패널조사에서 가중치 부여 방법 및 효과에 관한 연구

김규성*, 황영은**, 박진우***

가중치는 추출확률, 무응답, 사후 층화 등으로 인하여 표본에 포함된 조사단위가 서로 다른 대표성을 가질 때 이를 보정해 주기 위해서 조사단위별로 부여된다. 본 논문은 패널조사에서 가중치 부여 방식을 다루고 있으며 한국노동패널조사에서 가중치 분포가 어떻게 나타나는지를 분석하고 있다. 또한 가중치가 모평균 추정에 어떤 영향을 미치는지를 이론적으로 고찰하였으며, 한국노동패널의 일부 변수에 대한 실례를 통하여 가중평균이 모평균을 과대 혹은 과소 평가할 수 있음을 보이고 가중치와 관심변수간의 상관관계가 크면 분산이 증가할 수 있음을 보였다.

Ⅰ. 서 론

패널조사에서 가중치를 부여하는 1차적인 이유는 확률추출, 무응답, 사후층화 등으로 인한 불균등 추출확률을 보정해주기 위해서이다. 표본추출단계에서 여러 가지 이유로 인하여 표본추출확률이 달라질 수 있다. 예컨대, 층화임의표집에서 각 층에 최적배정을 하면 층내 표본추출확률은 같지만 층간에는 추출확률이 다르게 된다. 또한 조사과정에서 발생하는 무응답으로 인하여 실제 얻어지는 응답의 수는 배정된 표본의 수 보다 작게 되어 표본에 포함될 확률이 달라지게 된다. 이러한 경우는 가중치를 부여하여 상이한 추출확률로 인한 표본의 왜곡현상을 보정해줄 수 있다. 통상적으로 가중치를 부여하는 방법은 추출확률과 응답확률의 역수를 가중치로 사용하는 것이다. 한국노동연구원에서 실시하는 한국노동패널조사의 가중치는 비록 복잡한 과정을 거쳐 작성된 가중치이긴 하지만, 이와 같은 기준에 준하여 만들어진 가중치이다. 그러면 이렇게 만들어진 가중치가 표본조사 결과에 어떤 영향을 주는가에 하는 질문을 던져볼 필요가 있다. 왜냐하면 가중치 부여는 표본이론의 측면에서 볼 때 추정치의 비편향성(unbiasness)에 충실하기 위한 작업과정이기 때문이다. 일반적으로 비편향성은 표본조사에서 만족시켜야할 중요한 지표중의 하나임에 분명하지만 절대적인 기준은 아니다. 경우에 따라서는 편향 추정치가 비편향 추정치보다 더 신뢰도가 높은 경우도 있기 때문이다.

본 연구에서는 패널조사에서 사용하는 가중치 부여 방법과 가중치 효과에 관한 내용을 다루고자 한다. 이를 위해서 가중치 부여 과정을 고찰 하고 한국노동패널 데이터를 분석하여 수치적으로 그

* 서울시립대학교 통계학과 교수(kskim@uos.ac.kr)
** 서울시립대학교 통계학과 대학원 석사과정
*** 수원대학교 통계정보학과 교수(jwpark@suwon.ac.kr)

예를 보이고자 한다. 패널에 가중치가 부여되면 모집단 총계에 대한 추정량은 가중치를 이용하여 다음과 같이 만드는 것이 보통이다.

$$t = \sum_{i \in s} w_i y_i \quad (1)$$

여기에서 s는 표본, w_i는 가중치, y_i는 특성치를 의미한다. 가중치가 부여된 식 (1)과 같은 추정량은 표본추출 및 무응답 확률분포에 대하여 비편향성을 가진다. 그리고 모총계에 대한 추정치는 가중치뿐만 아니라 조사변수 Y_i에도 영향을 받기 때문에 가중치 및 조사변수의 관계에 의하여 추정량의 효율이 결정된다. 즉, 특정조사에서 가중치의 효과는 그 조사의 조사변수와의 관계에서 규명될 수 있다. 한국노동패널의 가중치는 노동패널의 조사결과에 어떤 영향을 미치는가? 이러한 질문이 본 연구의 두 번째 목적이다.

Ⅱ. 가중치의 기능 및 형태

1. 가중치의 기능

패널조사에서 가중치는 하나의 표본이 대표하는 모집단 단위의 수를 뜻한다. 즉, 가장 간단한 예에서 만일 100가구에서 10가구를 표집하였다면 한 가구에 부여되는 가중치는 추출확률 0.1의 역수인 10이 된다. 표본의 대표성의 관점에서 보면 가중치는 포함확률의 역수가 되는 것이 합리적이다. 그런데 가중치는 모집단을 대표해야 하는 기능과 더불어 모수 추론의 효율을 높여야 하는 기능을 동시에 갖는다. 즉, 가중치는 표본추출방법의 영향을 받지만, 추정량의 형태에도 영향을 받는다는 뜻이다. 예컨대 단순임의추출을 했을 때 표본평균에서의 가중치와 비추정량의 가중치는 서로 다르다. 따라서 가중치는 표본추출방법과 추정과정에서 추정량의 형태를 고려하여 결정하는 것이 바람직하다. 가중치 결정에 선호되는 기준은 추정량의 비편향성이다. 그러나 비추정량이나 사후층화 추정량에서와 같이 포함확률의 역수인 기본가중치는 이용 가능한 보조변수나 사후층화에 의해서 보정이 되기도 하는데, 이는 대체로 추정의 효율을 높이기 위해서이다. 즉, 가중치는 설계비편향성의 유지와 추정 효율의 극대화를 위한 관점에서 절충이 된다고 할 수 있다.

2. 가중치의 형태

가. 표집 가중치

문제를 단순하게 하기 위하여 다음과 같은 상황을 설정하자. 모집단의 크기가 N인 유한 모집단을 고려하고, 조사단위 i는 특성치 (X_i, Y_i)를 갖는다고 하자. 여기서 X_i는 조사 전에 알려진 상수

로 간주하고 Y_i는 미지의 조사 변수이다. 추정하고자 하는 모수는 모총계 $T=\sum_{i=1}^{N} Y_i$이며, 이를 위하여 표집설계 $p(.)$에 의하여 표본 s를 선정한 후, 조사변수 $(y_i,\ i\in s)$를 관측하였다. 조사단위 i의 추출확률은 $p_i,\ \sum_{i=1}^{N} p_i=1$, 라고 하자.

통계조사에서 나타나는 대부분의 추정량은 아래와 같은 동질선형추정량의 일종이다.

$$t=\sum_{i\in s} w_{si} y_i \tag{2}$$

여기에서 계수 w_{si}는 표본 s와 조사단위 i에 의존하는 값으로, 가중치에 해당한다. 가중치의 형태를 살펴보기 위하여 몇 가지 예제를 들어본다.

- 예제 1. 단순임의추출에서 모총계 추정량에 대한 확장추정량(expansion estimator)을 고려하자. 확장추정량은 $t_0=N\bar{y}_s,\ \bar{y}_s=\sum_{i\in s} y_i/n$, 이므로 가중치는 조사단위나 표본에 의존하지 않는 상수의 값이 된다.

$$w_{si}=w=\frac{N}{n} \tag{2}$$

- 예제 2. 단순임의추출에서 비추정량, $t_r=(\sum_{i\in s} y_i/\sum_{i\in s} x_i)\sum_{i=1}^{N} x_i$,을 고려하면 가중치는 다음과 같다.

$$w_{si}=w_s=\frac{N}{n}\left(\frac{\bar{X}}{\bar{x}_s}\right) \tag{3}$$

이 경우 가중치는 표본 s에 부여된다.

- 예제 3. 단순임의추출에서 회귀추정량, $t_{lr}=N(\bar{y}_s+b(\bar{X}-\bar{x}_s))$,을 고려하면 가중치는 다음과 같이 된다.

$$w_{si}=\frac{N}{n}+\frac{(\bar{X}-\bar{x}_s)}{\sum_s (x_i-\bar{x}_s)^2}(x_i-\bar{x}_s) \tag{4}$$

이 경우 가중치는 표본 s와 조사단위 i에 의존한다.

- 예제 4. 불균등확률추출에서 Horvitz-Thompson 추정량, $t_{HT}=\sum_{i\in s} y_i/\pi_i$,을 고려하면 이때 가중치는 포함확률의 역수가 된다. 즉,

$$w_{si}=w_i=\frac{1}{\pi_i},\ i=1,\ldots,N \tag{5}$$

여기에서 $\pi_i=\Pr(i\in s)$는 조사단위 i가 표본에 포함될 확률이다. 이 경우 가중치는 조사단위 i에만 의존한다.

- 예제 5. 불균등확률추출에서 비추정량을 고려하면 가중치는 다음과 같다.

$$w_{si} = \frac{1}{\pi_i} \frac{\sum_{i=1}^{N} x_i}{\sum_{i \in s} x_i / \pi_i} \tag{6}$$

이 경우 가중치는 표본과 조사단위에 의존한다.

위의 예제를 통하여 다음과 같은 사실을 알 수 있다. (i) 가중치 w_{si}는 상수이거나((2)번 식), 조사단위 i에만 의존하거나((5)번 식), 표본 s만 의존하거나((3)번 식), 아니면 조사단위 i와 표본 s에 동시에 의존한다((4), (6)번 식). (ii) 가중치는 기본적으로 포함확률의 역수이다((2), (5)번 식). 그러나 보조변수가 있는 경우는 보조변수에 의하여 보정이 된다((3), (4), (6)번 식).

여기에서 두 번째 사실에 주목할 필요가 있다. 가중치가 대표하는 조사단위의 수는 모집단에 대한 표본비율의 역수가 아니라 표집분포에서 생성되는 포함확률의 역수이다. 이러한 사실은 본질적으로 가중치가 표집분포에 근거함을 나타낸다. 또한 보조변수에 의해서 가중치가 보정된다. 이는 추정의 효율을 높이기 위해서 가중치를 수정할 수 있다는 뜻이다.

나. 무응답 보정 가중치

무응답이 발생했을 때 무응답 효과는 가중치 보정을 통하여 추정량에 반영될 수 있다. 무응답을 보정하기 위해서는 무응답 보정 그룹을 만드는 일이 선행되어야 한다. 각 그룹에서 무응답 보정의 기본 형태는 다음과 같다.

$$w_{1i} = \frac{\text{그룹내 가중치 합}}{\text{그룹내 응답자 가중치 합}} \tag{7}$$

이때 주의할 점은 위와 같은 보정은 무응답이 관심 특성치와 무관하게 확률적으로 발생했을 때 잘 적용된다는 점이다.

다. 사후층 보정 가중치

대부분 패널조사에서는 관심있는 여러 보조변수를 표집에 활용하기 힘들기 때문에 사후적으로 추정과정에서 반영하는 방법을 사용하고 있다. 사후 층화를 한 뒤, 모집단 수치로 보정하는 방법등이 그러하다. 보정 가중치를 부여하려면 무응답 가중치와 마찬가지로 가중치를 부여하기 위한 그룹이 필요하다. 그리고 각 그룹 내에서 보정가중치는 다음과 같이 부여한다.

$$w_{2i} = \frac{\text{그룹내 모집단수}}{\text{그룹내 모집단수 추정치}} \tag{8}$$

최종적으로 얻어지는 가중치는 표집가중치, 무응답 보정 가중치 그리고 사후층 보정 가중치의 곱

으로 얻을 수 있다.

$$w_i = w_{si} \times w_{1i} \times w_{2i} \quad (9)$$

3.『한국노동패널』에서 가중치 부여 방법

한국노동패널에는 한 시점에서 표집에 관련한 횡단면 가중치와 시점이 반복됨에 따라 나타나는 종단면 가중치의 두 가중치가 있는데 이 논문에서는 횡단면 가중치만을 다루기로 한다.

가. 모집단과 표본선정

한국노동패널은 도시지역만을 조사대상으로 하기 때문에 한국노동패널조사의 모집단은 우리나라의 도시지역이다. 현실적으로 우리나라 도시지역의 가구를 대상으로 하는 추출틀을 확보하기 어렵기 때문에 한국노동패널조사에서는 1995년 인구주택총조사의 10% 표본을 고려하고 이 중 도시지역의 조사구를 모집단으로 간주하였다. 이러한 10% 표본은 인구주택총조사 조사구에서 계통추출되었으므로 10% 표본은 이중추출(double sampling)에서 1차 표본으로 간주할 수 있다.

지역별로 10% 표본을 층화하고 1,000개의 조사구를 다시 계통추출하였다. 그리고 추출된 조사구에서 조사가구를 추출한 후, 선정된 표본가구를 대상으로 표본 섭외를 하였는데 최초가구 승낙률은 75.5%였으며, 목표 표본수를 채우기 위하여 나머지 대체 표본을 재선정한 후 접촉한 대체표본가구의 승낙률은 62.1%였다.

나. 가중치 부여

추출확률과 응답확률을 모두 고려한 가구가중치는 다음과 같이 계산되었다. 가중치는 서울 및 6대 광역시, 도의 동부 그리고 도의 읍면부로 나누어 부여되었으며 형태는 다음과 같다.

$$w = 0.1 \times \frac{\text{표본조사구수}}{\text{도시조사구수}} \times \frac{\text{97고특조사가구수}}{\text{조사구내 전체가구 수}} \times \frac{\text{총접촉가구수}}{\text{97고특조사가구수}} \times \frac{\text{최종조사가구수}}{\text{총접촉가구수}} \quad (10)$$

개인 가중치를 부여하기 위한 사후층화는 별도로 시행하지 않았기 때문에 동일한 가구내의 가구원에는 모두 동일한 가중치를 부여하였다.

다. 6차년도 패널 가중치 분포

한국노동연구원에서 제공한 6차년도 한국노동패널의 가중치를 분석하였다.

1) 가구 가중치

전국의 가구 가중치 평균은 2,654로서 한 가구는 대략 2,654가구를 대표한다. 가구 가중치의 최대값은 9,095이고 아래 사분위수는 2,020으로 나타났다. 서울과 경기는 전국과 유사한 분포를 보였다.

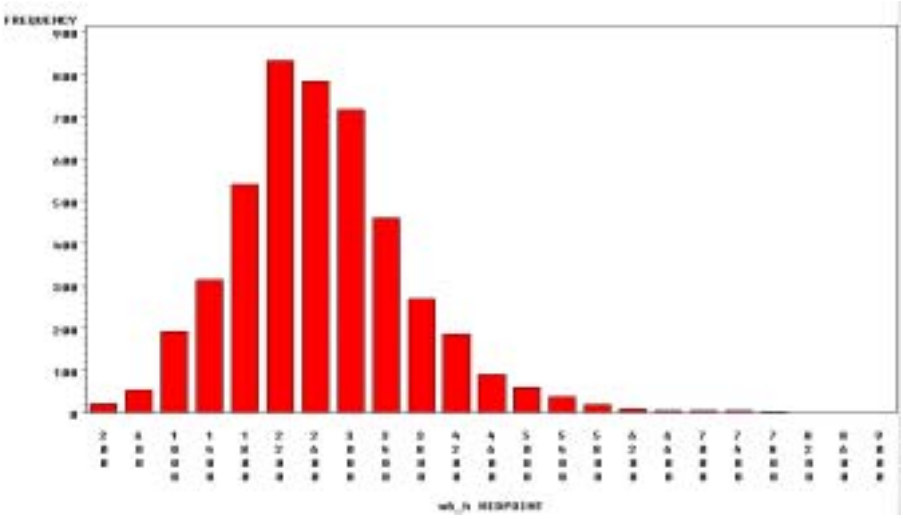
<표 1> 가구 가중치 분포(신규포함)

	전국	서울	경기
평균	2,654.2	2,792.7	2,693.9
표준편차	1,000.5	1,067.2	1,062.8
아래사분위값	2,020.0	2,038.4	2,051.9
중위수	2,571.3	2,751.5	2,706.0
위 사분위수	3,196.7	3,438.3	3,285.2
최대값	9,095.8	7,358.2	7,696.9

<표 2> 개인 가중치 분포(신규포함)

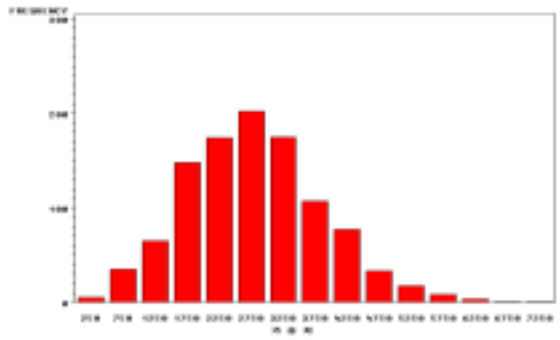
	전국		서울		경기	
	남자	여자	남자	여자	남자	여자
평균	2,824.2	2,796.2	2,987.7	2,959.5	2,838.4	2,835.5
표준편차	1,047.1	1,011.1	1,106.0	1,089.6	1,113.7	1,090.6
아래사분위값	2,161.2	2,156.9	2,181.3	2,180.9	2,157.9	2,186.0
중위수	2,750.7	2,730.3	2,926.6	2.939.8	2,850.3	2,873.6
위 사분위수	3,391.2	3,368.2	3,707.9	3,620.8	3,426.0	3,432.5
최대값	9,529.5	9,529.5	7,709.1	7,709.1	5,867.4	7,374.7

<그림 0> 전국 가구 가중치 분포

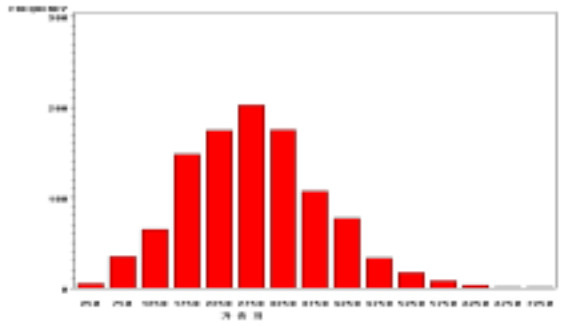


<서울 가구 가중치 분포>

<그림 1> 서울 가구 가중치 분포



<그림 1> 경기 가구 가중치 분포



2) 개인 가중치

전국의 개인가중치 평균은 남자가 2,824, 여자가 2,796으로 나타났다. 가구 가중치 평균과 유사하게 나타난 이유는 가구내 개인 가중치를 모두 동일하게 부여했기 때문이다. 전국과 서울, 경기의 분포도 유사하다.

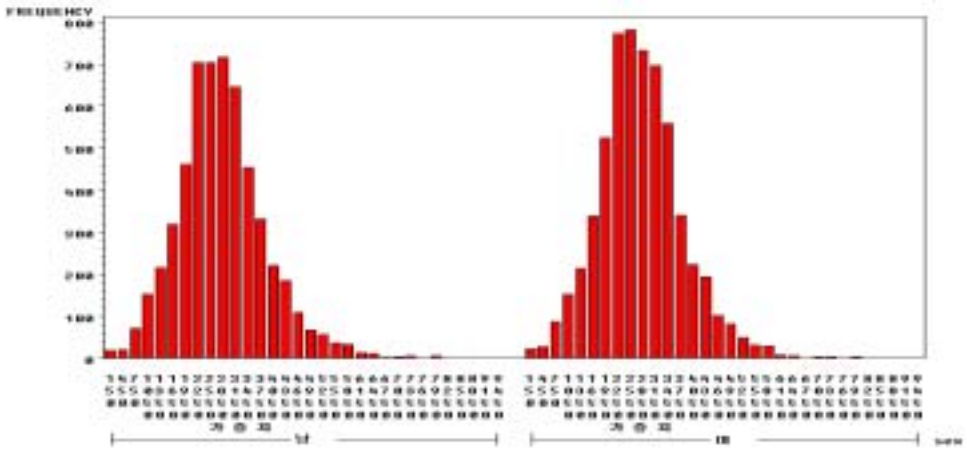
III. 가중치 효과

1. 가중치 효과

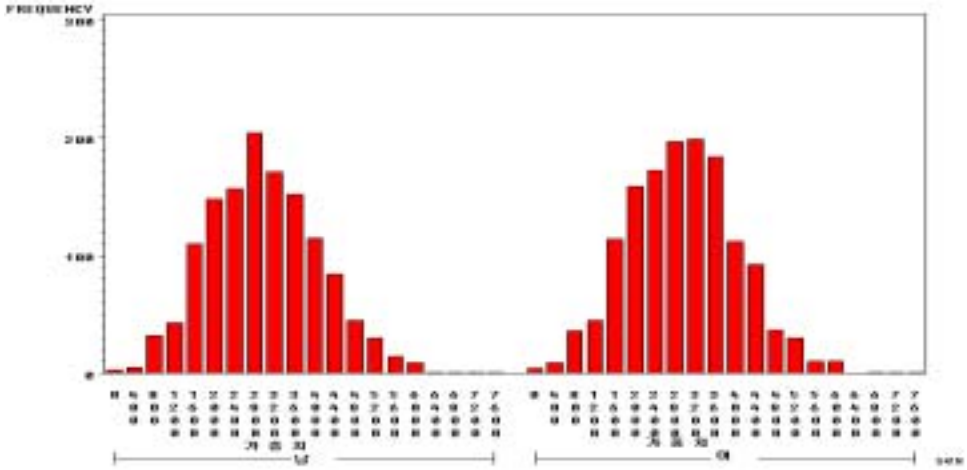
가. 모평균 추정

가중치를 이용한 모평균 추정은 다음의 식을 이용하는 것이 일반적이다.

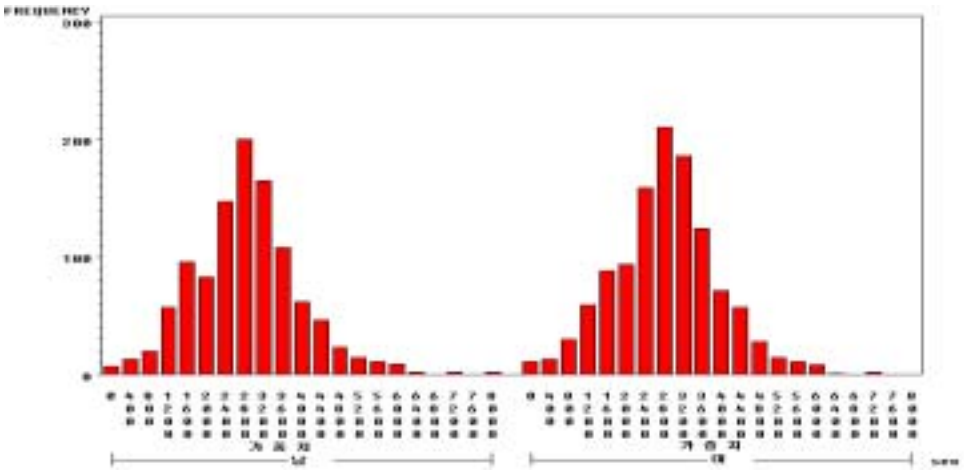
<그림 3> 전국 개인 가중치 분포



<그림 4> 서울 개인 가중치 분포



<그림 5> 경기 개인 가중치 분포



$$\bar{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i} \tag{11}$$

단순임의표집에서 위의 추정량은 가중치와 특성치의 상관관계에 따라 모평균을 과소추정하기도 하고 과대 추정하기도 한다. 만일 상관계수가 0이면 위의 추정량은 모평균의 일치추정량이다. 즉, 표본수가 증가하면 위의 가중평균은 모평균에 근접한다. 그러나 상관계수가 양수이면 위의 추정량은 모평균을 과대 추정하며, 반대로 상관계수가 음수이면 모평균을 과소 추정하게 된다 (김규성, 2004).

나. 분산 추정

가중치가 분산에 어떤 영향을 주는지는 아직 명료하게 밝혀지지 않고 있다. 그러나 특성치의 분포

가 동일한 평균과 동일한 분산을 갖는 분포에서 선정되었다면, 즉,

$$y_1^*, \ldots, y_n^* \sim i.i.d.\ (\bar{Y},\ \sigma_y^2)$$

이면 다음과 같은 결과를 얻는다.

$$Var\{\bar{y}_w|s\} = \frac{\sum_{i\in s} w_i^2}{(\sum_{i\in s} w_i)^2}\,\sigma_y^2 = (1+CV_w^2)\sigma_y^2 \qquad (12)$$

여기서 CV_w는 가중치의 변동계수이다. 이 결과가 시사하는 바는 가중치의 산포가 크면 클수록 가중평균의 변동계수는 증가하고, 결과적으로 가중평균의 신뢰도는 저하된다는 점이다.

2. 노동패널의 가중치 효과

한국노동패널조사는 많은 조사항목을 조사하고 조사항목마다 가중치의 효과가 다를 것이다. 본 연구에서는 예로써 가구 단위에서는 월 근로소득, 월평균 생활비, 월평균 저축액만을 고려했으며, 개인 단위에서는 세전 근로소득과 세후 근로소득만을 고려하였다.

가. 가구 가중치

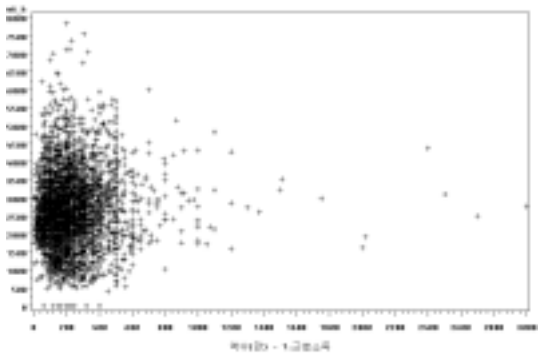
근로소득, 월평균 생활비, 월평균 저축액과 가중치와의 상관계수가 아래의 <표 3>에 주어져 있다. 근로소득과 월평균 저축액은 가중치와의 상관계수가 0에 가까운 반면 월평균 생활비는 가중치와의 상관계수가 0.12에 이르고 있다. 따라서 가중평균, 식(11),을 사용할 경우 월평균 생활비는 모평균을 다소 과대평가할 가능성이 있다.

또한 가구 가중치의 변동계수를 보면 37.6%를 나타내고 있다. 즉, 가구 가중치를 사용할 경우, 분산이 13.7%($0.376^2 = 0.137$) 증가할 수 있음을 보여주고 있다.

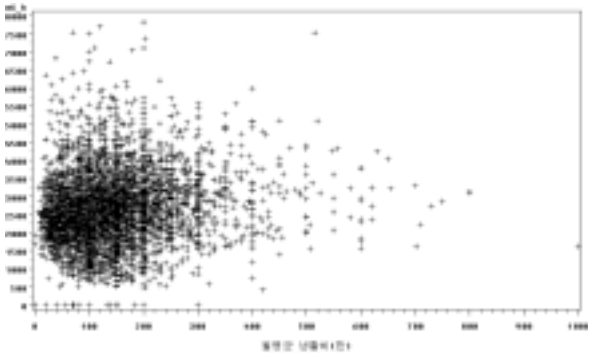
<표 3> 관심변수와 가구가중치의 기초통계량

변수명	라벨	표본수	평균값	변동계수(%)	최대값	상관계수
V06680	근로소득 (만원)	3,752	292.1	1,252.3	222,222.0	0.0048
V06691	월평균 생활비(만원)	4,567	146.2	68.0	1,600.0	0.1182
V06706	월평균 저축액(만원)	2,973	61.9	124.1	1,650.0	0.0039
w6_h	가구 가중치	4,592	2,654.2	37.6	9,095.8	

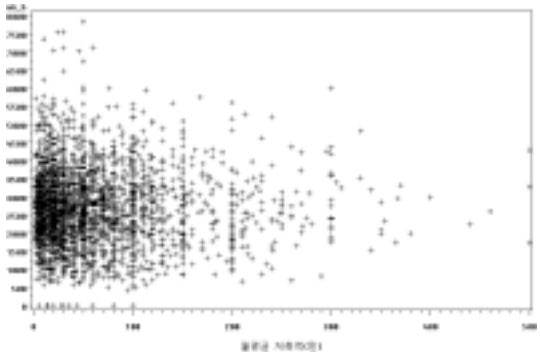
<그림 6> 근로소득과 가중치의 산점도



<그림 6> 월평균 생활비와 가중치의 산점도



<그림 8> 월평균 저축액과 가중치의 산점도



나. 개인 가중치

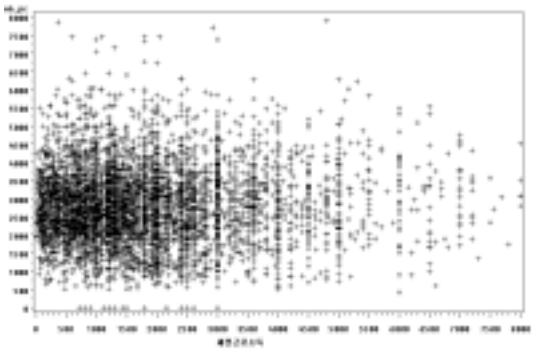
개인의 세전 근로소득, 세후 근로소득과 개인 가중치와의 상관계수가 아래의 <표 4>에 주어져 있다. 세전 근로소득과 가중치와의 상관계수가 0에 가까운 반면 세후 근로소득과 가중치의 상관계수가 0.287에 이르고 있다. 따라서 식 (11)의 가중평균은 세후 근로소득을 과대평가할 가능성이 있다.

또한 개인 가중치의 변동계수는 36.6%이다. 개인 가중치를 사용할 경우 분산이 13.4% ($0.366^2 = 0.134$) 증가할 수 있음을 보여주고 있다.

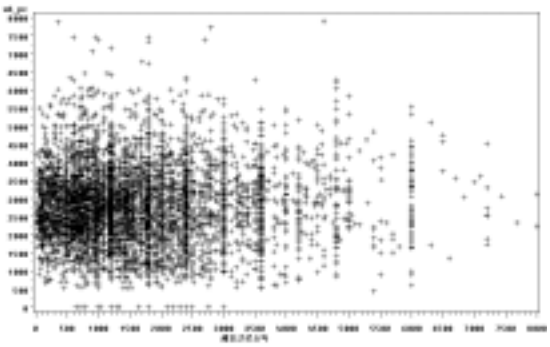
<표 4> 관심변수와 개인가중치의 기초통계량

변수명	라벨	표본수	평균값	변동계수(%)	최대값	상관계수
wbincome	세전 근로소득	5,947	1,981.5	170.3	170,016.0	0
waincome	세후 근로소득	5,983	1,744.4	83.9	24,000.0	0.287
w6_pc	횡단면 가중치	11,543	2,809.7	36.6	9,529.5	

<그림 9> 세전 근로소득과 개인 가중치의 산점도



<그림 9> 세후 근로소득과 개인 가중치의 산점도



4. 결론

가중치는 표본의 대표성을 확보해야 하는 기능과 추정의 효율을 높여야 하는 기능을 동시에 갖는다. 본 논문에서는 표집 가중치의 형태를 다섯 예제를 통하여 살펴보았고, 무응답을 보정하는 가중치와 사후층화 보정 가중치의 기본 형태를 살펴보았다. 또한 한국노동패널의 6차년 데이터 분석을 통하여 가구 가중치 및 개인 가중치의 분포를 살펴보았다. 가중치는 기본적으로 추정량의 비편향성을 유지하기 위하여 부여되기 때문에 경우에 따라서는 가중치의 산포가 커질 수 있다. 그런데 가중치의 산포가 크면 추정량의 분산의 증가로 추정의 효율이 저하될 수 있다.

한국노동패널 데이터 분석에 의하면 가구의 월평균 생활비는 모평균을 과대 평가할 소지가 있는 반면, 근로소득과 월평균 저축액은 모평균을 근사 비편향 추정하는 것으로 보인다. 또한 개인의 세전 근로소득은 모평균을 근사 비편향 추정하지만, 세후 근로소득은 모평균을 과대 추정할 소지가 있는 것으로 나타났다. 마지막으로 가구 가중치와 개인 가중치는 모두 가중치를 부여하기 전에 비하여 분산이 약 13% 증가할 수 있음을 살펴보았다.

참고문헌

강석훈(2003). KLISP의 가중치 부여방안 연구, 한국노동패널연구 2003-4, 한국노동연구원.

한국노동연구원(2004). 한국노동패널 자료 Userguide 1.

한국노동연구원(2004). 한국노동패널 자료 Userguide 2.

김기헌(2003). 한국노동패널조사(KLIPS)의 2003년 가중치. 노동통계개선시리즈 5.

김규성(2004). 표본조사에서 가중치부여 효과에 관한 연구. 한국조사연구학회 추계학술대회 논문집, 3-12.

박홍래(1989). 통계조사론, 영지문화사.