한국통계학회논문집
2010, 17권, 4호, 515–525

복합패널 데이터에 기초한 최소제곱 패널회귀추정량의 설계기반 성질

김규성[1,a]

[a]서울시립대학교 통계학과

요 약

본 논문에서는 패널회귀모형에서 회귀계수의 일반최소제곱추정량과 가중최소제곱추정량의 설계기반 성질을 살펴보았다. 복합표본이 주어진 경우에 두 추정량의 설계편향을 구하여 가중최소제곱추정량의 설계편향의 크기가 더 작음을 보였다. 또한 한국복지패널 데이터를 대상으로 모의실험을 실시하여 다음의 결과를 얻었다. 첫째, 일반최소제곱추정치의 상대편향이 가중최소제곱추정치의 상대편향보다 약 2배 정도 크게 나타났고 일반최소제곱추정치의 편향비가 더 크게 나타났다. 그리고 표본수가 증가하면 일반최소제곱추정치의 상대편향은 완만하게 줄어든 반면 가중최소제곱추정치의 상대편향은 급속도로 줄어들었다. 둘째, 표본수가 증가하면 일반초소제곱추정치와 가중최소제곱추정치의 분산과 평균제곱오차는 모두 줄어들었다. 그러나 평균제곱오차에서 차지하는 편향제곱의 비율은 표본수가 증가할 때 일반최소제곱추정치에서는 증가하는 반면 가중최소제곱추정치에서는 감소하는 경향이 나타났다. 마지막으로 거의 모든 경우에 일반최소제곱추정치의 분산이 가중최소제곱추정치의 분산보다 작게 나타났다. 그리고 많은 경우에 일반최소제곱추정치의 평균제곱오차가 가중최소제곱추정치의 평균제곱오차보다 작게 나타났다. 그러나 표본수가 증가할수록 일반최소제곱추정치의 평균제곱오차가 가중최소제곱추정치의 평균제곱오차보다 커지는 경우가 늘어났다.

주요용어: 가중최소제곱추정량, 상대편향, 일반최소제곱추정량, 편향비, 평균제곱오차 비.

1. 서론

패널회귀모형은 조사변수의 횡단면 특성 및 종단면 특성을 파악하기 위하여 구축된다. n개의 조사단위에 대하여 T번 조사를 수행한 패널조사에서 관심변수를 y라고 하고 p개의 설명변수를 $(x_1, \ldots, x_p)$라고 하면 일반적인 패널회귀모형은 다음과 같이 표현된다 (예를 들면 Hsiao, 2003; Hill 등, 2008; 등).

$$y_{jt} = \beta_{j0t} x_{j0t} + \beta_{j1t} x_{j1t} + \cdots + \beta_{jpt} x_{jpt} + \epsilon_{jt}, \quad j = 1, \ldots, n,\ t = 1, \ldots, T,$$

여기서 j는 패널 조사단위 그리고 t는 조사 시점을 나타낸다. 위 모형은 회귀계수가 시간에 따른 변화(첨자 t)와 조사단위에 따른 변화(첨자 j)를 반영하는 매우 일반적인 모형이다. 경우에 따라 시간 변화에 따른 회귀계수의 변동이 없다고 가정하거나(회귀계수에서 첨자 t제거), 조사단위에 따른 회귀계수의 변동이 없다는 가정하고(회귀계수에서 첨자 j제거), 혹은 더 나아가 시간과 조사단위에 따른 변화가 없다고 가정하면 (회귀계수에서 첨자 j와 t제거) 더 단순한 패널회귀모형을 가정할 수 있다. 그리고 가정된 패널회귀모형의 오차항 ϵ_{jt}에 서로 독립이면서 평균이 0이고 분산이 같은 동일한 분포를 갖는다

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

는 가정을 부여한 후 패널 데이터에 모형을 적합하고 회귀계수를 추정하는 것이 일반적인 패널회귀모형 분석법이다.

그러나 많은 경우에 패널조사에서는 위와 같은 오차 가정이 만족되지 않는다. 왜냐하면 대부분의 패널데이터는 층화, 집락화, 무응답 처리, 가중치 사후 조정 등을 통하여 생산된 복합데이터(complex data)이므로 조사단위의 추출확률이 서로 다르기도 하고 서로 독립도 아니기 때문이다. 따라서 이론적인 모형 가정에 기초한 패널데이터 분석보다는 복합데이터의 특성에 기초한 패널데이터 분석을 하여야 한다는 주장과 함께 이에 대한 연구가 활발하게 진행되고 있다 (대표적인 문헌은 Skinner 등, 1989; Chambers와 Skinner, 2003). 국내 연구로는 한국복지패널 데이터를 대상으로 패널회귀모형을 적합하였을 때 조사가중치를 이용하지 않고 회귀계수를 추정하면 조사가중치를 이용한 경우보다 상대편향이 더 크게 발생한다는 연구 결과가 있다 (김규성 등, 2009).

본 논문에서는 복합패널데이터에 기초한 회귀계수 추정량으로 조사가중치를 이용하지 않는 일반최소제곱추정량(Ordinary least square estimator; OLSE)과 조사가중치를 이용하는 가중최소제곱추정량(Weighted least square estimator; WLSE)의 설계기반(design-based) 성질을 비교 분석하고자 한다. 본 논문은 다음과 같이 구성되어 있다. 2절에서는 패널회귀모형에서 회귀계수의 일반최소제곱추정량과 가중최소제곱추정량을 소개하고 두 추정량의 설계편향을 고찰한다. 3절에서는 모의실험을 통하여 두 추정량의 설계편향, 설계분산, 설계평균제곱오차 등을 계산하여 설계기반의 관점에서 두 추정량의 성능을 비교한다. 4절에서는 논문의 내용을 요약하고 회귀계수 추정량을 선택할 때에 필요한 고려사항을 언급한다.

2. 최소제곱추정량의 설계 편향

2.1. 복합 패널 데이터와 회귀모형

크기 N인 유한모집단에서 크기 n인 표본을 확률추출한다고 하자. 그리고 표본 선정 후 응답 과정에서 발생한 무응답을 가중치 조정으로 처리하고 이어서 사후 층화 혹은 가중치 보정을 통하여 최종가중치를 얻는다고 하자. 그리고 분석자에게 다음과 같은 데이터를 제공한다고 하자.

$$\left\{(w_{jt},\ y_{jt},\ x_{jt}) : j \in s_r,\ \ t = 1, \ldots, T\right\}, \tag{2.1}$$

여기에서 w_{jt}는 t시점의 j번째 응답의 최종가중치, y_{jt}는 반응변수, $x_{jt} = (x_{j0t}, x_{j1t}, \ldots, x_{jpt})$는 설명변수, s_r는 크기 r인 응답 표본이다.

이제 앞에서 언급한 패널회귀모형을 고려하자. 각 시점에서 회귀모형을 적합하거나 혹은 T시점 데이터를 모두 병합한 후에 회귀모형을 적합한다고 하면 비록 시점에 따라 모형의 표현이 달라지기는 하지만 시점을 나타내는 첨자 t를 생략하면 두 종류의 모형은 아래와 같은 회귀모형으로 표현이 가능하다.

$$y_j = \beta_0 x_{j0} + \beta_1 x_{j1} + \cdots + \beta_p x_{jp} + \epsilon_j, \quad j \in s_r. \tag{2.2}$$

이때 각 시점의 응답 표본수는 r이고 T개 시점의 데이터를 병합한 응답 표본의 수는 Tr이다.

2.2. 최소제곱 회귀추정량

복합표본 (2.1)이 주어지고 패널회귀모형 (2.2)를 분석모형으로 한다고 하자. 이 모형의 회귀계수 추정량으로 먼저 생각할 수 있는 추정량은 일반최소제곱추정량이다.

$$\hat{\beta}_O = \left[\sum_{j \in s_r} x_j' x_j\right]^{-1} \left[\sum_{j \in s_r} x_j' y_j\right]. \tag{2.3}$$

위 추정량은 조사과정에서 발생한 표본추출확률 및 가중치를 고려하지 않은 추정량이기 때문에 조사가중치가 추정량에 반영되어 있지 않다. 잘 알려진 바와 같이 회귀모형 (2.2)에 대하여 일반적인 오차가정(오차 ϵ_j의 평균이 0이고 분산이 동일하며 서로 독립)이 만족되면 (2.3)의 회귀계수 추정량 $\hat{\beta}_O$는 선형 추정량 중에서 비편향이면서 분산이 가장 작은 최량 추정량이다(Gauss-Markov 정리, 예를 들면 Abraham과 Ledolter, 2006 등). 그런데 그런데 본 논문에서 다루는 복합표본은 이와 같은 가정을 충족시키지 않기 때문에 복합표본에 기초한 일반최소제곱추정량 $\hat{\beta}_O$은 최량추정량이라고 말하기 어렵다.

복합표본이 주어졌을 때 사용하기 적합한 추정량으로 가중최소제곱추정량을 생각할 수 있다. 가중최소제곱추정량은 유한모집단 회귀계수 $B_U = [\sum_{j=1}^N x_j' x_j]^{-1}[\sum_{j=1}^N x_j' y_j]$에서 B_U를 구성하는 유한모집단 모수 $\sum_{j=1}^N x_j' x_j$와 $\sum_{j=1}^N x_j' y_j$를 각각 설계 비편향 추정하고 각 추정량을 대입하여 만든다 (Sarndal 등, 1992, 193쪽).

$$\hat{\beta}_W = \left[\sum_{j\in s} \frac{x_j' x_j}{\pi_j}\right]^{-1} \left[\sum_{j\in s} \frac{x_j' y_j}{\pi_j}\right], \tag{2.4}$$

여기에서 s는 크기 n인 표본이고 π_j는 j번째 조사단위의 포함확률이다. 추정량 $\hat{\beta}_W$는 B_U의 설계일치추정량이기 때문에 (Sarndal 등, 1992, 194쪽) 응답률이 100%이고 사후 가중치 조정을 하지 않은 상태에서 포함확률이 알려진 경우에 활용하기 좋은 추정량이다. 그러나 대부분의 패널조사에서는 무응답이 발생하며 사후 가중치 조정을 하기 때문에 식 (2.4)의 추정량을 직접 사용하기는 현실적으로 어렵다. 대신 무응답 보정과 사후층화 보정을 거친 회귀추정량을 대안으로 생각할 수 있다 (예를 들면 Lohr, 1989, 354쪽).

$$\hat{\beta}_C = \left[\sum_{j\in s_r} w_j x_j' x_j\right]^{-1} \left[\sum_{j\in s_r} w_j x_j' y_j\right], \tag{2.5}$$

여기에서 w_j는 최종가중치이다. 식 (2.5)의 추정량은 식 (2.3)이나 (2.4)의 추정량과는 달리 무응답 보정, 사후 가중치 보정 등을 모두 포함하므로 실제적으로 사용 가능한 추정량이라고 할 수 있다.

2.3. 회귀추정량의 설계편향

일반 회귀분석에서는 서로 독립이고 동일한 분포를 갖는 표본을 선정하고 조사한다는 전제 아래 회귀모형에 대한 추론을 전개한다. 그러나 실제 많은 패널조사에서는 복합표본을 선정하므로 가정된 모형을 바탕으로 하는 모형기반 추론보다는 실제 표본추출 상황을 바탕으로 설계기반 추론이 더 현실적일 때가 있다. 본 소절에서는 설계기반의 관점에서 회귀추정량의 성능을 설계편향(design bias)을 구하여 비교하고자 한다. 두 추정량의 설계편향을 유도한 후 크기를 비교하자.

설계 편향을 유도하기 위하여 편의상 몇 가지 기호를 도입하다. 차원이 $(p+1)\times(p+1)$인 행렬 U와 $(p+1)\times 1$행렬 V를 다음과 같이 나타내자.

$$U = \sum_{j=1}^N x_j' x_j = (u_{kl})_{(p+1)\times(p+1)}, \qquad V = \sum_{j=1}^N x_j y_j = (v_k)_{(p+1)\times 1},$$

여기에서 $u_{kl} = \sum_{j=1}^N x_{jk} x_{jl},\ k,l = 0,1,\ldots,p$이고 $v_k = \sum_{j=1}^N x_{jk} y_j,\ k = 0,1,\ldots,p$이다. 그러면 유한 모집단 회귀계수는 $B_U = U^{-1}V$로 표현되고, U와 V를 각각 추정한 후 대입하면 회귀계수 추정량 $\hat{\beta}_r = [\hat{U}]^{-1}\hat{V}$을 얻을 수 있다.

회귀계수 추정량 $\hat{\beta}_r$는 비선형 추정량이므로 $\hat{\beta}_r$의 설계편향을 직접 유도하기는 쉽지 않다. 그러나 테일러 전개를 통하여 선형화를 하면 $\hat{\beta}_r$의 근사 설계편향은 어렵지 않게 구할 수 있다. 테일러 전개를

통하여 $\hat{\beta}_r$를 일차항까지 이용한 근사식은 다음과 같다 (Sarndal 등, 1992, 194쪽 Result 5.10.1.)

$$\hat{\beta}_r = B_U + U^{-1}\left(\hat{V} - \hat{U}B_U\right) + O_p\left(\|\hat{e} - e\|^2\right), \tag{2.6}$$

여기서 $\hat{e} = (\hat{u}_{kl},\ \hat{v}_k)$, $e = (u_{kl},\ v_k)$, $k \leqq l = 0, 1, \ldots, p$이다. 그리고 거리측도($\|\cdot\|^2$)는 제곱거리 ($\| x_j\|^2 = \sum_k x_{jk}^2$를 의미한다. 이제 추정량 $\hat{\beta}_r$의 설계편향은 다음의 식으로 표현할 수 있다.

$$\text{Bias}_C\left(\hat{\beta}_r\right) = E_C\left(\hat{\beta}_r\right) - B_U = U^{-1}\left[E_C\left(\hat{V}\right) - E_C\left(\hat{U}\right)B_U\right] + O\left(E_C\|\hat{e} - e\|^2\right), \tag{2.7}$$

여기에서 첨자 C는 복합설계, 응답 메카니즘, 사후 가중 과정을 종합한 복합 과정임을 나타낸다.

위의 식 (2.7)의 우변에서 벡터 $\hat{e}$의 원소를 $\hat{e}_l$로 나타내면 $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_l, \ldots, \hat{e}_L)'$로 표현할 수 있다. 이때 $L = (p+1)(p+2)/2 + (p+1) = (p^2 + 5p + 3)/2$은 벡터 $\hat{e}$를 구성하는 원소의 수이다. 대응하는 벡터 e도 $e = (e_1, \ldots, e_L)'$로 표현하자. 이제 식 (2.7)의 $E_C\|\hat{e} - e\|^2$은 다음과 같이 전개할 수 있다.

$$\begin{aligned} E_C\|\hat{e} - e\|^2 &= \sum_{l=1}^{L} E_C\left(\hat{e}_l - e_l\right)^2 \\ &= \sum_{l=1}^{L}\left[V_C(\hat{e}_l) + \text{Bias}_C^2(\hat{e}_l)\right] \\ &\leqq L \times \max_l\left[V_C(\hat{e}_l) + \text{Bias}_C^2(\hat{e}_l)\right]. \end{aligned} \tag{2.8}$$

만일 분산 항과 편향 항에 다음과 같은 가정이 성립한다고 하자.

$$\max_l V_C(\hat{e}_l) = O\left(\frac{1}{n}\right), \qquad \max_l \text{Bias}_C(\hat{e}_l) = O\left(\frac{1}{\sqrt{n}}\right) \tag{2.9}$$

그러면 식 (2.9)의 가정 아래 식 (2.7)의 편향은 더 간단하게 표현될 수 있다.

$$\text{Bias}_C\left(\hat{\beta}\right) = U^{-1}\left[E_C\left(\hat{V}\right) - E_C\left(\hat{U}\right)B_U\right] + O\left(\frac{1}{n}\right). \tag{2.10}$$

본 논문에서 언급한 회귀계수 추정량, $\hat{\beta}_O$, $\hat{\beta}_W$ 그리고 $\hat{\beta}_C$의 설계편향을 직접 비교하기 위해서는 표집설계, 응답 메카니즘, 사후 가중 과정 등이 상술되어야 한다. 이에 대한 포괄적인 연구는 나중으로 미루고, 본 논문에서는 확률표집(probability sampling)만을 고려한 경우를 다루기로 한다.

이제 모든 단위의 포함확률(π_j)이 양수라고 가정하자. 그러면 식 (2.10)으로 부터 일반최소제곱추정량 $\hat{\beta}_O$에 대한 설계편향은 다음과 같이 유도된다.

$$\begin{aligned} \text{Bias}_p\left(\hat{\beta}_O\right) &= U^{-1}\left[E_p\left(\hat{V}\right) - E_p\left(\hat{U}\right)B_U\right] + O\left(\frac{1}{n}\right) \\ &= U^{-1}\left[\frac{N}{n}\sum_{j=1}^{N} x_j' y_j \pi_j - \frac{N}{n}\sum_{j=1}^{N} x_j' x_j \pi_j B_U\right] + O\left(\frac{1}{n}\right) \\ &= U^{-1}\left[\frac{N}{n}\sum_{j=1}^{N} \pi_j x_j'\left(y_j - x_j B_U\right)\right] + O\left(\frac{1}{n}\right) \\ &= U^{-1}\left[\frac{N}{n}\sum_{j=1}^{N} \pi_j x_j' a_j\right] + O\left(\frac{1}{n}\right), \end{aligned} \tag{2.11}$$

여기에서 $a_j = y_j - x_j B_U$이다. 그런데 최소제곱추정량의 잔차 성질로부터 $\sum_{j=1}^{N} x_j' a_j = 0$가 성립하고, 또한 만일 고정크기 확률표집설계를 가정하면 $\sum_{j=1}^{N} \pi_j = n$이 되므로 위의 식 (2.11)은 아래의 식으로 간단하게 표현할 수 있다.

$$\text{Bias}_p\left(\hat{\beta}_O\right) = U^{-1}\text{Cov}_p\left(\frac{N}{n}\pi, x'a\right) + O\left(\frac{1}{n}\right), \tag{2.12}$$

여기에서 π는 $\pi_j,\ j = 1, \ldots, N$을 나타내는 변수이고 $x'a$는 $x_j' a_j,\ j = 1, \ldots, N$을 나타내는 변수이다. 따라서 포함확률 π_j가 모집단 특성값 $x_j' a_j$와 어떤 연관성이 있으면 공분산 항은 0이 되지 않으므로 식 (2.12)의 우변 중 첫번째 항은 소거되지 않는다. 공분산 항이 소거되는 경우는 단순확률표집이나 자기가중표집 같이 포함확률이 상수인 경우이다. 그러나 확률비례추출과 같이 포함확률이 x_j의 크기에 연관되는 경우 공분산 항은 0이 되지 않기 때문에 식 (2.12)의 우변의 첫 번째 항이 사라지지 않는다.

다음으로 가중최소제곱추정량 $\hat{\beta}_W$에서는 $\hat{U}_W = \sum_{j\in s} x_j' x_j/\pi_j$과 $\hat{V}_W = \sum_{j\in s} x_j' y_j/\pi_j$이 각각 U와 V의 설계 비편향 추정량이므로 식 (2.10)에 대입하면 다음과 같은 결과를 얻을 수 있다.

$$\text{Bias}_p\left(\hat{\beta}_W\right) = O\left(\frac{1}{n}\right). \tag{2.13}$$

따라서 식 (2.12)와 (2.13)을 비교하여 보면 일반적으로 가중최소제곱추정량 $\hat{\beta}_W$의 설계편향의 크기가 일반최소제곱추정량 $\hat{\beta}_O$의 설계편향의 크기보다 작음을 알 수 있다.

3. 모의실험

일반최소제곱추정량과 가중최소제곱추정량의 설계기반 성능을 수치적으로 비교하기 위하여 모의실험을 실시한다. 모의실험에서는 두 추정량의 설계편향뿐만 아니라 설계분산과 설계평균제곱오차도 계산하여 비교하기로 한다. 편향이 작더라도 분산이 커서 전체적으로 평균제곱오차가 더 크면 좋은 추정량이라고 하기 어렵기 때문에 두 추정량의 성능을 비교하기 위해서는 편향뿐 아니라 분산 및 평균제곱오차를 비교하는 것이 필요하다.

3.1. 모집단 및 표본추출

모의실험에 이용한 데이터는 한국복지패널의 3개년(2006년~2008년) 표본데이터이다. 표본데이터 중 3개년 모두 응답이 있는 가구 중에서 가구주가 일치하는 가구만을 모의실험의 대상으로 하였다. 한국복지패널은 저소득 가구의 추출률이 높기 때문에 저소득 가구의 표본이 상대적으로 많다. 따라서 한국복지패널 표본데이터를 그대로 모집단으로 간주하고 사용할 경우 저소득 가구의 특성이 더 많이 반영되게 된다. 이를 방지하기 위하여 추출확률을 역산하여 저소득 가구의 비율을 줄인 후 모의실험에 사용할 모집단을 새로이 구성하였다. 최종적으로 모집단에 포함된 패널가구는 3,336가구이다.

반응변수는 가구의 상용근로자 연간 총급여액, 고용주 및 자영업자의 연간 순소득, 농림축산업 순소득 등을 합하여 가구소득 변수를 만든 후 로그를 취하여 생성하였다. 설명변수로는 균등화에 따른 가구 구분, 가구원 수, 가구 형태, 가구주 교육 수준, 가구주 건강상태, 경제활동 구분, 주거 형태를 사용하였다 (한국보건사회연구원, 2006).

표본은 모집단으로부터 1차년도 가구 가중치를 크기 측도로 하여 크기비례 확률추출 하였다. 표본크기에 따른 측도의 변화를 보기 위하여 표본수를 50개, 100개, ..., 450개의 9가지로 하였다.

3.2. 회귀계수추정량의 성능 평가지표

연도별 데이터 3종류와 3개년 병합 데이터로부터 각각 크기비례 확률표본을 추출한 후 일반최소제곱추정치와 가중최소제곱추정치를 계산하고 이를 R = 1,000회 반복하여 추정치의 평균($\bar{\beta}_{kt}$), 분산(v_{kt}),

평균제곱오차(m_{kt})를 계산하였다.

$$\bar{\beta}_{kt} = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_{kt}^{(r)}, \qquad v_{kt} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{kt}^{(r)} - \bar{\beta}_{kt}\right)^2, \qquad m_{kt} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{kt}^{(r)} - \beta_{kt}\right)^2,$$

여기에서 $\hat{\beta}_{kt}^{(r)}$는 t시점의 회귀계수 β_{kt}를 r번째 반복표본으로 구한 회귀계수 추정치이다.

추정량의 성능 평가지표로는 상대편향(Relative bias), 편향비(Bias ratio), 상대 평균제곱오차 제곱근(Relative Root of MSE), 편향제곱 비율(Percentage of squared bias), 분산 비(Variance ratio), 평균제곱오차 비(MSE ratio)를 고려하였고 7개 회귀계수 ($k = 0, 1, \ldots, 7$, 절편제외), 4개 연도(t = 2006, 2007, 2008, 3개년 병합), 9개 표본수를 조합하여 총 252 경우에 대하여 지표를 계산하였다. 각각의 계산식은 다음과 같다.

- 상대편향(%) = $\dfrac{\bar{\beta}_{kt} - \beta_{kt}}{\beta_{kt}} \times 100$
- 편향비 = $\dfrac{\bar{\beta}_{kt} - \beta_{kt}}{\sqrt{v_{kt}}}$
- 상대 평균제곱오차 제곱근 = $\dfrac{\sqrt{m_{kt}}}{\beta_{kt}}$
- 편향제곱 비율(%) = $\dfrac{(\bar{\beta}_{kt} - \beta_{kt})^2}{m_{kt}} \times 100$
- 분산비 = $\dfrac{v_{kt}(\text{WLSE})}{v_{kt}(\text{OLSE})}$
- 평균제곱오차 비 = $\dfrac{m_{kt}(\text{WLSE})}{m_{kt}(\text{OLSE})}$

상대편향은 편향의 상대적인 수치로서 상대편향이 0에 가까울수록 좋은 추정량이라고 할 수 있다. 편향비는 신뢰구간의 포함확률을 평가할 때 유용한 측도로서 편향비가 클수록 신뢰구간의 명목확률과 실제 포함확률의 차이는 크게 나타난다. 예를 들어 편향비의 값이 1이면 명목확률이 95%인 신뢰구간의 실제 포함확률은 83%가 되어 명목확률과 실제 포함확률의 차이가 12%p가 된다 (Sarndal 등, 1992, 164쪽). 따라서 편향비가 0에 가까운 추정량이 신뢰구간 추정에 좋은 추정량이라고 할 수 있다. 추정량의 성능은 분산과 평균제곱오차의 관점에서도 평가할 수 있다. 평균제곱오차는 분산과 편향제곱의 합이므로 비록 편향이 크더라도 분산이 작아서 평균제곱오차가 더 작으면 더 나은 추정량이라고 할 수도 있다. 이를 평가하기 위하여 상대 MSE 제곱근, MSE 중에서 편향제곱이 차지하는 비율, 분산이 차지하는 비율 등을 살펴본다.

3.3. 모의실험 결과

3.3.1. 상대편향과 편향비

일반최소제곱추정치와 가중최소제곱추정치의 상대편향과 편향비의 산점도가 그림 1의 (가)와 (나)에 각각 나타나 있다. 산점도는 252가지 (회귀계수 7개(절편 제외) × 시점 4개 × 표본수 종류 9가지) 경우에 대하여 두 추정치의 상대편향과 편향비를 구하여 그린 것이다. 그림 1(가)에서 보면 일반최소

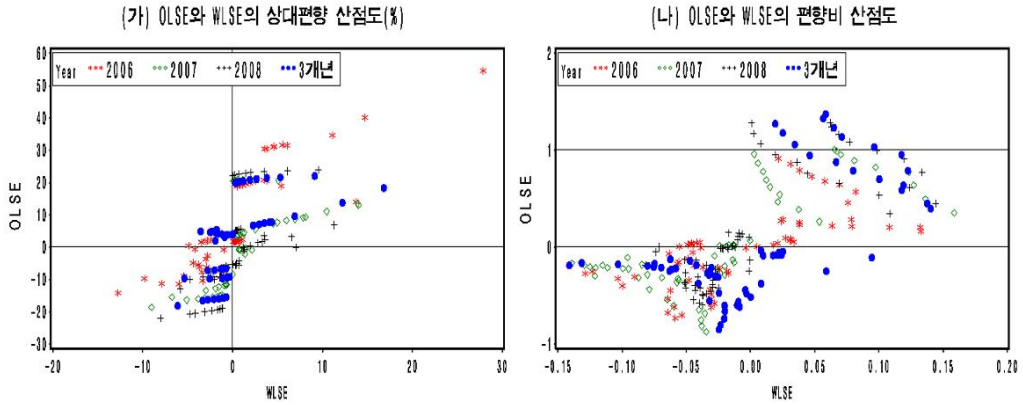

그림 1: *일반최소제곱추정치(OLSE)와 가중최소제곱추정치(WLSE)의 상대편향과 편향비의 산점도*

제곱추정치의 상대편향이 가중최소제곱추정치의 상대편향보다 더 넓게 산포함을 볼 수 있다. 이 그림에서 두 추정치의 상대편향의 관계를 평균적으로 살펴보면 다음과 같다.

$$(\text{일반최소제곱추정치의 상대편향}) \approx 2.11 \times (\text{가중최소제곱추정치의 상대편향}).$$

즉, 평균적으로 일반최소제곱추정치의 상대편향이 가중최소제곱추정치의 상대편향보다 약 2.1배 크게 나타난다는 의미로 해석할 수 있다. 그림 1(나)에서는 가중최소제곱추정치의 편향비는 252개 모두가 $(-0.15, 0.16)$ 사이에 위치하는 반면 일반최소제곱추정치의 편향비는 $(-0.9, 1.4)$사이에 위치하고 있다. 이는 편향으로 인한 신뢰구간 추정의 오류가 일반최소제곱추정치에서 더 크게 발생함을 의미한다.

또한 두 그림에서 병합 데이터와 개별 연도 데이터의 추정치를 비교하면, 병합 데이터의 상대편향은 상대적으로 작게 나타나는 반면(그림 1(가)), 편향비는 비슷하게 나타남을 볼 수 있다(그림 1(나)). 데이터를 병합하면 상대편향은 줄어들더라도 편향비는 줄어들지 않기 때문에 데이터를 병합하더라도 신뢰구간을 추정할 때에는 이 점을 유의하여야 한다.

그림 2는 두 추정치의 상대편향을 표본수 별로 그린 산점도이다. 식 (2.12)에 의하면 만일 포함확률(π)과 모집단 데이터의 특성($x'a$)이 연관이 있는 경우에는 표본수가 증가하더라도 일반최소제곱추정치의 편향은 사라지지 않는다. 그림 2의 (가)는 이러한 현상을 잘 보여준다. 그러나 식 (2.13)과 그림 2의 (나)에서 보듯이 가중최소제곱추정치의 상대편향은 표본수가 증가하면 현격히 줄어드는 추세를 보인다. 이러한 형상은 연도별 데이터나 병합 데이터 모두에서 동일하게 나타난다(그림 2의 (다), (라)).

3.3.2. 평균제곱오차와 편향제곱 비율

그림 3은 표본수에 따라 두 추정량의 상대 MSE 제곱근을 보여준다. 두 경우 모두 표본수가 증가하면 평균제곱오차는 줄어드는 추세를 보인다. 그림을 제시하지는 않았지만 분산도 비슷한 추세를 보인다. 비록 일반최소제곱추정치의 편향은 표본수가 증가하더라도 감소하지 않는 부분이 남아 있지만 분산이 줄어들기 때문에 편향제곱과 분산의 합인 평균제곱오차는 표본수가 증가하면 줄어들게 되는 것이다.

그림 4는 평균제곱오차 중에서 편향제곱이 차지하는 비율을 표본수별로 나타낸 그림이다. 앞에서 언급한 바와 같이 일반최소제곱추정치는 표본수가 증가하면 분산이 줄어들더라도 편향이 크게 줄어들지 않기 때문에 표본수가 증가할수록 평균제곱오차에서 편향이 차지하는 비율은 커지는 경향이 있다. 반면 가중최소제곱추정치는 표본수가 증가하면 편향 및 분산이 모두 줄어드는데 그 중에서 편향제곱이 줄어드는 속도가 더 빠르게 나타난다.

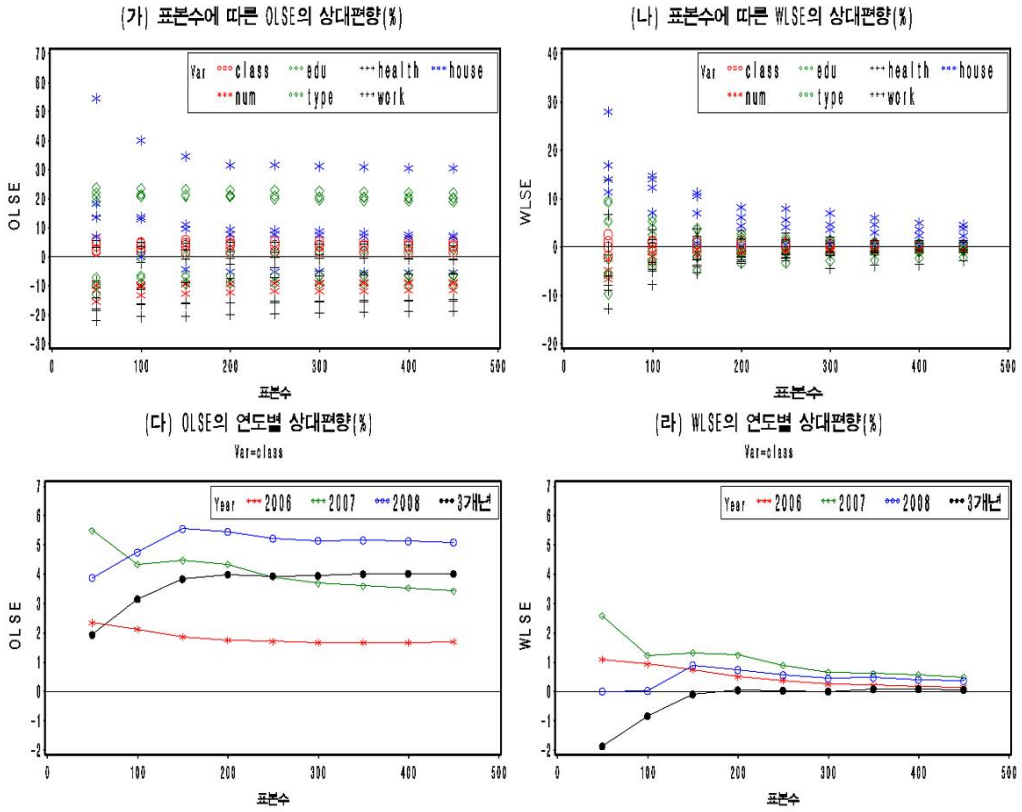

그림 2: *일반최소제곱추정치와 가중최소제곱추정치의 표본수별 상대편향 추이*

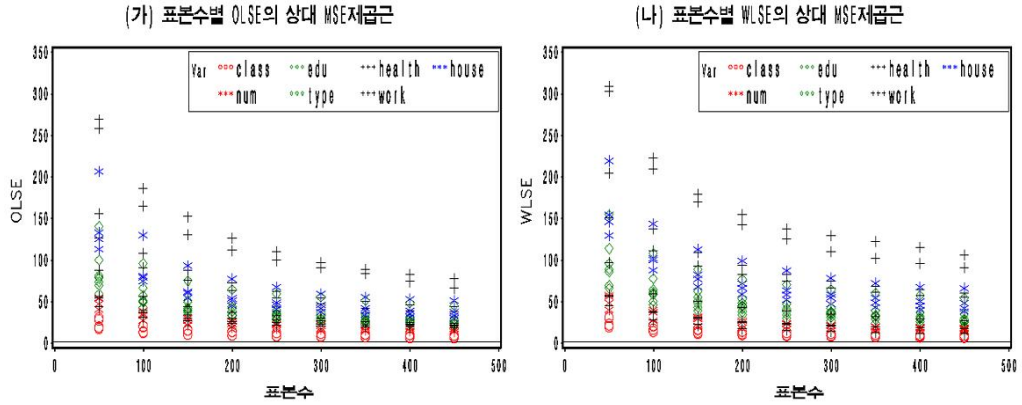

그림 3: *표본수에 따른 일반최소제곱추정치와 가중최소제곱추정치의 상대 MSE 제곱근*

3.3.3. 분산의 비와 평균제곱오차의 비

그림 5의 (가)는 일반최소제곱추정치와 가중최소제곱추정치의 분산의 비(분산비 = WLSE 분산 / OLSE 분산)를 나타낸 그림이고 (나)는 평균제곱오차의 비(MSE 비 = WLSE의 MSE / OLSE의

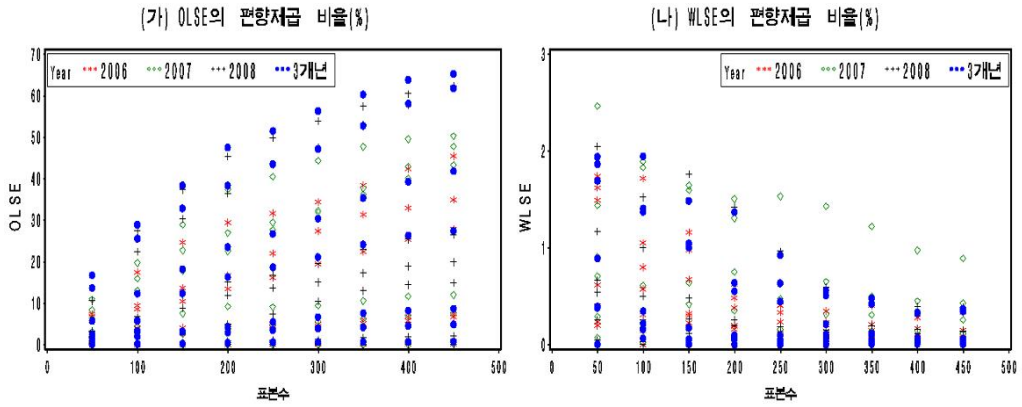

그림 4: *표본수에 따른 일반최소제곱추정치와 가중최소제곱추정치의 편향제곱 비율*

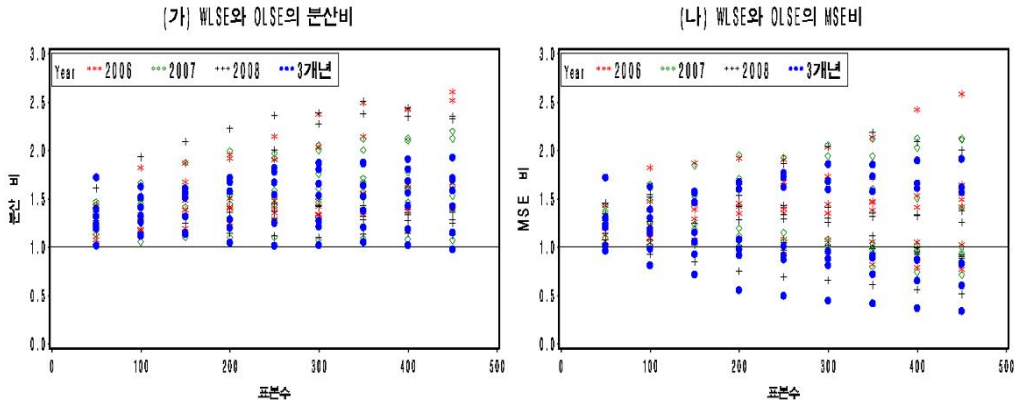

그림 5: *표본수에 따른 OLSE와 WLSE의 분산 비 및 MSE 비*

MSE)를 나타낸 그림이다. 그림 5(가)가 분산비에서는 거의 모든 경우에 1보다 큰 값을 보이는 반면, 그림 5(나)의 MSE 비에서는 이러한 사실이 적용되지 않는다. 252개 경우 중 187개(74.2%)에서 가중최소제곱추정치의 평균제곱오차가 일반최소제곱추정치의 평균제곱오차보다 더 크고, 나머지 65개(25.8%)에서는 반대로 일반최소제곱추정치의 평균제곱오차가 더 큰 것으로 나타났다. 또한 표본수가 증가할수록 평균제곱오차의 비가 1 이하로 되는 경우가 늘어나는 추세가 있다. 예를 들어 표본수가 50일 때에는 평균제곱오차의 비가 1 이하인 경우는 28개 중 2개(7.1%)였으나 표본수가 450일 때에는 28개 중 12개(42.9%)가 1 이하였다.

4. 결론

본 논문에서는 패널회귀모형에서 회귀계수의 일반최소제곱추정량과 가중최소제곱추정량의 설계기반 성질을 살펴보았다. 복합표본이 주어진 경우에 두 추정량의 설계편향을 구하여 가중최소제곱추정량의 설계편향의 크기가 더 작음을 보였다. 또한 한국복지패널 데이터를 대상으로 모의실험을 실시하여 다음과 같은 결과를 얻었다. 첫째, 일반최소제곱추정치의 상대편향이 가중최소제곱추정치의 상대편향보다 약 2배 정도 크게 나타났고 일반최소제곱추정치의 편향비가 더 크게 나타났다. 둘째, 표본수가 증가하면 일반최소제곱추정치의 상대편향은 완만하게 줄어든 반면 가중최소제곱추정치의 상대

편향은 급속도로 줄어들었다. 셋째, 표본수가 증가하면 일반최소제곱추정치와 가중최소제곱추정치의 분산과 평균제곱오차는 모두 줄어들었다. 넷째, 평균제곱오차에서 차지하는 편향제곱의 비율은 표본수가 증가할 때 일반최소제곱추정치에서는 증가하는 반면 가중최소제곱추정치에서는 감소하는 경향이 나타났다. 다섯째, 거의 모든 경우에 가중최소제곱추정치의 분산이 일반최소제곱추정치의 분산보다 크게 나타났다. 그리고 많은 경우에(약 74%) 가중최소제곱추정치의 평균제곱오차가 일반최소제곱추정치의 평균제곱오차보다 크게 나타났다. 그러나 표본수가 증가할수록 일반최소제곱추정치의 평균제곱오차가 가중최소제곱추정치의 평균제곱오차보다 커지는 경우가 늘어났다.

본 연구의 결과를 토대로 패널회귀모형 분석에서 일반최소제곱추정량과 가중최소제곱추정량 중 하나를 선택하여야 한다고 하자. 먼저 편향의 관점에서는 가중최소제곱추정량을 사용하는 것이 타당하다. 반면 평균제곱오차의 관점에서는 표본수가 작을 때에는 일반최소제곱추정량을 선택하고 표본수가 클 때에는 가중최소제곱추정량을 선택하는 것이 합리적이다. 그런데 대부분의 패널조사에서는 표본수가 상당히 크므로 가중최소제곱추정량을 선택하는 것이 더 실제적이다. 이에 더하여 현실적인 문제로 패널데이터 분석을 끝내고 회귀계수 추정치의 신뢰도를 공표하여야 하는 상황이 있다고 하자. 앞서 언급한 편향, 분산, 평균제곱오차는 모두 이론적인 값이기 때문에 실제 사용하기 위해서는 세 값을 추정하여야 한다. 그런데 분산을 제외하면 편향과 평균제곱오차는 추정이 곤란한 경우가 대부분이다. 따라서 편향 추정량인 일반최소제곱추정량을 사용할 경우 편향의 크기를 나타내기가 쉽지 않다. 이러한 이유로 추정치의 신뢰도를 공표해야 하는 상황에서는 설계일치 추정량인 가중최소제곱추정량을 선택하는 것이 더 현실적이다.

최근 들어 패널조사 데이터를 이용한 연구가 활발하게 이루어지고 있다. 그러나 많은 경우에 연구자들은 패널 데이터가 복합데이터임에도 불구하고 통상적인 데이터, 즉 서로 독립이고 동일한 분포에서 나온 데이터로 간주하고 패널회귀모형을 적합 한다. 본 논문에서는 복합표본의 추출확률을 무시하고 패널회귀모형에 적합하여 회귀계수를 추정하는 경우 편향이 발생하여 추정의 신뢰도 및 분석의 타당성이 저하됨을 보이고, 한국복지패널 데이터를 이용한 모의실험을 통하여 그 정도를 실증적으로 나타내었다. 향후 복합표본이 주어지는 경우 가중최소제곱 추정법을 이용하여 패널 데이터를 분석하는 사례가 늘어나기를 기대한다.

참고 문헌

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

한국보건사회연구원 (2006). <한국복지패널 1차년도 조사자료 User's Guide>, 한국보건사회연구원.

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

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

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

Hsiao, C. (2003). *Analysis of Panel Data*, 2nd Ed., Cambridge Press.

Lohr, S. L. (1989). *Sampling: Design and Analysis*, Brook.

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

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

2010년 2월 접수; 2010년 5월 채택

Design-Based Properties of Least Square Estimators of Panel Regression Coefficients Based on Complex Panel Data

Kyu-Seong Kim[1,a]

[a]Department of Statistics, University of Seoul

Abstract

We investigated design-based properties of the ordinary least square estimator(OLSE) and the weighted least square estimator(WLSE) in a panel regression model. Given a complex data we derive the magnitude of the design-based bias of two estimators and show that the bias of WLSE is smaller than that of OLSE. We also conducted a simulation study using Korean welfare panel data in order to compare design-based properties of two estimators numerically. In the study we found the followings. First, the relative bias of OLSE is nearly two times larger than that of WLSE and the bias ratio of OLSE is greater than that of WLSE. Also the relative bias of OLSE remains steady but that of WLSE becomes smaller as the sample size increases. Next, both the variance and mean square error(MSE) of two estimators decrease when the sample size increases. Also there is a tendency that the proportion of squared bias in MSE of OLSE increases as the sample size increase, but that of WLSE decreases. Finally, the variance of OLSE is smaller than that of WLSE in almost all cases and the MSE of OLSE is smaller in many cases. However, the number of cases of larger MSE of OLSE increases when the sample size increases.

Keywords: Bias ratio, ordinary least square estimator, MSE ratio, relative bias, weighted least square estimator.

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