연구논문

한국노동패널조사 자료의 분석을 위한 패널 가중치 산출 및 사용방안 사례 연구*

A Case Study on Construction and Use of Longitudinal Weights for Korea Labor Income Panel Survey

박민규[a] · 김사라[b]
Mingue Park · Sarah Kim

본 연구에서는 1998년부터 실시되어 현재까지 진행되고 있는 한국노동패널조사 자료의 분석을 위해 2013년도에 새롭게 구축된 종·횡단면 가중치의 산출과정 및 선형혼합모형의 적합을 위한 종단면 가중치의 새로운 사용방안이 소개된다. 또한 제안된 종단면 가중치 사용방안을 선형혼합모형과 일반화 선형혼합모형이 사용된 기존 사례 연구에 적용하여 종단면 가중치의 사용 여부에 따른 결과들을 비교하였다. 가중치를 적용한 분석을 위해서 본 연구에서는 추가적인 프로그래밍 작업 없이 기존의 통계 패키지를 직접 사용하였다. 제한적이기는 하나 본 연구에서 고려한 사례에서는 종단면 가중치의 사용 여부가 모형에 대한 통계적 검정 결과에 영향을 주지 않는 것으로 나타나고 있다. 그러나 종단면 가중치의 적용 여부에 따른 결과의 차이와 상관없이 유한모집단의 종단 분석을 위해서는 패널의 대표성과 모형 실패(model failure)에 대한 추론 결과의 강건성을 유지하기 위하여 종단면

* 이 연구는2013년도 기본연구지원사업(2013R1A1A2006363)과 2010년도 한국사회기반연구사업비(NRF-2010- 330-B00280)로 한국연구재단의 지원을 받아 이루어졌습니다.

a) 교신저자(corresponding author): 고려대학교 통계학과 교수 박민규.
E-mail: mpark2@korea.ac.kr

b) 고려대학교 통계학과

가중치를 이용한 분석을 하는 것이 바람직하리라 생각된다.

주제어 : 패널조사, 횡단면 가중치, 종단면 가중치, 선형혼합모형, 일반화 선형혼합모형

In this study, the method of constructing cross-sectional and longitudinal weights and a new way of using longitudinal weight for fitting a linear mixed and general linear mixed model for Korean Labor Income Panel Survey that started at 1998 are discussed. And we apply the suggested method of using a set of longitudinal weights to the existing two case studies to fit general linear mixed models, and compare the analyses results to the one to which longitudinal weights are not applied. In the limited case study, no significant change is found by applying a set of longitudinal weights. However, in all cases, it is recommended to use the longitudinal weights for maintaining the representativeness of the panel and the robustness of the result to model failure in longitudinal studies.

Key words : panel survey, cross-sectional weight, longitudinal weight, linear mixed model, general linear mixed model

Ⅰ. 서론

모집단에 대한 추론을 위해 모집단의 일부인 표본을 추출하여 조사하는 표본조사(sample survey) 중 추출된 표본이 일정한 주기를 갖고 반복적으로 측정되는 조사를 패널조사(panel survey)라고 한다. 즉 표본조사의 한 방법인 패널조사는 원년에 구성된 패널을 일정기간 동안 반복 추적하는 종단조사

(longitudinal survey)로서, 관심 있는 모집단의 동적인 변화에 대한 연구를 위하여 현재 많은 국가에서 이를 활용하고 있다. 대표적인 패널조사로는 우리나라의 한국노동패널조사(Korea Labor Income Panel Survey: KLIPS), 미국의 PSID(Panel Study of Income Dynamics) 그리고 독일의 SOEP(German Socioeconomic Panel) 등이 있다.

1998년부터 시행된 한국노동패널조사의 표본은 1995년 인구주택총조사 10% 표본조사구 중 제주도 및 군부지역을 제외한 도시지역 조사구 표집틀로부터 조사구를 추출하고, 이로부터 가구를 추출하는 층화이단계집락추출법을 통해 구성되었다. 조사구 추출을 위해서는 기본적으로 15개 지역을 층으로 이용한 층화추출법이 사용되었으며 각 조사구 내에서 5~6개의 가구가 무작위로 추출되었다. 이를 통해 총 951 조사구가 추출되었으며 이에 대응되는 표본 가구는 총 5,000 가구이다. 강석훈(2003)은 원년 패널 구성을 위한 표본설계 및 이에 기반을 한 가중치 산출방안을 소개하였다.

원년인 1998년 이후 한국노동패널조사는 가구와 해당 가구에 속한 가구원을 매 해 추적하여 이루어졌다. 원년에 구성된 가구 패널은 시간이 지남에 따라 결혼이나 분가 등으로 인해 신규 분가 가구와 비원표본 가구원을 포함하게 된다. 비원표본 가구원은 분가표본 가구원이라고도 한다. 예를 들면 원표본 가구원이 결혼을 통해 분가하거나 독립하면서 새롭게 가구를 형성할 때 추가되는 원표본 가구원의 배우자와 자녀가 이에 해당한다. 결혼 및 분가로 인한 패널의 변동 이외에 장기간에 걸친 연속조사 과정에서 여러 가지 원인에 의한 패널 마모(panel attrition) 역시 패널 변동의 한 원인이다.

이러한 원패널의 변동으로 인해 발생하는 패널의 모집단 대표성 문제를 해결하기 위하여 매 해 한국노동연구원은 한국노동패널조사 자료의 분석을 위한 두 종류의 가중치를 생성하였다. 첫 번째 가중치는 횡단면 가중치로, 해당 연도에 조사된 모든 가구 및 가구원에 부여되며 해당 연도의 모집단 특성치에 대한 비편향 추정량이 산출되도록 작성된다. 또한 이 과정에서 생성된 횡단면 가중치는 횡단면 자료를 이용한 인과관계 분석을 위해서도 사용될 수 있다.

두 번째 가중치는 종단면 가중치로, 원년도 모집단의 시간의 경과에 따른 동태적 분석을 위해 사용된다. 패널조사의 목적인 유한모집단의 동태적 분석을 위해서는 제공되는 종단면 가중치가 사용되는 것이 바람직하나 이의 사용방안에 대한 연구나 이를 위한 통계 프로그램은 매우 부족하다.

본 연구에서는 한국노동패널조사에서 원패널의 변동을 보정하기 위하여 2013년에 새롭게 구축된 가구와 가구원의 횡·종단면 가중치의 산출과정을 소개하고, 특별히 종단면 분석을 위한 가중치 사용방안을 혼합모형하에서 제시하고 가중치 사용 여부에 따른 결과를 기존 연구 사례들을 바탕으로 비교하였다.

Ⅱ. 패널자료의 분석을 위한 횡·종단면 가중치 산출 과정

패널조사 자료의 분석을 위한 가중치 역시 일반조사의 가중치 산출과 마찬가지로 추출확률 및 응답확률 그리고 캘리브레이션(calibration) 과정을 통해 산출된다. 패널조사를 위해서는 1회성 조사나 독립적인 반복조사에서와는 달리 자료 분석을 위하여 두 개의 서로 다른 개념의 가중치가 산출된다. 생성되는 2가지 가중치는 조사년도의 자료를 이용한 통계 산출 및 분석을 위한 횡단면 가중치, 그리고 시간의 흐름에 따른 유한모집단의 동태적 분석을 위한 종단면 가중치이다. 패널조사 자료를 위한 가중치 산출방안에 대한 대표적 연구로는 Kalton & Brick(1994)과 Duncan(1995)이 있다. 대부분의 관련 연구에서 가구 패널조사의 횡단면 가구 가중치는 이를 이용한 추정량의 기댓값이 근사적으로 대응되는 모수에 불편성을 만족하도록 그 산출과정을 구축하였다. Kalton & Brick(1994)의 연구가 이러한 가중치 산출을 위한 이론적인 부분에 초점을 맞추었다면, Duncan(1995)은 이러한 조건을 만족하며 그 산출이 수월한 가중치 추출과정을 제안하는 것으로 연구의 방향을 잡았다.

종단면 가구 가중치 산출을 위해서는 조사년도까지 계속적으로 조사에 참여한 가구원에 부여되는 가구원 종단면 가중치를 이용하는 방안과, 조사년도까지 조사에 응한 가구원을 적어도 한 명 이상 포함한 가구를 이용하여 가구

종단면 가중치를 직접 계산하는 두 가지 방안을 고려할 수 있다. 가구 패널조사의 특성상 시간에 따른 가구원의 변동이 빈번하게 일어나고 또한 가구 단위 분석보다는 가구원 분석이 주로 이루어지는 한국노동패널조사의 특성을 감안하여 가구원의 종단면 가중치의 가구 평균을 이용한 Duncan(1995)의 방안을 한국노동패널의 종단면 가구 가중치 산출을 위하여 사용하였다. 그러나 분석 단위가 가구원이 아닌 가구이거나 패널 유지 기간이 길지 않아 가구 변동이 심하게 발생하지 않은 경우에는 가구 단위 분석을 통한 가구 종단면 가중치의 산출 방안 역시 고려할 수 있을 것이다. 본 장에서는 Duncan(1995)의 연구를 소개하고 이를 바탕으로 계산된 한국노동패널조사 자료의 분석을 위한 가중치 산출 과정을 소개한다.

1. Duncan(1995)의 가구 패널조사를 위한 가중치 산출 방안

시간이 경과함에 따라 신규 및 분가 가구의 발생 및 패널 마모 등으로 발생하는 패널의 대표성 문제를 해결하기 위하여 Duncan(1995)이 제시한 종·횡단면 가구원과 가구 가중치 산출 절차는 다음과 같다.

① 추출율과 무응답을 고려한 통상적인 종단면 가중치 산출 방식으로 첫 시점 종단면 가구 가중치를 계산하여 각 패널가구에 부여한다.
② 첫 시점의 종단면 가구 가중치로 해당 가구원의 첫 시점 종단면 가구원 가중치를 부여한다. 따라서 첫 시점에 부여된 종단면 가구 및 가구원 가중치는 횡단면 가중치와 동일하다.
③ 전 단계에서 산출한 종단면 가구원 가중치에 무응답 보정 및 캘리브레이션을 적용하여 두 번째 시점 종단면 가구원 가중치를 산출한다.
④ 각 가구에 사는 모든 가구원의 종단면 가중치의 평균을 해당 가구의 두 번째 시점 종단면 가구 가중치로 부여한다.
⑤ 두 번째 시점의 종단면 가구 가중치를 두 번째 시점 해당 가구원의 횡단면 가중치로 부여한다.

각 단계별로 살펴보면 먼저 1단계에서는 횡단면 조사 설계에서 사용되는 전형적인 가중치 산출 방식인 추출확률의 역수로 1차 웨이브 즉, 첫 시점의 횡단면 가구 가중치를 산출하는 과정이다. 패널이 구성되는 첫 시점에서는 표본추출 과정에서 필요한 수준의 표본대체가 이루어져 추가적인 무응답 조정이 일반적으로 이루어지지 않는 것으로 가정한다. 그러나 필요에 따라서는 모집단 정보를 이용한 캘리브레이션이 추가적으로 이루어질 수 있다.

2단계에서는 산출된 첫 시점의 횡단면 가구 가중치로 해당 가구에 속한 가구원의 첫 시점 가구원 가중치를 부여한다. 이는 주어진 표본 설계하에서 1차 웨이브에서 같은 가구 안에 속해 있는 가구원의 추출확률은 가구 자체의 추출확률과 동일하다는 것에 그 이론적 근거를 둔다.

3단계는 전 단계에서 산출한 종단면 가구원 가중치를 가구원의 응답확률을 이용한 무응답 보정을 통해 두 번째 시점의 종단면 가구원 가중치를 산출하는 과정으로 이해할 수 있다. 무응답 보정은 응답 여부를 종속변수로 그리고 가구원의 특성변수들을 설명변수로 갖는 로지스틱 회귀분석을 통해 응답확률을 추정하여, 그 역수를 기존의 가구원 가중치에 곱하여 이루어진다. 결혼이나 동거로 인해 유입된 비원표본 가구원과 새로 태어난 자녀(가구원)의 경우에는 무응답 가중치 보정 과정에 포함되지 않는다.

4단계에서는 무응답 보정 과정을 통해 산출된 가구원 종단면 가중치를 가구단위로 평균을 구하여 이를 해당 가구의 가구 가중치로 부여한다. 세 번째 단계에서와 달리 평균 산출 시, 비원표본 가구원에게 가중치 0을 부여하고 가구원 수에 포함시켜 평균을 계산한다. 그러나 새로 태어난 가구원은 평균을 계산할 때 제외된다.

마지막으로 이전 단계에서 산출한 가구 가중치를 두 번째 시점 해당 가구원의 횡단면 가구원 가중치로 부여한다. 조사자가 가구원 차원에서 횡단면 가중치를 산출할 때에는 원표본 가구원과 비원표본 가구원의 최초 시점 추출확률이 가구 내에서 동일하다는 가정을 하였기 때문에, 원표본 가구원뿐 아니라 비원표본 가구원도 횡단면 가중치 부여 대상으로 포함한다. 따라서 횡단면 분석을 위해서는 조사년도에 응답한 모든 가구원과 이를 포함하고 있는 가구들이 사용된다.

Duncan(1995)이 제시한 방법을 통해 산출된 가구 가중치는 횡단면과 종단면의 구분 없이 분석에서 동일하게 사용된다는 특징이 있다. 그러나 가구원 차원의 분석에서는 연구자가 횡단면과 종단면 가중치를 분석의 목적에 맞게 선택하여 사용해야 한다.

2. 한국노동패널조사의 가중치 산출 방안

한국노동패널조사는 사회, 경제적인 변화가 급속히 진행되고 있는 가운데 노동, 가구 경제, 관광 수요 등 노동시장연구의 활성화를 위해 1998년 처음으로 시행된, 비농촌 지역에 거주하는 한국의 가구와 가구원을 조사한 종단면 조사(longitudinal survey)이다.

한국노동패널 자료 분석을 위한 가중치 산출은 설명된 Duncan(1995)이 제시한 방법을 기반으로 매 해 모집단의 변화를 표본이 반영할 수 있도록 조정되어 이루어졌다. 가중치 산출에서 기본이 되는 원리는 원패널이 구성된 시점에서 추출된 표본이 대표하는 모집단의 변화를 그대로 반영할 수 있도록 하는 것이다. 가중치 산출은 크게 4단계의 과정을 통해 산출되었다. 첫 번째 단계에서는 매 해 발생하는 표본 마모를 고려한 무응답 보정 가구원 종단면 가중치가 일차적으로 산출되고 이를 바탕으로 가구와 가구원의 횡단면 가중치가 산출된다. 두 번째 단계에서는 1단계에서 산출된 가구원 가중치의 가구별 평균을 이용하여 가구 가중치가 부여되고, 세 번째 단계에서는 이전 단계에서 산출된 가구 가중치가 가구원의 횡단면 가중치로 정의된다. 마지막 단계에서는 모집단 규모의 변화를 반영하기 위해 종·횡단면 가구원과 가구 가중치의 스케일 조정이 이루어진다. 각 단계에서 이루어진 과정에 대한 설명들은 다음과 같다.

첫 번째 단계의 무응답 보정을 위해서는 응답 여부를 종속변수로, 가구와 가구원의 정보를 설명변수로 정의한 로지스틱 회귀분석을 통해 응답경향 점수(propensity score)라고 불리는 응답확률을 추정하여 종단면 가중치 산출에 이용하였다. 제 t차 연도에 가구원 i의 응답 여부를 나타내는 변수를 R_{ti}라 정의하면 로지스틱 회귀모형은 다음과 같이 표현된다.

$$R_{ti} \sim Bernoulli(p_{ti}),$$

$$\log\left(\frac{p_{ti}}{1-p_{ti}} \;\middle|\; \boldsymbol{x_{t-1}}\right) = \boldsymbol{x}'_{t-1,i}\,\boldsymbol{\beta}_t.$$

위의 로지스틱 회귀모형에서 t시점 가구원의 응답확률은 전년도의 자료인 $\boldsymbol{x_{t-1}} = (x_{t-1,1},\ \cdots\ x_{t-1,k})'$ 을 이용하여 예측된다. 로지스틱 회귀모형의 통계적 타당성 확보를 위해서는 적절한 설명변수를 선택해야 하기 때문에, 결측 및 모형의 설명력을 고려하여 최종적으로 한국노동패널의 응답확률 추정을 위한 로지스틱 회귀분석에 사용된 설명변수로는 성별, 지역, 학력, 주된 활동 그리고 만 나이가 선택되었다.

이전 연도 $t-1$에 가구원 i에 부여된 종단면 가중치를 $w_{pl,t-1,i}$ 이라 하면 t차 연도에 가구원 i에게 부여되는 기본 종단면 가중치 $w_{pl,t,i}$는 다음과 같이 나타낼 수 있다.

$$w_{pl,t,i} = \frac{1}{\hat{p}_{ti}} w_{pl,t-1,i} \tag{1}$$

여기서 $\hat{p}_{ti} = [1+\exp(-\boldsymbol{x}_{ti}'\hat{\beta})]^{-1}$으로 로지스틱 회귀분석으로부터 추정된 응답확률을 나타낸다.

무응답이 조정된 가중치 (1)에 추계 가구원 수 정보를 이용한 스케일 조정을 통해 최종 가중치가 산출되게 된다. 스케일 조정을 통한 t차 연도의 최종 종단면 가구원 가중치 $w^*_{pl,t,i}$ 는 아래의 식(2)와 같다. 식(2)에서 $w^*_{pl,t-1,i}$ 는 $t-1$시점에 가구원 i에 부여된 최종 종단면 가중치를 나타낸다.

$$w^*_{pl,t,i} = w_{pl,t,i} \frac{\sum_j w^*_{pl,t-1,j}}{\sum_j w_{pl,t,j}} \times \text{추계인구증가율} \tag{2}$$

가구 가중치 역시 추계 가구 수 정보를 통해 산출된 스케일 조정 지수를 아래와 같이 적용하여 최종 가구 가중치가 산출된다.

$$w_{t,h}^{*} = w_{t,h}\frac{\sum_{l} w_{t-1,l}^{*}}{\sum_{l} w_{t,l}} \times \text{추계가구증가율} \quad (3)$$

즉 (3)에서 정의된 가구 가중치를 사용하여 추정되는 가구증가율은 사용된 추계가구증가율과 일치하게 된다. 횡단면 가구원 가중치는 원표본 가구원뿐 아니라 비원표본 가구원에게도 부여되기 때문에 횡단면 가중치의 합이 항상 종단면 가중치의 합보다 크게 된다. 두 종류의 가중치의 합을 동일하게 조정하기 위하여 횡단면 가구원 가중치는 종단면 가중치의 합을 이용하여 최종적으로 결정된다. 즉 시점 t에서 가구원 i에게 부여되는 최종 횡단면 가중치는 다음과 같다.

$$w_{pc,t,i}^{*} = w_{pc,t,i}\frac{\sum_{j} w_{pl,t,j}^{*}}{\sum_{j} w_{pc,t,j}} \quad (4)$$

여기서 $w_{pc,t,j}$ 는 시점 t의 가구원 j에 부여된 가중치로 가구원이 속한 가구에 부여된 가구 가중치인 식(3)의 $w_{t,h}$을 나타낸다. 마지막 단계에서 이루어진 스케일 조정은 일종의 캘리브레이션 과정으로 이해할 수 있다. 이는 실제로 캘리브레이션 방법들 중 가장 많이 사용되는 회귀추정방안으로 모집단 정보인 조사 시점 추계 모집단 인구 및 가구 수를 이용한 추정량이다. 또한 Deville & Sarndal(1992)과 Fuller(2002)에서 증명된 바와 같이 주어진 추정량은 근사 비편향성을 만족하는 강건한(robust) 추정량이다. 한국노동패널을 위한 가중치 산출에 대한 연구로는 강석훈(2003), 김규성 외(2005) 그리고 박민규(2013)가 있다.

Ⅲ. 종단 연구를 위한 가중치 조정

제2장에서 소개된 한국노동패널조사 자료의 분석을 위한 종·횡단면 가중치는 분석의 목적과 분석 방법에 맞추어서 사용되어야 한다. 횡단면 가중치는 조사 당해년도의 관심변수들에 대한 모평균 혹은 모비율 같은 모수 추정에 사용되는 것이 바람직하며 실제 대부분의 연구자들이 이러한 목적으로 횡단면 가구 및 가구원 가중치를 사용하고 있다. 또한 1년의 자료만을 이용한 횡단면 인과관계 분석을 위해서도 횡단면 가중치를 사용할 수 있다. 이를 위해서는 기존 통계 패키지의 조사자료 분석 모듈을 이용할 수 있다. SAS의 경우 Surveymeans, Surveyfreq 그리고 Surveyreg를 이용하여 층화나 집락정보 그리고 가중치를 포함하는 통계적으로 타당한 분석을 실시할 수 있다.

원칙적으로 패널조사의 연구 대상인 모집단의 시간의 경과에 따른 동태적 분석을 위해서는 종단면 가중치가 사용되는 것이 바람직하나, 횡단면 가중치와는 달리 분석을 위한 종단면 가중치의 사용 방안에 대한 연구나 가중치를 활용한 종단면 분석이 가능한 통계 패키지의 모듈 개발 등은 거의 이루어지지 않고 있다. 이는 종단면 분석의 내용과 방법이 매우 다양하고 복잡한 형태를 갖기 때문으로 생각된다.

본 장에서는 종단면 가중치 사용의 첫 걸음으로서 모집단의 동태적 분석을 위하여 흔히 사용되는 (일반화) 선형혼합모형하에서 종단면 분석을 위해 적용할 수 있는 종단면 가중치 보정 방법 및 이의 사용 방안을 제시한다. 종단면 자료 분석을 위하여 고려할 수 있는 선형혼합모형은 다음과 같다.

$$y_{ti} = \boldsymbol{x}_{ti}'\boldsymbol{\beta} + \eta_i + \epsilon_{ti} \tag{5}$$

여기서 y_{ti}는 t년도의 가구원 i로부터 관측된 관심변수를, $\boldsymbol{x}_{ti}$는 분석에 사용되는 설명변수 벡터를, η_i는 가구원 랜덤효과를 그리고 ϵ_{ti}는 가구원 내 반복측정된 값들의 랜덤효과를 나타난다. 식(5)의 선형혼합모형의 적합을 위한 자료의 형태는 <표 1>과 같이 표현될 수 있다.

<표 1> 종단면 분석을 위해 결합된 자료의 형태

t	설명변수	종속변수	가중치
1	$\boldsymbol{X}_1$	$\boldsymbol{y}_1$	$\boldsymbol{w}_1$
2	$\boldsymbol{X}_2$	$\boldsymbol{y}_2$	$\boldsymbol{w}_2$
⋮	⋮	⋮	⋮
T	$\boldsymbol{X}_T$	$\boldsymbol{y}_T$	$\boldsymbol{w}_T$

<표 1>에서 $\boldsymbol{X}_t$는 t년차의 설명변수 행렬을, $\boldsymbol{y}_t$는 종속변수 벡터를 나타내며 $\boldsymbol{w}_t$는 각 연도별로 계산된 종단면 가중치 벡터를 나타낸다. 따라서 종단면 분석을 위한 자료의 행의 수는 각 연차에 관측된 가구원 수의 총 T년 동안의 합이다. 반복된 조사를 통해 얻은 자료이므로 전체 자료 수는 실제 자료를 제공하는 서로 다른 가구원의 수보다 크게 나타나며, 따라서 각 연도별 종단면 가중치를 혼합선형모형의 적합을 위해 그대로 사용할 경우 실제 관심 모집단의 T배가 되는 불분명한 대상을 분석하는 오류가 발생하게 된다. 이러한 문제를 해결하기 위해서는 각 연도별로 산출된 가중치는 일종의 스케일링과 같은 조정을 통하여 분석에 적용해야 한다. 본 연구에서는 두 단계를 통해 산출되는 종단면 연구를 위한 가중치를 제시한다. 1단계에서는 먼저 T년 동안 축적된 자료의 수에 대한 표준화가 이루어진다. 즉 1단계에서 t차 년도의 i번째 가구원에게 부여되는 조정된 가중치 $W_{pl,t,i}^{(1)}$는 다음과 같다. 즉 $W_{pl,t,i}^{(1)}$은 T년도 중 특정 연도 t에 대응되는 모집단의 상대적 크기를 반영한 가중치로 이해할 수 있다.

$$W_{pl,t,i}^{(1)} = w_{pl,t,i}^{*} \frac{\sum_{j} w_{pl,t,j}^{*}}{\sum_{s=1}^{T} \sum_{j} w_{pl,s,j}^{*}} \tag{6}$$

식(6)에 제시된 가중치를 선형혼합모형 분석을 위해 사용할 경우, 모수 추정치에 대한 자유도는 영향을 받지 않으나 최대우도추정량의 값은 가중치를

적용한 경우와 그렇지 않은 경우에 차이가 발생한다. 이는 최대화를 위한 우도함수(likelihood function)의 정의가 서로 다르기 때문이다. 이를 살펴보기 위해 모형(5)에 주어진 혼합모형의 일반적인 형태로 다음의 모형을 고려한다.

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \epsilon \tag{7}$$

여기서 γ는 랜덤효과를, β는 고정효과, $\boldsymbol{X}$는 설명변수를 나타내며 $\boldsymbol{y}$는 관심변수를 의미한다. $\boldsymbol{Z}$는 지시변수로 이루어진 행렬이며 랜덤효과를 나타내는 ϵ와 γ는 정규분포를 따르고 이의 평균과 분산은 다음과 같다.

$$E\begin{bmatrix}\gamma \\ \epsilon\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix},\ \ VAR\begin{bmatrix}\gamma \\ \epsilon\end{bmatrix} = \begin{bmatrix}\boldsymbol{G}, & 0 \\ 0, & \boldsymbol{R}\end{bmatrix} = \begin{bmatrix}\sigma_\gamma^2\boldsymbol{I}, & 0 \\ 0, & \sigma_\epsilon^2\boldsymbol{I}\end{bmatrix}$$

주어진 모형하에서 우도함수는 $\boldsymbol{G}$와 $\boldsymbol{R}$의 함수로만 정의될 수 있으며 이에 근거한 로그 우도함수는 아래와 같이 나타난다.

$$l(\boldsymbol{G},\boldsymbol{R}) = -\frac{1}{2}log|\boldsymbol{V}| - \frac{1}{2}\tau'\boldsymbol{V}^{-1}\tau - \frac{n}{2}log(2\pi). \tag{8}$$

여기서 $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} = \sigma_\gamma^2\boldsymbol{Z}\boldsymbol{Z}' + \sigma_\epsilon^2\boldsymbol{I}$는 $\boldsymbol{y}$의 분산-공분산 행렬을 나타내며 τ는 일종의 잔차로서 $\tau = y - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$이다. 우도함수를 최대화 하는 두 모수 행렬 $\boldsymbol{G}$와 $\boldsymbol{R}$의 해는 Newton-Raphson의 방법을 통해서 얻을 수 있다. $\boldsymbol{G}$와 $\boldsymbol{R}$의 최대우도추정량을 $\hat{\boldsymbol{G}}$와 $\hat{\boldsymbol{R}}$로 표기할 때 고정효과를 나타내는 β와 τ의 추정량은 아래의 식의 해로 정의된다(Henderson 1984).

$$\begin{bmatrix}\boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{X} & \boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{Z} + \hat{\boldsymbol{G}}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\beta} \\ \hat{\tau}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{y}\end{bmatrix} \tag{9}$$

식(9)의 해로 정의되는 고정효과의 최대우도추정량(maximum likelihood estimator) 혹은 최량선형불편추정량(best linear unbiased estimator)은 다음과 같다.

$$\hat{\beta}=(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y, \tag{10}$$
$$\hat{\tau}=\hat{G}Z'\hat{V}^{-1}(y-X\hat{\beta})$$

가중치가 적용된 경우에는, 각 개체가 모수 추정에 미치는 영향이 다르기 때문에 이를 반영한 최대우도 추정량의 정의를 위해 식(8)의 V대신 $\tilde{V}=Z'GZ+W^{-1/2}RW^{-1/2}$이 사용된다. 여기서 W는 가중치를 대각원소로 갖는 대각행렬을 나타낸다. 즉 가중치가 적용된 경우의 최대화되는 우도함수는 가중치가 적용되지 않은 경우와 다르며 따라서 서로 다른 우도함수에 근거한 모수들의 최대우도추정량 역시 달리 나타나게 된다.

선형모형에서 Zyskind(1962)는 가중치 적용 여부에 관계없이 정규분포 하에서 회귀계수의 추정치가 일정해지는 조건을 제시했으나 이 조건은 혼합선형모형에는 적용되지 않는다. 특별히 최적화를 위해서 사용하는 목적함수 (8)로부터 계산되어지는 잔차제곱합은 가중치 적용 여부에 따라 그 값이 매우 다르게 나타난다. 대부분의 표본조사에서 모든 가중치가 1보다 크게 나타나며 가중치가 적용된 추정 방안에서는 잔차제곱합의 계산을 위해서 가중치가 적용된 가중합이 사용되고 따라서 σ_ϵ^2이 과대 추정되는 문제가 발생한다. 이러한 문제를 해결하기 위하여 본 연구에서는 종단면 분석을 위한 가중치 산출의 두 번째 단계로 (6)에서 정의된 가중치를 표본 크기 기준으로 재 표준화하는 방안을 제시한다. 즉 패널조사 자료의 혼합모형에 근거한 종단면 분석을 위하여 사용할 수 있는 최종 가중치로 본 연구에서는 식(11)을 고려하였다.

$$W_{pl,t,i}=W_{pl,t,i}^{(1)}\frac{n}{\sum_{j=1}^{T}\sum_{i}W_{pl,t,i}^{(1)}} \tag{11}$$

식(11)의 표준화된 가중치를 적용하여 혼합선형모형을 적합할 경우 가중치 (6)을 사용할 때 발생하는 각 가구원의 랜덤효과의 분산을 나타내는 σ_ϵ^2의

과대추정 문제를 해결할 수 있다. 이는 σ_ϵ^2의 추정을 위해서는 가중치가 적용된 $\sum w_i \hat{e_i^2}$ 형태의 가중 잔차제곱합이 사용되는데 가중치 (6)이 사용될 경우에는 가중치의 합이 모집단의 수와 일치하도록 구축되어 있어 가중 잔차제곱합을 부풀리기 때문이다.

Ⅳ. 사례연구

본 장에서는 한국노동패널조사의 자료를 활용하여 종단면 분석을 실시한 양정호(2005)의 논문과 문영만(2013)의 논문을 바탕으로 종단면 가중치 사용 여부에 따른 분석 결과를 비교하였다. 양정호(2005)는 각 가구가 지출한 월평균 사교육비에 영향을 미치는 변수들에 대한 연구를 하였으며, 문영만(2013)은 노동조합의 가입 여부와 정규직 여부의 관계를 분석하였다.

사례연구 1: 한국의 사교육비 지출에 대한 종단 연구

양정호(2005)는 사교육비 지출에 가구 및 지역 관련 변수가 미치는 영향을 분석하기 위하여 한국노동패널의 2000년부터 2002년까지의 자료를 사용하였다. 그의 논문에서는 사교육비 분석을 위하여 2000년 3차, 2001년 4차 그리고 2002년 5차 자료 중에 재수생을 포함한 고등학생 이하의 자녀를 둔 가구 중 자녀에게 학원, 개인/그룹과외, 학습지, 방과 후 교내 보충학습, 방과 후 교실 활동을 위해 사교육비를 지출한 가구를 선별하여 사용하였다.

양정호(2005)는 선형혼합모형의 적합을 위해 각 조사 연도의 월평균 명목 사교육비를 2000년도를 기준으로 조정한 실질 사교육비의 자연로그를 종속변수로 사용하였다. 분석을 위한 설명변수로는 가구 관련 변수와 지역 관련 변수가 사용되었다. 가구 관련 변수로는 가구주의 성별, 가구주의 교육수준, 가구 월평균 총소득, 사회적 자본 그리고 재수생 포함 고등학생 이하 자녀 수가 사용되었으며, 사회적 자본 변수로 3차 년도에 조사된 사회의 고위층(교수, 국회의원, 국장 이상 공무원, 대기업 임원, 장성급 이상 군인 등)에 속한 친척

유무가 고려되었다. 지역 관련 변수들은 각 가구 내 가구주의 교육수준을 평균한 지역 평균 가구주의 교육수준, 각 가구의 총소득을 지역별로 평균한 지역 평균 총소득이다.

분석을 위해 고려된 선형혼합모형은 다음과 같다.

$$Y_{tij} = \beta_0 + \beta_1 \alpha_t + \gamma_i + (\alpha\gamma)_{ti} + \eta_{j(i)} + (\alpha\eta)_{tj(i)} + \epsilon_{tij} \tag{12}$$

$$\begin{aligned} Y_{tij} = {} & \beta_0 + \beta_1 \alpha_t + \beta_3 S_i + \beta_4 T_i + \beta_5 (S\alpha)_{ti} + \beta_6 (T\alpha)_{ti} \\ & + \gamma_i + (\alpha\gamma)_{ti} + \eta_{j(i)} + (\alpha\eta)_{tj(i)} + \epsilon_{tij} \end{aligned} \tag{13}$$

$$\begin{aligned} Y_{tij} = {} & \beta_0 + \beta_1 \alpha_t + \beta_3 \mathrm{sex}_{j(i)} + \beta_4 \mathrm{study}_{j(i)} + \beta_5 \mathrm{total}_{j(i)} \\ & + \beta_6 \mathrm{num}_{j(i)} + \beta_7 p_{j(i)} + \beta_8 (\alpha \,\mathrm{sex})_{tj(i)} + \beta_9 (\alpha \,\mathrm{study})_{tj(i)} \\ & + \beta_{10} (\alpha \,\mathrm{total})_{tj(i)} + \beta_{11} (\alpha \,\mathrm{num})_{tj(i)} + \beta_7 (\alpha p)_{tj(i)} \\ & + \gamma_i + (\alpha\gamma)_{ti} + \eta_{j(i)} + (\alpha\eta)_{tj(i)} + \epsilon_{tij} \end{aligned} \tag{14}$$

여기서 종속변수 Y_{tij}는 i번째 지역(구/군/소도시)에 있는 j번째 가구가 측정연도 t에 소비한 월평균 총 사교육비를 의미한다. 변수 α_t은 2000년도는 0, 2001년도는 1 그리고 2002년도는 2의 값을 갖는 범주형 변수이다. γ_i는 지역을, 그리고 $(\alpha\gamma)_{ti}$은 지역과 연도의 교호작용을 나타내는 랜덤효과이며 각 각 평균 0과 분산 σ_γ^2 그리도 $\sigma_{\alpha\gamma}^2$을 갖는 정규분포를 따른다. $\eta_{j(i)}$은 지역 내 가구를 그리고 $(\alpha\eta)_{tj(i)}$은 지역 내 가구와 $\tilde{V} = Z'GZ + W^{-1/2}RW^{-1/2}$도의 교호작용을 나타내는 랜덤효과이며 각 각 평균은 0, 분산 σ_η^2 그리고 $\sigma_{\alpha\eta}^2$을 갖는 정규분포를 따른다. 표기 $j(i)$는 각 가구가 가구가 위치한 지역에 속해 있음(nested)을 나타낸다. ϵ_{tij}는 시간과 지역, 가구 내 즉, 패널가구가 갖는 랜덤효과를 나타내며 평균 0과 분산 σ_ϵ^2을 갖는 정규분포를 따른다.

<표 2> 가구 및 지역 수준 변수 설명

구분	변수		설명
가구 수준	가구주 성별	(sex)	여성=1, 남성=0
	가구주 교육수준	(study)	가구주의 총 교육연수
	가구 총 소득	(total)	월평균 가구 총 실질소득(ln)
	가구 자녀 수	(num)	가구 내 고교 이하 자녀 수
	사회적 자본	(p)	친척 중 사회고위층 유=1, 무=0
지역 수준	지역 평균 교육수준	(S)	구/군/소도시 평균 가구주 교육수준
	지역 평균 총 소득	(T)	구/군/소도시 평균 월평균 총 실질소득

모형(12)는 시간(연도) 이외에 고정효과를 갖는 아무런 설명변수를 투입하지 않았을 때 가구의 사교육비 지출에 대한 가구별 분산을 분석하였다. 모형(13)에서는 지역 수준 변수와 시간과 지역 수준 변수의 교호작용을 고정효과로 포함한 선형혼합모형을 고려하였고 모형(14)에서는 가구 수준 변수와 시간과 가구 수준 변수의 교호작용을 고정효과로 포함한 선형혼합모형을 고려하였다. 사용한 각 가구 수준 및 지역 수준 변수에 대한 설명은 <표 2>와 같다. 분석에 사용된 자료는 3,316가구이며, 종단면 분석을 위한 최종적인 가중치는 표본 크기 3,316을 이용하여 조정해 주었다. 모형 적합과 모수의 최대우도 추정량을 구하기 위해서는 SAS의 Mixed 프로시저가 사용되었다.

<표 3>은 모형(12)의 적용 결과 각 모수의 추정치와 이의 표본오차를 나타내고 있다. 최초 측정된 2000년도의 α_t을 0으로 부여했기 때문에 β_0는 2000년도에 지출한 총 사교육비를 의미한다. 고정효과를 살펴보면 <표 3>에서 나타나는 것과 같이 가중치를 적용하지 않았을 때에 2000년도 총 사교육비 지출은 2.781이며 매년 약 0.155씩 증가하고 있고, 가중치를 적용하였을 때에 2000년도 총 사교육비 지출은 2.791이며 매년 약 0.154씩 증가하고 있음을 알 수 있다. 추정치 모두 유의수준 α=0.01에서 통계적으로 유의하게 나타났다. 가중치를 적용하지 않았을 때와 적용하였을 때의 회귀계수 값을 비교해 보면 절편은 조금 커지고 시간의 계수 값은 조금 작아졌으나 전반적으로

<표 3> 모형(12)의 분석결과

		가중치 미적용		가중치 적용	
		회귀계수	표준오차	회귀계수	표준오차
고정효과	β_0	2.781***	0.03	2.791***	0.03
	$time$	0.155***	0.01	0.154***	0.01
랜덤효과	ϵ_{tij}의 분산	0.210***		0.202***	
	η_{ij}의 분산	0.304***		0.287***	
	$\eta_{ij}^{'}$의 분산	0.016**		0.023***	
	γ_i의 분산	0.038***		0.038***	
	$\gamma_i^{'}$의 분산	0.002~		0.003~	

~ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

<표 4> 모형(13)의 분석결과(지역변수 투입)

		가중치 미적용		가중치 적용	
		회귀계수	표준오차	회귀계수	표준오차
고정효과	β_0	1.061*	0.48	1.068*	0.48
	S	0.115***	0.02	0.111***	0.02
	T	0.041	0.07	0.047	0.07
	$time$	-0.572*	0.27	-0.614*	0.27
	$S \times time$	-0.007	0.01	-0.002	0.01
	$T \times time$	0.108**	0.04	0.106**	0.04
랜덤효과	ϵ_{tij}의 분산	0.212***		0.205***	
	η_{ij}의 분산	0.305***		0.287***	
	$\eta_{ij}^{'}$의 분산	0.015**		0.021***	
	γ_i의 분산	0.012*		0.013*	
	$\gamma_i^{'}$의 분산	0.001***		0.001	

~ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

<표 5> 모형(14)의 분석결과(가구변수 투입)

		가중치 미적용		가중치 적용	
		회귀계수	표준오차	회귀계수	표준오차
고정효과	β_0	1.481***	0.12	1.505***	0.12
	sex	-0.026	0.09	-0.043	0.10
	$study$	0.035***	0.01	0.035***	0.01
	$total$	0.045***	0.01	0.046***	0.01
	num	0.282***	0.03	0.274***	0.03
	p	0.136**	0.05	0.142**	0.05
	$time$	-0.123	0.09	-0.126	0.09
	$sex \times time$	0.134*	0.06	0.123~	0.07
	$study \times time$	0.005	0.01	0.007	0.00
	$total \times time$	0.015	0.01	0.012	0.01
	$num \times time$	0.054~	0.02	0.056**	0.02
	$p \times time$	-0.007	0.03	-0.011	0.03
랜덤효과	ϵ_{tij}의 분산	0.218***		0.201***	
	η_{ij}의 분산	0.226***		0.210***	
	$\eta_{ij}^{'}$의 분산	0.007~		0.014**	
	γ_i의 분산	0.027***		0.028***	
	$\gamma_i^{'}$의 분산	0.002		0.002~	

~ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$

크게 차이가 없음을 알 수 있다. 랜덤효과들의 분산추정량 역시 가중치 적용 여부에 따라 다르게 나타나나 그 차이는 통계적으로 유의하지 않음을 확인할 수 있다.

지역수준 변수를 투입한 모형(13)와 가구수준 변수를 투입한 모형(14)의 분석 결과는 각각 <표 4> 그리고 <표 5>와 같다. 고정효과를 살펴보면 특정 몇몇 설명변수들만이 사교육비 지출과 유의한 관계가 있는 것을 확인할 수 있다. <표 4>에서 제시된 랜덤효과를 살펴보면 지역에 따라 변하는 랜덤효

과 γ와 $(\alpha\gamma)$ 의 분산추정량은 지역 수준 변수를 투입하였을 때 가중치 적용 여부와 상관없이 모두 줄어들었다. 이것은 사교육비 지출 경향이 지역 간에 유의한 차이가 있음을 의미한다. 가구 변수를 포함하여 분석하였을 때 또한 η와 $(\alpha\eta)$의 분산추정량이 줄어들고 이것은 모형(14)을 통해서 가구 간 변동이 잘 설명되고 있으며 이로 인하여 가구 변수들과 사교육비 지출과의 관계 분석의 결과에 영향을 미침을 알 수 있다. 고정효과를 살펴보면 모형(12)의 결과와 마찬가지로 가중치를 적용하였을 때와 적용하지 않았을 때에 큰 차이가 없음을 <표 4>와 <표 5>를 통해 확인할 수 있다. 랜덤효과의 분산 추정량 역시 가중치 적용 여부에 크게 영향을 받지 않음을 확인할 수 있다.

사례연구 2: 패널데이터를 이용한 정규직과 비정규직의 노동조합 가입의향 결정요인

문영만(2013)은 장기간에 걸쳐 지속적으로 하락하고 있는 노조가입율의 구조적인 요인을 밝혀내기 위해 2006년 9차부터 2010년 13차까지의 한국노동패널조사 자료를 병합하여 종단면 분석을 시행하였다. 종단면 분석을 위해서 종속변수로는 노조가입 여부가 그리고 설명변수로는 학력, 성, 연령, 혼인 여부, 고용형태, 임금 그리고 노조 유무가 사용되었다. 분석에서 사용한 변수에 대한 설명은 <표 6>과 같다.

<표 6> 노조가입 관련 변수 설명

변수		설명
성별	sex	여성=0, 남성=1
연령	age	응답자의 연령
혼인 여부	$marriage$	기혼=0, 미혼=1
학력	$study$	연속형으로 사용
고용형태	job	비정규직=0, 정규직=1
임금	$money$	월평균 로그임금
노조 유무	nj	무=0, 유=1

Y_{ti}을 시점 t에서 가구원 i의 노조가입 여부를 나타내는 지시변수로 정의할 때, Y_{ti}는 서로 독립인 베르누이(Bernoulli) 확률변수로서 노조가입확률 $\Pr(Y_{ti}=1)=\pi_{ti}$를 모수로 갖는다. 이때 노조가입확률의 로짓(logit) 함수를 이용한 일반화 선형혼합모형을 다음과 같이 정의할 수 있다(Agresti 2002).

$$Y_{ti} \sim Bernoulli(\pi_{ti}) = \pi_{ti}^{y_{ti}}(1-\pi_{ti})^{1-y_{ti}} \tag{15}$$

$$logit(\pi_{ti}) = \log \frac{\pi_{ti}}{1-\pi_{ti}} = \alpha + \boldsymbol{\beta x}_{ti} + u_i$$

랜덤절편(random intercept) 모형이라고도 불리는 혼합모형(15)에서 u_i는 가구원의 랜덤효과를 나타내며 평균은 0, 분산은 σ_u^2인 정규분포를 따른다고 가정한다. 모수추정을 위해서는 최대우도추정량을 고려하였다. 분석에 사용된 가구원 수는 16,887명이며 모형적합을 위해서는 SAS의 Glimmix 프로시저가 사용되었다.

<표 7> 모형(15)의 분석결과

	가중치 미적용		가중치 적용	
	회귀계수	표준오차	회귀계수	표준오차
β_0	0.689***	0.18	0.853***	0.20
$time$	-0.099***	0.01	-0.099***	0.01
sex	0.327***	0.04	0.305***	0.04
age	-0.037***	0.00	-0.039***	0.00
$marriage$	-0.228***	0.05	-0.225***	0.05
$study$	-0.039***	0.01	-0.045***	0.01
job	0.114***	0.04	0.131**	0.04
$money$	0.019~	0.03	0.024	0.03
nj	-0.144~	0.06	-0.149*	0.07

~ $p \le 0.1$, * $p \le 0.05$, ** $p \le 0.01$, *** $p \le 0.001$

혼합 로짓모형을 통해 노조 가입의향 결정요인을 추정한 분석 결과는 <표 7>과 같다. 설명변수로 투입된 대부분의 변수가 통계적으로 유의한 결과를 보임을 알 수 있다. 가중치의 적용 여부에 따라 회귀계수들이 조금씩 커지거나 작아지긴 하나 계수들의 부호에는 변화가 없는 것으로 판단할 수 있으며 유의한 변수들 역시 가중치 적용 여부에 영향을 크게 받지 않는 것으로 파악된다.

양정호(2005)와 문영만(2013)의 두 논문을 바탕으로 3장에서 주어진 가중치 산출 방법에 따라 가중치를 산출하고 이를 적용한 경우와 그렇지 않은 경우에 (일반화) 선형혼합모형 결과를 비교하였다. 두 결과에서 회귀계수와 분산추정량이 조금씩 작아지거나 커지는 현상이 나타났지만 대체로 큰 차이가 없이 비슷한 결과를 제공하는 것으로 판단할 수 있다. 그러나 본 연구에서 고려한 두 사례를 바탕으로 이러한 결과를 일반화할 수는 없으며, 또한 이러한 결과를 가중치를 사용하지 않는 분석에 대한 타당성을 제공하는 사례로 확대하여 해석할 수도 없다.

Ⅴ. 결론

본 연구에서는 종단면 연구에서 흔히 사용되는 (일반화) 선형혼합모형에 적용할 수 있는 종단면 가중치 조정 방안을 제안하고 산출된 새로운 가중치를 적용하였을 때와 적용하지 않았을 때의 분석 결과를 한국노동패널 자료를 이용한 연구결과들을 이용하여 비교하였다.

두 실제 연구 사례를 바탕으로 (일반화) 선형혼합모형을 적합 후 비교한 결과, 사용된 자료와 모형에 따라 회귀계수와 분산추정량이 조금씩 변하는 현상이 나타났지만 대체로 큰 차이가 없이 비슷한 결과를 보여주고 있다. 패널조사 자료를 활용한 종단면 분석에서는 패널이 대표하는 모집단의 동적 변화에 대한 모형을 가정하고 이를 바탕으로 분석을 하게 된다. 즉 관심이 되는 연구의 주제가 일반적으로 패널에 의하여 대표되는 유한모집단의 동적 변화

에 있으며 따라서 이를 위해서는 패널조사 자료에 함께 제공되는 종단면 가중치를 사용하는 것이 바람직하다.

가중치의 사용 여부에 따라 그 통계의 분석 결과가 동일한 경우에도 가중치를 사용한 분석을 실시해야 하는 또 다른 이유는, 무한모집단의 동태적 분석 시에 패널에 포함된 각 분석 단위가 대표하는 모집단에서의 단위 수가 다르며 이를 반영한 가중치를 분석을 위해 사용할 경우 잘못된 모형의 사용에 대한 결과의 강건성을 보장할 수 있기 때문이다.

한편 기존의 통계 분석 프로그램이 가중치를 활용한 횡단면 분석을 위한 별도의 모듈을 제공하고 있으나 종단면 분석을 위한 이러한 독립된 모듈을 제공하지 않고 있는 상황에서, 본 연구가 제시한 조정된 종단면 가중치를 활용하면 조사 자료의 분석을 위해 별도로 작성된 프로그램이 아닌 기존 통계 프로그램을 별다른 조정 없이 직접 활용하여 종단면 가중치를 포함한 종단면 분석이 가능할 것으로 기대된다.

참고문헌

강석훈. 2003. “KLIPS의 가중치 부여방안 연구.” ≪한국노동패널연구≫ 한국노동연구원.

김규성 · 황영은 · 박진우. 2005. “패널조사에서 가중치 부여 방법 및 효과에 관한 연구.” ≪제6회 한국노동패널 학술대회 논문집≫ 한국노동연구원.

문영만. 2013. “패널데이터를 이용한 정규직과 비정규직의 노동조합 가입의향 결정요인.” ≪산업노동연구≫ 19(2): 127-159.

박민규. 2013. “한국노동패널 가중치 연구.” ≪한국노동패널연구≫ 한국노동연구원.

양정호. 2005. “한국의 사교육비 지출에 대한 종단적 연구: 한국노동패널의 위계적 선형모형 분석.” ≪제5회 한국노동패널 학술대회 논문집≫한국노동연구원.

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*(2nd ed.). Wiley.

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

Duncan, G.J. 1995. “A Simple Method for Weighing in Household Panel Surveys.” Working paper. Northwestern University.

Fuller, W.A. 2002. “Regression Estimation for Survey Samples.” *Survey Methodology* 28: 5-23.

Robinson, G.K. 1991. “That BLUP is a Good Thing: the Estimation of Random Effects.” *Statistical Science* 6: 15-51.

Zyskind, G. 1962. “On Conditions for Equality of Best and Simple Linear Least Squares Estimators.” *Annals of Mathematical Statistics* 33: 1502-1503.

<접수 2014/11/24 수정 2015/1/8 게재확정 2015/1/16>