연구논문

층화이단표집 데이터에 기초한 회귀계수의 설계기반분산추정량

Design−Based Variance Estimator of Regression Coefficients Based on Stratified Two−Stage Sampling Data

김규성[a)]
Kyu−Seong Kim

본 연구에서는 표본조사 데이터를 이용한 회귀분석에서 회귀계수의 분산추정 문제를 다루었다. 표본조사 데이터는 층화, 집락화, 그리고 불균등 추출확률 등의 표본선정 효과를 내포하므로 이를 무시하고 통상적인 회귀분석을 실시하면 심각한 오류를 범할 수 있다. 이러한 오류를 피하는 대안 중의 하나가 설계기반분산추정량을 이용하는 것이다. 본 연구에서는 설계기반분산추정량의 일반적인 형태를 소개하고 층화이단표집에서 구체적인 형태를 제시한다. 그리고 설계기반분산추정량을 사용하는 것이 일반최소제곱분산추정량이나 가중최소제곱분산추정량을 사용하는 것보다 더 우수함을 보이기 위하여 한국복지패널 데이터를 이용하여 수행한 모의실험의 결과를 소개한다.

모의실험의 결과는 다음과 같다. 첫째, 회귀계수에 대한 일반최소제곱분산추정량과 설계기반분산추정량의 상대편향은 평균 −10% 정도로 나타났고, 상대편향의 폭은 설계기반분산추정량의 경우가 더 작았다. 반면 가중최소제곱분산추정량은 약 −50% 정도의 심각한 과소추정을 보였다. 둘째, 신뢰계수가 95%인 신뢰구간 추정에서 설계기반추정량은 평균 92%의 포함률을 보였고 표본수가 증가하면 95%에 접근하였다. 반면, 일반최소제곱추정량은 평균 84%, 가중최소제곱추정량은 평균 82%의 포함률을 보여 과소포함 문제가 심각하게 있는 것으로 나타났다. 비록 제한적인 모의실험의 결과이기는 하지만 위 결과는 복합조사 데이터를 이용한 회귀분석에서는 설계기반분산추정량을 사용하는 것이 더 좋은 결과를 낼 수 있음을 시사한다.

* 이 논문은 2012년도 서울시립대학교 교내학술연구비에 의하여 연구되었음.

a) 서울시립대학교 통계학과 교수 김규성.
E−mail: kskim@uos.ac.kr

주제어 : 가중최소제곱추정량, 상대편향, 설계기반확률가중추정량, 신뢰구간 포함률, 일반최소제곱추정량

This paper deals with the problem of variance estimation for regression coefficients based on complex survey data. Since survey data contains the effect of sample selection such as stratification, clustering, unequal selection probability, etc., it will be misleading in case of ignoring such effects. An alternative way of avoiding such an error is to use the design-based variance estimator. A general form of design-based variance estimator as well as a specific form under the stratified two-stage sampling design are presented. In addition, a simulation study using Korean welfare panel data is conducted to show that the design-based variance estimator is much better than both the ordinary least square variance estimator and the weighted least square variance estimator.

The simulation results are as follows. First, both the ordinary least square variance estimator and the design-based variance estimator showed about -10% relative bias with much smaller variation of the design-based variance estimator. In contrast, the weighted least square variance estimator showed much severer under-estimation by about 50% relative bias. Second, in terms of confidence interval estimation with 95% level, the design-based estimator showed about 92% coverage rate with increase up to 95% as sample sizes increase. But the other two estimators showed under-coverage rates with 84% of the ordinary least square estimator and with 82% of the weighted least square estimator. Even in the basis of a limited simulation study, it may be said that the design-based variance estimator could give a better performance in regression analysis based on the complex survey data.

Key words: design-based p-weighted estimator, ordinary least square estimator, rate of confidence interval, relative bias, weighted least square estimator

Ⅰ. 서 론

표본조사 데이터를 이용한 회귀분석에서 회귀계수를 추정하는 문제를 고려하자. 표본조사 데이터는 층화, 집락화, 그리고 불균등 추출확률과 같은 표본선정 효과를 내포한다. 그러므로 이러한 효과를 무시하고 일반최소제곱추정량(Ordinary Least Square Estimator, OLSE)이나 가중최소제곱추정량(Weighted Least Square Estimator, WLSE)과 같은 통상적인 회귀계수 추정량을 사용하면 심각한 오류를 범할 수 있다. 이러한 오류를 피하는 대안방법 중에서 대표적인 것이 설계기반확률가중추정량(Design-Based p-weighted Estimator, PLSE)을 이용하는 것이다(Skinner et al. 1989: 154; Sarndal et al. 1992: 190; Lohr S. 2010: 442 등). 표본조사 데이터를 분석할 때에는 표본선정 효과를 고려한 분석법을 사용해야 한다는 것이 조사통계학(survey statistics) 연구자들의 일반적인 견해다(대표적으로 Skinner et al. 1989; Chambers et al. 2003; Heeringa et al. 2010 등).

복합조사 데이터에 기초한 회귀계수 추정에 관한 선행연구에 의하면 일반최소제곱추정량은 설계편향이 있고 이로 인하여 평균제곱오차가 설계기반확률가중추정량의 평균제곱오차보다 더 크게 나타나는 경향이 있다(김규성 2010). 이에 더하여 본 연구에서는 회귀계수 추정량의 분산추정에서도 설계기반추정량을 사용하는 것이 더 효과적일 수 있음을 보이고자 한다. 이를 보이기 위하여 한국복지패널 데이터를 이용한 모의실험을 실시한다.

본 연구는 다음과 같이 구성되어 있다. 제2절에서는 유한모집단 회귀계수를 정의하고 일반적인 확률표본에서 일반최소제곱추정량, 가중최소제곱추정량, 그리고 설계기반확률가중추정량의 분산추정량의 형태를 각각 수식으로 표현한다. 제3절에서는 복합표본조사에서 나타나는 설계기반분산추정량의 일반적인 형태를 소개하고, 층화이단표집에서 유도되는 구체적인 형태를 제시한다. 또한 제시된 분산추정량에 대하여 계산이 더 손쉬운 대안 추정량을 설명한다. 제4절에서는 앞서 언급한 세 종류의 분산추정량의 성질을 수치적으로 비교하기 위한 모의실험을 수행한다. 한국복지패널 3개년 표본데이터를 유한모집단으로 간주하고 이 모집단에서 확률비례표본을 표본수 50에서 600까지 바꾸어 가며 추출한 후 세 종류의 회귀계수 추정량과 분산추정량을 계산하고

비교한다. 마지막으로 제5장에서는 앞서의 연구 내용을 요약하고 정리한다.

II. 회귀계수 분산추정량

1. 유한모집단 회귀계수

유한모집단 $U=\{1,...,N\}$을 고려하고 조사변수를 $(y,x_1,...,x_p)$라고 하면 유한모집단 단위들의 값은 다음과 같이 된다.

$$\{(y_k,x_{k1},...,x_{kp}),k=1,...,N\} \tag{2.1}$$

이를 행렬 $X_0=(x_{kj})_{N\times p}$과 벡터 $\underline{y}=(y_k)_{N\times 1}$로 표현하자. 그리고 행렬 X_0에 $(N\times 1)$차원의 $\underline{1}=(1,...,1)'$ 벡터를 더한 행렬을 $X=(\underline{1}\ X_0)_{N\times(p+1)}$라고 하자. 그러면 유한모집단 회귀계수 B는 다음과 같이 정의된다.

$$B=\begin{pmatrix}B_0\\B_1\\...\\B_p\end{pmatrix}=(X'X)^{-1}X'\underline{y} \tag{2.2}$$

여기에서 행렬 $(X'X)$의 역행렬이 존재하는 것으로 가정한다.

식(2.2)의 유한모집단 회귀계수 B는 아래의 회귀모형

$$y=\beta_0+\beta_1x_1+...+\beta_px_p+\epsilon,\ \epsilon\sim(0,\sigma^2) \tag{2.3}$$

에서 식(2.1)의 데이터가 생성되었다고 할 때 무한모집단 회귀계수 $\underline{\beta}=(\beta_0,\beta_1,...,\beta_p)'$를 최적으로 추정하는 추정량으로 해석할 수 있다(예를 들면, Heeringa et al. 2010, p.184 등).

확률표집설계 $p(.)$에 의하여 크기 n인 확률표본 $s_n=\{i_1,...,i_n\}$을 선정한다고 하자. 그리고 조사 및 응답과정에서 생성된 가중치 w_k를 조사단위 k에 부여하여 다음과 같은 확률 데이터를 확보한다고 하자.

$$\{y_k, \underline{x}_{k,0}, w_k : k \in s_n\}$$

여기에서 $\underline{x}_{k,0} = (x_{k1}, ..., x_{kp})'$는 조사단위 k의 데이터 벡터이다. 확률 데이터를 행렬

$$X_{s_n,0} = (x_{kj}, k \in s_n)_{n \times p},\ X_{s_n} = (\underline{1}_n, X_{s_{n,0}})_{n \times (p+1)},\ \underline{y}_{s_n} = (y_k, k \in s_n)_{n \times 1}$$

로 표현하자. 그리고 가중치 $w_k,\ k \in s_n$,는 대각행렬 $W_{s_n} = diag\{w_k,\ k \in s_n\}_{n \times n}$ 로 나타내자.

2. 회귀계수 추정량의 분산추정량

식(2.3)의 회귀모형에서 표본이 독립적이고 동일한 확률로 얻어졌다고 가정하면 회귀계수 $\underline{\beta}$는 일반최소제곱추정량으로 추정될 것이다.

$$\hat{\beta}_o = (X_{s_n}{}' X_{s_n})^{-1} X_{s_n}{}' \underline{y}_{s_n}$$

그리고 이에 대응하는 분산추정량은

$$v_o(\hat{\beta}_o) = (X_{s_n}{}' X_{s_n})^{-1} \widehat{\sigma_o^2} \tag{2.4}$$

이다. 여기에서 $\widehat{\sigma_o^2} = \sum_{k \in s_n} (y_k - \hat{y}_k)^2 / (n-p-1)$이고, $\hat{y}_k = \underline{x}_k{}' \hat{\beta}_o$, $\underline{x}_k{}' = (1, \underline{x}_{k,0}{}')$ 이다(Abraham et al. 2006: 97 등).

만일 회귀모형에서 조사변수 y의 분산이 서로 다르다고 가정하면

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon,\ \epsilon \sim (0, \sigma_w^2)$$

여기에서 $\sigma_w^2 = \sigma^2 v$이고 σ^2는 미지의 상수, v는 알려진 값으로 가정, 회귀계수 $\underline{\beta}$는 가중최소제곱추정량

$$\hat{\beta}_w = (X_{s_n}{}' V_{s_n}^{-1} X_{s_n})^{-1} X_{s_n}{}' V_{s_n}^{-1} \underline{y}_{s_n}$$

으로 추정할 수 있다. 여기에서 $V_{s_n} = diag\{v_k,\ k \in s_n\}_{n \times n}$이고 v_k는 조사단위 k에 대응하는 v값이다. 그리고 추정량 $\hat{\beta}_w$에 대응하는 분산추정량은

$$v_w(\hat{\beta}_w) = (X_{s_n}{}' V_{s_n}^{-1} X_{s_n})^{-1} \widehat{\sigma_w^2} \tag{2.5}$$

이다. 여기에서 $\widehat{\sigma_w^2} = \sum_{k \in s_n} (y_k - \hat{y}_k)^2 / [v_k(n-p-1)]$이고, $\hat{y}_k = \underline{x}_k{}' \hat{\beta}_w$ 이다(Abraham et al. 2006, p.129 등).

마지막으로 가중값를 이용한 회귀계수 추정량을 고려하자. 회귀계수 $\underline{\beta}$의 설계기반 확률가중추정량은

$$\hat{\beta}_p = (X_{s_n}{}' W_{s_n} X_{s_n})^{-1} X_{s_n}{}' W_{s_n} \underline{y}_{s_n}$$

이고, 이에 대응하는 분산추정량은

$$v_p(\hat{\beta}_p) = (X_{s_n}{}' W_{s_n} X_{s_n})^{-1} v(\sum_{s_n} w_k \hat{u}_k)(X_{s_n}{}' W_{s_n} X_{s_n})^{-1} \tag{2.6}$$

가 된다. 이때 분산식 (2.6)의 가운데 항의 일반적인 표현은 다음과 같다.

$$v(\sum_{s_n} w_k \widehat{u_k}) = \sum_{s_n} \sum_{s_n} \frac{\Delta_{kl}}{\pi_{kl}} (w_k \hat{u}_k)(w_l \hat{u}_l{}') \tag{2.7}$$

여기에서 $\hat{u}_k = \underline{x}_k{}'(y_k - \underline{x}_k{}' \hat{\beta}_p)$이다(Sarndal et al. 1992:194 등).

이제까지 논의한 분산추정량 중 식(2.4)의 일반최소제곱분산추정량은 표본데이터만 주어지면 계산이 가능한 반면, 식(2.5)의 가중최소제곱분산추정량을 계산하기 위해서는 오차분산에 포함된 V_{s_n}을 사전에 알아야 한다. 오차분산의 V_{s_n}항에 대한 다양한 논의는 Valliant et al.(2000) 등에서 이루어지고 있다. 세 번째로 식(2.6)의 설계기반 분산추정량의 값을 얻기 위해서는 가운데 항인 식(2.7)의 분산추정량을 계산하여야 한다. 식(2.7)의 분산추정량의 형태는 매우 일반적인 형태이므로 개별 표집설계에서는 구체적으로 식을 더 전개하여야 한다.

다음 절에서는 대규모 표본조사에서 널리 쓰이는 층화이단표집을 전제한 후 설계기반분산추정량의 구체적인 형태를 소개한다.

III. 층화이단표집에서의 설계기반분산추정량

1. 층화이단표집

층화이단표집을 고려하자. 모집단 U는 H층으로 구성되고, $U=\bigcup_{h=1}^{H} U_h$, 여기에서 U_h는 h번째 층, U_h는 N_h개의 1차 추출단위(Primary Sampling Unit, PSU)로 구성된다고 하며, (hi)번째 PSU는 M_{hi}개의 원소로 구성된다고 하자. 층h에서 표집설계 $p_h(.)$에 의하여 뽑은 PSU 표본을 s_h라고 하고 크기는 n_h라고 하자. 또한 PSU (hi)에서 표집설계 $p_{hi}(.)$에 의해 뽑은 표본을 s_{hi}라고 하고 표본 원소의 수는 m_{hi}라고 하자. 그리고 총표본은 $s_m = \cup_{h=1}^{H} \cup_{i \in s_h} s_{hi}$, 모집단 총수는 $M=\sum_{h=1}^{H}\sum_{i=1}^{N_h} M_{hi}$, 표본 총수는 $m=\sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi}$로 나타내자.

층화이단추출에서 조사변수를 $(y,x_1,...,x_p)$라고 하면 유한모집단 단위들의 값은 다음과 같이 표현할 수 있다.

$$(y_{hij,}x_{hij,1},...,x_{hij,p}),\ h=1,...,H,\ i=1,...,N_h,\ j=1,...,M_{hi}$$

이를 행렬 $X_0=\left(x_{k,i}\right)_{M\times q}=(\underline{x}_1,...,\underline{x}_p)$, $\underline{y}=\left(y_{hij}\right)_{M\times 1}$로 표현하자. 절편이 있는 회귀모형을 고려하여 행렬 X_0에 $\underline{1}$벡터를 추가한 확장된 행렬을 X로 나타내고,

$$X=(\underline{1},X_0)=(\underline{1},\underline{x}_1,...,\underline{x}_p)_{M\times(p+1)}$$

그리고 $\underline{x}'_{hij,0}=(x_{hij,1},...,x_{hij,p})$이라고 할 때, $x_{hij,0}$에 1을 더한 벡터를 $\underline{x}'_{hij}=(1,\underline{x}'_{hij,0})$로 나타내자. 그러면 유한모집단 회귀계수 B는 다음과 같이 쓸 수 있다.

$$B=\begin{pmatrix}B_0\\B_1\\\cdots\\B_p\end{pmatrix}=(X'X)^{-1}X'\underline{y}=(\sum_U x_{hij}x'_{hij})^{-1}(\sum_U x_{hij}y_{hij})$$

2. 회귀계수 분산추정량

유한모집단 회귀계수의 설계기반확률가중추정량 $\hat{\beta}_p$는 선형화를 통하여 선형식 $\tilde{\beta}_p$로 근사할 수 있다(Sarndal et al. 1992: 194 등).

$$\hat{\beta}_p \approx \tilde{\beta}_p = B+ T^{-1}\left(\sum_{s_m}\frac{\underline{x}_{hij}E_{hij}}{\pi_{hij}}\right) \tag{3.1}$$

여기에서 $E_{hij}=y_{hij}-x_{hij}'B$는 유한모집단 잔차이다. 만일 가중치가 $w_{hij}=1/\pi_{hij}$라고 하면, $\tilde{\beta}_p = B+ T^{-1}\left(\sum_{s_m} w_{hij}\underline{x}_{hij}E_{hij}\right)$가 된다. $\tilde{\beta}_p$의 분산은

$$V(\tilde{\beta}_p) = T^{-1}V\left(\sum_{s_m} w_{hij}\underline{x}_{hij}E_{hij}\right)T^{-1}$$

이다. 이제 층화이단표집에서 분산 $V\left(\sum_{s_m} w_{hij}\underline{x}_{hij}E_{hij}\right)$의 형태를 구하자. 만일

$$d_{hij}=x_{hij}E_{hij}=x_{hij}(y_{hij}-x'_{hij}B),\ t_{d,hi}=\sum_{s_{hi}} d_{hij},\ V_{d,hi}=V(\hat{t}_{d,hi}|s_{hi})$$

라고 하면, $w_{hij}=1/\pi_{hij}$이고 $\pi_{hij}=\pi_{hi}\pi_{j|hi}$이므로

$$\sum_{s_m} w_{hij}d_{hij}=\sum_{h=1}^{H}\sum_{i\in s_h}\frac{1}{\pi_{hi}}\left(\sum_{s_{hi}}\frac{d_{hij}}{\pi_{j|hi}}\right)$$

가 된다. 따라서

$$V(\sum_{s_m} w_{hij} d_{hij}) = \sum_{h=1}^{H} \left(V\left(\sum_{i \in s_h} \frac{t_{d,hi}}{\pi_{hi}} \right) + E\left(\sum_{s_h} \frac{V_{d,hi}}{\pi_{hi}^2} \right) \right)$$

$$= \sum_{h=1}^{H} \left(\sum_{U_h} \sum_{U_h} \Delta_{hij} \frac{t_{d,hi}}{\pi_{hi}} \frac{t_{d,hj}}{\pi_{hj}} \right) + \sum_{h=1}^{H} \sum_{U_h} \frac{V_{d,hi}}{\pi_{hi}}$$

를 얻고, 결과적으로 추정량 $\hat{\beta}_p$의 근사분산(Approximate Variance, AV)을 다음과 같이 얻는다.

$$A\,V(\hat{\beta}_p) = V(\tilde{\beta}_p) = T^{-1} \left[\sum_{h=1}^{H} \left(\sum_{U_h} \sum_{U_h} \Delta_{hij} \frac{t_{d,hi}}{\pi_{hi}} \frac{t_{d,hj}}{\pi_{hj}} \right) + \sum_{h=1}^{H} \sum_{U_h} \frac{V_{d,hi}}{\pi_{hi}} \right] T^{-1} \quad (3.2)$$

근사분산 $V(\tilde{\beta}_p)$의 추정량은 행렬 T와 식(3.2)의 우변의 가운데 항에 해당하는 공분산 행렬 $V(\sum_{s_m} w_{hij} d_{hij})$를 각각 추정한 후, 식(3.2)에 대입하여 얻을 수 있다. 먼저 행렬 T는 다음의 추정량으로 비편향 추정한다.

$$\hat{T} = \sum_{hij \in s_m} \frac{x_{hij} x'_{hij}}{\pi_{hij}}$$

또한 공분산 행렬 $V(\sum_{s_m} w_{hij} d_{hij})$는 d_{hij}를 먼저 $\hat{d}_{hij} = x_{hij}(y_{hij} - x'_{hij}\hat{B})$로 추정한 후 다음과 같이 추정한다.

$$v(\sum_{s_m} w_{hij} \hat{d}_{hij}) = \sum_{h=1}^{H} \left(\sum_{s_h} \sum_{s_h} \frac{\Delta_{h(ij)}}{\pi_{h(ij)}} \frac{\hat{t}_{\hat{d},hi}}{\pi_{hi}} \frac{\hat{t}'_{\hat{d},hj}}{\pi_{hj}} \right) + \sum_{h=1}^{H} \sum_{s_h} \frac{v_{\hat{d},hi}}{\pi_{hi}}$$

여기에서 $v_{\hat{d},hi}$는 분산 $V_{\hat{d},hi}$의 비편향추정량이다. 마지막으로 설계기반확률가중추정량 $\hat{\beta}_p$의 근사분산추정량은 추정량 $\hat{T}$와 $v(\sum_{s_m} w_{hij} \hat{d}_{hij})$를 결합하여 얻는다.

$$v_p(\tilde{\beta}_p) = (\hat{T})^{-1} v(\sum_{s_m} w_{hij} \hat{d}_{hij}) (\hat{T})^{-1} \quad (3.3)$$

식(3.3)에 나타난 분산추정량 $v_p(\tilde{\beta}_p)$은 일반적인 층화이단표집에 적용할 수 있는 장점이 있는 반면 2차 포함확률 $\pi_{h(ij)}$을 알고 있어야 하는 제약이 있다. 만일 분석자가 2차 포함확률 $\pi_{h(ij)}$을 알 수 없다면, 식 (3.3)을 이용한 분산추정은 원천적으로 불가능하기 때문이다. 예를 들어 대부분의 패널조사에서 조사 후 패널데이터와 가중치는 이용자에게 제공되지만 2차 포함확률까지 제공되지는 않기 때문에 패널데이터 분석자는 식 (3.3)을 이용한 분산추정을 할 수 없는 것이다. 이런 경우 분산추정을 위한 대안 방법이 필요해진다.

다음 소절에서는 2차 포함확률을 활용할 수 없는 경우 분산추정량 $v_p(\tilde{\beta}_p)$의 대안으로 사용할 수 있는 방법을 소개한다.

3. 층화 복원집락추출

만일 집락표본이 복원으로 뽑혔다고 간주하면 2차 포함확률 문제에서 벗어날 수 있다. 즉, 실제로는 집락을 비복원 추출하였지만 분산추정값은 복원추출을 간주하고 계산하는 것이다. 이렇게 하면 분산을 과대추정하기 때문에 추정의 효율을 다소 과소평가하는 문제는 발생하지만, 추정의 효율을 과대평가하는 오류를 범하지는 않는다. 본 소절에서는 $v_p(\tilde{\beta}_p)$의 대안 추정량으로서 층화 복원이단표집에서 구한 추정량을 소개한다.

층h에서 N_h개의 PSU 중 n_h개 PSU를 추출확률 p_{hi} ($\sum_{i=1}^{N_h} p_{hi} = 1$)에 의하여 복원으로 뽑는다고 하자. 그리고 표본집락 (hi)에서 뽑은 2차 추출단위 표본 s_{hi}는 앞 소절에서 설명한 것과 같다고 하자. 만일 복원표본을 $s_h^* = (i_1, ..., i_\nu, ..., i_{n_h})$로 나타내면 추정량 $\hat{t}_{\hat{d}}$는 다음과 같이 표현할 수 있다.

$$\hat{t}_{\hat{d}} = \sum_{h=1}^{H} \hat{t}_{\hat{d},h} = \sum_{h=1}^{H} \left(\frac{1}{n_h} \sum_{i \in s_h^*} \frac{\hat{t}_{\hat{d},hi}}{p_{hi}} \right)$$

여기에서 $\hat{t}_{\hat{d},hi} = \sum_{s_{hi}} \hat{d}_{hij} / \pi_{j|hi}$이다. 그리고 분산추정량은

$$v(\hat{t}_{\hat{d}}) = \sum_{h=1}^{H} v(\hat{t}_{\hat{d},h}) = \sum_{h=1}^{H} \left(\frac{1}{n_h(n_h-1)} \sum_{i \in s_h^*} (\frac{\hat{t}_{\hat{d},hi}}{p_{hi}} - \hat{t}_{\hat{d},h})(\frac{\hat{t}_{\hat{d},hi}}{p_{hi}} - \hat{t}_{\hat{d},h})' \right) \quad (3.4)$$

이다(Sarndal et al. 1992: 151 등).

분산추정량 $v(\hat{t}_{\hat{d}})$을 가중치를 이용하여 표현하자. 앞의 식에서

$$\hat{t}_{\hat{d},hi} = \sum_{s_{hi}} \frac{\hat{d}_{hij}}{\pi_{j|hi}}, \ \hat{t}_{\hat{d},h} = \sum_{i \in s_h^*} \frac{\hat{t}_{\hat{d},hi}}{n_h p_{hi}} = \sum_{i \in s_h^*} \sum_{j \in s_{hi}} \frac{\hat{d}_{hij}}{n_h p_{hi} \pi_{j|hi}}$$

이므로 만일 가중치 w_{hij}를 $w_{hij} = 1/n_h p_{hi} \pi_{j|hi}$라고 하면 위의 $\hat{t}_{\hat{d},hi}/p_{hi}$과 $\hat{t}_{\hat{d},h}$는 가중치를 이용하여 다음과 같이 쓸 수 있다.

$$\frac{\hat{t}_{\hat{d},hi}}{p_{hi}} = \sum_{s_{hi}} \frac{\hat{d}_{hij}}{p_{hi}\pi_{j|hi}} = n_h \sum_{s_{hi}} w_{hij} \hat{d}_{hij}, \ \hat{t}_{\hat{d},h} = \sum_{i \in s_h^*} \sum_{j \in s_{hi}} \frac{\hat{d}_{hij}}{n_h p_{hi} \pi_{j|hi}} = \sum_{s_h^*} \sum_{s_{hi}} w_{hij} \hat{d}_{hij}$$

또한 새로운 기호 e_{hij}를 도입하여 $e_{hij} = w_{hij} \hat{d}_{hij}$라고 하고, $\hat{e}_{hi.} \equiv \sum_{s_{hi}} \hat{e}_{hij}$, $\bar{\hat{e}}_h = \sum_{s_h^*} \hat{e}_{hi.}/n_h$라고 하자. 그러면 분산추정량 (3.4)는

$$v(\hat{t}_{\hat{d}}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i \in s_h^*} (\hat{e}_{hi.} - \bar{\hat{e}}_{h..})(\hat{e}_{hi.} - \bar{\hat{e}}_{h..})' \quad (3.5)$$

이 된다.

식(3.5)의 분산추정량에 비복원 효과 $(1-f_h)$를 고려하고, 또한 설명변수의 수 p를 보정한 수정된 분산추정량은 다음과 같다.

$$v_p(\hat{t}_{\hat{d}}) = \frac{n-1}{n-p} \sum_{h=1}^{H} \left(\frac{n_h(1-f_h)}{n_h - 1} \sum_{i \in s_h^*} (\hat{e}_{hi.} - \bar{\hat{e}}_{h..})(\hat{e}_{hi.} - \bar{\hat{e}}_{h..})' \right)$$

최종적으로 얻은 추정량 $\hat{\beta}_p$의 대안 분산추정량은 아래와 같다.

$$\tilde{v}_p(\hat{\beta}_p) = \hat{T}^{-1} v(\hat{t}_{\hat{d}}) \hat{T}^{-1} \tag{3.6}$$

참고로 식(3.6)의 분산추정량 $\tilde{v}_p(\hat{\beta}_p)$는 SAS의 Surveyreg의 모평균의 분산추정량과 동일하다(SAS/STAT 9.2, Survey Procedure: 6557).

IV. 모의실험

1. 유한모집단 및 표본추출

회귀계수 분산추정량의 성질을 수치적으로 비교하기 위하여 모의실험을 실시하였다. 모의실험에는 한국복지패널조사의 3개년(2005년~2007년) 표본데이터 중 5,634가구를 사용하였다(한국보건사회연구원 2006).

회귀모형 구축에 활용한 조사변수는 반응변수(y)로는 가구소득(단위: 만 원)을 사용하고, 설명변수(x)로는 균등화 소득에 따른 가구 구분(x_1, 1.일반가구, 2. 저소득가구), 가구원 수(x_2), 주택가격(x_3, 단위: 천만 원), 세금(x_4, 단위:만 원) 그리고 총생활비(x_5, 단위:만 원)를 사용하였다. 정량변수인 가구소득, 주택가격, 세금, 총생활비는 로그를 취한 후 분석에 사용하였다.

원래 한국복지패널가구는 층화이단표집을 통하여 선정되었는데 이용자에게 데이터가 제공될 때에는 층화변수, 집락변수, 집락크기 등은 제공되지 않고 가중치만 제공되었기 때문에, 본 모의실험에서 원 표본추출과 동일한 방법으로 모의실험 표본을 선정하는 것이 불가능하였다. 대신에 제공되는 가중치를 이용하여 패널가구를 확률비례추출하는 것을 고려하였다. 크기측도로는 한국복지패널 데이터에 포함되어 있는 연도별 횡단면 가구 가중값을 사용하였다. 표본크기에 따른 측도의 변화를 관찰하기 위하여 표본크기를 50, 100,…, 600의 12가지로 하였다. 3개년 데이터이므로 표본의 가지 수는 총 $3 \times 12 = 36$개다. 표본추출은 SAS의 Proc surveyselect 절차를 이용하되 표본추출 선택사항으로 pps_sys를 사용하였다. 각각의 표본수에서 반복수 $R = 4{,}000$의 독립표본을 선정하여 통계량을 계산하였다.

2. 회귀계수 및 분산추정치 계산

3개 연도별, 12개 표본수별 4,000개의 확률비례표본에서 일반최소제곱추정값($\hat{\beta}_o$, OLSE), 가중최소제곱추정값($\hat{\beta}_w$, WLSE), 설계기반확률가중추정값($\hat{\beta}_p$, PLSE)를 계산하고 각각에 대하여 표준오차 추정값($ste(\hat{\beta}_o)$, $ste(\hat{\beta}_w)$, $ste(\hat{\beta}_p)$)를 계산하였다. 절편을 포함하며 6개의 회귀계수가 있으므로 하나의 독립표본으로 총 $3\times 6=18$개의 회귀계수 추정값과 표준오차 추정값을 계산하였다.

일반최소제곱추정값과 표준오차는 SAS의 Proc Reg 절차를 이용하여 계산하였고 가중최소제곱추정값과 표준오차는 Proc Reg의 절차를 이용하되 오차분산의 가중값으로 포함확률의 역수를 사용하였다. 또한 설계기반확률가중추정값은 Proc surveyreg 절차를 사용하여 계산하였다. 본 모의실험에서는 가중최소제곱추정값과 설계기반최소제곱추정값 계산에 동일한 가중값을 사용했기 때문에 두 회귀계수 추정값은 동일하다. 그러나 분산추정식이 다르기 때문에 분산추정값은 서로 다르다.

3. 분산추정량의 평가지표

1) 상대편향

앞에서 계산한 절편을 제외한 5개 회귀계수 추정값에 대하여 대응하는 분산추정값과 95% 신뢰구간의 상한값과 하한값을 다음과 같이 계산하였다.

$$v_j^{(r)}=(ste_j(\hat{\beta}^{(r)}))^2,\ L_j^{(r)}=\hat{\beta}_j^{(r)}-1.96\times\sqrt{v_j^{(r)}},\ U_j^{(r)}=\hat{\beta}_j^{(r)}+1.96\times\sqrt{v_j^{(r)}}$$

여기에서 $j=1,...,5$는 회귀계수 첨자이고 $r=1,...,R(R=4000)$은 반복표본 첨자이다.

반복표본 $R=4{,}000$개에서 각각 계산한 회귀계수 추정값과 분산추정값을 이용하여 회귀계수 추정량의 분산과 분산추정량의 기대값을 다음과 같이 계산하였다.

- $V_j=\frac{1}{R}\sum_{k=1}^{R}(\hat{\beta}_j^{(r)}-\overline{\beta}_j)^2$: 회귀계수 추정량의 분산. 여기에서 $\overline{\beta}_j=\frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_j^{(r)}$
- $\overline{v}_j=\frac{1}{R}\sum_{r=1}^{R}v_j^{(r)}$: 회귀계수 분산추정량의 기대값

그리고 이를 이용하여 평가지표인 분산추정량의 상대편향을 계산하였다.

- 분산추정량의 상대편향(%) $= \dfrac{\bar{v}_j - V_j}{V_j} \times 100,\ j = 1,...,5$

2) 신뢰구간 포함률

분산추정량의 성능을 평가하는 또 다른 지표는 신뢰구간의 포함률이다. 분산추정량이 일치추정량이고 회귀계수 추정량의 분포가 정규분포 형태를 따른다면 명목확률과 신뢰구간의 포함률은 비슷하게 나타난다. 따라서 95% 신뢰구간인 $(L_j^{(r)}, U_j^{(r)})$이 위 조건을 만족한다면 $R = 4{,}000$번의 반복표본 중 95%에 해당하는 3,800개 정도의 신뢰구간이 회귀계수를 포함할 것으로 기대된다. 명목확률 95%에 가까운 횟수를 포함하는 신뢰구간이 더 좋은 신뢰구간이라고 평가할 수 있다.

지시변수 $C_j^{(r)}$, $D_j^{(r)}$, $E_j^{(r)}$을 도입하자. 만일 회귀계수가 신뢰구간 하한값보다 작으면 $C_j^{(r)} = 1$(그렇지 않으면 $C_j^{(r)} = 0$)의 값을 부여하고, 혹은 신뢰구간에 포함되면 $D_j^{(r)} = 1$ (그렇지 않으면 $D_j^{(r)} = 0$), 혹은 신뢰구간의 상한값보다 크면 $E_j^{(r)} = 1$ (그렇지 않으면 $E_j^{(r)} = 0$)의 값을 부여한다. 반복수 $R = 4{,}000$개의 독립표본에서 계산한 각각의 평균

$$\overline{C}_j = \frac{1}{R}\sum_{r=1}^{R} C_j^{(r)},\ \overline{D}_j = \frac{1}{R}\sum_{r=1}^{R} D_j^{(r)},\ \overline{E}_j = \frac{1}{R}\sum_{r=1}^{R} E_j^{(r)},\ j = 1,...,5$$

은 각각 [0, 신뢰구간 하한), [신뢰구간], (신뢰구간 상한, ∞) 구간의 포함률을 의미한다. 즉,

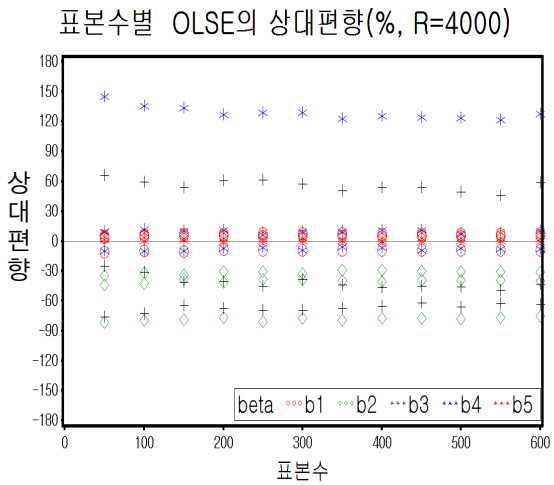
- $\overline{C}_j\ j = 1,...,5$: 신뢰구간 미만 구간 포함률
- $\overline{D}_j\ j = 1,...,5$: 신뢰구간 포함률
- $\overline{E}_j\ j = 1,...,5$: 신뢰구간 초과 구간 포함률

을 의미한다.

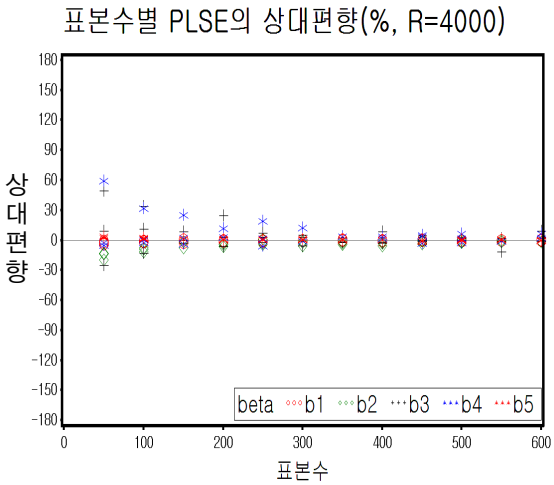
4. 모의실험 결과

1) 회귀계수 추정량의 상대편향

〈그림 1〉과 〈그림 2〉는 회귀계수의 일반최소제곱추정량과 설계기반확률가중추정량의 표본수별 상대편향을 보여준다. 일반최소제곱추정량의 상대편향이 설계기반확률가중추정량의 상대편향보다 크게 나타나고 있다. 특히 표본의 수가 증가하면서 일반최소제곱추정량의 상대편향은 완만하게 감소하는 반면, 설계기반확률가중추정량의 상대편향은 좀 더 빠르게 감소한다. 확률비례표본에서는 설계기반확률가중추정량을 사용하는 것이 더 타당하다는 하나의 수치적인 사례이며 기존의 연구결과와도 일치한다(김규성 2010).



〈그림1〉 OLSE의 상대편향



〈그림2〉 PLSE의 상대편향

2) 분산추정량의 상대편향

〈그림 3〉, 〈그림 4〉와 〈표 1〉은 3개년, 12가지 표본수, 그리고 5개 회귀계수의 조합인 총 180개의 경우에 대한 회귀계수 분산추정량의 상대편향을 보여주고 있다. 일반최소제곱분산추정량은 분산을 평균적으로 8% 과소추정하고 작게는 59% 과소추정, 많게는 135% 과대 추정하는 것으로 나타났다. 설계기반분산추정량은 평균적으로 분산을 10% 과소추정하고 작게는 40% 과소추정, 많게는 68% 과대추정하는 것으로 나타

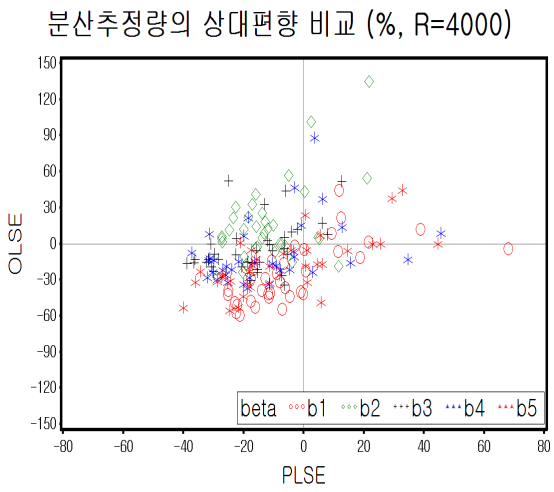
났다. 두 경우 모두 평균적으로 8%~10% 정도 분산을 과소추정하는 현상을 보이고 있고, 상대편향의 산포는 일반최소제곱분산추정량의 경우가 더 큰 것으로 나타났다. 그리고 가중최소제곱분산추정량은 분산을 심각하게 과소추정하는 것으로 나타났다. 180개 경우 모두 분산을 과소추정했고, 평균적으로 약 50%의 과소추정률을 보였다.

일반최소제곱분산추정량의 상대편향의 폭이 설계기반분산추정량의 폭보다 더 큰 것은 일반최소제곱분산추정량이 확률비례추출로 인한 표집의 효과를 설계기반분산추정량보다 덜 반영했기 때문인 것으로 풀이된다. 마찬가지로 가중최소제곱분산추정량도 확률비례추출 효과를 적절하게 반영하지 못했기 때문에 지나치게 분산을 과소추정하고 있는 것으로 보인다.

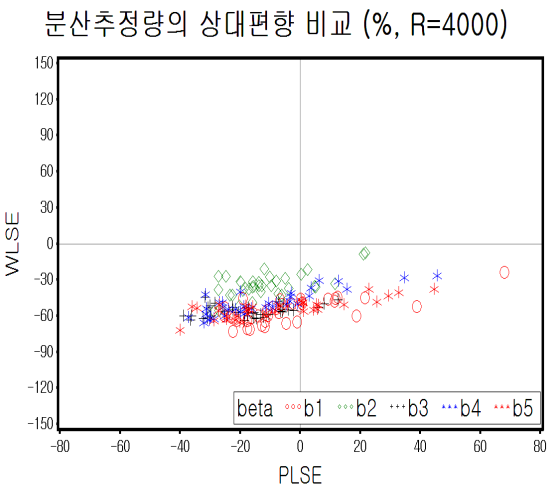
설계기반분산추정량은 다른 두 추정량보다는 양호한 수치를 보이고 있긴 하지만 평균적으로 약 10% 정도 분산을 과소추정하므로 이에 대한 검토가 추가적으로 필요하

〈표 1〉 분산추정량의 상대편향 분포

분산추정량	경우의 수	평 균	표준편차	최소값	최대값
OLSE	180	-8.080%	28.985	-59.420	134.948
WLSE	180	-49.987%	11.987	-73.017	-7.758
PLSE	180	-10.361%	17.497	-40.009	67.987



〈그림3〉 분산추정량의 상대편향1



〈그림4〉 분산추정량의 상대편향2

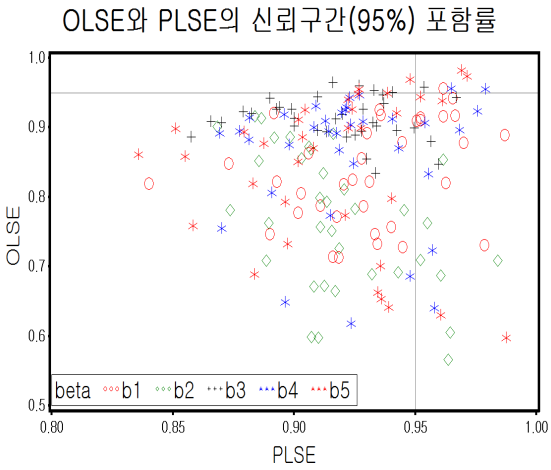
다. 설계기반분산추정량을 유도하는 과정 중 식(3.1)을 보면 설계기반확률가중추정량을 선형화하면서 잔여 항을 무시하고 있음을 알 수 있다. 이 과정에서 설계기반확률가중추정량의 변동성이 다소 축소된다. 만일 잔여 항을 제거하지 않았으면 잔여항의 변동성이 그대로 회귀추정량에 남아 있을 것이기 때문이다. 이 선형화 과정에서 설계기반 회귀추정량의 변동성이 축소되어 분산을 과소추정하는 현상이 나타나는 것으로 해석된다.

3) 신뢰구간 포함률

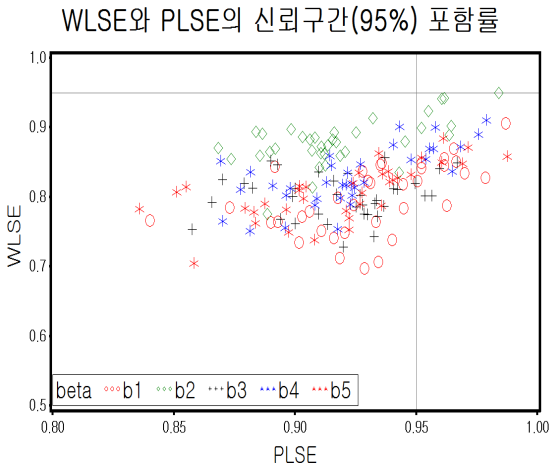
신뢰계수가 95%인 신뢰구간은 독립적으로 만든 R개의 표본 신뢰구간 중 약 95% 정도가 모수를 포함할 때 타당성을 인정받는다. 만일 95% 신뢰구간이 실제로는 95% 미만의 모수만을 포함한다면 이는 신뢰구간의 성능을 과대평가했다는 지적을 피할 수 없다. 본 모의실험에서도 이와 같은 신뢰구간 성능의 과대평가 현상이 나타났다. 일반최소제곱추정량의 신뢰구간은 84.4%, 가중최소제곱추정량의 신뢰구간은 82.3%, 그리고 설계기반확률가중추정량의 신뢰구간은 92.1%의 포함률을 보여 모두 명목 신뢰수준인 95%에 미치지 못하였다(<표 2> 참조). <그림 5>와 <그림 6>에서 보면 설계기반확률가중추정량의 신뢰구간은 95% 부근의 포함률을 보이고 있으나 90%가 안 되는 경우도 다수 있고, 일반최소제곱추정량과 가중최소제곱추정량의 신뢰구간은 과소포함

<표 2> 신뢰구간의 포함률

분산추정량	구 간	경우의 수	평균 포함률	표준편차	최소값	최대값
OLSE	신뢰구간 미만	180	0.068	0.099	0.000	0.402
	신뢰구간	180	0.844	0.098	0.566	0.982
	신뢰구간 초과	180	0.088	0.095	0.000	0.434
WLSE	신뢰구간 미만	180	0.089	0.031	0.023	0.176
	신뢰구간	180	0.823	0.050	0.697	0.950
	신뢰구간 초과	180	0.088	0.025	0.019	0.158
PLSE	신뢰구간 미만	180	0.037	0.023	0.000	0.130
	신뢰구간	180	0.921	0.031	0.836	0.988
	신뢰구간 초과	180	0.042	0.023	0.000	0.116



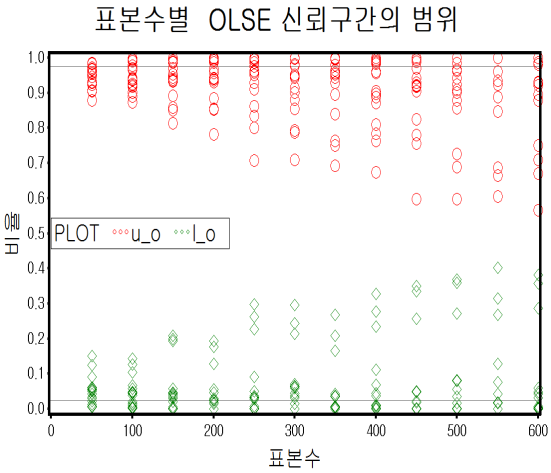
〈그림5〉 신뢰구간의 포함률1

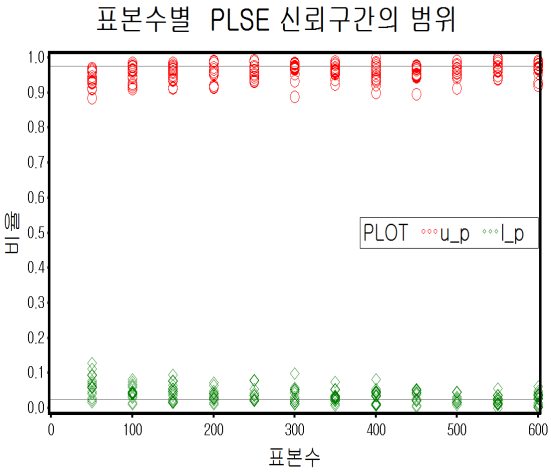


〈그림5〉 신뢰구간의 포함률2

률의 정도가 더 심하다. 또한 설계기반확률가중추정량의 신뢰구간과 비교하여 산포도 더 크다. 일반최소제곱추정량과 가중최소제곱추정량의 신뢰구간은 회귀계수를 과소포함할 뿐만 아니라 과소포함률의 산포도 더 큰 것으로 나타났다.

〈그림 7〉과 〈그림 8〉는 신뢰구간의 하한 미만 구간의 포함률과 상한 초과 구간의 포함률을 겹쳐 그린 신뢰구간 범위에 대한 그림이다. 그림에서 '1_o'는 일반최소제곱추정량의 하한 미만 구간의 포함률이고 'u_o'는 1에서 상한 초과 구간의 포함률을 뺀 값이다('1_p'와 'u_p'는 설계기반확률가중추정량에 대응하는 값이다). 신뢰구간의 포함률이 95%를 유지하고 회귀계수 추정량의 분포가 정규분포의 형태를 따른다면 하한 미만 구간의 포함률과 상한 초과 구간의 포함률은 2.5% 부근의 값을 갖는 그림이 그려질 것이다. 〈그림 7〉은 일반최소제곱추정량의 신뢰구간 범위에 대한 그림인데 표본수가 증가할수록 상한과 하한 모두 2.5%를 벗어나는 경우가 빠르게 증가하고 있다. 즉, 표본수가 증가할수록 신뢰구간의 포함률은 급격히 떨어지는 현상이 나타났다. 반면 설계기반확률가중추정량의 하한 미만 포함률과 상한 초과 포함률은 2.5% 부근에 있고, 표본수가 증가할수록 2.5%에 더 근접하는 것으로 나타났다. 두 그림을 비교해 보면 신뢰구간의 성능은 설계기반확률가중추정량의 경우가 일반최소제곱추정량의 경우보다 더 우수함을 알 수 있다. 가중최소제곱추정량의 신뢰구간의 범위에 대한 그림은 설계기반확률가중추정량의 그림과 유사하게 나타났지만 포함률이 설계기반최소제곱추정량보다 훨씬 더 떨어지는 것으로 나타났다(〈그림 8〉과 유사).





〈그림7〉 OLSE 신뢰구간 범위

〈그림8〉 PLSE 신뢰구간의 범위

V. 결론

표본조사 데이터를 이용한 회귀분석에서 조사통계학 연구자는 표집효과를 반영한 설계기반확률가중추정량 및 분산추정량을 사용하는 것이 더 타당하다는 견해를 가지고 있다. 회귀계수 추정에서는 기존의 연구에서 설계기반확률가중추정량의 상대편향이 일반최소제곱추정량의 상대편향보다 더 작음을 이론적인 입증과 함께 경험적인 수치를 보인 바 있다. 이에 더하여 본 연구에서는 회귀계수 분산추정에서도 설계기반분산추정량이 일반최소제곱분산추정량이나 가중최소제곱분산추정량보다 더 타당할 수 있음을 보이기 위하여 모의실험을 수행하였다.

모의실험이 보여주는 경험적인 결과는 다음과 같다. 첫째, 일반최소제곱분산추정량과 설계기반분산추정량의 상대편향은 평균적으로 10% 정도 과소추정하는 것으로 나타났고 상대편향의 폭은 설계기반분산추정량이 더 좁았다. 가중최소제곱분산추정량의 상대편향의 폭은 크지 않았으나 분산을 약 50% 정도 과소평가하여 과소추정의 정도가 지나친 것으로 나타났다. 둘째, 신뢰구간의 포함률에서 세 추정량의 포함률은 모두 명목 신뢰계수인 95%에 미치지 못했다. 그러나 설계기반추정량은 평균 92%의 포함률을 보였고 표본의 수가 증가할수록 명목계수 95%에 가까이 가는 현상이 나타난 반면, 일반최소제곱추정량의 신뢰구간 포함률은 평균 84%, 가중최소제곱추정량의 포

함률은 평균 82%로 나타났다. 또한 표본수가 증가하면 일반최소제곱추정량의 신뢰구간 포함률은 도리어 작아지는 현상이 나타났고, 가중최소제곱추정량의 신뢰구간 포함률은 12가지 경우의 표본수에서 지나치게 작게 나타나는 현상이 발견되었다. 이와 같은 사실을 종합하면 설계기반분산추정량을 사용하는 것이 일반최소제곱분산추정량이나 가중최소제곱분산추정량을 사용하는 것보다 더 효과적이라고 할 수 있다.

표본조사 데이터 분석자가 회귀분석을 실시하려고 할 때 일반최소제곱추정법, 가중최소제곱추정법, 혹은 설계기반확률가중추정법 중 어느 방법을 채택해야 하는지에 대한 명시적인 판단 기준이 있으면 매우 유용할 것이다. 향후 이 분야에 대한 연구가 더 진행되기를 기대한다.

참고문헌

김규성. 2010. “복합패널 데이터에 기초한 최소제곱추정 패널회귀추정량의 설계기반 성질.” ≪한국통계학회논문집≫ 17(4): 515-525.

한국보건사회연구원. 2006. ≪한국복지패널 1차 년도 조사자료 User's Guide≫.

Abraham, B. and J. Ledolter. 2006. *Introduction to Regression Modeling*. Thompson.

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

Heerings, S.G., B.T. West, and P.A. Berglund. 2010. *Applied Survey Data Analysis*. CRC Press.

Lohr, S. 2010. *Sampling: Design and Analysis*. 2nd Ed. Duxbury Press.

Sarndal, C.E., B. Swenson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer.

SAS Institute Inc. 2008. *SAS/STAT 9.2 User's Guide, the REG procedure(Book excerpt)*.

SAS Institute Inc. 2008. *SAS/STAT 9.2 User's Guide, the SURVEYREG Procedure(Book excerpt)*.

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

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference*. Wiley.

<접수 2013/10/31 , 수정 2014/1/8 , 게재확정 2014/1/24>