응용통계연구 (2009)
22(6), 1315–1329

가구 패널조사에서의 가중치 조정에 관한 연구

남궁 평[1] · 변종석[2] · 임찬수[3]

[1]성균관대학교 통계학과, [2]한신대학교 정보통계학과, [3]성균관대학교 통계학과

(2009년 11월 접수, 2009년 12월 채택)

요 약

일반적으로 계속조사는 매 조사시점마다 독립 표본설계로 진행되기도 하지만 대부분 표본을 고정하는 고정 패널 방식이거나 일정 표본을 주기적으로 교체하는 순환 패널(표본)로 설계되어 일정한 측정간격을 유지하면서 자료를 수집한다. 계속조사의 대표적인 조사방식은 고정된 패널을 대상으로 조사하는 패널조사이다. 패널조사는 패널 시작 시점에 추출된 패널을 고정하여 조사를 하므로 조사가 지속됨에 따라 표본의 이탈과 응답 부담으로 인해 무응답이 증가한다. 또한 모집단도 시간이 흐름에 따라 변화하게 되지만 변화되는 모집단의 속성 반영이 쉽지 않게 된다. 본 논문에서는 가구 패널조사에서의 발생되는 표본의 탈락과 추가, 혹은 새로 조사되는 표본을 순환표본설계의 교체방식으로 간주하여 가구 패널조사에서의 횡단면 및 종단면 추정에 효율적인 결과를 제공하는 가중치 부여방법을 제안하고자 한다. 제안된 가중치 조정법의 효율을 비교하기 위해 분산 추정량을 계산하였고, 상대효율이득을 비교하여 가중치 부여방법에 따른 효율을 살펴보았다.

주요용어: 순환표본, 가중값 분배 방법, 횡단면 가중값, 다시점 가중값, 분산추정.

1. 서론

최근 사회 변화, 경제현황 및 변동 등을 장기적이고, 지속적으로 파악하기 위해 다양한 형태의 표본조사가 수행되고 있으며, 대부분의 조사는 일정한 조사시점마다 연속적으로 자료를 수집하는 계속조사로 수행되고 있다. 계속조사에서 주로 관심을 갖는 사항은 특정한 조사시점에서의 수준에 관심을 두는 횡단면 분석(cross-sectional analysis), 측정 시점간 변화 및 몇 시점을 통합한 평균 수준(average level over a number of occasion) 분석에 관심을 둔 종단면 분석(longitudinal analysis) 등이다. 계속조사에서 특정 시점의 수준에 대한 추정 정도를 높이기 위해 매 조사시점마다 독립적인 표본을 추출하여 조사를 한다면 조사 비용이 많이 소요하게 될 것이며, 패널을 구성하여 조사한다면 연속 조사로 인해 응답자의 응답 부담이 가중됨에 따라 무응답이 많이 발생하게 되어 무응답으로 인한 편의가 발생하게 된다. 일반적으로 계속조사에서는 조사가 가능한 표본을 조사 시작시점에서 패널로 구축하는 표본설계를 하게 되는데, 패널에는 고정 패널과 순환 패널로 운영하게 된다. 표본이론의 관점에서 보면, 고정 패널은 최초 조사시점에서 추출된 표본을 대상으로 계속 조사하는 고정 표본설계를 의미하며, 순환 패널은 일정한 간격으로 패널을 교체하는 순환표본설계의 한 방식이다. 실제로 국내에서는 대부분 고정 패널 방식으로 설계하여 계속조사를 수행하지만 일정 시간이 흘러 표본의 탈락이나 마모가 심하거나 모집단 변화를 반영하기 위해 표본을 대체하고 추가하기도 한다.

이러한 문제점을 극복하기 위해 표본을 일정 기간 사용한 후 주기적으로 교체하는 순환표본설계(rotation sample design)방법을 대안 방법으로 활용할 수 있다. 순환표본설계에서는 연속적인 조사 시점

이 논문은 한신대학교 학술연구비 지원에 의하여 연구되었음.
[2]교신저자: (447-791) 경기도 오산시 양산동, 한신대학교 정보통계학과, 교수. E-mail: jsbyun@hs.ac.kr

표 **2.1.** 표본의 운용방식에 따른 추정 및 분석 관점

추정 관점	독립표본조사	패널을 이용한 계속조사	
		고정표본	순환표본
한 시점의 점추정	o	o	o
변수들의 관계	o	o	o
net change	o	o	o
trends	o	o	o
누적합계	x	o	o
gross change	x	o	o

마다 일정한 비율만큼의 표본은 조사를 하지 않거나 표본에서 탈락시키고 새로운 표본을 대체하거나 추가하는 방식으로 표본을 확보하여 조사하게 된다. 순환표본설계에 의한 계속조사는 표본의 응답 부담을 감소시켜 무응답을 줄일 수 있으며, 또한 일정한 표본을 교체함으로써 변화된 새로운 모집단의 특성을 일부 반영할 수 있다는 장점을 가지게 된다.

실제로 고정패널방식을 이용한 계속조사에서는 고정된 표본을 장기간 추적하므로 표본 마모로 인한 표본 교체(substitution) 및 새로운 표본의 추가, 단위 무응답 패턴에 따른 가중치 보정, 항목 무응답에 대한 대체(imputation) 방안, 모집단 변화를 반영한 종단면 가중치 및 분산 추정 등이 연구의 주된 관심사항이 된다.

본 논문에서는 국내에서 많이 실시되고 있는 가구패널조사에서 매 조사시점마다 발생되는 무응답 표본과 이전조사에서의 무응답 표본이 새로 응답하게 되는 표본이나 교체되어 새로 유입되는 표본의 조사형태에 대해 순환표본설계의 표본 교체 방식으로 가정하여 추정을 위한 효율적인 가중치 부여 방법을 검토해 보고자 하였다. 즉, 가구패널조사에서 매 조사시점마다 순환표본설계에서의 가중치 부여 방법을 적용하여 횡단면 분석 및 종단면 분석에 활용 가능한 효율적인 가중치 부여 방법을 제안하고, 이에 따른 적절한 분산 추정 방법을 검토하여 가구패널조사에서의 효율적이고 적절한 가중치 부여 방법을 제안하고자 한다.

2. 패널조사의 무응답과 평균 추정 관점

2.1. 패널을 이용한 계속조사의 무응답 실태

최근 국내에서는 고정된 표본을 선정하여 추적 조사하는 패널에 의한 계속조사가 급증하고 있다. 패널조사는 일회성조사, 독립표본을 이용한 계속조사 등에 비해 조사 자료의 변화나 총량 변화에 대한 원인규명 등 다양한 조사 목적을 달성할 수 있기 때문에 일회성조사나 독립표본의 계속조사에 비해 조사가 어렵지만 그 욕구가 높아지고 있는 것이다. 표 2.1은 일반적인 계속조사에서 표본의 운용방식에 따른 추정 및 분석 관점을 요약해 정리해 본 것이다.

고정 표본에 의한 계속조사가 어려운 점 중 하나는 고정 표본을 장기적으로 관리 및 유지하여 자료를 수집해야하므로 패널(표본)에 조사 부담을 가중한다는 점이다. 아래의 표를 보면, 국내외 주요 패널의 초기에 해당하는 3차년도 조사까지의 표본 유지 규모를 비교하여 나타내어 보았다 (표 2.2).

실제로 패널에 의한 계속조사의 표본 유지율을 보면, 처음 2–3년 동안 응답 거절 등으로 인한 표본의 탈락이 10%내외 씩 발생하고 있으므로 이러한 현상을 반영한 추정의 문제는 표본이론에서는 중요한 사항이 된다. 현재 국내에서는 패널조사에 대한 다양한 연구가 진행되고 있는데, 패널에 의한 계속조사에 대한 연구의 방향으로는 표본교체의 방안, 가중치 보정방법, 무응답 대체 및 종단면 가중치에 대해 연구들

표 **2.2.** 국내외 주요 패널의 표본 유지규모

패널조사명 (시작년도)	KLIPS (한국/1998)	복지패널 (한국/2005)	대우패널 (한국/1993)	PSID (미국/1968)	BHPS (영국/1991)	GSOEP (독일/1984)
2차년도	87.6	92.1	79.0	89.0	87.7	89.9
3차년도	80.9	86.7	66.0	86.3	81.5	86.0

이 각 패널조사마다 사례연구 형태로 다양하게 진행되고 있다. 또한 분산 추정에 대한 연구에서는 횡단면 가중치를 고려한 추정에 대해서는 연구가 일부 진행되고 있지만 종단면 가중치까지 고려한 추정의 문제는 충분히 다루어지지 않고 있다.

2.2. 순환표본설계에서의 평균 차이 추정

순환 패널을 이용한 계속조사는 횡단면 분석 및 종단면 분석이 가능하도록 설계되어 수행되는 조사이다. 고정 패널에 의한 계속조사는 일정한 기간이 지나면서 측정 변화와 더불어 표본추출오차의 증가, 응답부담의 증가, 모집단의 변화, 측정 오차 등의 증가가 발생하게 되므로 이에 대한 개선의 표본설계 방안이 필요하다. 이를 개선하기 위해 일정 기간 사용된 패널이나 마모된 패널을 주기적으로 교체하는 순환표본설계 방식의 순환 패널을 사용하는 방안을 고려할 수 있다. 순환표본설계에 의해 패널을 운영하면 분석에 사용하지 않는 기간에도 추적해야 하는 비용의 발생하거나 횡단면 및 종단면 분석을 위한 가중치 부여의 어려움 등이 존재하지만 모집단의 변화, 응답 부담의 감소 및 종단면 및 횡단면 분석이 가능한 이점이 있기 때문에 패널을 이용한 계속조사에서 충분한 이점이 있는 방법이 될 것이다.

순환표본설계에 의한 계속조사는 조사가 연속적으로 진행됨에 따라 패널 중 오래된 표본을 탈락시키고, 새로운 표본을 패널로 대체하여 추가함으로써 매 시점마다 크기가 n인 패널 표본을 유지하여 조사가 진행된다.

만약 매 측정시점마다 일정한 크기$(1/n)$의 패널을 주기적으로 교체하다고 가정하면 총 조사시점은 중복되는 패널의 수에 의존하게 된다. 즉, 계속조사에서 초기 조사시점을 t라 하고, 이후 조사시점을 $t+1, t+2, \ldots$로 표현하면, 최초 패널을 이용한 최종 완료 조사시점은 $t+n$이 된다.

크기가 n인 패널을 대상으로 매년 일정한 패널을 교체하여 매년 1회씩 조사를 하는 계속조사에서 t시점의 평균을 Y_t라 하고, 연속 또는 불연속적인 두 시점 사이의 평균 차이 $Y_{t,t+j\{j=1,2,\ldots,n\}}$의 추정에 관심을 둔다고 하자. 순환표본설계에 의한 계속조사에서 평균 차이에 대한 추정은 각각의 조사기간 동안 변화하는 모집단 특성을 반영한 횡단면적 접근의 추정과 비교하는 두 시점에 모두 고정되어 조사된 표본을 대상으로 비교하는 두 기간의 평균 차이를 분석하는 종단면 접근의 추정으로 구분하여 분석할 수 있다.

t시점의 모집단을 Ω_t, 표본 평균을 $\bar{Y}_t = \sum_{i\in\Omega_t} Y_i^t/n$, $(t+1)$시점의 모집단을 Ω_{t+1}, 표본 평균을 $\bar{Y}_{t+1} = \sum_{i\in\Omega_{t+1}} Y_i^{t+1}/n$라 하자. 여기서 Y_i^t, Y_i^{t+1}은 t와 $(t+1)$시점의 개별적인 i에 대한 관심대상 척도 변수이다. 먼저, 횡단면적 접근에 의한 연속하는 두 시점의 평균 차이는 다음과 같이 추정한다.

$$\triangle_{t,t+1}^{*} = \bar{Y}_{t+1} - \bar{Y}_t \tag{2.1}$$

그러나 순환표본에 의한 계속조사에서는 조사시점별로 표본의 탈락과 대체에 의한 추가 때문에 두 시점 사이의 평균 추정을 위해 탈락과 유입되는 표본을 제외하고 비교하는 두 시점 모두 조사된 표본을 이용하여 두 시점의 평균 차이를 추정할 수 있다. 이는 비교하는 두 시점의 모집단 특성을 반영하고, 표본의 탈락과 대체 추가에 의한 변동을 제거하여 추정 가능하기 때문에 두 시점 사이의 경향 변화나 순 변화(net change)를 고려한 평균 차이를 살펴볼 수 있게 된다. 비교하는 두 시점 모두 조사된 표본을 대상

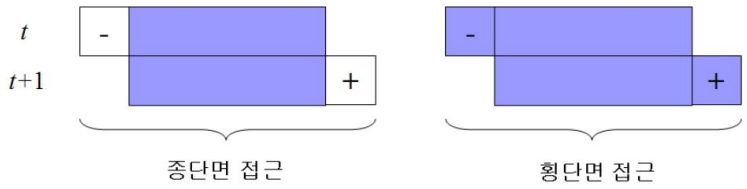


그림 **2.1.** 순환표본설계에 의한 계속조사에서 종단면 접근과 횡단면 접근의 개념

으로 평균 차이를 분석하는 종단면 접근의 평균 차이는 다음과 같이 추정한다.

$$\triangle_{t,t+1} = \sum_{i \in \Omega_{t,t+1}} \frac{Y_i^{t+1} - Y_i^t}{n_{\Omega_{t,t+1}}}, \tag{2.2}$$

여기서 $\Omega_{t,t+1}$는 조사시점 t와 $(t+1)$시점의 공통 모집단으로 $\Omega_{t,t+1} = \Omega_t \cap \Omega_{t+1}$이다.

순환표본설계에 의한 계속조사에서 두 시점의 평균 차이를 추정하는 개념을 알기 쉽게 이해하도록 그림으로 표현하면 그림 2.1과 같다. 그림 2.1의 좌측은 종단면 접근의 경우로서 t시점과 $(t+1)$시점에서 공통되게 조사된 가구의 차이추정이며, 우측은 횡단면 접근의 경우로서 t시점에서 조사된 모든 가구와 $(t+1)$시점에 조사된 모든 가구의 차이를 추정하는 것이다. 그림에서 $(-)$는 $(t+1)$시점에 표본에서 유출 및 탈락을 의미하고, $(+)$는 표본으로 대체 추가 및 유입을 의미한다.

3. 가중값 부여방법과 분산추정

본 논문에서는 최종 표본추출 단위를 가구로 하는 가구 대상 계속조사에서 적용되는 가중값 부여 방법과 분산추정에 대해 논의하고자 한다. 이 때, 표본으로 추출된 가구의 가구원은 모두 조사하는 것을 원칙으로 한다.

3.1. 가중값 부여 방법

1) 설계가중치

일반적으로 표본설계에서 기본적으로 사용하는 가중치로 표본설계 시점을 기반으로한 가중치 부여방법을 의미한다.

행정구역이나 지역의 규모 등을 층화변수로 하여 조사구를 추출하고, 추출된 조사구에서 표본가구를 추출하는 층화추출에 의한 표본설계를 가정하는 경우, 설계 가중치는 다음과 같다.

h: 층을 나타내는 첨자 $(h = 1, 2, \ldots, H)$

i: h번째 층의 i번째 표본 조사구를 나타내는 첨자 $(i = 1, 2, \ldots, n_h)$

j: h층의 i번째 조사구내 j번째 표본가구를 나타내는 첨자 $(j = 1, 2, \ldots, m_{hi})$

N_h: h층의 전체 조사구수

n_h: h층의 표본 조사구수

M_{hi}: h층의 i번째 조사구의 전체 모집단 가구 수

m_{hi}: h층의 i번째 조사구의 표본 가구 수

w_{hij}: h층의 i조사구 j번째 표본 가구에 대한 가중치

$$w_{hij} = \frac{N_h}{n_h} \cdot \frac{M_{hi}}{m_{hi}} \tag{3.1}$$

표본으로 추출될 때 부여된 설계 가중치는 표본가구가 조사되는 동안 계속 동일한 가중치를 가진다.

2) 균등가중치

균등 가중치는 현재조사시점에서 표본가구의 가중치는 이전시점의 표본가구 가중치로 사용하는 가중치이다 (강석훈, 2003). 우선 조사 첫 시점에 대하여 불균등 선택확률가중치∗무응답 조정을 위한 가중∗사후층화 가중치를 계산하여 가구 가중치를 구한다. 균등 가구가중치는 2차 조사시점에서부터 사용하는 가중치로 균등가구 가중치는 다음과 같다.

$$w_{hij}^{(t+1)} = \sum_{h}^{H} \sum_{i}^{n_h} \sum_{j}^{m_{hi}} \frac{w_{hij}^{(t)}}{J} \tag{3.2}$$

균등가중치는 현 시점$(t+1)$에서 조사된 표본가구의 가중치를 이전시점(t)에 부여되었던 가중치합을 현 시점에서 조사된 표본가구수로 나누어 균등하게 재분배하는 가중치를 의미한다. 예를 들어, 현 시점에서 10가구가 조사되었고 10가구가 무응답되었다면 조사된 10가구의 가중치는 이전 시점의 가중치합을 구하여 10으로 나누어 균등하게 가중치를 배분하게 된다.

3) Duncan 방법

Duncan방법은 최초 시점에서는 가구 가중치 및 가구원(개인) 가중치를 사용하지만 2차 시점이후 가구원가중치를 이용하여 가구가중치로 부여하는 방법으로 현재 한국노동패널(KLIPS)에서 현재 사용되고 있는 방법이다 (강석훈, 2003). Duncan방법의 특징은 사용하는 가정이 가장 약하고 또한, 계산과정이 간단하여 적용이 쉽다는 점이다. Duncan의 가중치 부여 방법에 대한 과정은 다음과 같은 절차로 부여한다 (Duncan, 2003).

① 조사 첫 시점에서 불균등 선택확률가중치∗무응답 조정을 위한 가중치∗사후층화 가중치를 계산하여 가구 가중치를 구한다.

② 조사 첫 시점에서 구한 가구 가중치를 그 가구에 속한 사람에 그대로 적용하여 개인 가중치로 사용한다.

③ 두 번째 조사시점에서의 개인가중치는 첫 번째 시점의 개인 가중치를 가구원의 응답률로 조정하여 사용한다.

④ 두 번째 조사시점에서 계산된 가구 내 구성원의 개인가중치의 평균을 두 번째 가구가중치로 사용한다.

⑤ 세 번째 이후의 가중치는 앞의 과정을 반복하여 구한다.

4) 가중치 분배 방법을 이용한 가중치 부여 방법

실제적으로 순환표본설계 (Ernst, 1989)에서 계속조사를 위한 가중치 부여 시 근본적으로 가장 어려운 점은 각각 서로 다른 시점에서 서로 다른 모집단으로부터 선택된 $(n-1)$개의 패널 부표본을 이용하여 조사시점 t의 모집단 Ω_t로 나타내는 것이다. 만약 $a_{t,k}$를 t조사시점에 k번째 선택된 부표본을 나타내고 $s_{t,t+1} = \bigcup_{k=1}^{n-1} a_{t,k}$이라고 하자. 고려할 점은 $a_{t+1,k+1} = a_{t,k}(\forall t, \forall k \neq n)$라는 것인데 이는 전체의

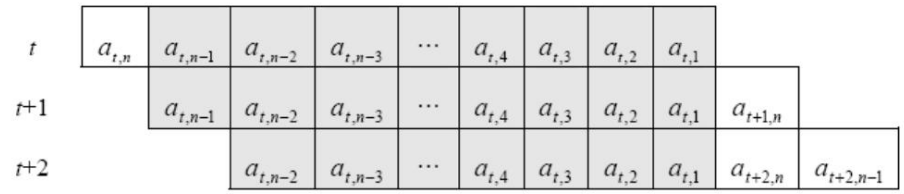


그림 3.1. 종단면 접근을 통한 표본프레임

조사시기마다 표본(패널)의 이탈이 없는 각각의 패널 부표본을 사용하여야 한다는 것이다 (Merkouris, 2001).

그림 3.1에서 회색부분은 $s_{t,t+1}$을 나타내며, 종단면 분석에 이용하는 표본을 의미한다. $s_{t,t+1}$내에 존재하는 각각의 개별적인 자료는 Y_i^t와 Y_i^{t+1}양쪽에서 얻게 되는데, 다른 의미로 각각 t, $(t+1)$시점에서 개별적인 가구대상 j에 대한 정보를 의미한다.

만약 $j \in a_{t,k}$라면, $W_j(t,k)$는 개별조사 대상의 기본 표본 가중치이다. 사실 $W_j(t,k)$는 패널 요소로서 선택되어진 개별조사 가구 j에 대한 조사시점에서 가구표본 가중치이다. $\sum_{j\in a_{t,k}} W_j(t,k)$는 모집단 Ω_{t-k+1}의 조사대상들에 대한 불편추정량을 제공한다고 알려져 있으며 (Lavallee, 1995), 종단면 가중치는 $s_{t,t+1}$내의 개별적인 조사개체 j에 배분되고 그 결과는 식 (3.3)과 같다 (Lavallee와 Ardilly, 2007).

$$w_j^{t,t+1} = \frac{1}{L_j} \sum_{k\in k_i} w_j(t,k), \tag{3.3}$$

여기서, $w_j^{t,t+1}$은 t, $(t+1)$시점에서 j번째 가구의 가중치를 의미하며, L_j는 t시점과 $(t+1)$시점의 공통되는 부표본의 수를 의미한다.

식 (3.3)은 원 종단면 가중치에 대한 가장 일반적인 공식으로 패널이 중복되는 경우가 없다고 가정하면 식 (3.4)와 같이 정의할 수 있다.

$$w_j^{t,t+1} = \frac{W_j}{L_j}. \tag{3.4}$$

식 (3.4)와 같이 계산된 종단면 가중치를 두 시점에서의 평균 차이에 대한 $\triangle_{t,t+1}$ 추정량의 추정에 사용하면 식 (3.5)와 같이 이용하게 된다.

$$\hat{\triangle}_{t,t+1} = \frac{\sum\limits_{s_{t,t+1}} W_j^{t,t+1}\left(Y_j^{t+1} - Y_j^t\right)}{\sum\limits_{s_{t,t+1}} W_j^{t,t+1}} \tag{3.5}$$

또한 횡단면적 관점에서 서로 다른 시점의 평균 차이 $\triangle_{t,t+1}^* = \bar{Y}_{t+1} - \bar{Y}_t$의 추정은 조사가 실제 안정적으로 진행된다는 가정 하에 식 (3.6)과 같이 횡단면 가중치를 계산하면 된다.

$$w_j^{t(c)} = \frac{\sum\limits_{k=1}^{n} \sum\limits_{j\in a_{t,k}} w_j(t,k)}{\sum\limits_{k=1}^{n} \sum\limits_{j\in\Omega_{t-k+1}} 1}, \tag{3.6}$$

여기서, $w_j(t,k)$은 $a_{t,k}$에서의 표본 가중치이다. 식 (3.6)을 계속조사에 적용한다면, 조사시점의 변화에 따라 분모부분을 바꿔가며 계산을 하면 된다. 분모부분은 실제로 조사 대상자수로 시점이 흘러감에 이

전에 조사된 가구와 새롭게 조사대상이 되는 가구의 합이 된다. 예를 들어, $t+2$시점의 횡단면 가중치는 식 (3.7)과 같다.

$$w_j^{t+2(c)} = \frac{\sum_{k=1}^{n} \sum_{j \in a_{t,k}} w_j(t,k)}{\left(\sum_{j \in \Omega_{t+2}} 1\right) + \left(\sum_{j \in \Omega_{t+1}} 1\right) + (n-2)\left(\sum_{j \in \Omega_t} 1\right)} \tag{3.7}$$

이를 이용하여 횡단면 관점에서의 $\triangle^*_{t,t+1} = \bar{Y}_{t+1} - \bar{Y}_t$의 추정값은 가중값 $w_J^{t(c)}$와 $w_J^{t+1(c)}$을 사용하여 식 (3.8)을 계산한다.

$$\triangle^*_{t,t+1} = \frac{\sum_{i \in \tilde{u}_{t+1}} w_i^{t+1(c)} \cdot Y_i^{t+1}}{\sum_{i \in \tilde{u}_{t+1}} w_i^{t+1(c)}} - \frac{\sum_{i \in \tilde{u}_t} w_i^{t(c)} \cdot Y_i^{t}}{\sum_{i \in \tilde{u}_t} w_i^{t(c)}}. \tag{3.8}$$

3.2. 분산추정

일반적으로 복합표본설계나 순환표본설계와 같은 표본조사에서의 분산 추정은 계산과정이 매우 복잡하다. 본 논문에서는 3.1절에서 검토한 가중치를 이용하여 횡단면 및 종단면 분석의 관점에서 두 시점의 평균 차이를 추정하여 가중치 부여 방법의 결과를 비교하고자 한다. 대부분의 복합표본설계에서는 추정치의 분산 추정을 위해 붓스트랩(Bootstrap)방법, 잭나이프방법 등을 이용하여 추정량의 분산 추정량을 계산하는 데, 붓스트랩방법은 복원추출을 이용하여 추정하고, 잭나이프방법은 표본을 하나씩 제외한 나머지 표본들로 이루어지는 표본집단을 사용하여 추정량 및 추정량에 대한 분산을 추정하게 된다. 본 논문에서는 Shao와 Tu (1995)가 소개한 delete-1 jackknife를 이용하여 분산추정량을 계산하였다.

1) 종단면 분석을 위한 잭나이프방법

잭나이프방법은 붓스트랩 방법의 특수한 형태로 표본 중 하나씩을 제거하여 새롭게 얻을 수 있는 표본의 모든 가능한 표본으로 사용하는 것이다. 일반적으로 잭나이프 방법을 통한 분산추정량은 식 (3.9)와 같다.

$$\hat{V}_{JK}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} \left(\hat{\theta}_{(j)} - \hat{\theta}\right)^2, \tag{3.9}$$

여기서 j는 제거되는 관측치를 의미하며 또한 n은 실제표본의 수를 의미한다. 모평균을 추정하는 경우, $\hat{\theta} = \bar{x}$로 실제 관측된 값들의 평균을 의미하며, $\hat{\theta}_{(j)} = \bar{x}_{(j)} = \sum_{i \neq j} x_i/(n-1)$로 제거된 j번째 관측치를 제외한 나머지 값들의 평균을 의미한다. 하지만 가중치가 존재하는 경우에는 잭나이프 방법을 이용하기 위해서는 가중치를 조정해야 한다. 잭나이프 분산추정량의 가중치 조정방법은 식 (3.10)과 같다.

$$w_{i(j)} = \begin{cases} 0, & \text{if } j^{th} \text{ observation,} \\ \dfrac{n}{n-1} w_i, & \text{if no } j^{th} \text{ observation.} \end{cases} \tag{3.10}$$

식 (3.10)은 j번째 관측치가 삭제된 경우 그 삭제되는 가중치는 0이고 삭제되지 않는 관측치들은 기존 가중값들에 $n/(n-1)$을 곱하여 가중치를 조정하여 이용하게 된다. 이를 통한 잭나이프 분산추정량은 다음과 같다.

$$\hat{V}_{JK}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} \left(\hat{\theta}_{i(j)} - \hat{\theta}\right)^2 \tag{3.11}$$

2) 횡단면 분석을 위한 잭나이프방법

종단면 분석을 위한 잭나이프 방법은 횡단면 분석에 적용하면 그 사용에 어려움이 존재하기 때문에 Roberts 등 (2001)이 제시한 횡단면 분석을 위한 잭나이프 방법을 이용하여 분산 추정량을 계산하기로 한다.

Roberts 등 (2001)은 서로 다른 두 집단의 차이 추정에 대한 분산추정량을 식 (3.12)와 같이 제안하였다.

$$\mathrm{Var}\left(\hat{\theta}_1 - \hat{\theta}_2\right) = \mathrm{Var}\left(\hat{\theta}_1\right) + \mathrm{Var}\left(\hat{\theta}_2\right) - 2\mathrm{Cov}\left(\hat{\theta}_1, \hat{\theta}_2\right). \tag{3.12}$$

식 (3.12)를 이용하여 계속조사에서 횡단면 분석에 적용 가능한 가중치로 조정하면 식 (3.13)과 같다.

$$w_{1i(j)} = \begin{cases} w_{1i(j)}, & i \in s_1,\ i \in s_2, \\ w_{1i}, & i \in s_1,\ i \notin s_2, \\ 0, & i \notin s_1, \end{cases} \qquad w_{2i(j)} = \begin{cases} \dfrac{w_{2i(j)} w_{1i(j)}}{w_{1i}}, & i \in s_1,\ i \in s_2, \\ w_{2i} & i \notin s_1,\ i \in s_2, \\ 0 & i \notin s_2. \end{cases} \tag{3.13}$$

식 (3.13)의 조정된 가중치를 이용하여 공분산을 계산하면 식 (3.14)와 같다.

$$\mathrm{Cov}_{JK}\left(\hat{\theta}_1, \hat{\theta}_2\right) = \frac{n-1}{n} \sum_{JK} \left(\hat{\theta}_{1(j)} - \hat{\theta}_1\right)\left(\hat{\theta}_{2(j)} - \hat{\theta}_2\right). \tag{3.14}$$

4. 사례연구

현재 국내에서는 패널을 이용한 계속조사가 횡단면 및 종단면 분석의 목적을 달성하기 위해 급증하고 있다. 하지만 고정 패널을 이용한 계속조사는 일정 시간이 흐른 뒤 표본의 마모 및 이탈, 응답자의 응답 부담 증가 등으로 무응답이 증가하거나 변화되는 모집단의 반영이 미비하여 포함오차 및 표본추출오차 등으로 인한 문제점이 나타나고 있다. 그 결과, 횡단면 및 종단면 분석을 위한 가중치 작업 및 과정도 복잡하고 어려워지고 있다. 이에 본 논문에서는 가구패널조사에서 매 시점 발생되는 무응답 표본과 새로 응답하는 표본에 대한 가중치로 일정 주기마다 표본을 교체하는 순환표본설계에서의 가중치 부여 방법을 이용하여 가구패널조사에서 적용 가능한 가중치 부여 방법에 대해 살펴보았다.

본 절에서는 국내 패널조사 중 한국노동패널을 이용하였으며, 조사 시점마다 조사에 불응하거나 추적 실패 및 응답 거절 등으로 표본 자료의 획득이 불가능한 상황이 발생하게 된다. 반면, 이전에 조사가 이루어지지 않았지만 조사시점에서 추적에 성공하거나 조사 협조를 통해 새로 자료가 획득되는 상황이 매번 발생하고 있다. 이에 본 논문에서는 매 조사시점마다 무응답으로 인한 표본 가구 및 가구원을 탈락으로, 다시 조사되는 가구나 가구원을 표본의 대체 및 유입으로 생각하고, 이를 매년 일정 표본이 교체된다고 보는 순환표본설계로 가정하여 횡단면 및 종단면 추정을 위해 제안된 가중치의 부여방법을 검토해 보았다.

한국노동패널조사는 한국노동연구원에서 도시지역에 거주하는 국내외 5,000가구와 가구원을 대표하는 구성원을 대상으로 1년에 1회씩 실시하는 조사로서 가구용 자료와 가구에 속한 15세 이상의 가구원을 대상으로 한 개인용 자료로 구별하고 있다. 본 논문에서는 1998년의 한국노동패널 1차 조사부터 2006년도 한국노동패널 9차 조사까지의 데이터를 이용하여 가구 소득 문항을 중심으로 제안된 가중치 부여방법을 이용하여 소득을 추정해 보았다. 한국노동패널조사는 고정패널을 이용하여 조사하는 계속조사로서 각 지역별로 조사구를 추출하고 추출된 조사구내에서 가구를 추출하는 패널조사이다. 이에 본 실증연구에서는 서울지역만을 선택하여 서울지역의 가구 소득변화만 살펴보았다. 그리고 서울지역의

표 **4.1.** 개별연도 추정량 및 분산추정량

	Duncan Weight Method	Equal Weight Method	Design Weight Method	Share Weight Method
	분산추정량	분산추정량	분산추정량	분산추정량
1998년	379.348	379.348	379.348	379.348
1999년	192.133	184.187	192.954	196.997
2000년	201.677	208.549	200.879	197.258
2001년	312.341	319.902	310.552	303.260
2002년	460.468	506.370	462.717	471.833
2003년	502.303	540.119	502.667	505.236
2004년	520.468	560.824	515.703	456.656
2005년	811.026	893.612	791.213	609.255
2006년	1057.576	1073.457	1030.619	785.715

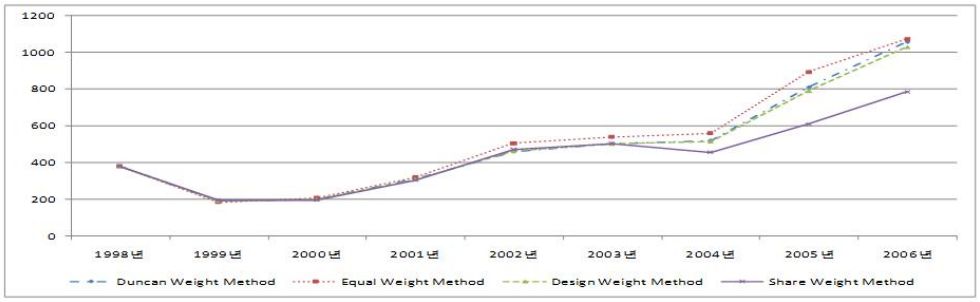


그림 **4.1.** 연도별 소득의 분산추정량

자료에서 고정된 패널이 아니라 순환표본으로 구성되어야 하기에 9차년도 조사에서 각 조사시점마다 유지되는 패널과 이탈 혹은 유입되는 패널을 포함하여 순환표본설계 형태로 가정하여 적용해 보았다. 한국노동패널조사 자료 중 근로소득, 금융소득, 부동산소득, 사회보험, 이전소득, 기타소득 등 소득 관련 자료를 결합하여 가구소득으로 계산하였다.

1) 연도별 가구소득

1차년도(1998년)에서부터 9차년도(2006년)까지의 한국노동패널조사 자료에 대해 제안된 가중치를 부여하여 연도별 소득을 추정하고, 잭나이프방법을 이용하여 분산추정량을 계산하였다. 표 4.1과 그림 4.1에서 보듯이 제안된 4가지 가중치 부여 방법 중 가중치 분배 방법이 다른 방법들에 비해 분산추정량이 적었으며, 균등 가중치 방법의 분산추정량이 큰 것으로 나타났다.

2) 횡단면 관점의 연도별 가구 소득 차이의 추정

횡단면 관점의 접근에서 연속한 두 시점의 연도별 소득의 차이 추정은 두 개의 상이한 조사시점별로 각각 추정을 하고, 추정된 개별 추정량의 차이를 두 시점의 차이에 대한 추정량으로 이용하는 것이다. 이에 한국노동패널조사에서 1차년도(1998년)에서부터 9차년도(2006년)자료를 1998년과 1999년, 1999년과 2000년처럼 인접한 조사시점별로 횡단면 가구 평균소득차이의 분산추정량을 계산하였다. 그 결과는 표 4.2, 그림 4.2와 같다. 각 방법들의 분산추정량을 살펴보면 가중치 분배 방법이 가장 크고, 다른 방법

표 **4.2.** 횡단면 관점의 시점별 소득 차이에 대한 횡단면 추정량 및 분산추정량

연도	Duncan Weight Method	Equal Weight Method	Design Weight Method	Share Weight Method
	분산추정량	분산추정량	분산추정량	분산추정량
1998–1999년	142.870	144.086	143.075	140.884
1999–2000년	98.452	98.560	498.458	98.184
2000–2001년	128.504	131.715	127.858	132.113
2001–2002년	193.202	203.972	193.317	206.568
2002–2003년	240.693	257.123	241.346	261.622
2003–2004년	255.693	253.129	254.593	275.236
2004–2005년	322.874	333.097	326.729	363.609
2005–2006년	467.151	435.928	455.458	491.767

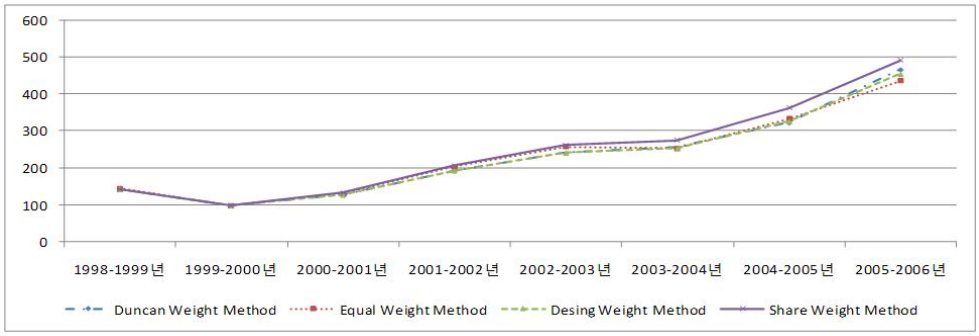


그림 **4.2.** 연도별 소득 차이 추정에 대한 횡단면 분산추정량

들은 대체적으로 비슷하게 나타났다.

3) 종단면 관점의 가구 소득 차이의 추정

종단면 관점의 접근은 두 개의 상이한 조사 시점에서 공통으로 조사된 대상만을 이용하여 두 조사시점의 차이를 추정하는 것이다. 이에 한국노동패널조사 자료에서 1998년과 1999년처럼 인접한 두 시점의 종단면 관점의 가구 소득 차이의 분산추정량을 계산하였고 그 결과는 표 4.3, 그림 4.3과 같다.

균등가중치 방법의 분산추정량이 전반적으로 큰 것으로 나타났고, 가중치 분배방법이 전체적으로 낮은 것으로 나타났다.

4) 효율 비교

우리는 한국노동패널조사 자료를 이용하여 순환표본설계에 의한 계속조사에서의 가중치 부여 방법으로 소득의 추정량과 분산추정량을 계산해 보았다. 가중치 부여 방법에 따른 추정량과 분산추정량을 이용하여 가중치 부여 방법에 대한 상대효율(relative efficiency)이득 (이계오 등, 2001)을 계산하여 비교해 보았다. 상대효율은 각 추정에 대한 상대표준오차(C.V)를 이용하여 식 (4.1)과 같이 계산하였다.

$$\frac{\text{설계 가중치 추정량의 C.V} - \text{비교 가중치 추정량의 C.V}}{\text{설계 가중치 추정량의 C.V}} \tag{4.1}$$

표 **4.3.** 종단면 관점의 시점별 소득 차이에 대한 종단면 분산추정량

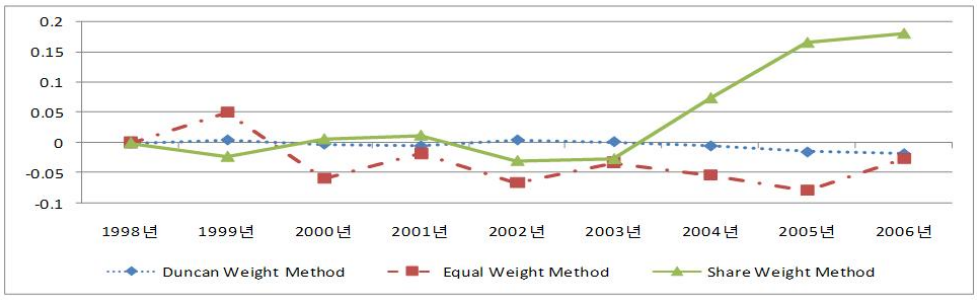
연도	Duncan Weight Method 분산추정량	Equal Weight Method 분산추정량	Design Weight Method 분산추정량	Share Weight Method 분산추정량
1998–1999년	442.868	442.868	442.868	442.868
1999–2000년	147.533	178.091	147.621	140.267
2000–2001년	213.195	214.463	210.506	193.331
2001–2002년	528.312	549.387	532.023	519.349
2002–2003년	645.802	726.577	649.716	632.724
2003–2004년	619.070	678.081	617.216	595.198
2004–2005년	486.349	540.552	475.604	364.306
2005–2006년	334.126	364.030	328.056	277.605

그림 4.3. 종단면 관점의 시점별 소득 차이에 대한 종단면 분산추정량

표 **4.4.** 설계 가중치 대비 가중치 부여 방법들의 상대효율

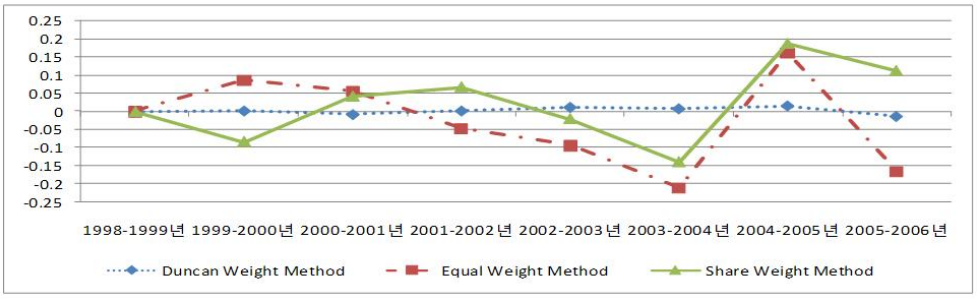
연도별 소득 추정의 상대효율				횡단면 관점의 소득 차이 추정에 대한 상대효율				종단면 관점의 소득 차이 추정에 대한 상대효율		
연도	Duncan Weight Method	Equal Weight Method	Share Weight Method	연도	Duncan Weight Method	Equal Weight Method	Share Weight Method	Duncan Weight Method	Equal Weight Method	Share Weight Method
1998년	0	0	0	98–99년	−0.009	0.060	−0.208	0	0	0
1999년	0.005	0.050	−0.023	99–00년	0.038	−0.381	0.170	0.001	0.087	−0.083
2000년	−0.002	−0.059	0.006	00–01년	−0.007	−0.043	0.086	−0.009	0.056	0.042
2001년	−0.005	−0.018	0.012	01–02년	−0.006	−0.047	0.081	0.001	−0.047	0.068
2002년	0.005	−0.066	−0.030	02–03년	0.011	−0.222	0.071	0.011	−0.094	−0.020
2003년	0.002	−0.033	−0.027	03–04년	0.022	−0.235	−0.111	0.007	−0.211	−0.139
2004년	−0.005	−0.054	0.074	04–05년	−0.622	0.713	0.128	0.015	0.162	0.188
2005년	−0.015	−0.079	0.166	05–06년	−0.021	0.002	−0.198	−0.014	−0.166	0.113
2006년	−0.018	−0.026	0.181							
평균	−0.004	−0.032	0.040		−0.074	−0.019	0.003	0.002	−0.027	0.021

식 (4.1)을 이용하여 개별연도별, 횡단면 및 종단면 관점의 소득 차이에 대한 추정치를 이용하여 구한 상대효율은 표 4.4, 그림 4.4와 같다.



(a) 연도별 소득 추정에 대한 상대효율

(b) 횡단면 관점의 소득 차이 추정에 대한 상대효율



(c) 종단면 관점의 소득 차이 추정에 대한 상대효율

그림 4.4. 설계가중값 대비 가중값 부여 방법들의 상대효율이득

식 (4.1)의 상대효율은 설계 가중치 보다 상대표준오차가 더 작다면 효율적이므로 (+)의 값이 클수록 가중치의 상대효율이 높음을 의미하게 된다. 표 4.4에 의하면 개별연도별 소득 추정 결과와 횡단면 및 종단면 관점에서 연속한 두 시점의 소득 차이에 대한 추정 결과의 상대효율을 비교해 보면, 1) 연도별 소득 추정의 경우 2003년 이전까지 상대효율 차이는 크지 않았지만 2004년 이후 가중치 분배 방법에 의한 가중치의 상대효율이 매우 크게 나타났으며, 2) 횡단면 관점의 소득 차이 추정에 대한 상대효율도 연도별 소득 추정에 근거하여 추정되므로 연도별 소득 추정과 비슷한 추세를 보이나 2004–2005년에 균등가중치의 상대효율이 높은 것으로 나타났다. 이는 표본가구의 가구원 수에 많은 영향을 받은 것으로 보인다. 3) 비교하는 두 시점 모두 측정된 자료만을 이용하여 추정하는 종단면 관점의 소득 차이 추정에 대한 상대효율은 2001–2002년 이후 가중치 분배 방법에 의한 가중치의 상대효율이 높게 나타나고 있다.

4) Duncan의 방법은 가구원 가중치를 이용하여 가구가중치를 사용하기에 설계 가중치와 큰 차이를 보이지 않았지만 바로 이전 시점의 가중치를 이용하는 균등 가중치나 순환표본설계를 위한 가중치 분배 방법에 의한 가중치는 조사 초기에는 상대효율이 약간 높았으나 특정 시점(2004년)이후에는 상대효율이 특히 높아지고 있음을 볼 수 있다.

이 결과로부터 본 논문에서 검토한 기간 동안 연도별 추정, 횡단면 및 종단면 관점의 추정에 전반적으로 모든 추정에서 효율적인 가중치는 존재하지 않지만 대체로 가중치 분배 방법을 이용한 가중치의 상대효율이 가장 높은 것으로 나타났다. 특히 패널조사의 초기보다 2004년 이후 변동이 크게 나타난 것은 고정 패널로 유지된 한국노동패널조사의 패널이 시간이 흐름에 따라 초기 패널의 유지에 큰 어려움이 있다는 것을 의미하는 것이다. 실제로 한국노동패널조사의 초기에는 표본의 탈락 등의 표본 변동이 작아 중복되는 표본의 비율이 높기 때문에 상대효율이 크지 않지만 시간이 흐름에 따라 표본의 탈락으로 인한 마모가 증가하게 되어 설계 가중치를 이용하는 방법보다 다른 3가지의 가중치의 상대 효율이 더 크게 나타난 것으로 보인다. 이 결과는 순환표본설계에 의한 계속조사에서 분배 가중치에 의한 방법의 상대효율이 높다는 이전 결과와도 어느 정도 일치한다고 생각한다.

5. 결론 및 향후과제

최근 증가하는 패널조사에서는 고정된 표본을 대상으로 조사하므로 표본의 탈락과 응답 부담으로 인한 무응답이 증가하여 편향의 원인이 되고 있다. 본 논문에서는 고정패널에 의한 계속조사에서 표본의 탈락으로 추가된 표본에 대한 가중치 부여 방법으로 순환표본설계에서의 가중치 부여 방법을 이용하여 횡단면 및 종단면 추정의 관점에 따른 효율적인 가중치 부여 방법을 제안해 보았다.

일반적으로 널리 사용하는 설계 가중치, 비교하는 두 시점의 분석을 위한 균등 가중치, 종단면 분석을 위한 Duncan 가중치 및 순환표본설계에 적용하는 분배 가중치 등을 제안하여 패널조사에 의한 계속조사에서의 적절한 가중치를 탐색하기 위해 한국노동패널조사 자료를 이용하여 모의실험으로 소득과 소득 차이에 대한 추정량과 분산 추정량을 구하여 상대표준오차를 이용하여 상대효율이득으로 비교해 보았다. 계속조사에서 횡단면 관점의 연도별 추정이나 소득 차이에 대한 추정은 표본의 유지율이 높게 유지될 때 가중치의 상대효율은 큰 차이가 없었으나 표본의 탈락이나 마모 등으로 무응답이 증가하는 경우에는 설계 가중치에 비해 순환표본설계를 위한 분배 가중치, 비교시점의 가중치를 사용하는 균등 가중치의 상대 효율이 높게 나타났다. 종단면 관점의 소득 차이 추정에서도 비슷한 경향을 보였으나 전반적으로 분배 가중치나 균등 가중치의 상대 효율이 높았으며, 특히 전반적으로 분배 가중치가 다른 방법들에 비하여 효율이 높다는 것을 알 수 있었다. 따라서 패널조사에서 매 조사시점마다 무응답 표본을 탈락으로, 새로 조사되는 표본을 추가되는 표본으로 가정하여 순환표본설계의 분배 가중치를 적용하여 본 결과 상대효율이 높아짐을 알 수 있었다.

향후 패널조사에서는 고정된 패널을 장기적으로 운영하는 방안과 일정 주기마다 일부를 교체하는 순환표본설계로 운영하는 방안의 비교 연구, 무응답 자료에 대한 대체 방법 및 본 논문에서 검토한 종단면 가중치와 이를 반영한 분산 추정 등에 대한 다양한 연구가 필요하다고 생각한다.

참고문헌

강석훈 (2003). KLIPS 가중치 부여방안 연구, <한국노동패널연구>, 2003–2004.

이계오, 류제복, 김영원, 김영근 (2001). <소지역 실업률 추정기법 및 전산프로그램 개발>, 공군사관학교, 항공우주연구소.

Duncan, G. (2003). A simple method for weighting in Household panel survey, woking paper, Nothwestern University.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates, *Research Reports of U.S Census Bureau*, **23**.

Lavallee, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight Share method, *Survey Methodology*, **21**, 25–32.

Lavallee, P. and Ardilly, P. (2007). Weighting in rotating samples: The SILC survey in France, *Survey Methodology*, **33**, 131–137.

Merkouris, T. (2001). Cross-sectional estimation in multiple-panel house surveys, *Survey Methodology*, **27**, 171–181.

Roberts, G., Kovacevi, M., Mantel, H. and Phillips, O. (2001). Cross-sectional inference based on longitudinal surveys: Some experiences with statistics Canada surveys, *Federal Committee for Statistical Method Conference*, Washington.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer-Verlag, New York.

A Study on the Weight Adjustment Method for Household Panel Survey

Pyong Namkung[1] · Jong-Seok Byun[2] · Chan-Soo Lim[3]

[1]Department of Statistics, Sungkyunkwan University
[2]Department of Statistics and Information, Hanshin University
[3]Department of Statistics, Sungkyunkwan University

(Received November 2009; accepted December 2009)

Abstract

The panel survey is need to have a more concern about a response due to a secession and non-response of a sample. And generally a population is not fixed and continuously changed. Thus, the rotation sample design can be used by the method replacing the panel research. This paper is the study of comparison to equal weight method, Duncan weight, Design weight method, weight share method in rotation sample design. More specifically, this paper compared variance estimators about the existing each method for the efficiency comparison, and to compare the precision using the relative efficiency gain by the Coefficient of Variance(CV) after getting the design weight from the actual data.

Keywords: Rotation sample design, shared weight method, longitudinal weight, cross-sectional weight.

This work was supported by Hanshin University research grant.
[2]Corresponding author: Professor, Department of Statistics and Information, Hanshin University, Osan, Gyeonggi-Do 447-791, Korea. E-mail: jsbyun@hs.ac.kr