연구논문

패널회귀모형에서 회귀계수 추정량의 설계기반 성질*

Design−based Properties of Least Square Estimators in Panel Regression Model

김규성**
Kyu−Seong Kim

본 논문에서는 패널회귀모형에서 회귀계수 추정량으로 일반최소제곱추정량과 가중최소제곱추정량의 설계기반 성질을 고찰한다. 회귀계수의 최소제곱추정량을 선형화하여 일반최소제곱추정량의 근사편향, 근사분산, 그리고 근사평균제곱오차의 수식과, 가중최소제곱추정량의 근사분산 수식을 유도한 후, 모의실험을 통하여 두 추정량의 근사분산 및 근사평균제곱오차의 크기를 수치적으로 비교한다.

모의실험에서는 한국복지패널 3개년 데이터를 모집단으로 간주하고, 가구소득 변수를 관심변수로 하며 가구와 가구주 관련 7개 변수를 설명변수로 하는 유한모집단 회귀계수를 고려한다. 두 추정량의 설계기반 성질을 비교하기 위하여 표본수를 50에서 1,000까지 50 간격으로 설정하여 일반최소제곱추정량의 근사편향, 근사분산 그리고 가중최소제곱추정량의 근사분산을 계산한다. 모의실험을 통하여 다음과 같은 경향을 확인하였다. 첫째, 표본의 크기가 커지면 일반최소제곱추정량의 평균제곱오차가 가중최소제곱추정량의 분산보다 커진다. 둘째, 일반최소제곱추정량의 평균제곱오차를 가중최소제곱추정량의 분산으로 나눈 비(ratio)는 설명변수에 따라 크기가 다르게 나타나고, 일반최소제곱추정량의 편향이 클수록 큰 값을 보인다. 셋째, 분산만 비교하면 일반최소제곱추정량의 분산이 가중최소제곱추정량의 분산보다 대부분의 경우에 더 작게 나타난다.

주제어: 가중최소제곱추정량, 근사분산, 근사편향, 일반최소제곱추정량, 패널회귀모형

In this paper we investigate design−based properties of both the ordinary least square estimator and the weighted least square estimator for regression coefficients in panel regression model. We derive formulas of approximate bias, variance and mean square error for the ordinary least square estimator and approximate variance for the

* 이 논문은 2010년도 서울시립대학교 연구년교수 연구비에 의하여 연구되었음.

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

weighted least square estimator after linearization of least square estimators. Also we compare their magnitudes each other numerically through a simulation study.

We consider a three years data of Korean Welfare Panel Study as a finite population and take household income as a dependent variable and choose 7 exploratory variables related household as independent variables in panel regression model. Then we calculate approximate bias, variance, mean square error for the ordinary least square estimator and approximate variance for the weighted least square estimator based on several sample sizes from 50 to 1,000 by 50. Through the simulation study we found some tendencies as follows. First, the mean square error of the ordinary least square estimator is getting larger than the variance of the weighted least square estimator as sample sizes increase. Next, the magnitude of mean square error of the ordinary least square estimator is depending on the magnitude of the bias of the estimator, which is large when the bias is large. Finally, with regard to approximate variance, variances of the ordinary least square estimator are smaller than those of the weighted least square estimator in many cases in the simulation.

key words: approximate bias, approximate variance, ordinary least square estimator, panel regression model, weighted least square estimator.

Ⅰ. 서 론

패널조사에서 패널회귀모형(panel regression model)을 고려하면서 조사단위가 n개이고 t_0차 웨이브까지 조사를 수행하였다고 하자. 그리고 관심변수를 y라고 하고 설명변수를 $(x_1,...,x_p)$라고 하자. 패널조사에서는 시점, 설명변수에 따라 회귀계수가 다를 수 있으므로 패널회귀모형의 일반적인 형태는 다음과 같이 표현될 것이다(Hsiao 2003; Hill et al. 2008 등 참조).

$$y_{kt} = \beta_{0t} + \beta_{1t}x_{1kt} + ... + \beta_{pt}x_{pkt} + \epsilon_{kt}, \quad k = 1,\dots,n, \quad t = 1,...,t_0 \tag{1}$$

여기에서 ϵ_{kt}는 평균이 0이고 분산이 σ^2인 오차항이라고 하자. 만일 t_0개 차수의 웨이브를 병합한 데이터에 회귀모형을 적합하고자 한다면 다음과 같은 좀 더 간단한 모형을 고려할 수 있다.

$$y_{kt} = \beta_0 + \beta_1 x_{1kt} + \ldots + \beta_p x_{pkt} + \epsilon_{kt}, \quad k = 1,\ldots,n, \;\; t = 1,\ldots,t_0 \qquad (2)$$

예를 들어 한국복지패널조사(한국보건사회연구원 2008)에서 조사시점이 2006년, 2007년, 2008년인 데이터를 대상으로 패널가구의 경제활동 소득을 가구 변수와 가구주 변수를 활용하여 설명한다고 하자. 회귀모형으로는 모형 (1)과 모형 (2)를 사용하고 3개년도 모두 응답이 있으며 가구주가 일치하는 가구를 분석대상으로 제한하자. 또한 모형 적합을 위하여 극단 이상치(extreme outliers)는 분석에서 제외한다고 하자. 조사회수가 3회이므로 $t_0 = 3$이 된다. 이러한 과정을 거쳐서 통상적인 회귀분석을 하여 다음의 결과를 얻었다. 분석에서 적합도를 높이기 위하여 원점을 지나는 회귀모형을 선택하였고, 가구주 태어난 해는 원 데이터에서 1,900을 뺀 뒤 12로 나눈 값을 분석 데이터로 사용하였다. 분석에 사용된 데이터 수는 3개년도 각각 n=3,964이고, 병합데이터의 수는 n=11,892이다. 계산에는 통계패키지 SAS의 프러시져 중 PROC REG를 이용하였다.

2006년도 데이터 분석:
가구소득 = −1,112.0 × (균등화에 따른 가구구분) + 324.9×(가구원 수) + 130.1 × (가구형태) +197.5 × (가구주 태어난 해) + 309.0 × (가구주 교육수준) − 41.2 × (경제활동 구분) − 34.5 × (주거형태)

3개년도(2006년−2008년) 병합 데이터 분석:
가구소득 = −1,140.5 × (균등화에 따른 가구구분) + 378.8 × (가구원 수) + 135.8 × (가구형태) + 205.9 × (가구주 태어난 해) + 339.5 × (가구주 교육수준) − 56.6 × (경제활동 구분) − 57.2 × (주거형태)

네 개 모형 모두 적합이 잘 되었고 회귀계수도 모두 통계적으로 유의하였다. 2007년과 2008년 데이터를 대상으로 한 분석에서도 통계적으로 유의한 회귀모형을 얻었다.

통상적인 회귀분석 절차에 의하면 잔차 분석 등을 통하여 모형 가정의 타당성을 검토한 뒤, 일부 수정 혹은 보완하여 최종 모형을 얻는다. 그리고 연구자는 최종 모형을 설명하는 단계로 넘어간다. 예를 들어 회귀계수 크기를 보면 균등화에 따른 가구 구분이 −1,112.0 (2006년), −1,140.5 (3개년 병합)으로 가장 크다. 이 변수는 일반가구가 1의 값을 가지며 저소득층 가구가 2의 값을 가지므로 그 차이는 각각 1,112.0과 1,140.5이다. 일반가구와 저소득층 가구의 가구소득 차이가 대략 1,100만원이라는 설명이 가능하다.

다른 회귀계수들도 유사한 방법으로 설명이 가능하다.

위 분석은 일반최소제곱추정법(ordinary least square estimation method, OLSE)을 사용하여 구한 회귀추정치를 근거로 하고 있는데, 잘 알려진 바와 같이 일반최소제곱추정법은 서로 독립이고 분포가 동일한 오차를 갖는 데이터를 가정하고 이루어지는 분석법이다(예를 들어, Abraham & Ledolter 2006, 26쪽). 그런데 한국복지패널데이터는 사용자 설명서에 자세히 설명되어 있듯이 층화, 집락화, 무응답, 가중치 사후 조정 등을 통하여 얻어진 복합데이터(complex data)이므로 이러한 모형가정이 성립하지 않는다. 따라서 한국복지패널데이터에 일반 회귀모형을 그대로 적용하는 것은 올바른 분석이라 하기 어렵다. 이러한 복합데이터에는 가중최소제곱추정법(weighted least square estimation method: WLSE)을 사용하는 것이 더 타당하다는 주장이 있다(예를 들어, Skinner et al. 1989; Lohr 1999 등). 회귀추정량의 편향은 근사적으로 유도한 결과가 있으며(Sarndal 1994; 김규성 2010 등), 실증적으로 한국복지패널데이터를 분석한 결과 일반최소제곱추정량의 상대편향이 가중최소제곱추정량의 상대편향보다 크게 나타난다는 보고가 있다. 여기에서 상대편향은 독립표본을 3,000번 반복 추출하여 계산하였다 (김규성 외 2인 2009).

$$\text{표본수 } 100: \text{(OLSE의 상대편향)} \approx 4.61 \times \text{(WLSE의 상대편향)}$$
$$\text{표본수 } 200: \text{(OLSE의 상대편향)} \approx 9.71 \times \text{(WLSE의 상대편향)}$$

위의 결과에 의하면 표본의 수가 증가할수록 일반최소제곱추정량의 상대편향은 증가하는 것으로 나타났다. 따라서 가중최소제곱추정량이 일반최소제곱추정량보다 더 타당하다는 결론에 도달하기 쉽다. 그런데 추정량의 타당성을 검토할 때에는 편향의 크기와 더불어 분산의 크기를 살펴보아야 한다. 왜냐하면 편향이 존재하더라도 분산의 크기가 작으면 추정량의 평균제곱오차는 더 작아질 수 있기 때문이다. 즉, 편향추정량이라 하더라도 평균제곱오차가 더 작으면 사용 가능한 추정량으로 채택될 수 있기 때문이다.

본 연구에서는 앞의 연구(김규성 2010)에 이어 일반최소제곱추정량의 분산과 평균제곱오차를 가중최소제곱추정량의 분산과 비교한다. 이를 통하여 일반최소제곱추정량과 가중최소제곱추정량의 객관적인 비교를 한다. 본 논문은 다음과 같이 구성된다. 제 2절에서는 일반최소제곱추정량과 가중최소제곱추정량의 근사편향, 근사분산, 근사평균제곱오차의 수식을 이론적으로 유도한다. 제 3절에서는 한국복지패널데이터를 실증 분석하여 두 추정량의 표본수에 따른 편향, 분산, 평균제곱오차의 크기를 구하여 비교한다.

마지막으로 제 4장에서는 본 연구의 내용을 요약한다.

II. 회귀계수 추정량의 근사편향 및 근사분산

1. 유한 모집단 회귀계수 추정량의 선형화

유한모집단에서 회귀모형을 고려하자.

$$y_k = \beta_1 z_{k1} + ... + \beta_q z_{kq} + \epsilon_t, \quad k = 1,.,.,n$$

만일 원점을 지나는 회귀모형을 고려하면 $z_{k1} = 1$로 하면 된다. 그러면 유한모집단 회귀계수는 다음과 같이 정의 된다 (예를 들면, Sarndal et al. 1994, 191쪽).

$$B = T^{-1} t \tag{3}$$

여기에서 $T = \sum_U \underline{z}_k \underline{z}_k'$, $t = \sum_U \underline{z}_k y_k$, 그리고 $\underline{z}_k = (z_{k1}, z_{k2}, ..., z_{kq})'$이다.

유한모집단 회귀계수의 추정량은 회귀계수를 이루는 T와 t를 각각 추정한 후 식 (3)에 대입하여 얻을 수 있다.

$$\hat{B} = \hat{T}^{-1} \hat{t} \tag{4}$$

만일 위의 식 (4)에서 T와 t 추정에 설계비편향 추정량을 사용하면 통상적으로 알려진 식 (5)와 같은 가중최소제곱추정량을 얻는다.

$$\hat{B}_W = \widehat{T_W}^{-1} \widehat{t_W} \tag{5}$$

여기에서 $\widehat{T_W} = \sum_s \underline{z}_k \underline{z}_k' / \pi_k$, $\hat{t}_W = \sum_s \underline{z}_k y_k / \pi_k$, 그리고 π_k는 단위 k의 표본 포함확률이다. 반면 식 (4)에서 T와 t의 추정에 균등 표본 포함확률 $\pi_k = f = n/N$을 대입하면 일반최소제곱추정량을 얻을 수 있다. 여기에서 N은 모집단 크기이다.

$$\hat{B}_O = \widehat{T_O}^{-1} \widehat{t_O} \tag{6}$$

여기에서 $\widehat{T_O}=\sum_s \underline{z}_k\underline{z}_k'/f$, 그리고 $\hat{t}_O=\sum_s \underline{z}_k y_k/f$ 이다.

유한모집단 회귀계수의 추정량 $\widehat{B_W}$과 $\hat{B}_O$은 역행렬을 포함하고 있기 때문에 두 추정량의 편향 및 분산을 직접 구하기는 어렵다. 대신 선형화를 통하여 근사편향 및 근사분산을 구하는 것이 더 편리하다. 테일러 전개를 통하여 식 (4)의 최소제곱추정량을 선형화하면 다음과 같이 된다(Sarndal 1994, 194쪽).

$$\hat{B}\approx B+T^{-1}(\hat{t}-\hat{T}B)$$

위의 식에 일반최소제곱추정량과 가중최소제곱추정량을 대입하면 다음과 같은 결과를 얻는다.

$$\hat{B}_O\approx B+T^{-1}(\hat{t}_O-\hat{T}_O B) \tag{7}$$

$$\hat{B}_W\approx B+T^{-1}(\hat{t}_W-\hat{T}_W B) \tag{8}$$

2. 최소제곱추정량의 근사편향 및 근사분산

먼저 두 추정량의 편향을 구하자. $\hat{T}_W$와 $\widehat{t_W}$는 각각 T와 t의 비편향추정량이므로 추정량 $\widehat{B_W}$는 근사적으로 회귀계수B를 비편향 추정한다. 반면 $\hat{T}_O$와 $\widehat{t_O}$는 T와 t의 비편향추정량이 아니므로 추정량 $\widehat{B_O}$는 회귀계수 B의 비편향추정량이 아니다. 즉, $\widehat{B_O}$는 B의 편향추정량이다. 그리고 근사편향의 크기는 다음과 같이 구할 수 있다(김규성 2011).

$$B(\hat{B}_O)\approx\left(\frac{1}{N}\sum_U \underline{z}_k\underline{z}_k'\right)^{-1}\sum_U \underline{z}_k E_k p_k$$

여기에서 $E_k=y_k-\underline{z}_k'B$, $p_k=\pi_k/n$이다. 만일 $\overline{T}=\sum_U \underline{z}_k\underline{z}_k'/N$, $\underline{u}_k=\underline{z}_k E_k$, 그리고 $\tilde{u}_W=\sum_U u_k p_k$라고 하면 위의 근사편향은 다음과 같이 간단하게 표현된다.

$$B(\hat{B}_O)\approx(\overline{T}')^{-1}\tilde{u}_W \tag{9}$$

위의 식에서 보는 바와 같이 근사편향은 표본수 n의 영향을 받지 않는다.

이제 두 추정량의 근사분산을 구하자. 먼저 가중최소제곱추정량의 근사분산은 다음과 같다.

$$Var(\hat{B}_W) \approx \frac{1}{n}(\overline{T})^{-1}\left\{\sum_U \frac{1}{N^2}\left(\frac{1}{p_k}-(n-1)\right)\underline{u}_k\underline{u}_k'\right\}(\overline{T})^{-1} \quad (10)$$

그리고 일반최소제곱추정량의 근사분산은 다음과 같다.

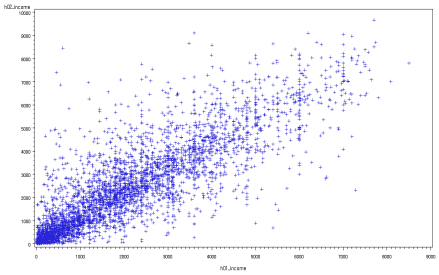
$$Var(\hat{B}_O) \approx \frac{1}{n}(\overline{T})^{-1}\left\{\sum_U p_k(1-(n-1)p_k)\underline{u}_k\underline{u}_k' - \tilde{u}_W\tilde{u}_W'\right\}(\overline{T})^{-1} \quad (11)$$

또한 일반최소제곱추정량의 평균제곱오차는 일반최소제곱추정량의 분산에 편향의 제곱을 더하여 얻는다.

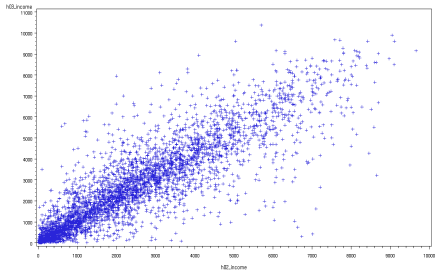
$$\begin{aligned} MSE(\hat{B}_O) &\approx V(\hat{B}_O) + B(\hat{B}_O)B(\hat{B}_O)' \\ &= \frac{1}{n}(\overline{T})^{-1}\left\{\sum_U p_k(1-(n-1)p_k)\underline{u}_k\underline{u}_k' + (n-1)\tilde{u}_W\tilde{u}_W'\right\}(\overline{T})^{-1} \end{aligned} \quad (12)$$

III. 모의실험

1. 유한모집단 회귀계수



〈그림 1〉 2006년 소득과 2005년 소득의 산점도



〈그림 2〉 2007년 소득과 2006년 소득의 산점도

모의실험에 사용한 데이터는 한국복지패널 3개년 데이터이다. 분석의 편의를 위하여 3개년 모두 응답이 있는 데이터만을 대상으로 하여 3개년 데이터를 병합하였다. 그리고 병합한 데이터에서 가구주가 일치하지 않는 데이터는 제외하였다. 관심변수로는 가구의 상용근로자 연간 총급여액, 고용주 및 자영업자의 연간 순소득, 농림축산업 순소득 등을 더하여 가구소득 변수를 생성하였다. 최종적으로 분석에 사용된 데이터는 3,964가구이다. 병합된 데이터는 3개년 데이터 11,892개이다. 3개년도 가구소득에 대한 산점도가 〈그림 1〉과 〈그림 2〉에 주어져 있다.

가구소득을 연도별로 7개 변수로 설명하는 회귀모형은 아래와 같다. 3개년 병합데이터를 사용한 모형은 아래 모형에서 $\beta_{jt} = \beta_j,\ j = 0,\ldots,7$인 경우이다.

$$\text{가구소득} = \beta_{1t} \times (\text{균등화에 따른 가구구분}) + \beta_{2t} \times (\text{가구원 수}) + \beta_{3t} \times (\text{가구형태}) + \beta_{4t} \times (\text{가구주 태어난 해}) + \beta_{5t} \times (\text{가구주 교육수준}) + \beta_{6t} \times (\text{경제활동 구분}) + \beta_{7t} \times (\text{주거형태}) + \epsilon, \quad t = 1, 2, 3$$

앞에서 언급한 바와 같이 설명변수 중 가구주가 태어난 해는 원 데이터에서 1,900을 뺀 후 12로 나눈 값을 사용하였다. 웨이브별로 구한 유한모집단 회귀계수가 아래 〈표 1〉에 나타나 있다.

〈표 1〉 유한모집단 회귀계수

(단위: 가구)

설명변수	변수명	웨이브1 (2005년)	웨이브2 (2006년)	웨이브3 (2007년)	웨이브 병합
	데이터 수	n=3,964	n=3,964	n=3,964	n=11,892
균등화 가구 구분	x1	-1,112.013	-1,111.429	-1,140.387	-1,140.482
가구원 수	x2	324.935	373.803	427.429	378.783
가구 형태	x3	130.116	152.275	122.064	135.831
가구주 태어난 해	x4	197.500	186.017	225.078	205.943
가구주 교육 수준	x5	309.036	347.318	368.478	339.452
경제활동 구분	x6	-41.157	-58.396	-73.424	-56.594
주거 형태	x7	-34.507	-65.892	-75.792	-57.226
결정계수		0.8242	0.8256	0.8312	0.8231

2. 일반최소제곱추정량의 근사편향

일반최소제곱추정량의 근사편향과 상대편향을 구한다. 웨이브별, 설명변수별 상대편향을 식 (9)에 의하여 계산한 값이 〈표 2〉에 나타나 있다. 또한 아래의 식으로 구한 근사상대편향이 〈표 3〉에 있다.

$$RB_j(\%) = \frac{\overline{\beta}_{jt} - \beta_{jt}}{\beta_{jt}} \times 100, \quad j = 1,...,7$$

여기에서 t는 웨이브를 나타내고 $\overline{\beta}_{jt}$는 t시점의 j번째 설명변수의 회귀계수에 대한 근사기댓값이다.

〈표 3〉에서 볼 수 있듯이 상대편향의 크기는 작게는 5.3%까지 크게는 120%까지 나타난다. 변수별로 보면 주거 형태의 상대편향이 가장 크고(병합 데이터, 63%) 가구원 수의 상대편향이 가장 작다(병합 데이터, 9.8%). 7개 설명변수 균등화 가구구분과 가구원 수를 제외한 5개 변수가 30% 이상의 높은 상대편향을 갖는다.

웨이브별로는 차이가 있기는 하지만 변수별 차이에 비하면 크지 않은 것으로 나타났다. 패널조사의 속성상 웨이브에 따라 관심변수인 소득은 변하지만 설명변수인 가구원 수, 가구 형태 등은 크게 바뀌지 않기 때문에 나타난 현상으로 풀이된다.

〈표 2〉 일반회귀추정량의 근사편향

(단위: 가구)

설명변수	웨이브 1 (2005년)	웨이브 2 (2006년)	웨이브 3 (2007년)	웨이브 병합
	n=3,964	n=3,964	n=3,964	n=11,892
균등화 가구 구분	-132.219	-149.003	-171.386	-143.227
가구원 수	42.020	37.591	22.882	37.331
가구 형태	43.273	60.571	53.907	50.014
가구주 태어난 해	-90.801	-112.608	-98.713	-101.170
가구주 교육 수준	149.709	154.743	160.402	154.421
경제활동 구분	-29.544	-27.606	-21.999	-27.464
주거 형태	-41.366	-30.721	-33.432	-36.523

〈표 3〉 일반회귀추정량의 근사 상대편향(%)

(단위: 가구)

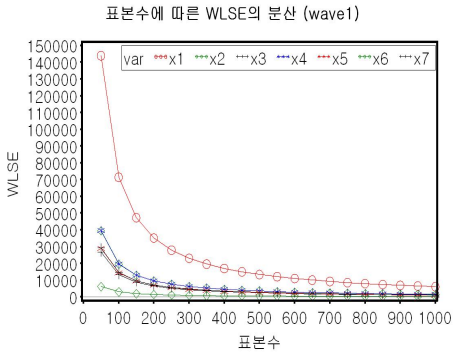
설명변수	웨이브 1 (2005년)	웨이브 2 (2006년)	웨이브 3 (2007년)	웨이브 병합
	n=3,964	n=3,964	n=3,964	n=11,892
균등화 가구 구분	11.890	13.407	15.029	12.558
가구원 수	12.932	10.057	5.353	9.856
가구 형태	33.258	39.777	44.163	36.821
가구주 태어난 해	-45.975	-60.536	-43.858	-49.125
가구주 교육 수준	48.444	44.554	43.531	45.491
경제활동 구분	71.785	47.275	29.962	48.529
주거 형태	119.878	46.624	44.112	63.821

3. 최소제곱추정량의 근사분산과 평균제곱오차 비교

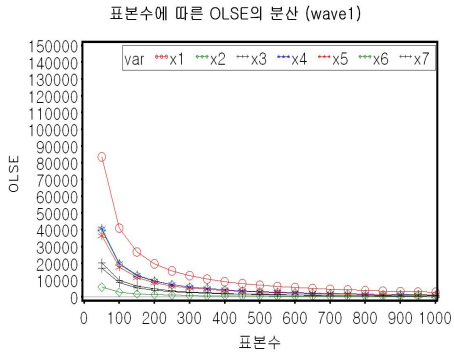
7개 변수의 회귀계수 추정량에 대한 분산을 구한다. 가중최소제곱추정량 $\widehat{B_W}$의 분산은 식(10)을 이용하여 계산하고, 일반최소제곱추정량 $\widehat{B_O}$의 분산은 식(11), 그리고 $\widehat{B_O}$의 평균제곱오차는 식(12)를 이용하여 계산한다. 이들 세 개의 식에서 알 수 있는 바와 같이 분산과 평균제곱오차는 표본수가 커지면 감소한다. 〈그림 3〉과 〈그림 4〉에 1차 웨이브에서 구한 가중최소제곱추정량과 일반최소제곱추정량의 분산이 나타나 있다. 두 경우 모두 표본수가 증가하면 분산이 감소한다. 다른 웨이브와 병합데이터에서도 유사한 패턴을 보인다.

그런데 일반최소제곱추정량의 평균제곱오차는 표본수가 증가하더라도 그 크기가 줄어들기는 하지만 0으로 수렴하지는 않는다(〈그림 5〉). 왜냐하면 평균제곱오차는 분산과 편향제곱의 합이고 일반최소제곱추정량은 편향이 존재하는 편향 추정량이기 때문이다.

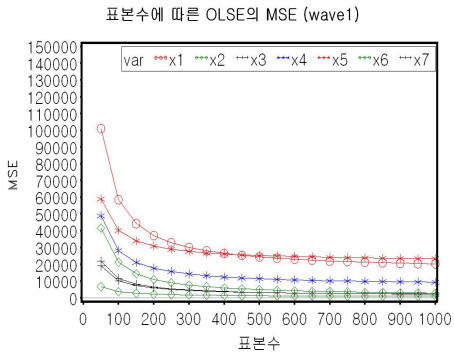
〈그림 6〉은 평균제곱오차를 100%로 했을 때 분산의 비율(rOLS_01, 분산 / 평균제곱오차 × 100)과 평균제곱오차의 비율(rbias_01, 평향 제곱 / 평균제곱오차 × 100)을 표본수에 따라 나타낸 그림이다. 표본수가 증가할수록 평균제곱오차에서 분산이 차지하는 비율은 작아지는 반면 평균제곱오차가 차지하는 비율은 커진다. 표본수가 대략 $n = 250$일 때 분산의 크기와 편향제곱의 크기가 비슷해진다.



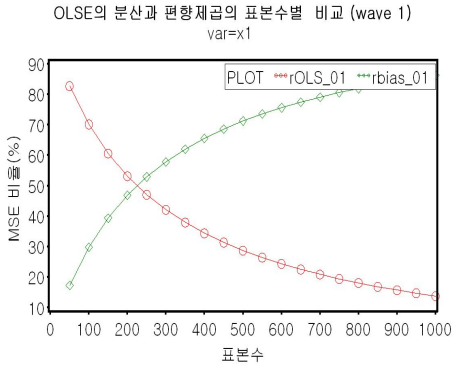
〈그림 3〉 가중최소제곱추정량의 회귀계수별 분산 (wave 1)



〈그림 4〉 일반최소제곱추정량의 회귀계수별 분산 (wave 1)



〈그림 5〉 일반최소제곱추정량의 회귀계수별 평균제곱오차(wave 1)



〈그림 6〉 일반최소제곱추정량의 분산과 편향제곱 비율(wave 1)

이제 가중최소제곱추정량의 분산과 일반최소제곱추정량의 분산 및 평균제곱오차의 크기를 비교해 본다. 비교 지표는 다음과 같은 분산비를 사용한다.

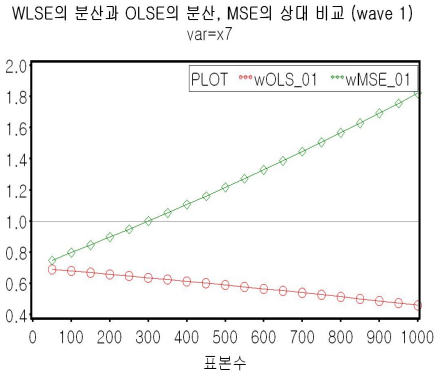
$$\text{(i) 분산비}(wOLS) = \frac{\text{일반최소제곱추정량의 분산}}{\text{가중최소제곱추정량의 분산}}$$

$$\text{(ii) 평균제곱오차비}(mOLS) = \frac{\text{일반최소제곱추정량의 평균제곱오차}}{\text{가중최소제곱추정량의 분산}}$$

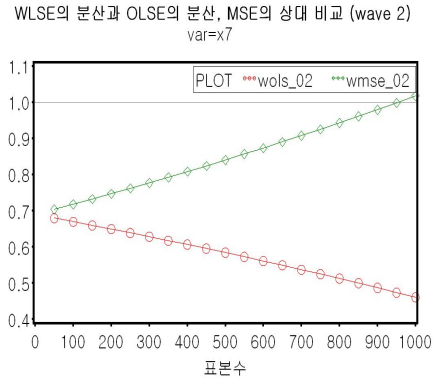
한국복지패널데이터의 경우 몇 가지 패턴이 발견된다.

첫째, 표본의 크기가 커지면 일반최소제곱추정량의 평균제곱오차가 가중최소제곱추정량의 분산보다 커진다. 4개 시점(3개 웨이브 + 병합시점)에서 7개 회귀계수 추정량을 계산한 28가지 경우에서 표본수가 1,000에 이르자 28가지 경우 모두 가중최소제곱추정량의 분산이 일반최소제곱추정량의 평균제곱오차보다 더 작아졌다. 표본수가 1,000 이하일 때에는 가중최소제곱추정량의 분산이 일반최소제곱추정량의 평균제곱오차보다 더 큰 경우도 나타났다. 예를 들어 x7 변수는 웨이브 1에서는 표본수가 300 이하일 때에는 일반최소제곱추정량의 평균제곱오차가 더 작다가 표본수가 300을 넘어서면서 가중최소제곱추정량의 분산이 더 작아졌다(<그림 7>). 반면 웨이브2에서는 표본수가 거의 1,000에 이르러야 가중최소제곱추정량의 분산이 더 작아지기 시작한다(<그림 8>). 모집단 크기가 3,964(병합시점은 11,892)이므로 어떤 회귀계수는 상당히 많은 표본(예컨대 n =1,000)을 뽑아야 가중최소제곱추정량의 분산이 일반최소제곱추정량의 평균제곱오차보다 작아짐을 의미한다.

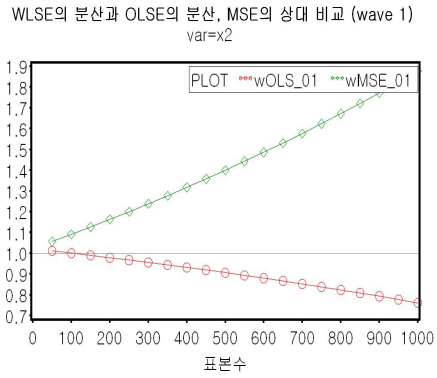
둘째, 평균제곱오차비의 크기는 설명변수에 따라 크기가 다르게 나타나며, 편향이 클수록 평균제곱오차비는 큰 값을 보인다.



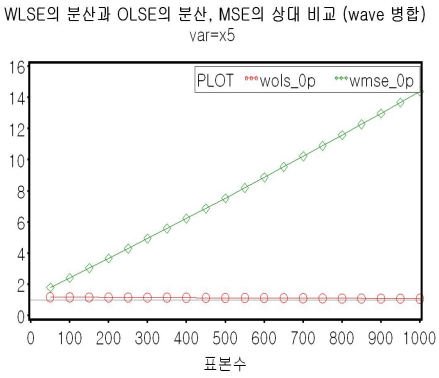
〈그림 7〉 분산과 평균제곱오차 비교 1



〈그림 8〉 분산과 평균제곱오차 비교 2



〈그림 9〉 분산과 평균제곱오차 비교 3



〈그림 10〉 분산과 평균제곱오차 비교 4

셋째, 분산만 비교하면 일반최소제곱추정량의 분산이 가중최소제곱추정량의 분산보다 대부분의 경우 더 작게 나타났다. 〈그림 9〉에서와 같이 표본수가 50일 때 일반최소제곱추정량의 분산이 가중최소제곱추정량의 분산보다 더 큰 경우도 있기는 하지만 대부분의 경우 일반최소제곱추정량의 분산이 더 작았다.

VI. 결론

본 논문에서는 복합패널데이터를 이용하여 패널회귀계수 추정량으로 일반최소제곱추정량과 가중최소제곱추정량의 설계기반 성질을 고찰하였다. 회귀계수의 최소제곱추정량을 선형화하여 일반최소제곱추정량의 근사편향, 근사분산, 그리고 근사평균제곱오차의 수식과 가중최소제곱추정량의 근사분산의 수식을 유도하였다. 또한 두 추정량의 근사분산 및 근사평균제곱오차를 수치적으로 비교하기 위하여 모의실험을 실시하였다.

모의실험에서 한국복지패널 3개년 데이터를 유한모집단으로 간주하였고, 가구소득변수를 관심변수로 하고 가구와 가구주 관련 7개 변수를 설명변수로 하는 유한모집단 회귀계수를 고려하였다. 그리고 두 추정량의 설계기반 성질을 비교하기 위하여 표본수를 50에서 1,000까지 50씩 바꾸어 가며 일반최소제곱추정량의 근사편향, 근사분산 그리고 가중최소제곱추정량의 근사분산을 계산하였다. 모의실험을 통하여 다음과 같은 경향을 발견하였다. 첫째, 표본의 크기가 커지면 일반최소제곱추정량의 평균제곱오차가 가중최

소제곱추정량의 분산보다 커진다. 둘째, 일반최소제곱추정량의 평균제곱오차를 가중최소제곱추정량의 분산으로 나눈 비(ratio)는 설명변수에 따라 크기가 다르게 나타나고, 일반최소제곱추정량의 편향이 클수록 큰 값을 보인다. 셋째, 분산만 비교하면 일반최소제곱추정량의 분산이 가중최소제곱추정량의 분산보다 대부분의 경우에 더 작게 나타난다.

일반적인 경향으로 가중최소제곱추정량은 설계비편향이며 표본수가 증가할수록 일반최소제곱추정량의 평균제곱오차보다 작아지는 성향이 있다고 하더라도 그 정도는 설명변수의 종류와 표본수에 따라 큰 차이를 보인다. 위의 모의실험의 경우 대부분의 경우에서 가중최소제곱추정량의 분산이 일반최소제곱추정량의 평균제곱오차보다 작아지는 표본수는 1,000 정도였다. 바꿔 말하면 1,000보다 작은 표본수에서는 일반최소제곱추정량의 평균제곱오차가 가중최소제곱추정량의 분산보다 더 작은 경우도 많았다.

모의실험의 결과에 근거하여 생각하면 패널회귀모형에서 회귀계수 추정량으로 가중최소제곱추정량을 사용할지 혹은 일반최소제곱추정량을 사용할지의 판단에는 표본수가 중요한 역할을 한다.

참고문헌

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

김규성. 2011. “포함확률비례추출에서 회귀계수 최소제곱추정량의 근사분산.” ≪한국통계학회논문집≫ 게재 예정.

김규성·이영민·전병돈. 2009. “회귀패널모형에서 가중치를 활용한 회귀계수 추정.” ≪2009년 제2회 한국복지패널 학술대회 논문집≫ 413−426.

한국보건사회연구원. 2008. ≪한국복지패널 3차년도 조사자료 : User's Guide≫.

Abraham, G. and J. Ledolter. 2006. *Introduction to Regression Modeling*. Thomson.

Hsiao, C. 2003. *Analysis of Panel Data*. Cambridge Press.

Hill, R.C., W.E. Griffiths, and G.C. Lim. 2008. *Principles of Econometrics*. Wiley.

Lohr, S. 1999. *Sampling: Design and Analysis*. Duxbury press.

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

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

<접수 2011/10/07, 수정 2011/11/15, 게재확정 2011/11/18>