한국통계학회논문집
2012, 19권, 1호, 23–32

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.1.023

포함확률비례추출에서 회귀계수 최소제곱추정량의 근사분산

김규성[1,a]

[a]서울시립대학교 통계학과

요 약

본 논문은 유한모집단에서 회귀계수추정량의 근사편향과 근사분산을 다루고 있다. 유한모집단에서 고정크기 포함확률비례표본을 추출하고 이 표본에서 조사된 데이터에 기초하여 회귀계수를 일반최소제곱추정량과 가중최소제곱추정량으로 추정할 때 두 추정량의 편향, 분산 그리고 평균제곱오차의 근사식을 유도하였다. 그리고 두 추정량의 효율을 비교하기 위하여 두 추정량의 분산을 비교하는 필요충분조건을 제시하였다. 또한 수치적인 비교를 위하여 간단한 예제를 소개하였다.

주요용어: 가중최소제곱추정량, 근사분산, 근사편향, 일반최소제곱추정량, 포함확률비례추출.

1. 서론

표본조사 데이터에 회귀모형을 적합하는 문제를 고려하자. 관심변수를 y라고 하고 설명변수를 $(x_1, \ldots, x_p)$라고 하면 고려하는 회귀모형은 다음과 같다.

$$y_k = \beta_0 + \beta_1 x_{1k} + \cdots + \beta_p x_p + \epsilon_k, \quad k = 1, \ldots, n.$$

일반 회귀분석에서는 오차가 서로 독립이고 분포가 동일하다는 가정을 부여한 후 회귀계수 추정량으로 일반최소제곱추정량(ordinary least square estimator; OLSE)을 사용하는 것이 보통이다. 모형 가정이 타당하다면 일반최소제곱추정량은 비편향이면서 최소분산을 갖는 매우 좋은 추정량임이 알려져 있다 (Abraham과 Ledolter, 2006). 그런데 많은 경우 표본조사 데이터는 층화, 집락화 등을 통하여 생산된 복합데이터(complex data)이기 때문에 통상적인 회귀모형에서의 독립성, 동일 분포 가정을 만족시키지 않는다. 따라서 표본조사데이터를 모형 가정이 어긋나는 일반 회귀모형에 곧바로 적합하는 것은 올바른 분석이라고 하기 어렵다. 대신에 표본데이터에는 표본추출확률을 가중치에 활용하는 가중최소제곱추정법(weighted least square estimation method; WLSE)을 사용하는 것이 더 타당하다는 주장이 있다 (Skinner 등, 1989; Lohr, 1999; 등). 표본추출시 사용한 표본추출확률을 표본데이터 분석에 어떻게 사용해야 하는가 하는 문제는 표본조사 데이터 분석 방법론에서 매우 중요한 논쟁거리이며 이 분야에 대한 연구가 활발히 진행되고 있다 (Chambers와 Skinner, 2003).

한국복지패널데이터를 실증적으로 분석한 결과를 보면 일반최소제곱추정량의 상대편향이 가중최소제곱추정량의 상대편향보다 크게 나타나고 있다. 특히 표본의 수가 증가할수록 일반최소제곱추정량의 상대편향은 더 증가하는 것으로 나타났다 (김규성 등, 2009). 편향의 관점에서 보면 표본조사데이터 분석에서 가중최소제곱추정량을 사용하는 것이 일반최소제곱추정량을 사용하는 것보다 더 타당하다는 결론에 도달한다. 그러나 추정량의 타당성을 검토할 때에는 편향의 크기와 더불어 분산의 크기를 살펴보아야 하므로 두 추정량의 분산을 구하여 편향과 함께 비교하여야 한다. 이에 관하여 모의실험을

[1] (130-743) 서울시 동대문구 시립대길 13, 서울시립대학교 통계학과, 교수. E-mail: kskim@uos.ac.kr

통하여 표본수에 따른 두 추정량의 편향과 분산을 비교한 연구는 있으나 (김규성, 2010; 등) 이론적으로 두 추정량의 편향과 분산을 구하여 구체적으로 비교한 연구는 거의 없다. 그 이유는 두 추정량의 설계분산의 일반적인 표현은 가능하지만 구체적인 표본설계에서는 그 표현이 복잡하기 때문에 연구자들이 더 이상은 수리적인 접근을 하지 않기 때문으로 보인다.

본 연구는 유한모집단의 회귀계수 추정량으로 최소제곱추정량의 편향과 분산 그리고 평균제곱오차의 근사식을 유도하고자 한다. 이때 사용된 표집설계는 고정크기 포함확률비례추출법이다. 두 추정량의 편향, 분산 그리고 평균제곱오차의 근사식을 유도하고, 두 추정량의 분산의 크기를 비교하는 필요충분조건을 제시한다. 또한 간단한 예제를 통하여 두 추정량의 분산과 평균제곱오차를 수치적으로 비교한다.

본 연구는 다음과 같이 구성되어 있다. 제 2절에서는 포함확률비례추출에서 선형추정량의 분산의 근사식을 유도하고 비편향 분산추정량을 제안한다. 제 3절에서는 회귀계수의 일반최소제곱추정량의 근사편향과 근사분산, 그리고 가중최소제곱추정량의 근사분산을 유도한다. 또한 두 추정량의 분산을 비교하는 필요충분조건을 제시하고, 수치 비교를 위하여 간단한 예제를 소개한다. 마지막으로 제 4절에서는 논문의 내용을 요약하고 향후 연구 과제를 언급한다.

2. 포함확률비례추출에서 선형추정량의 근사분산

크기가 N인 유한모집단 $U = \{1, \ldots, N\}$를 고려하자. 모집단 U로부터 크기가 n인 표본 s를 비복원 추출하려고 한다. 표본추출법으로 포함확률 π_k가 개별단위 추출확률 p_k에 비례하면서 모든 포함확률은 양수가 되는 포함확률비례 확률추출법(Inclusion probability proportional to size probability sampling)을 고려하자. 여기에서 추출확률 p_k는 양수이며 $\sum_{k=1}^{N} p_k = 1$을 만족한다. 그러면 개개의 모집단 단위에서 다음의 식이 성립한다.

$$\pi_k = np_k > 0, \quad k = 1, \ldots, N, \ \sum_U \pi_k = n.$$

이제 포함확률비례표본 s에서 구한 선형추정량 $\hat{\theta}_s = \sum_s a_k$의 기댓값과 분산, 그리고 분산추정량을 구하자. 먼저 $\hat{\theta}_s$의 기댓값은 다음과 같다.

$$E\left(\hat{\theta}_s\right) = \sum_U a_k \pi_k = n \sum_U a_k p_k$$

그리고 $\hat{\theta}_s$의 분산은 다음과 같이 된다.

$$V\left(\hat{\theta}_s\right) = \sum_U \sum_U \Delta_{kl} a_k a_l. \tag{2.1}$$

또한 비편향 분산추정량으로 아래의 식을 고려하자.

$$v\left(\hat{\theta}_s\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} a_k a_l, \tag{2.2}$$

여기에서 $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l \ (k \neq l)$, $\Delta_{kk} = \pi_k(1 - \pi_k)$이다.

그런데 위의 분산과 분산추정량을 계산하기 위해서는 2차 포함확률 π_{kl}을 계산하여야 하는데, 이 확률을 실제 수치로 계산하는 것은 대부분의 경우에 간단한 일이 아니다. 표본을 포함확률비례추출로 선정할 때 추정치를 계산한다 하더라도 2차 포함확률 π_{kl}를 계산하지 못하여 분산 추정치를 구하

지 못하면 추론을 계속하기 어렵다. 이 경우 정확한 분산 대신에 근사분산과 근사분산 추정치를 구하는 것이 하나의 대안이 될 수 있다. 2차 포함확률 π_{kl}에 대한 다음과 같은 근사를 고려해보자 (Asok와 Sukhatme, 1976).

$$\pi_{kl} \approx n(n-1)p_k p_l.$$

만일 이 같은 근사를 하면 $\Delta_{kl}\,(k \neq l)$과 Δ_{kl}/π_{kl}은 매우 간단하게 표현된다.

$$\Delta_{kl} = \pi_{kl} - \pi_k\pi_l \approx n(n-1)p_k p_l - n^2 p_k p_l = -np_k p_l,$$
$$\frac{\Delta_{kl}}{\pi_{kl}} \approx \frac{-np_k p_l}{n(n-1)p_k p_l} = -\frac{1}{n-1}.$$

이를 이용하여 다음의 결과를 얻는다.

정리 1. 크기 N인 유한모집단에서 크기가 n인 표본을 고정크기 포함확률비례 확률추출한다고 하고, 2차 포함확률은 $\pi_{kl} \approx n(n-1)p_k p_l\,(k \neq l)$로 근사한다고 하자. 그리고 선형추정량 $\hat{\theta}_s = \sum_s a_k$을 고려하자. 여기에서 a_k는 조사단위 k에서 계산되는 값이다.

(i) 선형추정량 $\hat{\theta}_s = \sum_s a_k$의 분산은 아래의 식으로 근사할 수 있다.

$$V\left(\hat{\theta}_s\right) \approx \sum_U \pi_k\left(1 - \frac{n-1}{n}\pi_k\right)a_k^2 - \frac{1}{n}\left(\sum_U \pi_k a_k\right)^2. \tag{2.3}$$

(ii) 위의 근사분산은 아래의 식으로 비편향 추정할 수 있다.

$$v\left(\hat{\theta}_s\right) = \sum_s \left(\frac{n}{n-1} - \pi_k\right)a_k^2 - \frac{1}{n-1}\left(\sum_s a_k\right)^2. \tag{2.4}$$

증명:

(i) 식 (2.1)에 $\pi_{kl} \approx n(n-1)p_k p_l$를 대입하고 계산하여 정리하면 $V(\hat{\theta}_s)$의 근사식을 얻을 수 있다.

$$\begin{aligned} V\left(\hat{\theta}\right) &= \sum_U \Delta_{kk}a_k^2 + \sum_{k\neq l}\Delta_{kl}a_k a_l \\ &\approx \sum_U \pi_k(1-\pi_k)a_k^2 - n\sum_{k\neq l} p_k p_l a_k a_l \\ &= \sum_U \pi_k\left(1 - \frac{n-1}{n}\pi_k\right)a_k^2 - \frac{1}{n}\left(\sum_U \pi_k a_k\right)^2. \end{aligned}$$

(ii) 식 (2.2)의 분산추정량 $v(\hat{\theta}_s)$에 2차 포함확률의 근사값 $\pi_{kl} \approx n(n-1)p_k p_l$을 대입한 후 식을 정리하면 결과를 얻을 수 있다.

$$\begin{aligned} v\left(\hat{\theta}\right) &= \sum_s \frac{\Delta_{kk}}{\pi_{kk}}a_k^2 + \sum_{k\neq l}\frac{\Delta_{kl}}{\pi_{kl}}a_k a_l \\ &\approx \sum_s (1-\pi_k)a_k^2 + \sum_{k\neq l}\left(1 - \frac{n^2 p_k p_l}{n(n-1)p_k p_l}\right)a_k a_l \\ &= \sum_s \left(\frac{n}{n-1} - \pi_k\right)a_k^2 - \frac{1}{n-1}\left(\sum_s a_k\right)^2. \end{aligned}$$

□

3. 회귀계수 최소제곱추정량의 근사분산

3.1. 최소제곱추정량의 근사

크기 N인 유한모집단에서 각각의 모집단 단위는 관심변수 y와 설명변수 $(x_1, \ldots, x_p)$의 값을 가지고 있다고 하자. $U = \{(y_k, x_{k1}, \ldots, x_{kp}) : k = 1, \ldots, N\}$. 설명변수 $(x_1, \ldots, x_p)$로 관심변수 y를 적합하는 회귀모형을 생각하면 유한모집단 회귀계수는 다음과 같이 정의된다 (Sarndal 등, 1994, 191쪽).

$$B = T^{-1}t, \tag{3.1}$$

여기에서 $T = \sum_U \underline{z}_k \underline{z}_k'$, $t = \sum_U \underline{z}_k y_k$, $\underline{z}_k = (z_{k1}, z_{k2}, \ldots, z_{kq})'$이다. 만일 절편이 있는 회귀모형을 적합할 경우에는 $z_{k1} = 1$, $z_{kj+1} = x_{kj}$, $j = 1, \ldots, p$, $q = p + 1$이 되고, 절편이 없는 회귀모형을 적합할 경우는 $z_{kj} = x_{kj}$, $q = p$가 된다. 여기에서 T의 역행렬이 존재한다고 가정하자.

이제 유한모집단 회귀계수는 T와 t를 각각 추정한 후 식 (3.1)에 대입하여 구한 플러그인 추정량(plug-in estimator)을 고려하자.

$$\hat{B} = \hat{T}^{-1}\hat{t}. \tag{3.2}$$

만일 위의 식 (3.2)에서 T와 t를 추정할 때 설계비편향 추정량을 사용하면 통상적으로 알려진 가중최소제곱추정량을 얻을 수 있다.

$$\hat{B}_W = \hat{T}_W^{-1}\hat{t}_W, \tag{3.3}$$

여기에서 $\hat{T}_W = \sum_s \underline{z}_k \underline{z}_k'/\pi_k$, $\hat{t}_W = \sum_s \underline{z}_k y_k/\pi_k$이다.

만일 T와 t를 추정할 때 표본추출 과정을 고려하지 않고 조사 단위가 동일한 확률로 뽑혔다고 생각하면 T와 t를 $\hat{T}_O = N\sum_s \underline{z}_k \underline{z}_k'/n$, $\hat{t}_O = N\sum_s \underline{z}_k y_k/n$로 추정할 것이다. 이 값을 식 (3.2)에 대입하면 통상적으로 알려진 일반최소제곱추정량을 얻는다.

$$\hat{B}_O = \hat{T}_O^{-1}\hat{t}_O. \tag{3.4}$$

유한모집단 회귀계수 추정량 $\hat{B}$은 행렬 T의 역행렬을 포함하고 있기 때문에 추정량 $\hat{B}$의 정확한 편향 및 분산을 직접 구하기는 쉽지 않다. 대신 선형화를 통하여 근사 편향 및 근사 분산을 구하는 방법이 알려져 있다 (Sarndal 등, 1994, 194쪽).

$$\hat{B} \approx B + T^{-1}\left(\hat{t} - \hat{T}B\right). \tag{3.5}$$

위의 식 (3.5)에서 $\hat{T}$와 $\hat{t}$자리에 $\hat{T}_O$와 $\hat{t}_O$를 대입하면 일반최소제곱추정량의 근사값을 얻을 수 있다.

$$\begin{aligned}
\hat{B}_O &\approx B + T^{-1}\left(\sum_s \frac{\underline{z}_k y_k}{f} - \sum_s \frac{\underline{z}_k \underline{z}_k'}{f} B\right) \\
&= B + T^{-1}\sum_s \frac{\underline{z}_k\left(y_k - \underline{z}_k' B\right)}{f} \\
&= B + \bar{T}^{-1}\frac{1}{n}\sum_s \underline{u}_k,
\end{aligned} \tag{3.6}$$

여기에서 $u_k = \underline{z}_k E_k$, $E_k = y_k - \underline{z}_k' B$ 그리고 $\bar{T} = T/N$이다. E_k는 유한모집단에 회귀모형을 적합할 때 발생하는 잔차이다.

또한 식 (3.5)에 $\hat{T}_W$와 $\hat{t}_W$를 대입하면 가중최소제곱추정량의 근사값을 얻는다.

$$\begin{aligned}\hat{B}_W &\approx B + T^{-1}\left(\sum_s \frac{\underline{z}_k y_k}{\pi_k} - \sum_s \frac{\underline{z}_k \underline{z}_k'}{\pi_k} B\right) \\ &= B + T^{-1}\left(\sum_s \underline{z}_k \left(y_k - \underline{z}_k' B\right)\frac{1}{\pi_k}\right) \\ &= B + T^{-1}\sum_s \frac{\underline{u}_k}{\pi_k}. \end{aligned} \tag{3.7}$$

3.2. 일반최소제곱추정량의 근사편향과 근사분산

일반최소제곱추정량 $\hat{B}_O$의 근사편향과 근사분산, 그리고 분산추정량을 구해보자. 먼저 식 (3.6)을 이용하면 $\hat{B}_O$의 기댓값은 다음과 같이 근사됨을 알 수 있다.

$$\begin{aligned} E\left(\hat{B}_O\right) &\approx B + \bar{T}^{-1}\frac{1}{n}E\left(\sum_s \underline{u}_k\right) \\ &= B + \bar{T}^{-1}\frac{1}{n}\left(\sum_U \underline{u}_k \pi_k\right) \\ &= B + (\bar{T})^{-1}\bar{u}_W, \end{aligned}$$

여기에서 $\bar{u}_W = \sum_U \underline{u}_k p_k$로 $\underline{u}_k$의 추출확률 p_k에 대한 가중평균이다. 따라서 $\hat{B}_O$의 근사편향으로 다음을 얻는다.

$$E\left(\hat{B}_O\right) \approx \left(\bar{T}\right)^{-1}\bar{u}_W. \tag{3.8}$$

이제 $\hat{B}_O$의 근사분산을 구해보자. 근사식 $\hat{B}_O \approx B + T^{-1}(\sum_s \underline{u}_k/f)$으로부터 다음의 근사분산을 얻는다.

$$V\left(\hat{B}_O\right) \approx T^{-1} V\left(\sum_s \frac{\underline{u}_k}{f}\right) T^{-1}.$$

정리 1에서 $a_k = \underline{u}_k/f$로 치환한 후 위의 일반최소제곱추정량의 분산에 적용하면 다음의 결과를 얻을 수 있다.

$$\begin{aligned} V\left(\hat{B}_O\right) &\approx T^{-1}\left[\sum_U \pi_k\left(1 - \frac{n-1}{n}\pi_k\right)\left(\frac{\underline{u}_k \underline{u}_k'}{f^2}\right) - \frac{1}{n}\left(\sum_U \pi_k \frac{\underline{u}_k}{f}\right)\left(\sum_U \pi_k \frac{\underline{u}_k'}{f}\right)\right] T^{-1} \\ &= \frac{1}{n}\bar{T}^{-1}\left[\sum_U p_k\left(1 - (n-1)p_k\right)\left(\underline{u}_k \underline{u}_k'\right) - \bar{u}_W \bar{u}_W'\right]\bar{T}^{-1}. \end{aligned} \tag{3.9}$$

또한 일반최소제곱추정량 $\hat{B}_O$의 평균제곱오차의 근사식은 근사분산에 근사편향의 제곱을 더하여 얻는다.

$$\begin{aligned} \mathrm{MSE}\left(\hat{B}_O\right) &\approx v\left(\hat{B}_O\right) + B\left(\hat{B}_O\right) B\left(\hat{B}_O\right)' \\ &= \frac{1}{n}\bar{T}^{-1}\left[\sum_U p_k\left(1 - (n-1)p_k\right)\left(\underline{u}_k \underline{u}_k'\right) + (n-1)\bar{u}_W \bar{u}_W'\right]\bar{T}^{-1}. \end{aligned} \tag{3.10}$$

이제까지의 내용을 요약하면 다음과 같다.

보조정리 1. 크기 N인 유한모집단에서 크기가 n인 표본을 고정크기 포함확률비례 확률추출한다고 하자. 그리고 2차 포함확률은 $\pi_{kl} \approx n(n-1)p_k p_l$, $(k \neq l)$로 근사한다고 하자.

(i) 일반최소제곱추정량 $\hat{B}_O$은 편향추정량이며 근사편향은 식 (3.8)과 같이 주어진다.

(ii) $\hat{B}_O$의 근사분산은 식 (3.9)와 같다.

(iii) $\hat{B}_O$의 근사평균제곱오차는 식 (3.10)과 같다.

이제 근사분산 $V(\hat{B}_O)$을 추정하는 문제를 생각해보자. 만일 모집단 단위 k에서 $\underline{u}_k$가 측정 가능하다면 일반최소제곱추정량 $\hat{B}_O$의 근사분산의 비편향 추정량은 식 (3.9)와 정리 1을 이용하여 다음과 같이 구할 수 있다.

$$\begin{aligned} v^*\left(\hat{B}_O\right) &= T^{-1} v\left(\sum_s \frac{\underline{u}_k}{f}\right) T^{-1} \\ &= T^{-1}\left[\sum_s \left(\frac{n}{n-1} - \pi_k\right)\left(\frac{\underline{u}_k \underline{u}_k'}{f^2}\right) - \frac{1}{n-1}\left(\sum_U \frac{\underline{u}_k}{f}\right)\left(\sum_U \frac{\underline{u}_k'}{f}\right)\right] T^{-1} \\ &= \bar{T}^{-1} \frac{1}{n}\left[\frac{1}{n}\sum_s \left(\frac{1}{n-1} - p_k\right)\left(\underline{u}_k \underline{u}_k'\right) - \frac{1}{n-1}\left(\bar{u}_s\right)\left(\bar{u}_s'\right)\right] \bar{T}^{-1}, \end{aligned}$$

여기에서 $\bar{u}_s = \sum_s \underline{u}_k / n$이다.

그런데 v^*를 구성하는 항인 $\underline{u}_k = \underline{z}_k(y_k - \underline{z}_k' B)$는 미지의 모집단 회귀계수 B때문에 표본단위 k에서 측정 가능한 값이 아니다. 따라서 표본 s에서는 $\underline{u}_k$를 먼저 추정하여야 한다. $\underline{u}_k$의 추정은 일반최소제곱추정량 $\hat{B}_O$를 이용하여 $\hat{\underline{u}}_k = \underline{z}_k(y_k - \underline{z}_k' \hat{B}_O)$를 이용하는 것이 논리적으로 자연스럽다. 또한 $\bar{T}$도 표본 s에서 추정한 $\hat{\bar{T}}_O = \sum_s \underline{z}_k \underline{z}_k' / n$으로 추정한다. 따라서 실제 사용 가능한 $\hat{B}_O$의 근사분산의 추정량은 아래와 같다.

$$v\left(\hat{B}_O\right) = \hat{\bar{T}}_O^{-1} \frac{1}{n}\left[\frac{1}{n}\sum_s \left(\frac{1}{n-1} - p_k\right)\left(\hat{\underline{u}}_k \hat{\underline{u}}_k'\right) - \frac{1}{n-1}\left(\bar{\hat{u}}_s\right)\left(\bar{\hat{u}}_s'\right)\right] \hat{\bar{T}}_O^{-1}, \tag{3.11}$$

여기에서 $\bar{\hat{u}}_s = \sum_s \hat{\underline{u}}_k / n$이다.

3.3. 가중최소제곱추정량의 근사분산

가중최소제곱추정량 $\hat{B}_W$의 근사편향, 근사분산 그리고 분산추정량을 구해보자. 먼저 식 (3.7)로부터 $\hat{B}_W$는 $\hat{B}_W \approx B + T^{-1}\sum_s \underline{u}_k / \pi_k$로 근사됨을 알 수 있다. 이를 이용하면 $\hat{B}_W$의 근사 기댓값은 다음과 같이 표현된다.

$$E\left(\hat{B}_W\right) \approx B + T^{-1} E\left(\sum_s \frac{\underline{u}_k}{\pi_k}\right) = B + T^{-1}\sum_U \underline{u}_k.$$

그런데 여기에서 $\underline{u}_k = \underline{z}_k E_k = \underline{z}_k(y_k - \underline{z}_k' B)$이고 회귀모형에서 최소제곱추정량의 잔차 성질에 의하여 $\sum_U \underline{u}_k = \underline{0}$이므로 $E(\hat{B}_W) \approx B$가 된다. 즉, 가중최소제곱추정량 $\hat{B}_W$은 근사적으로 회귀계수 B를 비편향 추정한다.

식 (3.7)로부터 가중최소제곱회귀추정량 $\hat{B}_W$의 근사분산을 다음과 같이 얻는다.

$$V\left(\hat{B}_W\right) \approx T^{-1} V\left(\sum_s \frac{\underline{u}_k}{\pi_k}\right) T^{-1}.$$

이제 정리 1에서 $a_k = \underline{u}_k/\pi_k$로 치환한 후 가중최소제곱추정량의 분산계산에 적용하면 다음의 결과를 얻을 수 있다.

$$\begin{aligned} V\left(\hat{B}_W\right) &\approx T^{-1}\left[\sum_U \pi_k\left(1-\frac{n-1}{n}\pi_k\right)\left(\frac{\underline{u}_k\underline{u}_k'}{\pi_k^2}\right) - \frac{1}{n}\left(\sum_U \underline{u}_k\right)\left(\sum_U \underline{u}_k'\right)\right] T^{-1} \\ &= \frac{1}{n}\bar{T}^{-1}\left[\sum_U \frac{1}{N^2}\left(\frac{1}{p_k}-(n-1)\right)\underline{u}_k\underline{u}_k'\right]\bar{T}^{-1}. \end{aligned} \tag{3.12}$$

보조정리 2. 크기 N인 유한모집단에서 크기가 n인 표본을 고정크기 포함확률비례 확률추출한다고 하자. 그리고 2차 포함확률은 $\pi_{kl} \approx n(n-1)p_k p_l$, $(k \neq l)$로 근사한다고 하자. 그러면 가중최소제곱추정량 $\hat{B}_W$의 성질은 아래와 같다.

(i) 가중최소제곱추정량 $\hat{B}_W$는 회귀계수 B를 근사적으로 비편향 추정한다.

(ii) $\hat{B}_W$의 근사분산은 식 (3.12)와 같다.

앞서의 일반최소제곱추정량의 근사분산 추정량을 유도할 때와 마찬가지로 가중최소제곱추정량에서도 분산추정을 하기 위해서는 $\underline{u}_k$를 추정한 후 분산추정량을 구해야 한다. 가중최소제곱추정량 을$\hat{B}_W$ 이용하면 $\underline{u}_k$는 다음과 같이 추정할 수 있다.

$$\hat{\underline{u}}_k = \underline{z}_k\left(y_k - \underline{z}_k'\hat{B}_W\right)$$

또한 T는 추정량 $\hat{T}_W = \sum_s \underline{z}_k\underline{z}_k'/\pi_k$으로 추정 가능하다. 이제 정리 1의 분산추정식과 위의 $\hat{\underline{u}}_k$를 이용하여 분산추정식을 정리하면 다음을 얻는다.

$$\begin{aligned} v\left(\hat{B}_W\right) &= \hat{T}_W^{-1} v\left(\sum_s \frac{\hat{\underline{u}}_k}{\pi_k}\right)\hat{T}_W^{-1} \\ &= \hat{T}_W^{-1}\left[\sum_s\left(\frac{n}{n-1}-\pi_k\right)\frac{\hat{\underline{u}}_k\hat{\underline{u}}_k'}{\pi_k^2} - \frac{1}{n-1}\left(\sum_s \frac{\hat{\underline{u}}_k}{\pi_k}\right)\left(\sum_s \frac{\hat{\underline{u}}_k'}{\pi_k}\right)\right]\hat{T}_W^{-1}. \end{aligned} \tag{3.13}$$

3.4. 비교

가중최소제곱추정량의 근사분산을 일반최소제곱추정량의 근사분산, 근사평균제곱오차와 비교해보자. 먼저 식 (3.12)의 가중최소제곱추정량의 근사분산에서 식 (3.9)의 일반최소제곱추정량의 근사분산의 차이를 계산하면 다음과 같다.

$$V\left(\hat{B}_W\right) - V\left(\hat{B}_O\right) \approx \frac{1}{n}\bar{T}^{-1}\left(\sum_U \delta_k\underline{u}_k\underline{u}_k' + \bar{u}_W\bar{u}_W'\right)\bar{T}^{-1}. \tag{3.14}$$

또한 일반최소제곱추정량의 근사평균제곱오차에서 가중최소제곱추정량의 분산을 빼면 다음을 얻는다.

$$M\left(\hat{B}_O\right) - V\left(\hat{B}_W\right) \approx \frac{1}{n}\bar{T}^{-1}\left((n-1)\bar{u}_W\bar{u}_W' - \sum_U \delta_k\underline{u}_k\underline{u}_k'\right)\bar{T}^{-1}, \tag{3.15}$$

표 1: 가상 모집단

단위	1	2	3	4	5	6
x_k	3	4	5	6	7	8
y_k	3	12	14.5	18	28.5	32.5
p_k	0.066	0.066	0.167	0.167	0.267	0.267

여기에서 $\delta_k = (1/Np_k - Np_k)/N^2 + (n-1)(p_k^2 - 1/N^2)$이다.

위의 식에서 $\bar{T}$는 비음정치 행렬(non-negative definite matrix)이므로 두 추정량의 분산 차이는 식 (3.14)와 식 (3.15)의 우변의 중간 항에 의존한다.

정리 2. 크기 N인 유한모집단에서 크기가 n인 표본을 고정크기 포함확률비례 확률추출한다고 하자. 그리고 2차 포함확률은 $\pi_{kl} \approx n(n-1)p_k p_l$, $(k \neq l)$로 근사한다고 하자. 그러면 다음의 결과를 얻는다.

(i) $V(\hat{B}_W) - V(\hat{B}_O)$가 비음정치행렬일 필요충분조건은

$$\sum_U \delta_k \underline{u}_k \underline{u}_k' + \bar{u}_W \bar{u}_W' \tag{3.16}$$

가 비음정치행렬이 되는 것이다.

(ii) $M(\hat{B}_O) - V(\hat{B}_W)$가 비음정치행렬일 필요충분조건은

$$(n-1)\bar{u}_W \bar{u}_W' - \sum_U \delta_k \underline{u}_k \underline{u}_k' \tag{3.17}$$

가 비음정치행렬이 되는 것이다.

3.5. 예제

앞 절의 정리 2의 내용을 간단한 예제를 통하여 수치적으로 확인해 보자. 크기가 6인 유한모집단에서 (x_k, y_k, p_k)의 값이 표 1과 같다고 하자. 여기에서 x_k는 설명변수, y_k는 조사변수이고 p_k는 추출확률로서 y_k의 값이 클 때 큰 값을 갖으며 $\sum_{k=1}^6 p_k = 1$을 만족한다.

절편이 없는 회귀모형을 적합하면 추정하고자 하는 회귀계수는 $B = T^{-1}t = 3.502$이다. 이제 크기 n인 표본을 추출확률 p_k에 비례하는 포함확률비례 확률추출을 한다고 하자. 그리고 일반최소제곱추정량 $\hat{B}_O$과 가중최소제곱추정량 $\hat{B}_W$으로 회귀계수 B를 추정한다고 하자. 정리 2에 의하여 식 (3.16)과 식 (3.17)의 필요충분조건을 계산하면 다음과 같다.

(i) 식 (3.16): $A = \sum_U \delta_k \underline{u}_k \underline{u}_k' + \bar{u}_W \bar{u}_W' = \sum_U \delta_k u_k^2 + (\sum_U u_k p_k)^2$.

(ii) 식 (3.17): $B = (n-1)\bar{u}_W \bar{u}_W' - \sum_U \delta_k \underline{u}_k \underline{u}_k' = (n-1)\bar{u}_W^2 - \sum_U \delta_k u_k^2$.

여기에서 $u_k = x_k(y_k - x_k B)$이다. 표본수에 따라 A와 B를 계산하면 표 2를 얻을 수 있다.

표 2에서 알 수 있는 바는 다음과 같다. 첫째, 가중최소제곱추정량의 근사분산과 일반최소제곱추정량의 근사분산의 차이는 표본수에 따라 크기가 다르게 나타난다. 표본수가 2, 3일 때에는 가중최소제곱추정량의 근사분산이 일반최소제곱추정량의 근사분산보다 작고, 표본수가 4와 5일 때에는 반대현상이 나타난다. 둘째, 표본수에 관계없이 일반최소제곱추정량의 근사평균제곱오차가 가중최소제곱추정량의 근사분산보다 크게 나타난다.

표 2: 표본에 따른 $\hat{B}_W$와 $\hat{B}_O$의 분산 비교

표본수	A	B
2	−234.3	413.2
3	−24.5	292.9
4	118.5	239.3
5	235.9	212.4

4. 결론

본 연구에서는 유한모집단의 고정크기 포함확률비례 표본이 뽑혔을 때 회귀계수 추정량으로 일반최소제곱추정량의 편향과 분산 그리고 평균제곱오차의 근사식과, 가중최소제곱추정량의 분산의 근사식을 유도하였다. 유도된 근사식을 이용하여 두 추정량의 분산의 크기를 비교하는 필요충분조건을 제시하였다. 또한 간단한 예제를 통하여 두 추정량의 분산과 평균제곱오차를 수치적으로 비교하였다. 예제에서는 표본수가 작으면 ($n = 2, 3$) 가중최소제곱추정량의 근사분산이 일반최소제곱추정량의 근사분산보다 작게 나왔고, 표본수가 크면 ($n = 4, 5$) 가중최소제곱추정량의 근사분산은 일반최소제곱추정량의 근사분산보다 커지지만 근사평균제곱오차보다는 작게 나타났다. 이러한 현상은 포함확률비례표본에 회귀모형을 적합했기 때문에 생기는 것이다. 표본수에 따른 경향을 요약하면 표본수가 작으면 가중최소제곱추정량을 사용하는 것이 더 타당하고, 표본수가 커지면 가중최소제곱추정량의 분산은 일반최소제곱추정량의 분산보다는 커지지만 평균제곱오차보다는 작게 되는 경향이 있다.

본 논문에서는 가중최소제곱추정량과 일반최소제곱추정량의 근사분산에 대한 추정량을 제안하였다(식 (3.11)과 식 (3.13)). 그러나 제안된 분산추정량에 대한 성질은 고찰하지 못하였다. 분산추정량의 성질에 대한 고찰은 향후 연구과제로 남긴다.

참고 문헌

김규성 (2010). 복합패널 데이터에 기초한 최소제곱 패널회귀추정량의 설계기반 성질, <한국통계학회 논문집>, **17**, 515–525.

김규성, 이영민, 전병돈 (2009). 패널회귀모형에서 가중치를 활용한 회귀계수 추정, <2009년 제2회 한국복지패널 학술대회 논문집>, 413–426.

Abraham, G. and Ledolter, J. (2006). *Introduction to Regression Modeling*, Thompson.

Asok, C. and Sukhatme, A. K. (1976). On Sampford's procedure of unequal probability sampling without replacement, *Journal of the American Statistical Association*, **71**, 912–918.

Chambers, R. L. and Skinner, C. J. (2003). *Analysis of Survey Data*, Wiley.

Lohr, S. (1999). *Sampling: Design and Analysis*, Duxbury Press.

Sarndal, C. E., Swensson, B. and Wretman, J. (1994). *Model Assisted Survey Sampling*, Springer.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1989). *Analysis of Complex Surveys*, Wiley.

2011년 10월 접수; 2011년 11월 수정; 2011년 11월 채택

Approximate Variance of Least Square Estimators for Regression Coefficient under Inclusion Probability Proportional to Size Sampling

Kyu-Seong Kim[1,a]

[a]Department of Statistics, University of Seoul

Abstract

This paper deals with the bias and variance of regression coefficient estimators in a finite population. We derive approximate formulas for the bias, variance and mean square error of two estimators when we select a fixed-size inclusion probability proportional to the size sample and then estimate regression coefficients by the ordinary least square estimator as well as the weighted least square estimator based on the selected sample data. Necessary and sufficient conditions for the comparison of the two estimators in terms of variance and mean square error are suggested. In addition, a simple example is introduced to numerically compare the variance and mean square error of the two estimators.

Keywords: Approximate bias, approximate variance, inclusion probability proportional to size sampling, ordinary least square estimator, weighted least square estimator.

[1] Professor, Department of Statistics, The University of Seoul, Seoul 130-743, Korea. E-mail: kskim@uos.ac.kr