응용통계연구 제 14권 1호, 2001년, pp. 81-90

P-수준교체표본에서 교체그룹내 상관관계를 고려한 일반화 복합추정량

박유성[1] 배경화[2] 김기환[3]

요 약

모집단의 변화를 효과적으로 추정하기 위한 반복조사 방법으로 교체표본조사를 고려할 수 있다. 교체표본조사는 크게 일수준교체표본조사와 다수준교체표본조사로 나누어지며 모집단의 특성, 특성의 변화를 추정하기 위하여 복합추정량을 사용하고있다. 본 논문에서는 다수준교체표본조사의 경우 교체그룹내 표본개체들의 상관관계를 고려한 일반화복합추정량과 추정량의 분산을 최소화시키는 최적계수를 제시하였다. 또한 수치예에서는 교체그룹내 표본개체의 수, 교체그룹내 상관정도의 변화에 따라 표본개체들의 상관관계를 고려한 일반화 복합추정량의 효율성을 제시하였다.

주요용어: 다수준교체표본조사, 일반화 복합추정량, 급내상관.

1. 서 론

시간에 따라 변화하는 모집단의 특성을 조사하기 위한 방법인 반복조사(repeated survey)는 동일한 모집단으로부터 표본을 반복추출하는 것으로 고정표본조사(fixed sample survey), 독립표본조사(independent sample survey), 교체표본조사(rotation sample survey)가 있다. 고정표본조사는 동일한 표본을 계속적으로 조사하는 것으로 표본 간에 발생하는 상관관계를 이용하여 표본오차를 줄일 수 있고 시계열을 유지할 수 있다는 장점이 있다. 반면 조사대상자가 같은 질문과 절차에 익숙해지고 응답에 대한 부담감을 갖게되어 비표본오차가 발생하므로 표본효율이 떨어진다는 점과 모집단에 대한 표본의 대표성이 떨어진다는 단점이 있다. 독립표본조사는 각 조사시점마다 새로운 표본을 조사하는 것으로 모집단에 대한 대표성이 확보된다는 장점이 있으나 표본추출비용이 많이 들고 수집된 자료에 의한 시계열 구성 및 유지가 어렵다는 단점이 있다. 교체표본조사는 고정표본조사와 독립표본조사의 장점을 결합한 형태로 t시점의 표본을 구성할때 $t-1$시점의 일부 표본개체(old elements)를 없애고 새로운 표본개체(new elements)로 교체하는 방법이다. 교체표본조사는 보고하는 형태에 따라 일수준교체표본조사(one-level rotation sample survey)와 다수준교체표본 조사(multi-level rotation sample survey)로 나눌 수 있다. 다수준교체표본조사는 t시점

1) (130-070)서울특별시 성북구 안암동 5-1, 고려대학교 통계학과, 교수
E-mail: yspark@mail.korea.ac.kr
2) (130-070)서울특별시 성북구 안암동 5-1, 고려대학교 통계학과 대학원
E-mail: khbae@kustat.korea.ac.kr
3) (130-070)서울특별시 성북구 안암동 5-1, 고려대학교 통계연구소, 연구원
E-mail: korpen@ahanet.co.kr

의 표본이 $t-1$시점의 표본과 중복되는 부분이 없고 과거시점의 데이터를 동시에 보고하는 형태이다. 다수준교체표본조사에서는 표본을 같은 크기의 부표본(subsample)들로 나눈 후 매 조사시점마다 하나의 부표본을 조사하고 이 부표본은 다른 모든 부표본이 조사된 후 표본으로 돌아와 다시 조사된다. 이러한 이유로 이 부표본을 교체그룹(rotation group)이라고 한다. 다수준교체표본조사의 대표적인 경우는 미국 통계국에서 실시하는 Survey of Income and Program Participation(SIPP)이다. SIPP는 미국내 가정이나 구성원들의 인구통계학적 성격, 경제상태, 정부 프로그램의 참여정도를 측정하는 조사로 매해 2월 새로운 표본가구들로 구성된 4개의 교체그룹이 들어오며 교체그룹들은 4개월에 한번씩 조사되고 각 조사에서 과거 4개월의 정보를 동시에 보고하는 4수준교체표본조사를 실시한다.

교체표본조사에서 모집단의 특성치 추정을 위한 방법은 크게 시계열적 방법과 고전적인 방법이 있다. 시계열적 방법은 Blight와 Scott(1973)에 의하여 연구되었다. 고전적인 방법에는 최소분산불편추정량(*minimum variance unbiased estimator*)과 복합추정량(*composite estimator*)이 있다. 최소분산불편추정량에 대한 연구는 Eckler(1955), Gurney와 Daly(1965), Wolter(1979), Jones(1980)에 의해 수행되었고 복합추정량은 Rao와 Graham(1964), Gurney와 Daly(1965), Graham(1973), Breau와 Ernst(1983), Cantwell(1990)에 의해 연구되었다. 복합추정량은 t시점의 단순추정량과 $t-1$시점의 복합추정량의 선형결합이므로 최소분산불편추정량보다 계산이 쉬운 장점이 있으며 t시점의 정보뿐 아니라 과거시점 자료들을 이용하여 t시점 추정량의 분산을 줄일 수 있다.

본 논문에서는 2절에서 Cantwell이 제안한 p-수준교체표본설계에서 교체그룹내 최종표본추출단위(final sampling unit)간의 급내상관(intraclass correlation)을 고려한 일반화 복합추정량(generalized composite estimator)을 제시하고 이 추정량의 분산 및 분산을 최소화하는 가중치를 유도하였다. 3절에서는 수치 예를 통하여 일반화 복합추정량에 대한 교체그룹내 상관관계를 고려한 일반화 복합추정량의 효율성을 비교하였다(이후 일반화 복합추정량은 GCE, 교체그룹내 상관관계를 고려한 일반화 복합추정량은 EGCE로 쓰기로 한다) .

2. p-수준교체표본설계에서의 일반화 복합추정량

본 논문에서는 다음 조건(Cantwell, 1990)들을 만족하는 p-수준교체표본설계를 가정하였다.

(i) 표본은 p개의 교체그룹으로 구성되며, 각 조사시점에서 하나의 교체그룹이 조사된다.

(ii) p개의 교체그룹이 돌아가면서 매 p번째 조사시점에서 조사된다.

(iii) 조사시점 t에서의 보고기간은 $t-1$에서 $t-p$이다.

위의 가정에 의하여 조사시점 t에서 교체그룹의 특성 추정치 $x_{t,i'}$,$i'=1,\cdots,p$를 얻게된다. 여기서 t는 조사시점을 나타내고 i' 은 회상시간(recall time)을 나타낸다. 즉, 시점 t에서 교체그룹 1이 조사되었다면, 교체그룹 1은 $(t-1),(t-2),\cdots,(t-p)$시점에 해당하는 p개의 보고를 하게 되며 교체그룹 2는 시점 $t+1$에서 조사되고 $t,(t-1),\cdots,(t-p+1)$시점에 해당하는 p개의 보고를 하게된다. 같은 방법으로 나머지 $p-2$개의 교체그룹에 대해서

그림 2.1: $p = 4$인 경우의 p-수준교체표본설계

조사시점(월)	교체그룹			
	1	2	3	4
⋮	⋮	⋮	⋮	⋮
$t-5$	$x_{t-5,2}$	$x_{t-5,3}$	$x_{t-5,4}$	$x_{t-5,1}$
$t-4$	$x_{t-4,1}$	$x_{t-4,2}$	$x_{t-4,3}$	$x_{t-4,4}$
$t-3$	$x_{t-3,4}$	$x_{t-3,1}$	$x_{t-3,2}$	$x_{t-3,3}$
$t-2$	$x_{t-2,3}$	$x_{t-2,4}$	$x_{t-2,1}$	$x_{t-2,2}$
$t-1$	$x_{t-1,2}$	$x_{t-1,3}$	$x_{t-1,4}$	$x_{t-1,1}$
t	$x_{t,1}$	$x_{t,2}$	$x_{t,3}$	$x_{t,4}$
$t+1$	$x_{t+1,4}$	$x_{t+1,1}$	$x_{t+1,2}$	$x_{t+1,3}$
$t+2$	$x_{t+2,3}$	$x_{t+2,4}$	$x_{t+2,1}$	$x_{t+2,2}$
$t+3$	$x_{t+3,2}$	$x_{t+3,3}$	$x_{t+3,4}$	$x_{t+3,1}$
$t+4$	$x_{t+4,1}$	$x_{t+4,2}$	$x_{t+4,3}$	$x_{t+4,4}$
⋮	⋮	⋮	⋮	⋮

도 조사를 실시하고 이런 과정을 반복한다. 그림 2.1은 $p = 4$인 경우의 교체표본설계를 나타내고 있다. 이 경우를 예로 보면 교체그룹 1은 시점 t에서 $(t-1), \cdots, (t-4)$에 해당하는 자료를 보고하게 되며 , $\cdots$, 교체그룹 4는 시점 $t+3$에서 $(t+2), \cdots, (t-1)$에 해당하는 자료를 보고하게 되고, 이러한 과정이 반복된다.

이제 Y_t를 t월에서 조사하려는 특성치라 하고 y_t를 Y_t의 추정량이라 한다면, 일반화 복합추정량(Breau와 Ernst, 1983)은 다음과 같이 정의된다.

$$y_t = \sum_{i=1}^{p} a_i x_{t,i} - k \sum_{i=1}^{p} b_i x_{t-1,i} + k y_{t-1} , \quad \sum_{i=1}^{p} a_i = 1, \ \sum_{i=1}^{p} b_i = 1.$$

여기에서 k는 $0 \leq k < 1$을 만족하는 상수이다. $x_{t,i}$는 조사시점 t에서의 i번째 교체그룹의 특성추정치로, 교체그룹내 표본개체들의 단순평균이다. 이것은 회상시간을 기준으로 표현한 $x_{t,i'}$과는 차이가 있다. 그러나 조사월 t에서 얻어진 $x_{t,i}$가 서로 다른 회상시간을 갖는 경우 그림 2.1에서 알 수 있듯이 서로 다른 교체그룹에 속하게 되므로 조사월 t에서 회상시간 i'이 결정되면 교체그룹은 유일하게 결정된다. 조사월 t에서 교체그룹 i의 특성치가 회상시간 i'을 갖는 것을 $x_{t,i}(i')$ $i, i' = 1, \cdots, p$ 이라 하자. 이 교체그룹 i가 조사월 t에서 조사되었다면 $t-p$, $t-2p$시점에서도 이미 보고를 하였으므로 $[x_{t,i}(1), x_{t-1,i}(2), \cdots, x_{t-p,i}(p), x_{t-p-1,i}(1), \cdots, x_{t-2p,i}(p), \cdots]$들은 모두 같은 교체그룹 i에 속하게된다. 따라서 주어진 시차 $t_\circ$에 대하여 $j' = mod_p(i' + t_\circ - 1) + 1$이면 조사월 t에서 회상시간 i'을 갖는 교체그룹 특성추정치와 조사월 $t - t_\circ$에서 회상시간 j'을 갖는 교체그룹 특성추정치는 같은 교체그룹에 속하게된다. Cantwell(1990)은 GCE의 분산을 유도하기 위하여 사용한 가정은

다음과 같다.

$$Cov(x_{t,i'}, x_{t-t_\circ,j'}) = \begin{cases} d_{i'}^2\sigma^2 & t_\circ = 0, i' = j' \text{ 모든 } t, i' \text{에 대하여, } d_{i'} > 0 \\ 0 & t_\circ = 0, i' \neq j' \\ \rho_{t_\circ,i'} d_{i'} d_{j'} \sigma^2 & j' = mod_p(i' + t_\circ - 1) + 1,\ t_\circ > 0 \end{cases} \tag{2.1}$$

여기서 i', j'은 회상시간을 나타내며, $d_{i'}$은 보고하는 시점과 보고되는 자료의 시점차이로 발생하는 응답변동(response variability)을 조정해주는 가중치이다. Cantwell(1990)이 제시한 분산식의 정확한 형태는 그의 논문을 참조하기 바란다.

3. 교체그룹내 상관관계를 고려한 일반화 복합추정량

대부분의 p-수준 교체표본추출에서는 집락표집(cluster sampling)을 이용하여 교체그룹을 구성하며 SIPP의 경우도 동일하다. 따라서 교체그룹은 동질적인 성격을 갖는 표본개체들로 구성되며, 표본개체들간에 급내상관이 내재하게 된다. 그러나 GCE에서는 최소단위로 교체그룹의 추정량을 사용하며 이 추정량은 교체그룹내 표본개체들의 단순평균이므로 교체그룹에 내재하는 상관관계를 반영할 수 없다. 그러므로 이를 반영하기 위해서는 교체그룹내 표본개체를 최소단위로 하는 GCE를 새로이 정의하여야 한다.

교체그룹내 표본개체의 수를 m이라고 할때 EGCE는 다음과 같이 정의된다.

$$y_t^* = \sum_{i=1}^{p}\sum_{j=1}^{m} a_{i,j} x_{t,i,j} - k\sum_{i=1}^{p}\sum_{j=1}^{m} b_{i,j} x_{t-1,i,j} + k y_{t-1}^* , \quad \sum_{i=1}^{p}\sum_{j=1}^{m} a_{i,j} = 1, \quad \sum_{i=1}^{p}\sum_{j=1}^{m} b_{i,j} = 1 \tag{3.1}$$

여기에서 $x_{t,i,j}$는 조사월 t에서 교체그룹 i내의 표본개체 j를 나타낸다.

논의를 간단히 하기 위하여 식(3.1)을 벡터형태로 표시하면 다음과 같다.

$$y_t^* = \mathbf{a}'\mathbf{x}_t - k\mathbf{b}'\mathbf{x}_{t-1} + k y_{t-1}^*, \ \mathbf{1}'\mathbf{a} = 1, \ \mathbf{1}'\mathbf{b} = 1 \tag{3.2}$$

여기에서 $\mathbf{x}_t = (x_{t,1,1}, x_{t,1,2}, \cdots, x_{t,1,m}, \cdots, x_{t,p,1}, \cdots, x_{t,p,m})'_{pm\times 1}$인 벡터이다. $\mathbf{a}$와 $\mathbf{b}$는 각각 크기가 $pm \times 1$ 인 벡터로 $(a_{1,1}, \cdots, a_{i,j}, \cdots, a_{p,m})'$, $(b_{1,1}, \cdots, b_{i,j}, \cdots, b_{p,m})'$ 으로 정의되며 $\mathbf{1}$은 크기가 $pm \times 1$인 단위벡터이다. k는 $0 \leq k < 1$을 만족하는 상수이다.

3.1. 교체그룹내 상관관계를 고려한 일반화 복합추정량의 분산

EGCE의 분산을 구하기 위해 가정하는 표본개체간의 분산 공분산구조는 식(2.1)로 부터 다음과 같이 확장된다.

$$Cov(x_{t,i,j}, x_{t-t_\circ,i',j'}) = \begin{cases} d_i^2\sigma^2 & t_\circ = 0, i = i', j = j' \\ d_i^2\sigma^2\rho_{I,j,j'} & t_\circ = 0, i = i', j \neq j' \\ d_i d_{i'}\sigma^2\rho_{t_\circ,j,j'} & t_\circ > 0, i' = mod_p(i + t_\circ - 1) + 1, j = j' \\ d_i d_{i'}\sigma^2\rho_{t_\circ,j,j'}^{(b)} & t_\circ > 0, i' = mod_p(i + t_\circ - 1) + 1, j \neq j' \\ 0 & \text{그 이외의 경우} \end{cases} \tag{3.3}$$

이제 EGCE의 분산을 구하기 위하여, 교체그룹내의 상관구조를 R_I로, 조사시차 $t_\circ$간에 발생하는 교체그룹간의 상관구조를 $R_{t_\circ,l}$, $l=1,2,\cdots,p$으로 정의한다.

$$(R_I)_{i,j}=\begin{cases}1 & i=j\text{인 경우}\\ \rho_{I_{i,j}} & i\neq j\text{인 경우}\end{cases}\quad (R_{t_\circ,l})_{i,j}=\begin{cases}\rho_{t_\circ,l} & i=j\text{인 경우}\\ \rho^{(b)}_{t_\circ,i,j} & i\neq j\text{인 경우}\end{cases},\quad i,j=1,2,\cdots,m$$

조사시점 t에서의 응답조정 가중치 벡터 $D=diag(d_1,d_2,\cdots,d_p)$로 정의하고 $\mathbf{x}_t=(\mathbf{x}_{t,1}^{\circ'},\mathbf{x}_{t,2}^{\circ'},\cdots,\mathbf{x}_{t,p}^{\circ'})'$, $\mathbf{a}_t=(\mathbf{a}_1^{\circ'},\mathbf{a}_2^{\circ'},\cdots,\mathbf{a}_p^{\circ'})'$, $\mathbf{b}_t=(\mathbf{b}_1^{\circ'},\mathbf{b}_2^{\circ'},\cdots,\mathbf{b}_p^{\circ'})'$ 는 모두 크기가 $pm\times 1$인 벡터로 $\mathbf{x}_{t,i}^{\circ'},\mathbf{a}_i^{\circ'},\mathbf{b}_i^{\circ'}$은 각각 $(x_{t,i,j}),(a_{i,j}),(b_{i,j})$, $j=1,\cdots,m$인 $m\times 1$ 벡터를 원소로 갖는다.

$$Var(\mathbf{a}'\mathbf{x}_t)=\sum_{i=1}^{p}\left(\mathbf{a}_i^{\circ'}Var(\mathbf{x}_{t,i}^{\circ})\mathbf{a}_i^{\circ}\right)=\sum_{i=1}^{p}\left(\mathbf{a}_i^{\circ'}(D^2R_I)\mathbf{a}_i^{\circ}\right)=\sigma^2\mathbf{a}'\left(D^2\otimes R_I\right)\mathbf{a}\qquad(3.4)$$

$$\begin{aligned}Cov(\mathbf{a}'\mathbf{x}_t,\mathbf{b}'\mathbf{x}_{t-1})&=Cov\left(\sum_{i=1}^{p}\mathbf{a}_i^{\circ'}\mathbf{x}_{t,i}^{\circ},\sum_{i=1}^{p}\mathbf{b}_i^{\circ'}\mathbf{x}_{t-1,i}^{\circ}\right)=\sum_{i=1}^{p}\mathbf{a}_i^{\circ'}Cov(\mathbf{x}_{t,i}^{\circ},\mathbf{x}_{t-1,i}^{\circ})\mathbf{b}_i^{\circ}\\&=\sigma^2\sum_{i=1}^{p-1}\mathbf{a}_i^{\circ'}\left(d_id_{i+1}R_{1,i}\right)\mathbf{b}_i^{\circ}+\sigma^2\mathbf{a}_p^{\circ'}\left(d_pd_1R_{1,p}\right)\mathbf{b}_p^{\circ}\\&=\sigma^2\mathbf{a}'\left(D\otimes I\right)R_1\left(PD\otimes I\right)\mathbf{b}\end{aligned}$$

여기서 $\otimes$는 크로네커곱을 나타낸다. $R_1=diag\left(R_{1,1},R_{1,2},\cdots,R_{1,p}\right)$이며, I는 $m\times m$ 단위행렬이다. P는 크기가 $p\times p$인 순열행렬(permutation matrix)로

$$(P)_{i,j}=\begin{cases}1 & j-i=1\text{ 인 경우와 }i=1,j=p\text{인 경우}\\ 0 & \text{그 이외의 경우}\end{cases}$$

로 정의되며, P행렬의 역할은 D행렬의 원소인 $d_i,i=1,\cdots,p$, 즉 응답조절 가중치의 배열을 조정해 주는 것이다. 응답조절 가중치의 배열은 각 조사시점에서 회상시간 i의 배열과 일치한다. 순열행렬 P의 성질(Graybill, 1983)에 의하여 조사월 $t-1$에서의 응답조절 가중치 벡터는 $(d_2,d_3,\cdots,d_p,d_1)'=PD$, 조사월 $t-2$에서는 $(d_3,\cdots,d_p,d_1,d_2)'=P\cdot PD=P^2d_t$, $\cdots$, 조사월 $t-t_\circ$에서는 $P^{t_\circ}D$가 성립한다. 이러한 성격은 그림 2.1을 통하여 쉽게 확인할 수 있다. 그러므로

$$Cov\left(\mathbf{a}'\mathbf{x}_t,\mathbf{b}'\mathbf{x}_{t-t_\circ}\right)=\sigma^2\mathbf{a}'\left(D\otimes I\right)R_{t_\circ}\left(P^{t_\circ}D\otimes I\right)\mathbf{b}\qquad(3.5)$$

으로 정의된다. 여기서 $R_{t_\circ}=diag\left(R_{t_\circ,1},R_{t_\circ,2},\cdots,R_{t_\circ,p}\right)$이다.

정리 3.1 p-수준교체표본추출에서 교체그룹내 상관관계를 고려한 일반화 복합추정량의 분

산은 분산 공분산 조건 (3.3)하에서 다음과 같다.

$$Var(y_t^*) = \frac{\sigma^2}{1-k^2}\left[\mathbf{a}'(Z_1+2Z_2)\mathbf{a} - 2\mathbf{a}'(k^2Z_1+Z_2+k^2Z_2')\mathbf{b} + k^2\mathbf{b}'(Z_1+2Z_2)\mathbf{b}\right]$$

$$Var(y_t^* - y_{t-1}^*) = \frac{\sigma^2}{k^2}\left[\mathbf{a}'Z_1\mathbf{a} - 2k\mathbf{a}\big(D\otimes I\big)R_1\big(PD\otimes I\big) + k^2\mathbf{b}'Z_1\mathbf{b}\right] - \frac{(1-k)^2}{k}Var(y_t^*)$$

여기에서 $\mathbf{Z}_{1_{(pm\times pm)}} = D^2\otimes R_I$, $\mathbf{Z}_{2_{(pm\times pm)}} = \big(D\otimes I\big)\Big[\sum_{i=1}^{\infty}k^iR_i\big(P^i\otimes I\big)\Big]\big(D\otimes I\big)$ 이다.

증명: $Var(y_t^*)$을 증명하기 위해 교체그룹내 상관관계를 고려한 일반화 복합추정량을 다음과 같이 재표현한다.

$$y_t^* = \mathbf{a}'\mathbf{x}_t + (\mathbf{a}-\mathbf{b})'\sum_{i=1}^{\infty}k^i\mathbf{x}_{t-i};$$

그러므로

$$\begin{aligned} Var(y_t^*) &= Var\Big(\mathbf{a}'\mathbf{x}_t + (\mathbf{a}-\mathbf{b})'\sum_{i=1}^{\infty}k^i\mathbf{x}_{h-i}\Big) \\ &= Var(\mathbf{a}'\mathbf{x}_t) + \sum_{i=1}^{\infty}k^{2i}Var\Big((\mathbf{a}-\mathbf{b})'\mathbf{x}_{t-i}\Big) \\ &\quad +2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}k^{i+j}Cov\Big((\mathbf{a}-\mathbf{b})'\mathbf{x}_i,(\mathbf{a}-\mathbf{b})'\mathbf{x}_j\Big) + 2Cov\Big(\mathbf{a}'\mathbf{x}_t,\sum_{i=1}^{\infty}k^i(\mathbf{a}-\mathbf{b})'\mathbf{x}_{t-i}\Big) \end{aligned}$$

이 식에 식 (3.4), 식 (3.5)을 대입하면 $Var(y_t^*)$를 얻을 수 있다.

$Var(y_t^* - y_{t-1}^*)$을 구하기위해 $\omega_t = \mathbf{a}'\mathbf{x}_t - k\mathbf{b}'\mathbf{x}_{t-1}$라고 하면 $Var(\omega_t)$는

$$Var(\omega_t) = Var(\mathbf{a}'\mathbf{x}_t) - 2kCov(\mathbf{a}'\mathbf{x}_t,\mathbf{b}\mathbf{x}_{t-1}) + k^2Var(\mathbf{b}'\mathbf{x}_{t-1}) \tag{3.6}$$

이다. $y_t^* = \omega_t + ky_{t-1}^*$이므로

$$Var(y_t^*) = Var(\omega_t) + k^2Var(y_{t-1}^*) + 2kCov(\omega_t\ ,\ y_{t-1}^*) \tag{3.7}$$

이다. 따라서 $Var(y_t^* - y_{t-1}^*) = Var(y_t^*) - 2Cov(y_t^*,y_{t-1}^*) + Var(y_{t-1}^*)$에 대하여 식(3.7)를 대입하여 정리하면

$$Var(y_t^* - y_{t-1}^*) = \frac{1}{k}\Big(Var(\omega_t) - (k-1)^2Var(y_t)\Big)$$

를 얻게되고, 다시 식 (3.6)을 대입하면 $Var(y_t^* - y_{t-1}^*)$가 얻어진다. □

3.2. 교체그룹내 상관관계를 고려한 일반화 복합추정량의 최적계수

식 (3.2)에 제시된 EGCE를 실제로 이용하기 위해서는 계수 $\mathbf{a}$와 $\mathbf{b}$를 결정하여야 한다. $\mathbf{a}$와 $\mathbf{b}$는 추정량의 분산을 최소화 하도록 하여 결정하게 된다. 계수 $\mathbf{a}$와 $\mathbf{b}$는 $\mathbf{1}'\mathbf{a}=1$, $\mathbf{1}'\mathbf{b}=1$이라는 제약식이 존재하므로 라그랑지 승수법을 이용하여 구하면 된다. 라그랑지 승수법 적용을 위한 목적함수를 다음과 같이 정의한다.

$$O_1 = Var(y_t^*) - 2\lambda_1\big(\mathbf{1}'\mathbf{a}_L - 1\big) - 2\lambda_2\big(\mathbf{1}'\mathbf{b}_L - 1\big) \tag{3.8}$$

$$O_2 = Var(y_t^* - y_{t-1}^*) - 2\lambda_1\big(\mathbf{1}'\mathbf{a}_c - 1\big) - 2\lambda_2\big(\mathbf{1}'\mathbf{b}_c - 1\big) \tag{3.9}$$

$\mathbf{a}$와 $\mathbf{b}$의 유도과정은 생략하고 결과만을 제시하도록 하겠다.

$$Var(y_t^*) = \frac{\sigma^2}{1-k^2}\left[\mathbf{a}_L' A_L \mathbf{a}_L - \mathbf{a}_L' B_L \mathbf{b}_L + \mathbf{b}_L' C_L \mathbf{b}_L\right]$$

라 하자. 여기서 $A_L = Z_1 + 2Z_2$, $B_L = -2(k^2 Z1 + Z_2 + k^2 Z_2')$, $C_L = k^2 A_L$이다.

$G_L = (A_L + A_L')^{-1}$이고 $J = \mathbf{1}'\mathbf{1}$ 일때 $A_L^* = G_L(JG_L - I)$로 놓으면 식 (3.8)을 라그랑지 승수법으로 푼 결과는 다음과 같이 정리할 수 있다.

$$\widehat{\mathbf{a}}_L = \left[I - \frac{1}{k^2}A_L^* B_L A_L^* B_L'\right]^{-1}\left[I + A^* B_L\right]G_L\mathbf{1},\ \ \widehat{\mathbf{b}}_L = \frac{1}{k^2}A_L^* B_L' \widehat{\mathbf{a}}_L + G_L\mathbf{1} \tag{3.10}$$

$$Var(y_t^* - y_{t-1}^*) = \frac{\sigma^2}{1-k^2}\left[\mathbf{a}_C' A_C \mathbf{a}_C - \mathbf{a}_C' B_C \mathbf{b}_C + \mathbf{b}_C' C_C \mathbf{b}_C\right]$$

라 하자. 여기서 $A_C = \big((1-k^2)/k^2\big)Z_1 - \big((1-k)^2/k\big)A_L$, $B_C = \big(-2(1-k^2)/k\big)(D \otimes I)R_I(PD \otimes I) + \big((1-k)^2/k\big)B_L$, $C_C = k^2 A_C$이다. 식 (3.9)을 라그랑지 승수법으로 풀어 $\widehat{\mathbf{a}}_C$와 $\widehat{\mathbf{b}}_C$를 구하는 것은 $A_C^* = G_C(PG_C - I)$, $G_C = (A_C + A_C')^{-1}$로 놓을 때 식 (3.10)의 결과에서 A_L, B_L, C_L을 A_C, B_C, C_C로 A_L^*, G_L을 A_C^*, G_C로 바꾸어 넣은 것과 같다.

4. 수치 예

GCE(Cantwell,1990)의 분산과 EGCE의 분산을 비교하기 위해서 SIPP에서 사용하는 4수준교체표본을 사용하였다. 하나의 표본을 4개의 교체그룹으로 나누고 각 교체그룹은 동일한 갯수의 표본개체가 있음을 가정하였다. 교체그룹내의 상관구조는 다음과 같이 정의하였다.

$$(R_I)_{i,j} = \begin{cases} 1 & i=j\text{인 경우} \\ \rho_I^{|i-j|} & i\neq j\text{인 경우} \end{cases} \quad i,j = 1,2,\cdots,m$$

각 교체그룹내의 표본개체의 분산을 σ^2로 동일하게 놓았으며 응답조절 가중치는 $d_i = i/10, i = 1,2,\cdots,4$로 하였다. 그리고 시간상관관계는 교체그룹에 상관없이 모두 동일하다고 하였다. 즉, 시차 $t_\circ$에서의 교체그룹간의 시간상관구조를 다음과 같이 정의하였으며,

가중치 ω은 0.9를 사용하였다.

$$(R_{t_\circ})_{i,j} = \begin{cases} \rho^{t_\circ} & i = j\text{인 경우} \\ (\omega\rho + (1-\omega)\rho_I)^{t_\circ}\rho_I^{|i-j|} & i \neq j\text{인 경우} \end{cases} \quad i,j = 1,2,\cdots,m$$

GCE와 EGCE의 효율성을 비교하기 위하여 각 추정량의 분산비를 계산하였다. 효율성을 $Var(EGCE)/Var(GCE)$로 정의하였으므로 효율이 1 보다 작다는 것은 EGCE가 더 나은 추정량임을 의미하게 된다. 각 추정량에 포함되어 있는 가중치 k를 0.1부터 0.9까지 0.1

표 4.1: 4-수준교체표본추출에서 표본개체수에 따른 추정량의 효율성 비교

표본		eff of y_t^*			eff of $y_t^* - y_{t-1}^*$		
개체	ρ_I	$\rho = .4$	$\rho = .6$	$\rho = .8$	$\rho = .4$	$\rho = .6$	$\rho = .8$
4	.3	0.9841(.3)	0.9818(.3)	0.9700(.4)	0.9837(.9)	0.9810(.9)	0.9735(.9)
	.5	0.9705(.3)	0.9677(.3)	0.9486(.4)	0.9714(.9)	0.9671(.9)	0.9544(.9)
	.7	0.9676(.3)	0.9676(.3)	0.9527(.4)	0.9722(.9)	0.9686(.9)	0.9559(.9)
5	.3	0.9826(.3)	0.9803(.3)	0.9688(.4)	0.9822(.9)	0.9794(.9)	0.9722(.9)
	.5	0.9639(.3)	0.9610(.3)	0.9419(.4)	0.9648(.9)	0.9604(.9)	0.9476(.9)
	.7	0.9557(.3)	0.9555(.4)	0.9398(.4)	0.9607(.9)	0.9566(.9)	0.9431(.9)
6	.3	0.9826(.3)	0.9804(.3)	0.9697(.4)	0.9822(.9)	0.9796(.9)	0.9727(.9)
	.5	0.9608(.3)	0.9580(.3)	0.9399(.4)	0.9617(.9)	0.9574(.9)	0.9451(.9)
	.7	0.9471(.3)	0.9468(.4)	0.9312(.4)	0.9521(.9)	0.9478(.9)	0.9343(.9)

단위로 변화시켜서 분산을 최소화하는 값으로 선택하였다. 표 4.1은 교체그룹내 표본개체의 수를 $m = 4,5,6$으로 변화시키고 시차에 의해 발생하는 상관계수를 $\rho = 0.4, 0.6, 0.8$ 교체그룹내 내재하는 상관계수를 $\rho_I = 0.3, 0.5, 0.7$로 변화시키면서 GCE에 대한 EGCE y_t^*와 $y_t^* - y_{t-1}^*$의 효율성을 계산한 것이다. 결과에서 알 수 있듯이 모든 경우에 대하여 EGCE의 효율이 좋게 나타나고 있으며 약 1.2% ~ 6.8%의 효율성 개선이 있는 것으로 나타나고 있다. 표의 결과는 y_t^*와 $y_t^* - y_{t-1}^*$에서 모두 ρ의 값이 증가할수록, 교체그룹내의 표본개체의 수가 증가할수록, 그리고 ρ_I의 값이 증가할수록 높은 효율 개선이 얻어짐을 보여주고 있다.

5. 결론

p-수준교체표본에서의 GCE는 교체그룹내 내재하는 표본개체사이의 상관관계를 고려하지 않고 교체그룹의 추정량을 사용하였다. 그러나 대부분의 다수준교체표본이 최종표본개체를 집락표집하여 교체그룹을 구성하므로 교체그룹내 표본개체사이의 상관관계를 고려하지 않은 것은 옳지 않다. 본 논문에서는 교체그룹내의 표본개체들 사이의 교체그룹내 상관관계를 고려한 일반화복합추정량(EGCE)을 제시하고 현월 추정량과 전월대비 추정량의 분산식과 이들에 대한 최적계수를 제시하였다. 또한 GCE에 비하여 EGCE가 더 나은

추정량임을 수치예를 통하여 확인하였다. 이러한 결과는 교체그룹의 추정치로 단순평균을 사용하는 것보다 교체그룹내 상관관계를 고려한 가중평균을 사용하는 것이 더 좋다는 것을 의미한다. 실제로 3.2절의 결과에 의해 얻어진 $\hat{\mathbf{a}}$와 $\hat{\mathbf{b}}$의 값을 확인하면 각 교체그룹내의 표본개체에 서로 다른 가중치가 부여되게 된다.

참고문헌

[1] Blight, B.J.N., and Scott, A.J. (1973). A Stochastic Model for Repeated Surveys, *Journal of the Royal Statistical Society* Series B 35, 61-66.

[2] Breau, P., and Ernst, L.R. (1983). Alternative Estimators to the Current Composite Estimator, *Proceedings of the American Statistical Association, Section on Social Statistics* 397-402.

[3] Cantwell, P.J. (1990). Variance Fomulae for Composite Estimators Rotation Designs, *Survey Methodology* 16, 153-163.

[4] Eckler,A.R. (1955). Rotation Sampling, *Annals of Mathematical Statistics* 26, 664-685.

[5] Graybill, F.A. (1983). *Matrices with Applications in Statistics*, Second edition, Wadsworth. Inc.

[6] Graham, J.E. (1973). Composite Estimation in Two Cycle Sampling Designs, *Communications in Statistics* 1, 419-431.

[7] Gurney, M., and Daly, J.F. (1965). A Multivariate Approach to Estimation in Periodic Sample Survey, *Proceedings of the American Statistical Association, Section on Social Statistics* 242-257.

[8] Huggins, V.J. and Fischer, D.P. The Redesign of the Survey of Income and Program Participation, Working Paper for 1994 American Statistical Associations Meetings.

[9] Jones, R.G. (1980). Best Linear Unbiased Estimators for Repeated Surveys, *Journal of the Royal Statistical Society* Series B 42, 221-226.

[10] Rao,J.N.K., and Graham,J.E. (1964). Rotation Design for Sampling on Repeated Occasions, *Journal of the American Statistical Association* 59, 492-509.

[11] Wolter, K.M. (1979). Composite Estimation in Finite Populations, *Journal of the American Statistical Association* 74, 604-613.

[1999년 10월 접수, 2000년 12월 채택]

Generalized Composite Estimator with Intraclass Correlation in p-level Rotation Sampling

YouSung Park[1] Kyoung Hwa Bae[2] Kee Whan Kim[3]

ABSTRACT

One of the Repeated survey which estimates variability of population, we can be consider rotation sample survey. There are two kinds of rotation sample survey - one-level rotation sample survey and multi-level rotation sample survey. In rotation sample survey, Composite estimator is used to measure level or level change of the population. This study suggests Generalized Composite estimator as considering intraclass correlation in multi-level rotation sample survey, and optimal weight minimizing variance of estimator. Numerical example shows efficiency of Generalized Composite estimator as considering intraclass correlation according to the sample unit and change degree of intraclass correlation in the rotation group.

Keywords: Multi-level rotation sample survey; Generalized composite estimator; Intraclass correlation

1) Associate Professor, Dept. of Statistics, Korea University.
E-mail: yspark@mail.korea.ac.kr
2) Graduate Student, Dept. of Statistics, Korea University.
E-mail: khbae@kustat.korea.ac.kr
3) Researcher, Institute of Statistics, Korea University.
E-mail: korpen@ahanet.co.kr