응용통계연구 제 11권 2호, 1998년, pp. 461-475

패널표본조사에서 층간변동을 고려한 추정방법 *

김영원[1] 오명신[2]

요 약

우리나라 주요기관의 표본조사에서 폭넓게 활용되고 있는 패널조사에서는 시간의 경과에 따른 모집단의 변동을 추정과정에 적절히 반영하는 것이 필요하다. 본 연구에서는 층화추출에 의한 패널조사에서 모집단에 층간변동이 발생하는 경우 이를 추정과정에 반영하지 않으면 심각한 편의가 발생한다는 사실을 밝히고, 층별 가중값을 조정하여 이런 편의를 제거할 수 있는 보정된 불편추정량들을 제시하였다. 또한 제시된 추정량에 대한 분산의 식을 유도하고 이에 대한 추정방법을 제안하였다. 아울러 모의실험을 통하여 제시한 추정량과 이에 따른 분산 추정방법이 매우 효율적이라는 사실을 보였다.

1. 서론

현재 우리나라 주요기관에서 실시되고 있는 많은 통계조사에서는 일정시점에 관한 통계자료에 관심을 가질 뿐만 아니라 월별, 분기별, 또는 연도별 통계자료를 생산하여 관심변수의 시간에 따른 변동을 파악하는 것을 주요한 조사목적으로 하고 있다. 이와 같이 경시적(longitudinal) 통계분석을 효과적으로 수행하기 위해서는 표본으로 추출된 조사단위를 일정한 시간간격을 두고 반복적으로 조사하는 것이 필요하며 따라서 표본설계시 이에 대한 충분한 고려가 필수적이다. Kalton과 Citro(1993)은 이런 연속조사를 위해 활용되고 있는 반복조사(repeated survey), 패널조사(panel survey), 교체패널조사(rotation panel survey) 등의 장단점을 체계적으로 정리하였다. 또한 패널조사에서 복합(composite) 추정량을 활용하여 추정의 정도(precision)를 향상시키는 방법에 대한 연구가 진행되고 있다(Binder와 Hidiroglou, 1988; Cantwell과 Ernst, 1993).

우리나라의 경우 통계청의 경제활동인구조사, 도시가계조사, 농림부의 농가경제조사, 축산물생산비조사, 한국은행의 기업경영분석조사, 주택은행의 주택가격동향조사 등이 모두 패널조사에 해당한다. 이들 조사에 있어서는 표본설계시 한번 선정된 표본을 일정 기간 동안 계속 유지하며 반복적으로 조사하는 방법을 사용하고 있다. 특히 우리나라의 경우 경시적인 분석의 효율성보다는 표본설계나 표본관리의 편리함을 이유로 패널조사를 활용하고 있다.

패널조사에서는 독립 표본조사의 일반적인 오차 이외에도 표본마모(sample attrition), 패널 조건응답(panel conditioning) 등 패널조사의 특성에 의한 독특한 형태의 오차가 발생하게 된다(Bailor, 1989; Kalton과 Citro, 1993). 특히 패널조사에서는 표본설계시 정의된 조사모집단에서 추출된 표본을 계속 사용하므로 시간이 경과함에 따라 발생하는 모집단의

* 본 연구는 숙명여자대학교 1996년도 교비연구비 지원에 의해 수행되었음

1) (140-742) 서울 용산구 청파동, 숙명여자대학교 통계학과, 부교수
2) (140-742) 서울 용산구 청파동, 숙명여자대학교 통계학과

변동과 관련된 오차가 주요한 관심대상이 된다. 패널조사에서는 모집단 변동에 따른 오차를 해결하기 위해 시간의 경과에 따라 발생하는 모집단 구성의 변화를 적절하게 표본에 반영하는 것이 필요하다. 만일 이런 모집단 변동이 반영되지 않은 경우 프레임 오차를 초래하게 되는데 Colledge(1989)는 연속조사에서 프레임 관리문제를 다루고 있다. 이와 관련하여 박진우(1997)는 패널조사에서 모집단의 일부 조사단위가 사라지고 또 다른 조사단위는 새롭게 유입되는 경우, 이를 반영한 추정방법에 관한 연구결과를 제시하고 있다.

한편 대부분의 패널조사에 있어서 표본설계시 층화추출법을 활용하여 층별 특성값과 아울러 전체 모집단의 특성값을 추정하고 있다(박홍래 1989, 이기재 등 1991, 조신섭 등 1995, 한국은행 1997). 이 경우에 있어서 각 시점에서 전체 모집단 모수의 추정을 위해서는 모집단의 층별 구성비를 파악하는 것이 필요하다. 일반적으로 이런 패널조사에서 표본설계시 층별 구성비는 알 수 있지만, 시간이 경과함에 따라 조사단위들의 층간이동이 발생하여 다음 조사시점에는 모집단의 층별 구성비에 변동이 발생하게 된다. 하지만 대규모 표본조사에 있어서 이런 변동은 파악이 곤란하고 따라서 패널조사의 추정에 있어서 많은 경우 표본설계시의 층별 구성비를 그대로 사용하고 있는 것이 현실이다. 만일 이런 변동이 미미하여 추정값에 큰 영향을 주지 않는 경우에는 별 문제가 없겠지만 그렇지 않은 경우 층간 변동을 추정과정에 고려해 주는 것이 절대적으로 필요하다. 본 연구에서는 패널조사에서 층화추출법을 사용하는 경우 모집단의 층간 변동을 고려하지 않는 추정량을 사용할 때 발생하는 편의(bias)를 검토하고, 이런 편의를 제거할 수 있는 추정방법을 제시하고자 한다.

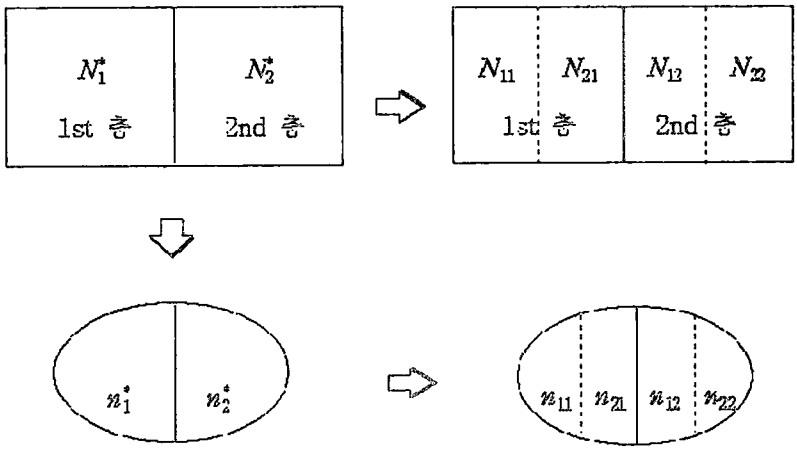
2절에서는 먼저 본 논문에서 필요한 기본적인 개념과 기호에 대해 설명하고, 층화추출법에 의한 패널조사의 추정에 있어서 모집단의 층간 변동을 고려하지 않은 층별 가중값을 사용한 기존의 추정량은 편의가 있음을 보였다. 3절에서는 이런 문제를 해결하기 위해 기존의 추정량을 보완한 불편추정량과 이에 따른 분산 추정방법을 제시하였다. 또한 4절에서는 모의 실험을 통하여 제시된 추정방법의 효율성을 검토하였다.

2. 층간 변동에 의한 편의

2.1. 기본개념

층화추출법을 이용한 패널조사에서 모집단내 조사단위들은 시간이 경과함에 따라 원래의 층에 그대로 남아 있기도 하지만 층화지표의 변화로 다른 층으로 옮겨가기도 한다. 모집단의 변화를 표본에 반영하기 위한 노력으로 표본설계시에 여러 가지 방법으로 보정을 해주는 것도 연구되고 있지만 그에 못지 않게 시간의 경과에 따라 생기는 모집단 구성의 변화를 적절히 추정량에 반영하는 것도 매우 중요하다.

본 연구에서는 시간이 경과함에 따라 모집단에 새롭게 유입되거나 사라지는 조사단위는 없다고 가정한다. 이런 경우 모집단이 변화하기 이전을 가리키는 1차년도와 시간의 경과함에 따라 모집단의 변동이 예상되는 2차년도의 모집단과 표본의 층별 구성을 고려해 보자. 패널조사에서 1차년도에 두 층에 속했던 조사단위 중 일부는 설정된 층화기준에 따라 2차년도에 다른층으로 이동하는 경우가 흔히 발생하게 된다. 그림 2.1은 이런 층간 변동에 따른 모집단과 표본의 변화를 나타낸 것이다.



(가) 1차년도의 모집단과 표본 (나) 2차년도의 모집단과 표본

그림 2.1: 모집단과 표본의 변화

그림 2.1에서 (가)의 표본은 모집단의 각 층에서 랜덤하게 추출된 표본으로서 각 층을 대표하는 표본에 해당한다. 이 표본을 그대로 다음 조사에 이용하게 되면 모집단과 표본은 (나)와 같이 층화지표의 변화로 층이 바뀌어 세분화된 4개의 층으로 구성된다.

1차년도의 첫 번째 층에 속했던 N_1^*개의 조사단위 중 2차년도에 그대로 첫 번째 층에 남게되는 N_{11}개의 조사단위와 두 번째 층으로 이동한 N_{12}개의 조사단위로 나뉘어지게 된다. 또한 1차년도에 두 번째 층에 속했던 N_2^*개의 조사단위는 그대로 두 번째 층에 남게되는 N_{22}개의 조사단위와 첫째 층으로 이동하게 되는 N_{21}개의 조사단위로 구분될 수 있다.

한편 패널조사를 구성하는 표본에 대해 살펴보면 1차년도에 첫 번째 층을 대표하는 n_1^*개로 이루어진 표본은 2차년도에는 더 이상 첫 번째 층을 대표하는 표본으로 볼 수 없다. 즉 2차년도에는 $n_{11}+n_{21}=n_1$으로 이루어진 새로운 표본이 첫 번째 층을 대표하는 표본으로 간주되는 것이 타당하다. 따라서 층의 변화가 랜덤하게 이루어진다고 할 때 그림 2.1과 같이 층을 4부분으로 나누어 보면 각 층은 변화된 모집단의 각 세분화된 층에서 랜덤하게 표본을 뽑는 것과 같은 효과를 내어 각 세분화된 층을 대표하는 표본을 구성하게 된다. 이와 같은 시간의 경과에 따른 층간변동은 앞으로 고려할 여러 가지 추정방법을 연구하는데 중요한 역할을 하게 된다.

우선 본 연구에서 사용할 구체적인 기호를 정리하면 다음과 같다.

N : 전체 모집단의 크기
N_{ij} $(i,j=1,2,\ldots,L)$: i번째 층에서 j번째 층으로 변화한 모집단의 크기
N_i^* : 1차년도에서의 i 번째 층의 모집단의 크기
N_i : 2차년도에서의 i 번째 층의 모집단의 크기
$W_i^*=N_i^*/N$: 1차년도 모집단의 층별 구성비

$W_i = N_i/N$: 2차년도 모집단의 층별 구성비
$W_{ij} = N_{ij}/N \ (i,j = 1,2,\dots,L)$: 2차년도 모집단의 세분화된 층별 구성비
n : 표본의 크기
n_{ij} : i 번째 층에서 j 번째 층으로 변화한 표본의 크기
n_i^* : 1차년도에서의 i 번째 층의 표본의 크기
n_i : 2차년도에서의 i 번째 층의 표본의 크기
$f_i^* = n_i^*/N_i^*$: 1차년도에서의 각 층에서의 표본 추출 비
y_{ijk} : i 번째 층에서 j 번째 층으로 변화한 표본의 k 번째 관찰값
$\bar{y}_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}/n_{ij}$: i 번째 층에서 j 번째 층으로변화한 표본의 평균
$\bar{y}_i = \sum_{j=1}^{L}\sum_{k=1}^{n_{ij}} y_{jik}/n_i$: 2차년도의 i 번째 층에 속하는 표본의 평균
$\mu = \sum_{i=1}^{L}\sum_{j=1}^{L} W_{ij}\bar{Y}_{ij} = \sum_{i=1}^{L} W_i\bar{Y}_i$: 모평균

우선 n_{ij}, n_i와 $\bar{y}_{ij}$, $\bar{y}_i$의 성질에 대해 살펴보자. n_{ij}는 1차년도에 N_i^*에서 추출된 n_i^* 중 어떤 속성(2차년도에서는 j층으로 변하는)을 갖는 단위의 개수이다. 따라서 층간변동이 랜덤하게 발생한다고 가정하면 n_{ij}는 초기하분포를 따르므로, 다음 결과를 얻을 수 있다.

$$E(n_{ij}) = n_i^* N_{ij}/N_i^* = f_i^* N_{ij}, \quad i,j = 1,\dots L$$
$$V(n_{ij}) = \frac{n_i^* N_{ij}}{N_i^*}\left(\frac{N_i^* - N_{ij}}{N_i^*}\right)\left(\frac{N_i^* - n_i^*}{N_i^* - 1}\right) \approx \frac{n_i^* N_{ij}}{N_i^*}\left(\frac{N_i^* - N_{ij}}{N_i^*}\right) \tag{2.1}$$

2차년도에서의 각 층의 표본수의 기대값과 공분산은 다음과 같다.

$$E(n_i) = \sum_{j=1}^{L} E(n_{ji}) = \sum_{j=1}^{L} \frac{n_j^* N_{ji}}{N_j^*} = \sum_{j=1}^{L} f_j^* N_{ji}$$
$$Cov(n_{ij}, n_{ik}) = n_i^*(n_i^* - 1)\frac{N_{ij}N_{ik}}{N_i^*(N_i^* - 1)} - n_i^{*2}\frac{N_{ij}N_{ik}}{N_i^{*2}} \approx -n_i^*\frac{N_{ij}N_{ik}}{N_i^{*2}} \tag{2.2}$$

패널조사는 일반적으로는 대규모의 모집단을 조사대상으로 고려하고 있다. 따라서 본 연구에서는 유한모집단 수정계수(fpc)를 생략한 (2.1)과 (2.2)의 근사적인 분산과 공분산을 이용한다. 여기서 (2.2)는 $Cov(n_{ij}, n_{lk})$에서 $i = l$인 경우이고, $i \neq l$인 경우는 n_{ij}와 n_{jk}가 독립임으로 0이 된다.

한편 테일러 급수전개를 이용하여 다음 결과를 얻을 수 있다.

$$E\left(\frac{1}{n_{ij}}\right) \approx \frac{1}{E(n_{ij})} + 0 + \frac{Var(n_{ij})}{E(n_{ij})^3} + \cdots \approx \frac{N_i^*}{n_i^* N_{ij}} + \left(\frac{N_i^*}{n_i^* N_{ij}}\right)^2\left(\frac{N_i^* - N_{ij}}{N_i^*}\right)$$
$$E\left(\frac{1}{n_i}\right) \approx \frac{1}{E(n_i)} + 0 + \frac{Var(n_i)}{E(n_i)^3} + \cdots \approx \frac{1}{\sum_{j=1}^{L} f_j^* N_{ji}}$$

본 연구에서는 편의상 급수전개의 나머지 항들은 무시한 위에 제시된 근사적인 기대값을 사용하도록 한다.

또한 n_{ij}가 확률변수이므로 $\bar{y}_{ij}$와 $\bar{y}_i$의 기대값은 다음과 같다.

$$E(\bar{y}_{ij}) = E(E(\bar{y}_{ij}|n_{ij})) = \bar{Y}_{ij}$$

$$E(\bar{y}_i) = E(\sum_{j=1}^{L} \frac{n_{ji}}{n_i}\bar{y}_{ji}) \approx (\sum_{j=1}^{L} \frac{f_j^* N_{ji}}{(\sum_{k=1}^{L} f_k^* N_{ki})}\bar{Y}_{ji}) \neq \bar{Y}_i$$

만약, 각 층에서의 표본 추출비 f_j^*들이 같다면 즉, 비례배정(proportional allocation)인 경우에는 $E(\bar{y}_i) \approx \bar{Y}_i$이 근사적으로 성립한다. 또한 $\bar{y}_{ij}$의 분산은 fpc를 무시하면 다음과 같다.

$$V(\bar{y}_{ij}) = E(V(\bar{y}_{ij}|n_{ij})) + V(E(\bar{y}_{ij}|n_{ij})) \approx \sigma_{ij}^2 E(\frac{1}{n_{ij}})$$

이와 같은 기본 성질들을 이용해 추정량들의 기대값, 편의, 분산 등을 유도하기로 한다.

2.2. 편 의

층화추출에 의한 패널표본조사에서는 우선 관심대상인 층별 평균을 추정하고, 이들 평균에 가중값을 적용하여 전체 모집단 평균을 추정하고 있다. 하지만 일반적으로 현재 모집단에 대한 정보가 부족하여 2차년도의 모집단의 구성비인 층별 가중값을 알기가 어렵다. 따라서 많은 패널조사에서는 모평균 추정을 위해 1차년도의 표본설계시에 쓰였던 층별 가중값을 그대로 적용하고 있다.이렇게 모집단의 층간 변화를 고려하지 않은 1차년도의 가중값을 그대로 추정에 이용한 추정량과 편의는 다음과 같다.

$$\hat{\mu}_1 = \sum_{i=1}^{L} W_i^* \bar{y}_i \tag{2.3}$$

$$Bias(\hat{\mu}_1) \approx \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{N_{ij}}{N}(\frac{f_i^* N_j^*}{\sum_{k=1}^{L} f_k^* N_{kj}} - 1)\bar{Y}_{ij} \tag{2.4}$$

또한 변화된 2차년도 모집단에 대한 층별 구성비를 알고 있을 경우에도 표본의 층간 이동을 고려하지 않은 추정방법을 그대로 사용하면 편의가 발생하게 된다. 이런 결과가 나오게 되는 이유는 패널조사의 성격상 표본설계시 선정된 동일한 표본으로 조사를 반복하기 때문에 모집단에 층간 변동이 발생하게 되면 고정된 표본은 더 이상 각 층에서 랜덤하게 추출된 것으로 볼 수 없기 때문이다. 즉, 2차년도의 모집단에 있어서 일반적인 층화추출에서 성립하는 $E(\bar{y}_i) = \bar{Y}_i$ 이 성립하지 않기 때문이다.

2차년도의 모집단 평균 추정을 위해 표본의 변동을 적절하게 반영하기 않고 단순히 2차년도의 층별 가중값 W_i을 이용하여 구한 모평균의 추정량은 다음과 같다.

$$\hat{\mu}_2 = \sum_{i=1}^{L} W_i \bar{y}_i \tag{2.5}$$

(2.5)에 제시된 추정량을 사용하는 경우 발생하는 편의는 다음과 같다.

$$Bias(\hat{\mu}_2) \approx \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{N_{ij}}{N}(\frac{f_i^* N_j}{\sum_{k=1}^{L} f_k^* N_{kj}} - 1)\bar{Y}_{ij} \tag{2.6}$$

여기서 표본설계시 각 층의 추출률에 해당하는 f_j^*들이 같다면, 즉 층화추출에서 비례배정법을 사용하면 $E(\bar{y}_i) \approx \bar{Y}_i$이 성립하여 $\hat{\mu}_2$는 근사적인 불편추정량이 된다. 하지만 대부분의 패널조사를 위한 표본설계에서는 네이만(Neyman)배정법을 사용한다는 점에 유의할 필요가 있다.

한편 2차년도에 세분화된 각 층의 구성비를 알고 있는 경우를 고려해 보자. 이 경우 기존의 층을 그림 2.1에서 나타낸 것과 같이 세분화된 층으로 구분해 원래의 L개 층에서 L^2개 층으로 나누어 세분화된 각 층의 구성비를 알고 있다면 추정량은 다음과 같고 기대값을 구해 보면 불편추정량이 된다.

$$\hat{\mu}_3 = \sum_{i=1}^{L}\sum_{j=1}^{L} W_{ij}\bar{y}_{ij} \tag{2.7}$$

$$E(\hat{\mu}_3) = \sum_{i=1}^{L}\sum_{j=1}^{L} W_{ij}E(\bar{y}_{ij}) = \sum_{i=1}^{L}\sum_{j=1}^{L} W_{ij}\bar{Y}_{ij} = \mu$$

이 추정량의 분산은 사후층화 개념을 도입하여 구하면 다음과 같다.

$$V(\hat{\mu}_3) \approx \sum_{i}^{L}\sum_{j}^{L} \frac{N_{ij}^2}{N^2}\left(\frac{N_i^*}{n_i^* N_{ij}} + \left(\frac{N_i^*}{n_i^* N_{ij}}\right)^2\left(\frac{N_i^* - N_{ij}}{N_i^*}\right)\right)\sigma_{ij}^2 \tag{2.8}$$

추정량 $\hat{\mu}_3$을 활용하기 위해서는 모집단 변동을 반영한 세분화된 층들의 구성비 W_{ij}를 정확하게 파악하는 것이 필요하다. 하지만 일반적인 패널조사에서 W_{ij}를 파악하여 이를 추정량에 반영하는 것은 현실적으로 거의 불가능하다. 본 논문에서 추정량 $\hat{\mu}_3$을 제시한 이유는 4절에서 제시되는 추정량의 효율성 비교에 그 목적이 있다.

3. 층간 변동을 고려한 추정방법

본 절에서 제안할 두 추정량은 표본설계시 파악하고 있는 1차년도의 층별 구성비와 현재 표본에서 얻을 수 있는 층간 변동과 관련된 정보를 적절히 변환해 L^2개 층의 표본평균에 보정된 가중값을 적용한 불편추정량이다. 새로이 제시된 추정량 중 $\hat{\mu}_4$는 2차년도의 층별 구성비에 대한 정보를 전혀 갖고 있지 못하는 경우에 활용 가능한 추정량으로 (2.3)의 추정량 $\hat{\mu}_1$을 보완한 추정방법이다. 또한 $\hat{\mu}_5$는 2차년도의 층별 구성비에 대한 정보는 갖고 있지만 층간 변동에 따른 세분화된 층별 구성비는 알 수 없는 경우에 활용 가능한 추정량으로 (2.5)의 추정량 $\hat{\mu}_2$를 보완한 추정방법이다.

3.1. 층별 구성비를 모르는 경우

2차년도 층별 구성비에 관한 정보가 없는 경우 1차년도의 층별 가중값에 표본에서 얻은 층간 변동률을 반영하여 $\hat{\mu}_1$을 보정한 추정량은 다음과 같다.

$$\hat{\mu}_4 = \sum_{i=1}^{L}\sum_{j=1}^{L} W_i^* \frac{n_{ij}}{n_i^*}\bar{y}_{ij}$$

2.1절의 내용을 활용하여 $\hat{\mu}_4$의 기대값을 구하면, 다음과 같이 불편추정량임을 알 수 있다.

$$E(\hat{\mu}_4) = \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{N_i^*}{N}\frac{n_i^* N_{ij}}{N_i^*}\frac{1}{n_i^*}\bar{Y}_{ij} = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_{ij}}{N}\bar{Y}_{ij} = \mu$$

한편, $\hat{\mu}_4$ 의 분산은 2.1절의 내용을 활용하면 다음과 같은 과정을 통하여 유도될 수 있다.

$$\begin{aligned} V(\hat{\mu}_4) =& E(Var(\hat{\mu}_4|n_{ij})) + Var(E(\hat{\mu}_4|n_{ij})) \\ =& (\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i^{*2}}{N^2 n_i^{*2}}\frac{n_i^* N_{ij}}{N_i^*}\sigma_{ij}^2) + (\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i^{*2}}{N^2\ n_i^{*2}}\bar{Y}_{ij}^2 Var(n_{ij})) \\ &+ \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L} 2\frac{N_i^{*2}}{N^2}\frac{\bar{Y}_{ij}\bar{Y}_{ik}}{n_i^{*2}}Cov(n_{ij}, n_{ik}) \end{aligned}$$

따라서 이를 정리하면 다음 결과를 얻을 수 있다.

정리 3.1 층별 구성비를 모르는 경우, 2차년도 모집단의 평균에 대한 불편추정량과 그 분산은 다음과 같다.

$$\hat{\mu}_4 = \sum_{i=1}^{L}\sum_{j=1}^{L} W_i^*\frac{n_{ij}}{n_i^*}\bar{y}_{ij} \tag{3.1}$$

$$V(\hat{\mu}_4) = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i^* N_{ij}}{N^2 n_i^*}\sigma_{ij}^2 + \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L}\frac{N_{ij}N_{ik}}{N^2 n_i^*}(\bar{Y}_{ij} - \bar{Y}_{ik})^2 \tag{3.2}$$

한편 (3.2)의 분산을 추정하기 위해 모수 대신 표본에서 얻은 추정값을 대입한 다음의 추정량을 고려할 수 있다.

$$\begin{aligned} V(\hat{\mu}_4) =& \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i^*}{N^2 n_i^*}\frac{N_i^* n_{ij}}{n_i^*}s_{ij}^2 \\ &+ \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L}\frac{1}{N^2 n_i^*}\frac{N_i^*(N_i^*-1)n_{ij}n_{ik}}{n_i^*(n_i^*-1)}\{(\bar{y}_{ij} - \bar{y}_{ik})^2\} \end{aligned} \tag{3.3}$$

여기서 $V(\hat{\mu}_4)$의 기대값을 구해 보면 다음과 같다.

$$\begin{aligned} E(V(\hat{\mu}_4)) =& \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i^*}{N^2 n_i^*}\frac{N_i^*}{n_i^*}\frac{n_i^* N_{ij}}{N_i^*}\sigma_{ij}^2 \\ &+ \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L}\frac{1}{N^2 n_i^*}\frac{N_i^*(N_i^*-1)}{n_i^*(n_i^*-1)}\frac{n_i^*(n_i^*-1)N_{ij}N_{ik}}{N_i^*(N_i^*-1)}\{(\bar{Y}_{ij} - \bar{Y}_{ik})^2 + Var(\bar{y}_{ij}) + Var(\bar{y}_{ij})\} \end{aligned}$$

따라서 제시된 분산의 추정량을 사용할 때 발생하는 편의는 다음과 같다.

$$Bias(V(\hat{\mu}_4)) = \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L} \frac{N_{ij}N_{ik}}{N^2 n_i^*}\{Var(\bar{y}_{ij}) + Var(\bar{y}_{ij})\}$$

3.2. 층별 구성비를 아는 경우

2절에서 살펴본 바와 같이 2차년도 층별 구성비 W_i를 알고 있더라도 이를 그대로 층화추출 추정량의 가중값으로 사용하면 편의가 발생하게 된다. 세분화된 층별 구성비 W_{ij}는 모르고 W_i만을 알고 있을 때 $\hat{\mu}_2$를 보정한 불편추정량은 다음과 같다.

$$\hat{\mu}_5 = \sum_{i=1}^{L}\sum_{j=1}^{L}(W_j - \sum_{k\neq i}^{L} W_k^* \frac{n_{kj}}{n_k^*})\bar{y}_{ij}$$

2.1절의 내용을 활용하여 $\hat{\mu}_5$의 기대값을 구하면, 다음과 같이 불편추정량임을 알 수 있다.

$$E(\hat{\mu}_5) = \sum_{i=1}^{L}\sum_{j=1}^{L}(\frac{N_j}{N} - \sum_{k\neq i}^{L}\frac{N_{kj}}{N})\bar{Y}_{ij} = \sum_{i=1}^{L}\sum_{j=1}^{L}(\frac{N_{ij}}{N})\bar{Y}_{ij} = \mu$$

한편, $\hat{\mu}_5$의 분산은 2.1절의 내용을 활용하면 다음 과정을 통하여 유도될 수 있다.

$$\begin{aligned} V(\hat{\mu}_5) =& E(Var(\hat{\mu}_5|n_{ij})) + Var(E(\hat{\mu}_5|n_{ij})) \\ =& \sum_{i=1}^{L}\sum_{j=1}^{L}\{(W_j - \sum_{k\neq i}\frac{N_{kj}}{N})^2 + \sum_{k\neq i}^{L}\frac{N_{kj}(N_k^* - N_{kj})}{N^2 n_k^*}\}E(\frac{1}{n_{ij}})\sigma_{ij}^2 \\ &+ \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L}(\frac{N_{ij}N_{ik}}{N^2 n_i^*}(\sum_{m\neq i}^{L}\bar{Y}_{mj} - \sum_{m\neq i}^{L}\bar{Y}_{mk})^2) \end{aligned}$$

따라서 이를 정리하면 다음 결과를 얻을 수 있다.

정리 3.2 층별 구성비를 아는 경우, 2차년도 모집단의 평균에 대한 불편추정량과 그 근사적인 분산은 다음과 같다.

$$\hat{\mu}_5 = \sum_{i=1}^{L}\sum_{j=1}^{L}(W_j - \sum_{k\neq i}^{L} W_k^* \frac{n_{kj}}{n_k^*})\bar{y}_{ij} \tag{3.4}$$

$$\begin{aligned} V(\hat{\mu}_5) \approx& \sum_{i=1}^{L}\sum_{j=1}^{L}\{(W_j - \sum_{k\neq i}\frac{N_{kj}}{N})^2 + \sum_{k\neq i}^{L}\frac{N_{kj}(N_k^* - N_{kj})}{N^2 n_k^*}\}\{\frac{N_i^*}{n_i^* N_{ij}} + (\frac{N_i^*}{n_i^* N_{ij}})^2(\frac{N_i^* - N_{ij}}{N_i^*})\}\sigma_{ij}^2 \\ &+ \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L}(\frac{N_{ij}N_{ik}}{N^2 n_i^*}(\sum_{m\neq i}^{L}\bar{Y}_{mj} - \sum_{m\neq i}^{L}\bar{Y}_{mk})^2) \end{aligned} \tag{3.5}$$

한편 (3.5)의 분산을 추정하기 위해 모수 대신 표본에서 얻은 추정값을 대입한 다음의 추정량을 고려할 수 있다.

$$V(\hat{\mu}_5) = (\sum_{i=1}^{L}\sum_{j=1}^{L}(W_j - \sum_{k\neq i}^{L} W_k^* \frac{n_{kj}}{n_k^*})^2 \frac{s_{ij}^2}{n_{ij}}) + \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L}(\frac{1}{N^2 n_i^*}\frac{N_i^*(N_i^*-1)n_{ij}n_{ik}}{n_i^*(n_i^*-1)}(\sum_{m\neq i}^{L}\bar{y}_{mj} - \sum_{m\neq i}^{L}\bar{y}_{mk})^2) \quad (3.6)$$

또한 $V(\hat{\mu}_5)$의 기대값을 구해보면 이 추정량의 편의는 근사적으로 다음과 같다.

$$Bias(V(\hat{\mu}_5)) \approx \sum_{i=1}^{L}\sum_{j<}^{L}\sum_{k}^{L}\frac{N_{ij}N_{ik}}{N^2 n_i^*}\{Var(\sum_{m\neq i}^{L}\bar{y}_{mj}) + Var(\sum_{m\neq i}^{L}\bar{y}_{mk})\}$$

4. 모의실험을 통한 효율성 비교

이제까지 살펴본 5가지 추정량을 모의실험 자료에 적용시켜 효율성을 비교해 보고자 한다. 모의실험과정을 개략적으로 살펴보면 다음과 같다.

(절차 1) 모집단 생성

(절차 2) 층화추출에 의한 표본 생성

(절차 3) 모집단의 층간 변동을 반영한 후 표본에서 각각의 추정값을 구함

(절차 4) 충분한 횟수만큼 (절차 2)와 (절차 3) 반복

(절차 5) 추정량들의 편의와 MSE 비교, 분산의 추정량을 활용한 신뢰구간 검토

모의실험은 SAS/IML 을 사용하여 간단한 형태인 층이 2개일 경우에 적용해 보았다. 원래는 변화이전 모집단에서 표본을 추출한 후 모집단을 임의로 변화시킬 때 모집단 및 표본의 변화를 관찰하여야 하지만, 일정한 모집단에서 표본을 여러 번 추출해 보아야 효율성 비교가 가능하므로 본 모의실험에서는 관찰하려는 변수와 함께 1차년도의 층과 2차년도의 층을 표시해 이미 변화된 모집단을 생성하되 표본은 변화이전 층을 기준으로 추출하였다.

4.1. 모집단의 생성과 표본 추출

경제 관련 조사에서 흔히 발생하는 비대칭 모집단 분포를 고려하기 위해 관심 대상이 되는 변수 Y를 모집단의 크기($N = 5,000$)만큼 로그정규분포(log-normal distribution)에서 랜덤하게 생성시킨 후, 관심변수와 층화변수의 상관관계를 고려하기 위해 층화지표가 되는 X 변수를 다음 관계를 만족하도록 생성하였다.

$$X_i = aY_i + U_i \quad , \quad i = 1, \ldots, N$$

$$a = \sqrt{(\frac{\rho^2}{1-\rho^2})(\frac{S_{uu}}{S_{yy}} - \frac{S_{yu}^2}{S_{yy}})} - \frac{S_{yu}}{S_{yy}}$$

ρ : X 와 Y의 상관계수
U_i : 균일분포에서 생성한 난수

다음 단계로 모집단에 층을 설정하는데 있어서 비대칭 분포에서 실제조사와 비슷한 층화형태를 감안하기 위해 X값을 기준으로 상이한 크기로 층을 2개($N_1^* = N_{11} + N_{12} = 4,000, N_2^* = N_{22} + N_{21} = 1,000$)로 분할하였다. 아울러 두 조사시점간에 층화지표의 변화에 따라 모집단에서 발생하는 층간 변동을 반영하기 위해 각 층의 조사단위 중 일정 개수(N_{12}, N_{21})를 랜덤하게 선정하여 층화변수 X를 랜덤하게 변화시키고, 변화된 X에 따라서 위의 식을 변형해 다시 관심변수 Y를 생성해 층간 변동이 반영된 두 조사시점의 모집단을 설정하였다. 표본은 표본의 크기($n = n_1^* + n_2^* = 500$)를 각각 비례배정법과 네이만배정법을 적용해 1차년도 모집단의 각 층(N_1^*, N_2^*)에서 추출한다. 1차년도와 2차년도에서의 모집단 변동에 따라 추출된 표본 중 일부는 조사시점이 바뀜에 따라 층간 이동을 하게 된다.

4.2. 추정량의 효율성 분석

각 추정량의 효율성 비교를 위해 동일한 모집단에서 표본을 2000번 반복적으로 뽑아 각 표본에서 앞에서 제시된 추정방법들에 의한 2차년도 모평균에 대한 편의와 MSE를 구하였다. 아울러 3절에서 제시한 $\hat{\mu}_4$와 $\hat{\mu}_5$의 분산의 추정량이 유효한지 알아보기 위해 각 표본에서 분산 추정값에 의한 95 % 신뢰구간을 구해 모평균이 신뢰구간 안에 들어가는 비율을 계산하였다.

본 모의실험에서는 층화변수 X와 관심변수 Y의 상관계수 ρ의 변화에 따른 영향을 검토하기 위해 ρ=0.4, 0.6, 0.8인 경우를 고려하였고, 또한 층간변동이 심해짐에 따라 발생하는 변화를 검토하기 위해 층간변동 비율 $p = N_{12}/N_1^* = N_{21}/N_2^*$ 이 0.1 , 0.2 , 0.3 인 경우를 고려하였다.

모의실험 결과는 표 4.1과 같다. 물론 층별로 층간변동 비율이 다른 경우도 관심대상이 될 수 있으나, 이 경우 모의실험 결과 표 4.1와 큰 차이가 없어 그 결과를 수록하지 않았다. 표 4.1에서 비례배정의 경우 $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_4$, $\hat{\mu}_5$이 약간의 차이는 있지만 예상과 같이 모두 편의가 발생하지 않는 것을 확인할 수 있다. 또한 이들 추정량들의 MSE에 있어서 큰 차이는 없지만 대체적으로 $\hat{\mu}_2$와 $\hat{\mu}_4$가 효율적임을 볼 수 있다. 특히 2차년도 모집단의 층별 구성비를 아는 경우에만 활용이 가능한 $\hat{\mu}_2$와 비교하여 이런 층별 구성비를 모르는 경우에도 활용이 가능한 $\hat{\mu}_4$의 효율성이 큰 차이가 없다는 점에 유의할 필요가 있다.

반면, 층간변동을 고려하지 않고 흔히 사용되고 있는 1차년도의 층별 구성비를 이용한 추정량 $\hat{\mu}_1$를 살펴보면 추정상에 큰 오차가 발생한다는 사실을 확인할 수 있다. 특히 ρ와 p가 커짐에 따라 이런 현상은 점점 더 심각해진다. 즉 층화변수와 관심변수의 상관계수가 크고 모집단에서 층간 변동이 많이 일어난다면 이 추정량은 매우 잘못된 추정결과를 산출한다는 사실을 확인할 수 있다.

또한 패널조사의 표본설계에서 흔히 적용되는 네이만배정의 경우, $\hat{\mu}_2$이 더 이상 근사적인 불편추정량이 되지 않는다. $\hat{\mu}_1$, $\hat{\mu}_2$을 제외한 나머지 추정량은 거의 편의가 발생하지 않고 있다. $\hat{\mu}_3$, $\hat{\mu}_4$, $\hat{\mu}_5$의 MSE를 살펴보면 본 연구에서 제시한 $\hat{\mu}_4$가 가장 효율적인 추정방법

표 4.1: 편의(Bias)와 MSE

ρ	P		비례 배정					네이만배정				
	μ		$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
0.4	0.1	Bias	-3.82	0.14	0.09	-0.03	0.24	-2.70	1.44	0.20	0.08	0.35
	112.98	MSE	43.83	38.14	40.97	38.27	41.46	36.42	33.98	40.61	38.03	41.08
	0.2	Bias	-6.46	0.55	0.53	0.00	0.83	-4.49	2.73	0.73	0.18	0.01
	114.76	MSE	78.76	55.42	62.42	54.03	57.13	63.99	55.65	63.02	54.55	59.82
	0.3	Bias	-8.85	1.59	1.37	0.04	2.79	-6.94	3.97	1.47	0.12	2.91
	116.96	MSE	117.89	72.98	83.66	65.18	81.08	98.27	77.35	90.66	69.90	90.93
0.6	0.1	Bias	-5.90	0.34	0.27	0.07	0.46	-3.61	2.87	0.31	0.10	0.50
	115.03	MSE	51.35	21.55	22.22	22.34	23.57	30.82	27.70	21.97	21.82	23.18
	0.2	Bias	-9.86	0.96	0.73	0.10	1.46	-6.16	5.37	0.75	0.10	1.45
	117.19	MSE	116.29	29.43	31.10	28.89	32.73	62.38	55.69	31.61	29.28	33.58
	0.3	Bias	-13.77	2.49	2.16	0.07	4.07	-8.67	8.44	2.16	0.06	4.09
	122.89	MSE	210.10	40.66	42.30	33.25	54.86	103.44	104.24	44.96	35.48	57.71
0.8	0.1	Bias	-7.86	0.36	0.28	0.01	0.52	-4.43	4.12	0.31	0.04	0.54
	115.58	MSE	73.62	15.91	16.04	17.52	18.08	31.40	30.23	15.02	16.50	17.06
	0.2	Bias	-13.83	1.23	1.00	0.03	1.82	-7.19	8.79	1.01	0.03	1.84
	121.05	MSE	203.49	21.21	21.41	21.77	25.89	67.34	95.42	21.64	21.87	26.84
	0.3	Bias	-17.57	3.09	2.62	0.01	5.23	-10.30	11.58	2.66	0.04	5.27
	123.26	MSE	323.14	33.29	32.59	25.15	55.72	128.36	158.13	35.05	26.96	59.00

임을 확인할 수 있다. 여기서 경우에 따라 $\hat{\mu}_3$가 $\hat{\mu}_4$보다 효율적인 것으로 나타나고 있으나, $\hat{\mu}_3$는 2차년도의 세분화된 층별 구성비를 알아야 적용이 가능한 추정방법이기 때문에 현실적으로 활용이 거의 불가능한 추정방법이란 사실을 유의할 필요가 있다.

4.3. 분산 추정량을 활용한 신뢰구간

본 연구에서 제시한 분산의 추정량 (3.3)과 (3.6)은 편의추정량이다. 본 모의실험에서는 이런 편의에도 불구하고 이들 분산 추정량을 적용하여 얻은 신뢰구간이 실제 모수를 포함하는 비율을 검토하여 분산 추정량의 유효성을 확인하였다. 모의실험 표본에서 모평균에 대한 95% 신뢰구간 $\hat{\mu} \pm 1.96\sqrt{V(\hat{\mu})}$ 을 구해 실제 모평균을 포함하는 비율을 구한 결과는 표 4.2와 같다.

모의실험 결과 $\hat{\mu}_5$의 분산 추정량으로 구한 95 %신뢰구간은 경우에 따라 원하는 신뢰수준과 상당한 차이를 보여주고 있지만, $\hat{\mu}_4$의 분산 추정량으로 구한 95%신뢰구간은 원하는

표 4.2: 신뢰구간의 모수 포함 비율

	추정 방법	ρ=0.4			ρ=0.6			ρ=0.8		
		p=0.1	p=0.2	p=0.3	p=0.1	p=0.2	p=0.3	p=0.1	p=0.2	p=0.3
비례 배정	$\hat{\mu}_4$	0.947	0.946	0.950	0.949	0.951	0.952	0.942	0.941	0.949
	$\hat{\mu}_5$	0.942	0.944	0.938	0.951	0.946	0.913	0.945	0.940	0.845
네이만 배정	$\hat{\mu}_4$	0.948	0.952	0.947	0.949	0.952	0.949	0.944	0.946	0.946
	$\hat{\mu}_5$	0.945	0.949	0.937	0.948	0.946	0.909	0.947	0.937	0.855

신뢰수준을 거의 만족시키고 있다. 따라서 본 연구에서 제시한 $\hat{\mu}_4$와 $V(\hat{\mu}_4)$를 활용한 신뢰구간은 패널조사의 통계적 분석에 있어서 효과적으로 활용될 수 있을 것으로 판단된다.

5. 결 론

패널조사는 표본설계 당시에 선정된 동일한 표본을 가지고 일정한 시간 간격으로 반복하여 조사하는 표본조사의 한 방법이다. 이런 조사에서 모집단의 구성은 시간이 지남에 따라 달라지게 되는데 모집단의 변화에 적절히 대응하지 않는다면 표본조사 결과는 편의를 초래하게 된다. 본 연구에서는 층화추출을 이용한 패널조사에서 모집단의 층화지표의 변화로 층간 변동이 있는 경우 모집단의 변화를 고려하지 않고 종래의 추정량을 그대로 사용한다면 편의가 발생한다는 것을 밝혔다. 또한, 이런 층간 변동에 의한 편의를 제거해주는 보정된 불편 추정량을 제시하고 그 분산의 식을 유도하였다. 아울러 추정량의 분산을 추정하는 방법을 제시하였다.

한편 본 연구에서 고려한 선형추정량들은 다음과 같이 정리될 수 있다.

$$\hat{\mu} = \sum_{i=1}^{L}\sum_{j=1}^{L}\sum_{k=1}^{n_{ij}} w_{ijk} y_{ijk}$$

제시된 추정량 $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$, $\hat{\mu}_4$, $\hat{\mu}_5$을 정리하면 w_{ijk}는 각각 $N_i^*/(Nn_i)$, $N_i/(Nn_i)$, $N_{ij}/(Nn_{ij})$, $N_i^*/(Nn_i^*)$, $(W_j - \sum_{k\neq i}^{L} W_k^* \frac{n_{kj}}{n_k^*})/n_{ij}$ 인 경우에 해당한다. 여기서 추출확률을 고려해 보면 불편추정량이 되는 조건은 $E[w_{ijk}] = N_i^*/(Nn_i^*)$이다. 결과적으로 $\hat{\mu}_3$, $\hat{\mu}_4$, $\hat{\mu}_5$의 경우는 모두 불편성 조건을 만족시킨다.

모의실험 결과 본 연구에서 제시한 불편추정량 $\hat{\mu}_4$이 매우 효율적이며, 분산의 추정량을 이용해 신뢰구간을 구해 모수를 포함하는 비율을 검토해 본 결과 제시된 추정량에 의한 신뢰구간이 설정된 유의수준을 만족시켜 주는 것으로 나타났다.

한편 본 연구에서는 층간변동이 랜덤하게 발생하는 것으로 가정하고 있으나, 실제적으로는 층간 경계에 위치한 조사단위들이 다른 것들에 비하여 층간변동이 발생할 가능성이 높을 것이다. 따라서 이런 층간 유입 현상을 확률모형화하여 추정에 반영하는 방법에 관한 연구가 향후에 수행되기를 기대하여 본다.

참고문헌

[1] 박진우 (1997). 패널조사에서 표본 변경을 고려한 추정. 〈응용통계연구〉. 제10권. 367-374.

[2] 박홍래 (1989). 면적조사 및 생산량조사 표본설계 . 〈박홍래교수 회갑기념논총〉. 55-74.

[3] 이기재, 박진우, 박홍래 (1991). 전국 도시 주택가격 동향조사를 위한 표본설계 연구, 〈응용통계 연구〉. 제4권. 137-148.

[4] 조신섭, 박홍래, 김영원, 최대우, 김규성 (1995). 〈축산물 생산비조사 표본설계〉. 서울대학교 통계연구소.

[5] 한국은행 (1997). 〈1997 기업경영분석〉. 한국은행.

[6] Bailor, R. A. (1989). *Information Needs, Surveys, and Measurement Errors, Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh). New York: John Wiley. 1-24.

[7] Binder, D. A. , Hidiroglou, M. A. (1988). *Sampling in time, Handbook of Statistics.* (Vol.6) (Eds. P. R. Krishnaiah and C. R. Rao). New york: North Holland, 187-211.

[8] Cantwell, P. J. , Ernst, L. R. (1993). New Developments in Composite Estimation for the Current Population Survey. *Proceedings: Symposium 92. Design and Analysis of Longitudinal Surveys.* Statistics Canada. 121-130.

[9] Colledge, M. J. (1989). Coverage and Classification Maintenance Issues in Economic Surveys. *Panel Surveys.* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh). New York, John Wiley. 80-107.

[10] Kalton, G. and Citro, C. F. (1993). Panel Surveys : Adding the Fourth Dimension. *Survey Methodology.* vol.19. 205-215.

[1998년 4월 접수, 1998년 7월 최종수정]

An Estimation Procedure Using Updated Stratification Sample in Panel Survey *

Young-Won Kim [1] Myung-Shin Oh [2]

ABSTRACT

In panel survey in which the sample is selected by stratified random sampling, if the sampling units shift from a stratum to others in time, then the movement should be incorporated in the estimation procedures. Dealing with the problem caused by the movement of units across stratum in the updated stratification sample, the bias of the conventional estimator neglecting the movement is investigated, and the bias-adjusted estimators are proposed. The variance estimator of the suggested estimators is also derived. It is illustrated via a simulation study that the proposed estimators beat the conventional estimator in the sense of bias and mean squared error. In particular, when the Neyman allocation is applied in stratified sampling, the proposed estimator is shown much more effective to this end.

* This Research was supported by the Sookmyung Women's University Research Grants in 1996.

1) Associate Professor, Department of Statistics, Sookmyung Women's University, Yongsan-ku, Seoul 140-742 Korea

2) Department of Statistics, Sookmyung Women's University, Yongsan-ku, Seoul 140-742, Korea