응용통계연구 제10권 2호
1997년 9월 pp. 367-374

패널조사에서 표본 변경을 고려한 추정

박 진 우1)

요 약

우리나라 정부나 주요 기관에서 실시하는 표본조사들 중 많은 조사가 패널조사이다. 패널조사에서는 시간의 경과에 따른 모집단구성의 변화를 표본에 적절히 반영해 주어야 하며 이러한 변화를 추정과정에서도 고려하여야 한다. 본 논문에서는 패널조사에서 표본에 일부 변화가 생길 경우 이 변화를 고려하지 않은 일반적인 추정량은 편의를 갖게 되며, 일반적으로 사용하고 있는 추정량의 분산의 식도 적절하지 않음을 보였다. 아울러 표본의 변화를 고려한 불편추정량과 그 분산을 제시하였다.

1. 서 론

표본조사들 중에는 한번의 조사(single time survey)로 끝나지 않고 월별, 분기별 또는 연별로 비슷한 조사를 반복하는 경우가 많다. 이러한 조사들은 특성에 따라서 반복조사(repeated survey)나 패널조사(panel survey)등으로 구분된다. 반복조사는 동일한 모집단에 대해 매번 조사할 때마다 독립적인 표본을 추출하여 조사하는 것을 일컫는다. 이에 반해 패널조사란 동일한 표본에 대해 일정 시간간격으로 반복하여 조사를 실시하는 조사이다(Bailor, 1989; Kalton 과 Citro, 1993). 우리나라의 통계청, 농림수산부, 보건복지부 등 정부기구나 한국은행, 주택은행 등 기타기관에서 실시하는 여러 종류의 표본조사들을 살펴보면 동일한 표본을 반복하여 조사하는 패널조사가 많음을 알 수 있다. 패널조사인 경우 동일한 조사단위들의 특성값이 시간의 경과에 따라 어떻게 달라지는 지를 조사할 수 있으므로 시계열적 분석(longitudinal analysis)이 가능하다는 장점을 갖는다. 그러나 우리나라의 여러조사들을 살펴보면 시계열적 분석을 위한 측면보다는 단순히 표본설계의 용이성 때문에 패널조사가 사용되고 있음을 알 수 있다.

모집단내 조사단위들은 시간이 경과함에 따라 새롭게 생기는 것이 있는가 하면 또한 사라지는 것들이 생기기도 한다. 따라서 패널조사에서는 시간의 경과에 따라 생기는 모집단구성의 변화를 적절히 프레임에 반영하는 것이 중요하다. 만일 이러한 변화를 고려하지 않은 채 과거의 표본에 의해서만 조사를 실시한다면 미포함(under-coverage)에 의한 프레임오차를 초래하게 된다. Colledge(1989)는 이러한 프레임의 관리문제를 다루고 있다. 그러나 표본설계에 관한 국내의 연구들(박흥래, 1987 : 이기재외, 1993; 김규성외, 1994)을 보면 시간경과에 따라 생기는 프레임 및 표본의 변동에 대한 관리에는 관심을 가지나 이 변동을 추정과정에서는 고려하지 않고 있음을 볼 수 있다. 모집단 구성의 변화가 미미하여 전체추정량에 별 영향을 미치지 못한다고 판단

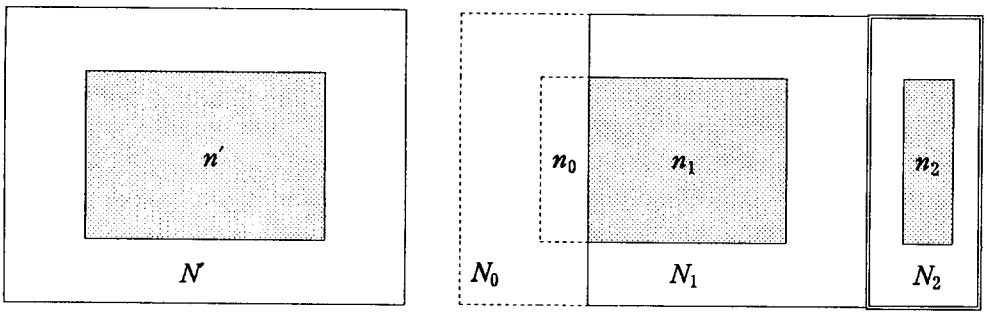
1) (445-743) 경기도 화성군 봉담면 와우리 산2-2, 수원대학교 응용통계학과 조교수

될 때에는 별 영향이 없겠지만 그렇지 않을 때에는 추정과정에서도 프레임 및 그에 따르는 표본의 변경을 고려해야 한다.

이 논문의 목적은 패널조사에서 프레임의 변경이 발생하여 표본에도 일부 변화가 생기게 될 때 이를 고려하지 않은 추정량은 편의(bias)를 야기시킨다는 사실을 지적하고, 이에 대한 적절한 추정법을 연구하는 것이다. 2절에서는 프레임의 변화에 따라 표본에도 변화가 생겼을 경우 기존의 추정량을 사용하게 될 때 생기는 편의를 구하였다. 3절에서는 수정된 불편추정량과 그 추정량의 분산을 제시하였다.

2. 표본의 변화를 고려하지 않은 추정

시간이 경과함에 따라 모집단 단위들의 구성에 변동이 있는 경우를 고려하자. 변화되기 전 모집단내의 단위들의 수를 N', 표본의 크기를 n'이라 하고 변화된 이후의 모집단 단위들의 수를 N, 표본의 크기를 n 이라고 하자. 또한 변화된 이후의 모집단에서 사라진 단위들의 수를 N_0, 모집단이 사라짐으로 인해 사라지게 된 표본의 크기를 n_0 라고 하자. 반면 지난번 모집단 중 다음 조사 때에도 남아있는 단위의 수를 N_1, 남아있는 표본의 수를 n_1이라고 하면 $N' = N_0 + N_1$ 이고 $n' = n_0 + n_1$ 이 된다. 또한 새로 생긴 모집단 단위들의 수를 N_2, 이에 따라 새로 추가된 표본의 수를 n_2라고 하면 $N = N_1 + N_2$, $n = n_1 + n_2$ 가 된다. 아래의 <그림1>은 이러한 사항을 그림으로 나타낸 것이다. 아래 그림에서 짙게 칠해진 부분은 표본을 나타낸다.



(a) 처음의 모집단과 표본 (b) 변화된 이후의 모집단과 표본

<그림 1> 모집단과 표본의 변화

단순임의추출법을 가정할 때 처음 표본설계 당시의 모집단을 기준으로 한 모평균의 추정량과 그 분산은 다음의 식 (2.1), (2.2)와 같이 나타낼 수 있다.

$$\bar{y}' = \frac{1}{n'} \sum_{i=1}^{n'} y_i \qquad (2.1)$$

$$Var(\overline{y'}) = \frac{N-n'}{N}\frac{S'^2}{n'} \quad , \quad S'^2=\frac{\sum_{i=1}^{N}(y_i-\overline{Y'})^2}{N-1} \tag{2.2}$$

위의 <그림 1>처럼 시간이 지남에 따라 모집단 상황이 변화되었다고 하자. 아울러 이러한 변화에 따라 프레임을 수정, 보완하였다고 하면 사라진 모집단에 해당되는 표본의 일부가 없어질 것이고 새로 생긴 모집단에 해당되는 조사단위들 중 일부가 새로운 표본으로 추가될 것이다. 만일 n 개의 전체표본을 모두 다시 랜덤하게 추출한다면 위의 식을 그대로 사용하여도 아무런 문제가 없을 것이다. 하지만 변화가 없는 모집단의 부분에 대해서는 과거의 표본을 그대로 이용하고 새로 생긴 모집단에 대해서만 N_2개 중에서 단순랜덤추출법으로 n_2개의 표본을 추가적으로 뽑는 경우 만일 위의 추정량의 식을 그대로 사용한다면 (2.1)은

$$\overline{y'}=\frac{1}{n_1+n_2}[\sum_{i=1}^{n_1} y_{1i}+\sum_{j=1}^{n_2} y_{2j}] \tag{2.3}$$

으로 표현되어지며 그 기대값을 구하면 다음과 같게 된다.

$$E(\overline{y'})=\frac{n_1}{n_1+n_2}\overline{Y_1}+\frac{n_2}{n_1+n_2}\overline{Y_2}.$$

한편 변화된 모집단의 모평균 $\overline{Y}$는

$$\overline{Y}=\frac{N_1\overline{Y_1}+N_2\overline{Y_2}}{N_1+N_2}$$

로 표현되므로 기존의 추정량을 그대로 사용할 경우의 편의는 다음과 같다.

$$\begin{aligned} Bias(\overline{y'}) &= E(\overline{y'})-\overline{Y} \\ &= [\frac{n_1}{n_1+n_2}-\frac{N_1}{N_1+N_2}]\cdot\overline{Y_1} + [\frac{n_2}{n_1+n_2}-\frac{N_2}{N_1+N_2}]\cdot\overline{Y_2} \quad . \end{aligned}$$

만일 $\overline{Y_1}=\overline{Y_2}$ 가 성립한다면, 즉 새로 모집단에 추가된 모집단 단위들의 모평균과 이전 모집단 단위들의 모평균이 같을 때에는 $Bias(\overline{y'})=0$ 이 되므로 위의 식 (2.1)이 불편추정량이 되지만 그렇지 않으면 불편추정량이 되지 못함을 알 수 있다. 그런데 일반적으로는 새로 추가된 모집단 부분의 모평균과 이전 모집단의 모평균이 다르므로 위의 편의의 값이 0이 아니게 된다. 따라서 모집단의 변화에 따른 표본의 변화를 고려할 경우 추가된 부분을 따로 하나의 독립된 층으로 간주하지 않은 채 처음 설계에 따른 추정량을 그대로 사용하는 것은 추정에 있어서의 편의를 야기시키게 된다. 한편 (2.3)의 추정량의 분산으로 (2.2)식 형태를 그대로 사용한다면 분산의 식은 아래의 (2.4)식으로 표현되어진다.

$$Var(\overline{y'}) = \frac{N-n}{N}\frac{S^2}{n} \quad , \quad S^2=\frac{\sum_{i=1}^{N}(y_i-\overline{Y})^2}{N-1} \tag{2.4}$$

이 경우 $\overline{Y}$가 $\overline{y'}$의 불편추정량이 아니므로 (2.4)식 자체는 올바른 분산의 식이라고 볼 수가 없다. 만일 분산의 식을 올바르게 표현하려고 한다면 다음과 같이 나타내어야 할 것이다.

$$Var(\overline{y'}) = Var(\frac{n_1}{n}\overline{y_1}+\frac{n_2}{n}\overline{y_2})$$
$$=(\frac{n_1}{n})^2 Var(\overline{y_1})+(\frac{n_2}{n})^2 Var(\overline{y_2}).$$

위의 식에서 $Var(\overline{y_2})=\frac{N_2-n_2}{N_2}\frac{S_2^2}{n_2}$로 표현할 수 있는데 반해 $Var(\overline{y_1})$는 사후층화의 개념을 이용하여 유도하여야 한다. $Var(\overline{y_1})$의 식의 유도는 다음 절의 [정리 1]에서 다루기로 한다.

3. 표본의 변화를 고려한 추정

앞 절의 <그림 1>에서와 같이 모집단 구성에 변화가 생겼고 그에 따라 표본에도 일부 변화가 생긴 경우의 추정에 대해 생각해보자. 이 경우 과거의 모집단단위들 중에서 소멸된 부분은 전체에서 랜덤하게 발생한다고 가정하고 또한 소멸된 표본단위들은 소멸된 모집단내 조사단위들 중 랜덤하게 뽑힌 것으로 가정한다. 그렇다면 현재의 변화된 모집단은 과거의 모집단중 현재 남아있는 부분과 새로이 추가된 부분의 두 층으로 구성된 것으로 볼 수 있다. 따라서 표본은 남아있는 층에서 n_1개, 새로이 추가된 층에서 n_2개씩 추출하여 얻은 것으로 볼 수 있다. 먼저 종래의 표본으로 모집단중 변화되지 않은 부분의 평균을 구한 추정량과 그 분산의 식을 나타낸 것이 아래의 [정리 1]이다.

[정리 1] 위의 <그림 1>에서 보는 바와 같이 원래 N개의 단위들로 구성된 모집단에 변화가 생겨 N_1개의 단위만 남고 N_0개는 사라졌다. 또한 종래의 표본 n'개 중에서는 n_0개가 사라지고 n_1개만 남았다. 남은 n_1개 표본들을 통해 모집단 N_1개에 대한 평균의 추정량을 구하고 그 분산을 유도하면 다음과 같다.

$$\overline{y_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} \tag{3.1}$$

$$Var(\overline{y_1}) = \frac{S_1^2}{N_1}[\frac{1}{f}(1+\frac{1-f}{n'}\frac{N_0}{N_1})-1] \tag{3.2}$$
$$, \ f=\frac{n'}{N} \quad , \quad S_1^2=\frac{\sum_{i=1}^{N_1}(y_{1i}-\overline{Y_1})^2}{N_1-1}$$

<증명> 원래 n'개의 표본에 의해 변화되지 않은 모집단의 모평균을 추정하는 문제에서 사후층화(post-stratification)를 했을 때 N_1과 N_0 두 개의 층으로 구분하였다고 생각하자. 이 때 위의 식 (3.1)이 불편추정량이 되는 것은 쉽게 알 수 있다. 한편 식 (3.2)는 $\overline{y_1}$의 식에서 n_1도 하나의 확률변수로 볼 수 있다는 사실에 기인하여 다음과 같이 유도될 수 있다(Hansen et al.(1952)). 먼저

$$
\begin{aligned}
Var(\overline{y_1}) &= EE(\overline{y_1} - \overline{Y_1})^2 \\
&= E_{n_1}\left(\frac{N_1 - n_1}{N_1}\frac{S_1^2}{n_1}\right) \\
&= S_1^2 E_{n_1}\left(\frac{1}{n_1}\right) - \frac{S_1^2}{N_1}
\end{aligned}
$$

인데 여기서 $1/n_1$을 테일러식으로 전개한 후 기대값을 취하면

$$
\begin{aligned}
E_{n_1}\left(\frac{1}{n_1}\right) &\approx \frac{1}{E_{n_1}(n_1)} + \frac{1}{E_{n_1}(n_1^3)} \cdot Var(n_1) \\
&= \frac{N}{n'N_1} + \left(\frac{N}{n'N_1}\right)^2\left(\frac{N-N}{N}\right)\left(\frac{N-n'}{N-1}\right) \\
&\approx \frac{1}{f}\frac{1}{N}\left(1 + \frac{1-f}{n'}\frac{N_0}{N_1}\right)
\end{aligned}
$$

가 된다. 이를 위의 식에 대입하여 구하면 다음과 같이 표현된다.

$$
\begin{aligned}
Var(\overline{y_1}) &= EE(\overline{y_1} - \overline{Y_1})^2 \\
&= E_{n_1}\left(\frac{N_1 - n_1}{N_1}\frac{S_1^2}{n_1}\right) \\
&\approx \frac{S_1^2}{N_1}\left[\frac{1}{f}\left(1 + \frac{1-f}{n'}\frac{N_0}{N_1}\right) - 1\right].
\end{aligned}
$$

여기서 $f = \frac{n'}{N}$, $S_1^2 = \frac{\sum_{i=1}^{N_1}(y_{1i} - \overline{Y_1})^2}{N_1 - 1}$ 이다. ∎

위의 [정리1]을 이용하여 변화된 모집단에 대해 수정된 평균의 추정량과 그 분산의 식을 구한 것이 다음의 [정리2]이다.

[정리2] 앞 절의 <그림1>에서 보는 바와 같이 원래 N개의 단위들로 구성된 모집단에 변화가 생겨 N_1개의 단위만 남고 N_0개는 사라졌다. 또한 새로이 N_2개의 단위들이 모집단에 추가되었다. 이에 따라 표본에도 변화가 생겨 원래 n' 개 중에서 n_0개는 사라지고 n_1개만 남았으며 그 밖에 새로 N_2개의 모집단 단위들 중에서 단순랜덤추출법으로 n_2개의 표본을 추가하였다. 이 때 다음의 식 (3.3)은 변화된 모집단의 평균에 대한 불편추정량의 식이며 식 (3.4)는 그 분산의 식이다.

$$
\overline{y_{adj}} = \sum_{i=1}^{2} W_i \overline{y_i} \tag{3.3}
$$

$$
Var(\overline{y_{adj}}) = \sum W_i^2 Var(\overline{y_i})^2 \tag{3.4}
$$

여기서 $\overline{y_1}$와 $V(\overline{y_1})$는 식 (3.1)과 (3.2)의 식들이며,

$$Var(\overline{y_2}) = \frac{N_2-n_2}{N_2}\frac{S_2^2}{n_2} \quad , \quad S_2^2=\frac{\sum_{i=1}^{n_2}(y_{2i}-\overline{Y_2})^2}{n_2-1}$$

$W_i=N_i/N \quad , i=1,2 \; ; N=N_1+N_2$ 이다.

위의 정리에 대한 증명은 $\overline{y_1}$와 $\overline{y_2}$의 성질과 그 분산의 식을 이용하면 쉽게 이루어질 수 있다. 패널조사에서 모집단중 일부의 구성이 변경될 때 그것을 프레임과 표본에 대해서도 각각 반영해준다면 위의 식을 사용하는 것이 효과적인 대처방안이 될 수 있을 것이다.

4. 결 론

대부분의 패널조사에서 모집단의 구성은 시간이 지남에 따라 달라지게 된다. 물론 모집단의 성격에 따라 그 변화가 거의 미미한 경우도 있고 그렇지 않은 경우도 있다. 변화가 어느 정도 심각한 경우 그에 따라 적절히 대응하지 않는다면 과거의 프레임에 의해 얻어진 표본을 통해 계속조사를 하는 것은 편의를 초래하게 된다. 실제 많은 조사에서 모집단의 변화를 표본에도 반영시켜 가고 있다. 그러나 그러한 표본의 변경이 추정과정에서는 제대로 반영되지 못하고 있다는 사실은 주목할 만한 일이다.

본 연구에서는 패널조사에서 모집단구성의 변화에 따라 적절히 프레임 및 표본을 변화시켜갈 경우 이러한 표본의 변경을 고려하지 않고 종래의 추정량을 그대로 사용한다면 편의가 초래될 수 있음을 밝혔다. 한편 이러한 편의를 제거해주는 수정된 불편추정량을 제시하였고 그 분산의 식을 유도하였다.

참고문헌

[1] 김규성, 전종우, 박홍래 (1994). 어가경제조사 표본설계에 관한 연구, 「응용통계연구」, 제8권 2호, 43-54.

[2] 박홍래 (1989). 면적조사및 생산량조사 표본설계, 「박홍래교수 회갑기념논총」, 55-74.

[3] 이기재, 박진우, 박홍래 (1991). 전국 도시 주택가격 동향조사를 위한 표본설계 연구, 「응용통계연구」, 제4권 2호, 137-148.

[4] Bailor, R. A.(1989). Information needs, surveys, and measurement errors. *Panel Surveys*, (Eds. Kasprzyk, D. , Duncan, G., Kslton, G. and Singh, M.P.), New York, John Wiley, 1-24.

[5] Colledge, M. J.(1989). Coverage and classification maintenance issues in economic surveys, *Panel Surveys*, (Eds. Kasprzyk, D. , Duncan, G., Kslton, G. and Singh, M.P.), New York, John Wiley, 80-107.

[6] Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample survey methods and theory*, Vol. 2, Wiley, New York.

[7] Kalton, G. and Citro, C. F. (1993). Panel surveys: adding the fourth dimension, *Survey Methodology*. vol. 19, No. 2, 205-215.

[8] Lessler, J. T. and Kalsbeek, W. D. (1992). *Nonsampling errors in surveys*, John Wiley & Sons, New York.

An Estimation Procedure with Updated Sample in Panel Survey

Park, Jin-Woo[2)]

Abstract

In panel surveys it is necessary to manage both sampling frame and sample units across time. When sample is updated according to the change of its frame, it should be incorporated in the estimation procedure. This paper derives the bias of the conventional estimator caused by neglecting the change of sample, and provides a bias-adjusted estimator with its variance.

2) Department of Applied Statistics, Suwon University, Hwasung, Kyunggi, 445-743, Korea.