널이 안정화 된 이후에는 무응답률이 크게 우려할 만한 수준은 아니라는 점을 알 수 있었다. 이는 패널조사의 경우 무응답에 따른 편향을 줄이기 위해서는 조사 초기 시점에 가능한 무응답으로 표본에서 이탈하는 가구의 비율을 줄이는 것이 표본의 대표성 유지에 있어서 매우 중요하다는 점을 확인시켜 주고 있다.

한편 표본 가구(원)의 특성에 따른 무응답률을 검토해 본 결과 가구원 특성 측면에서는 성, 연령 범주에 따른 무응답률 차이보다는 학력 범주에 따라 무응답률에 있어서 상대적으로 차이가 큰 것으로 나타났다. 이는 무응답에 따른 표본 마모 현상으로 학력 관점에서 표본의 대표성에 문제가 발생할 소지가 있다는 것을 시사한다고 볼 수 있다. 한편 가구주 특성에 따른 가구 무응답률을 검토해 본 결과 가구주의 연령 범주에 따라 무응답률에 있어서 차이가 발생하는 것으로 파악되었다.

제Ⅲ장에서는 다른 신뢰할 수 있는 외부 정보(통계)를 이용하여 KLIPS 가중치를 보정(calibration)하는 방안을 제시하고, 이에 따라 산출된 보정 가중치의 특성을 살펴보고 동시에 보정 가중치를 적용하는 경우 주요 조사 변수에 대한 추정결과에 있어서 어떤 차이가 발생하는지 검토해 보았다. 검토결과 대체적으로 큰 차이를 보이고 있지는 않지만 어떤 외부 통계를 기준으로 보정 가중치를 산출하는지에 따라 일부 변수의 경우 추정결과에 있어서 차이를 보이고 있다. 이 결과를 보고 각 보정 가중치을 적용함에 따라 추정결과에 어떤 영향을 주는지 일관성을 갖고 설명할 수는 없다. 하지만 이런 보정 기법의 적용 필요성 및 보정 방법의 선택을 위해서는 관련 전문가들의 논의가 필요하다고 생각된다. 또한 어떤 보정 가중치가 효율적인지 검토해 보기 위해서는 보정 가중치를 적

용하는 경우 얻어지는 추정결과에 대한 표본오차를 추정하는 작업이 추가적으로 필요하다.

본 연구에서는 캐나다 통계청에서 이용하는 SAS Macro인 CALJACK을 사용하여 보정(calibration) 가중치를 산출하는 과정에서 편의상 GREG 추정과 동일한 결과를 산출하는 선형방법(linear method)를 사용하고 있다. 하지만 이 방법을 적용하게 되면 가중치가 음수가 발생하는 경우를 배제할 수 없다. 따라서 보정 과정에서 다른 거리함수를 적용하여 그 효율성을 검증해 보는 작업이 향후 필요하다고 판단된다. 아울러 어떤 보조정보들을 벤치마크 통계로 활용하는 것이 효과적인지에 대한 검토도 필요하다. 또한 성, 연령, 교육수준, 지역 등을 모두 동시에 고려한 raking ratio 방법의 활용에 대해서도 고려해 볼 필요가 있다.

이와 관련하여 KLIPS 자료의 분석을 위해 현재 한국노동연구원에서는 무응답에 따른 가중치 조정방법을 적용한 가중치를 제공하고 있는데, 현실적으로 이런 가중치 조정만으로는 모집단 변동을 제대로 반영하는 데 한계가 있다고 생각된다. 따라서 대부분의 외국 통계작성 기관에서 사용하고 있는 사후 가중치 보정(calibration) 기법의 도입 필요성에 대해 관련 전문가들의 논의가 필요하다고 판단된다.

마지막으로 여기서는 KLIPS의 가구자료를 갖고 분석작업을 수행하였는데, 실제 경제활동 관련 통계분석을 위해서는 가구자료 보다는 개인(가구원)자료를 사용하는 것이 필요하기 때문에 개인자료 및 해당 가중치를 중심으로 본 연구에서 다룬 무응답 및 보정 추정문제를 다루어 보는 것이 필요하다고 판단된다. 본 연구에서 개인자료 대신 가구자료를 이용한 이유는 각 변수에 대한 추정결과 자체 보다는 추정치 보정방법의 적용에 초점을 두고 있기 때문에 사이즈 측면에서 훨씬 다루기가 쉬운 자료를 사용한다는 점을 고려한 것임을 밝혀둔다. 아울러 CALJACK에서는 잭나이프 기법을 사용하여 추정결과에 대한 표본오차를 계산할 수 있다. 하지만 이런 작업을 위해서는 표본가구들의 집락 구성 상황을 파악하는 것이 필요한데, 현행 KLIPS 자료에서는 표본 가구들이 어떻게 집락으로 구성되어 있는지 알 수가 없기 때문에 본 연구에서는 분산추정 문제는 논외로 하고 있다.

참고문헌

강석훈(2003),「KLIPS 가중치 부여방안 연구」,『한국노동패널연구 2003-04』2003. 7.

김기헌(2003),「한국노동패널조사의 2003년 가중치」,『노동통계개선시리즈5』, 한국노동연구원

김대일・남재량・류근관(2000),「한국노동패널 표본의 대표성과 패널조사 표본 이탈자의 특성 연구」,『노동경제논집』, 제23권 특별호, 한국노동경제학회, p.1-33

김재광, 한근식, 윤연옥 (2004),「가계조사 무응답 처리기법 연구」,『통계연구』, 9, p79-102.

Deville, J.C. and Sarndal, C.E.,(1992), “Calibration Estimators in Survey Sampling”, Journal of American Statistical Association, 87, p.376-382

Deville, J.C., Sarndal, C.E. and Sautory, O.(1993), “Generalized Raking Procedures in Survey Sampling”, Journal of the American Statistical Association, 88, p.1013-1020

Duncan, G.(1995), “A Simple Method for Weighting in Household Panel Surveys”, Working Paper, Northwestern University.

Bernier, N. and Lavallee, P.,(1994), “The SAS Macro: CALJACK Version 2.04”, Social Survey Methods Division, Statistics Canada, p.1-9

<부표 1> 가구당 연평균 근로소득

(단위 : 만원)

	KLIPS 가중치	전국 성-연령 보정 가중치	시도 성-연령 보정 가중치	교육수준 보정 가중치
서울	3,050.8307	3,087.7429	3,057.9352	2,928.7876
부산	2,492.1624	2,516.5302	2,527.5018	2,363.9958
대구	2,244.5511	2,257.7934	2,235.0441	2,114.9425
대전	2,720.7458	2,693.8897	2,672.7544	2,568.1539
인천	2,505.1034	2,534.0252	2,522.9680	2,430.1920
광주	2,470.2939	2,538.3592	2,591.9776	2,452.2074
울산	3,519.5767	3,530.7803	3,442.6021	3,287.8714
경기	2,956.6790	2,997.3594	2,999.1819	2,818.0818
강원	2,285.4055	2,354.8508	2,291.3445	2,285.6480
충북	2,635.4587	2,670.2100	2,662.3262	2,676.3946
충남	1,999.3361	2,087.6506	2,142.5464	2,252.6532
전북	2,324.5389	2,355.8130	2,425.3394	2,288.0113
전남	2,484.1847	2,538.3832	2,287.4487	2,516.6709
경북	1,766.1804	1,820.9086	1,871.7632	1,883.1932
경남	2,531.1358	2,554.8605	2,575.9412	2,536.0018
전국 (단순평균)	2,678.8306 (2,633.2421)	2,716.8274	2,683.8964	2,603.0388

<부표 2> 가구당 연평균 총소득

(단위 : 만원)

	KLIPS 가중치	전국 성-연령 보정 가중치	시도 성-연령 보정 가중치	교육수준 보정 가중치
서울	3,433.6500	3,499.2891	3,429.9116	3,246.7970
부산	2,449.5018	2,492.2736	2,521.0329	2,390.5009
대구	2,361.0704	2,385.9852	2,332.6813	2,260.2094
대전	2,820.7242	2,813.9587	2,737.3162	2,644.7731
인천	2,411.6303	2,472.1767	2,494.0618	2,340.7691
광주	2,490.4316	2,584.8550	2,674.7209	2,546.7225
울산	3,529.4053	3,565.7709	3,503.0788	3,380.7538
경기	3,071.1444	3,120.5000	3,134.8675	3,018.1591
강원	3,077.1988	3,072.4940	2,826.1308	3,954.4393
충북	2,889.0970	2,948.3970	2,943.0159	2,937.4120
충남	2,151.8963	2,279.6963	2,464.4738	2,545.0816
전북	2,300.8856	2,343.1353	2,393.4431	2,360.3398
전남	2,413.9218	2,485.7264	2,218.9946	2,499.6014
경북	1,743.9979	1,800.6794	1,848.8422	1,913.7318
경남	2,476.6886	2,528.8040	2,499.2804	2,592.4444
전국 (단순평균)	2,808.7094 (2,716.5251)	2,866.3416	2,816.1427	2,792.2566

<부표 3> 가구당 월평균 교육 및 보육비 (자녀 1인당)

(단위 : 만원)

	KLIPS 가중치	전국 성-연령 보정 가중치	시도 성-연령 보정 가중치	교육수준 보정 가중치
서울	34.2189	33.4027	32.2053	34.5864
부산	21.2874	21.0981	21.3452	21.6628
대구	22.2261	22.1653	21.9547	21.8101
대전	21.2495	20.9644	21.0789	20.7498
인천	21.4181	21.1100	21.0621	21.8699
광주	23.4311	23.4301	27.7586	23.5230
울산	23.2235	22.5903	22.2609	22.0434
경기	27.2814	26.7522	26.7489	26.7503
강원	31.0077	31.2483	35.3065	34.4765
충북	20.5096	20.3812	20.3295	20.5616
충남	15.8629	15.8515	16.1357	17.5905
전북	19.1057	19.1233	19.5908	18.2654
전남	18.7883	18.8276	17.9270	19.2848
경북	17.3074	17.3354	17.1589	18.2157
경남	19.1128	19.1757	19.0881	18.3176
전국 (단순평균)	25.8635 (25.0098)	25.4388	24.8326	25.9152

<부표 4> 가구당 월평균 저축액

(단위 : 만원)

	KLIPS 가중치	전국 성-연령 보정 가중치	시도 성-연령 보정 가중치	교육수준 보정 가중치
서울	68.3904	69.1671	69.2605	65.8678
부산	50.6925	50.7303	51.0698	49.2597
대구	51.1690	51.8954	52.9807	46.7437
대전	50.6801	50.0440	47.4748	45.6603
인천	41.7659	41.5896	40.4605	40.8373
광주	74.7421	78.9736	76.6983	70.4580
울산	84.5945	85.7450	84.6552	82.0135
경기	63.4659	63.6223	64.2918	61.1155
강원	141.7764	141.4636	181.1590	143.9379
충북	65.6593	67.4374	66.8754	65.9816
충남	50.1318	50.3642	52.5549	49.0981
전북	58.7411	60.5470	62.6112	55.8182
전남	64.9704	66.8800	57.0344	65.7733
경북	50.9808	50.3488	50.5309	53.5657
경남	57.9529	58.4270	57.1507	57.8685
전국 (단순평균)	62.0869 (61.9654)	62.7837	63.3868	59.9839

<부표 5> 가구당 평균 금융자산 총액

(단위 : 만원)

	KLIPS 가중치	전국 성-연령 보정 가중치	시도 성-연령 보정 가중치	교육수준 보정 가중치
서울	3,639.2797	3,482.5384	3,385.6777	3,484.6360
부산	3,072.7066	2,912.3086	3,000.1153	3,019.5010
대구	2,365.7461	2,220.5960	2,210.7703	2,175.4427
대전	2,222.3158	2,102.0625	1,862.7479	1,883.0629
인천	1,522.3331	1,500.8056	1,460.4327	1,461.0269
광주	2,595.7108	2,510.6421	2,506.8497	2,411.9685
울산	3,688.1551	3,619.0499	3,779.5924	3,488.7282
경기	2,432.5537	2,328.8282	2,295.0570	2,457.2907
강원	2,683.6417	2,474.1519	2,176.6968	2,813.7751
충북	2,381.2234	2,290.1077	2,309.4479	2,287.4564
충남	1,841.9454	1,886.9539	2,012.4610	2,099.2256
전북	2,194.6895	2,135.6452	2,428.2425	2,299.4931
전남	3,220.5316	3,178.6621	3,119.1254	2,870.3525
경북	1,202.9245	1,197.5780	1,183.9628	1,089.5869
경남	1,636.2721	1,579.8065	1,673.3524	1,609.8764
전국 (단순평균)	2,632.4914 (2,536.2840)	2,527.8393	2,477.2017	2,547.7518

<부표 6> 가구당 평균 부채 잔액

(단위 : 만원)

	KLIPS 가중치	전국 성-연령 보정 가중치	시도 성-연령 보정 가중치	교육수준 보정 가중치
서울	6,918.9157	6,727.7279	6,592.4129	6,382.7716
부산	3,275.3869	3,281.9763	3,436.4954	3,037.5979
대구	3,265.9687	3,308.2022	3,298.5412	2,981.0627
대전	3,468.7883	3,484.7931	3,468.6587	3,004.6853
인천	3,978.1079	3,987.9796	3,923.1373	3,951.0109
광주	3,320.3520	3,127.5886	3,414.9356	2,824.2404
울산	4,142.6084	3,964.5152	3,417.0656	3,033.4561
경기	6,111.5170	5,847.4159	5,768.4984	6,340.5836
강원	5,920.6202	6,219.7884	6,609.5066	7,104.7046
충북	4,819.8455	4,798.7920	5,205.5803	4,351.0331
충남	3,274.1968	3,220.1448	3,091.7556	3,971.7438
전북	5,539.3175	5,663.0655	8,190.0753	3,656.8149
전남	4,196.0723	4,218.8321	3,416.1572	3,980.0610
경북	3,138.5733	3,188.5088	3,114.5262	3,338.7288
경남	4,351.6789	4,359.2587	3,922.5794	4,894.8878
전국 (단순평균)	5,105.5840 (5,053.7604)	5,021.1335	4,983.1778	4,907.1885