응용통계연구 제 **20**권 2호, 2007년, pp. 281–290

가중치 보정 추정량에 대한 일반적인 분산 추정법 연구*

김재광[1)]

요 약

유한 모집단에서 총계 추정에는 표본의 각 관측값으로 만들어지는 선형 추정량이 사용되는데 이 때 사용되는 가중치는 표본 추출 확률의 역수를 사용한 기본 가중치를 모집단 전체에서 얻어지는 보조 정보를 이용하여 보정한 형태로 종종 사용된다. 이렇게 보정된 가중치를 사용한 추정량은 그렇지 않은 추정량보다 효율이 더 좋아질 수 있는 장점이 있으나 이러한 경우 분산 추정은 더 어려워지게 된다.

본 연구에서는 보정된 가중치를 사용한 추정량의 분산 추정을 다룬다. 가중치 보정의 일반적인 형태를 밝히고 이 경우 가중치 보정항은 유한개의 장애 모수(nuisance parameter)의 함수로 나타낼 수 있으므로 이 장애 모수에 대한 테일러 전개를 사용한 분산 추정식을 구한다. 이렇게 구현된 분산 추정식은 기존의 가중치 보정 추정량 뿐만 아니라 보다 일반적인 경우에서도 적용될 수 있다는 장점이 있다. 몇가지 응용 사례와 모의 실험 결과를 소개한다.

주요용어: 보정, 테일러 전개, 회귀 추정량.

1. 서론

크기가 N인 유한 모집단의 총계 $Y = \sum_{i=1}^{N} y_i$를 추정하는 문제에 대하여 생각하기로 하자. 집합 A를 추출된 표본 원소들의 집합이라고 하고 추출된 표본에서 관심 항목 y를 모두 관측할 수 있고 무응답은 존재하지 않는다고 가정하자. 이러한 경우 다음과 같이 정의된 Horvitz-Thompson 추정량은 Y에 대해 비편향(unbiased)임이 알려져 있다.

$$\hat{Y}_d = \sum_{i \in A} d_i y_i. \tag{1.1}$$

여기서 d_i는 원소 i가 표본으로 선택될 확률인 일차 표본 포함확률(first-order inclusion probability)의 역수로 정의된다. 이 가중치 d_i는 표본설계로부터 결정되므로 종종 설계가중치(design weight) 또는 기본가중치(base weight)로 불리운다. 만약 모집단 전체에 대해 관측 가능한 변수 $\mathbf{x}_i$가 존재하는 경우 주어진 표본 하에서

$$\sum_{i \in A} d_i \mathbf{x}_i \neq \sum_{i=1}^{N} \mathbf{x}_i \tag{1.2}$$

* 2004년 연세대학교 교내 연구비 지원으로 이루어진 연구임.

1) (120-749) 서울시 서대문구 신촌동 134, 연세대학교 응용통계학과, 부교수
E-mail: kimj@yonsei.ac.kr

이라면 보조 정보 $\sum_{i=1}^{N}\mathbf{x}_i$를 이용하여 (1.1)의 추정량이 개선될 여지가 있다. 이 경우 가중치 보정추정량은

$$\hat{Y}_w = \sum_{i\in A} w_i y_i \tag{1.3}$$

의 형태로 표현할 수 있으며, 이때 최종 가중치 w_i는

$$\sum_{i\in A} w_i \mathbf{x}_i = \sum_{i=1}^{N} \mathbf{x}_i \tag{1.4}$$

을 만족한다. 이러한 가중치 보정추정량에 대한 보다 자세한 내용은 Deville과 Särndal(1992)이나 Fuller(2002)를 참고할 수 있다.

2. 분산 추정

가중치 보정추정량에 대한 분산 추정을 논하기 위하여 먼저 기본가중치를 사용한 추정량 (1.1)의 비편향 분산추정량이 관측치들의 이차 형식의 형태로 존재함을 가정하자. 즉,

$$\hat{V} = \sum_{i\in A}\sum_{j\in A} \Omega_{ij} y_i y_j \tag{2.1}$$

으로 정의된 분산추정량이

$$E\left(\hat{V}\right) = V\left(\hat{Y}_d\right) \tag{2.2}$$

을 만족한다고 하자. 여기서 Ω_{ij}는 이차항 $y_i y_j$의 계수로써 Horvitz-Thompson 분산추정량의 경우에는 $\Omega_{ij} = \left(\pi_{ij} - \pi_i\pi_j\right)/\pi_{ij}$으로 계산되고 여기서 π_{ij}는 이차표본 포함확률(second-order inclusion probability)이다. 또한 추정량의 근사적 성질을 다루기 위하여 Isaki와 Fuller(1982)에서처럼 유한 모집단과 그 표본들의 수열을 가정하고 그 경우 다음이 성립한다고 가정하자.

$$N^{-1}\sum_{i\in A} d_i\left(\mathbf{x}_i', y_i\right) - N^{-1}\sum_{i=1}^{N}\left(\mathbf{x}_i', y_i\right) = O_p\left(n^{-1/2}\right). \tag{2.3}$$

여기서 $O_p\left(a_n\right)$은 그 확률 변수를 a_n으로 나눈 것이 확률적으로 유계(bounded in probability)임을 의미한다. 마지막으로 최종 가중치 w_i는

$$w_i = d_i g\left(\mathbf{x}_i; \hat{\boldsymbol{\lambda}}\right)$$

으로 표현되고 여기서 $\hat{\boldsymbol{\lambda}}$는 (1.4)을 만족하도록 결정된다. 또한 $g\left(\mathbf{x}_i;\cdot\right)$는 두번 미분 가능하고 이차 미분 계수가 연속임을 가정하자.

이러한 경우 가중치 보정 추정량은

$$\hat{Y}_w = \sum_{i\in A} d_i g\left(\mathbf{x}_i; \hat{\boldsymbol{\lambda}}\right) y_i \tag{2.4}$$

으로 표현된다. 여기서 $\boldsymbol{\lambda}$는 그 자체로서는 관심이 없지만 관심 모수 Y의 추정에 도움이 되는 장애 모수(nuisance parameter)이다. 만약 $\hat{\boldsymbol{\lambda}}$가 상수라면 $g(\mathbf{x}_i;\hat{\boldsymbol{\lambda}})y_i$가 상수이고 따라서 (2.1)의 분산 추정식에 y_i 대신 g_iy_i를 대입하여 구현할 수 있을 것이다. 그러나 실제로는 $\hat{\boldsymbol{\lambda}}$는 (1.4)로부터 결정되는 확률 변수이고 이 확률 변수의 변동성을 반영하여 분산 추정식을 구현하여야 할 것이다. 장애 모수(nuisance parameter) $\boldsymbol{\lambda}$의 추정을 위하여

$$U(\boldsymbol{\lambda}) \equiv \sum_{i\in A} d_i g(\mathbf{x}_i;\boldsymbol{\lambda})\,\mathbf{x}_i - \sum_{i=1}^{N}\mathbf{x}_i$$

으로 정의하면 $U(\boldsymbol{\lambda})=\mathbf{0}$의 해가 $\hat{\boldsymbol{\lambda}}$이 될 것이다.

테일러 전개를 위하여 $U_0(\boldsymbol{\lambda})=E\{U(\boldsymbol{\lambda})\}$으로 정의하고 $U_0(\boldsymbol{\lambda})=\mathbf{0}$의 해를 $\boldsymbol{\lambda}_0$이라 하자. 이 $\boldsymbol{\lambda}_0$은 $\sqrt{n}\left(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0\right)=O_p(1)$ 을 만족함을 보일수 있다. 따라서 (2.4)의 추정량을 $\boldsymbol{\lambda}_0$를 중심으로 테일러 전개를 실시하면

$$\hat{Y}_w = \sum_{i\in A} d_i g_i(\boldsymbol{\lambda}_0)\,y_i + \left\{\sum_{i\in A} d_i\mathbf{h}_i(\boldsymbol{\lambda}_0)\,y_i\right\}'\left(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0\right) + O_p\left(n^{-1}N\right) \tag{2.5}$$

이고 여기서

$$\mathbf{h}_i(\boldsymbol{\lambda}) = \frac{\partial g(\mathbf{x}_i;\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}$$

으로 정의된다. 마찬가지로 $U(\hat{\boldsymbol{\lambda}})=\mathbf{0}$을 $\boldsymbol{\lambda}=\boldsymbol{\lambda}_0$에서 전개하면

$$\mathbf{0} = U(\boldsymbol{\lambda}_0) + \left\{\sum_{i\in A} d_i\mathbf{h}_i(\boldsymbol{\lambda}_0)\,\mathbf{x}_i'\right\}\left(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0\right) + O_p\left(n^{-1}N\right) \tag{2.6}$$

이 성립하고 따라서 (2.6)을 (2.5)에 대입하면

$$\begin{aligned}\hat{Y}_w &= \sum_{i\in A} d_i g_i(\boldsymbol{\lambda}_0)\,y_i - \left\{\sum_{i\in A} d_i\mathbf{h}_i(\boldsymbol{\lambda}_0)\,y_i\right\}'\left\{\sum_{i\in A} d_i\mathbf{h}_i(\boldsymbol{\lambda}_0)\,\mathbf{x}_i'\right\}^{-1}\cdot U(\boldsymbol{\lambda}_0) + O_p\left(n^{-1}N\right)\\ &= \sum_{i=1}^{N}\mathbf{x}_i'B + \sum_{i\in A} d_i g_i(\boldsymbol{\lambda}_0)\,(y_i-\mathbf{x}_i'B) + O_p\left(n^{-1}N\right)\end{aligned} \tag{2.7}$$

으로 표현될 수 있는데 여기서

$$B = \left\{\sum_{i\in A} d_i\mathbf{h}_i(\boldsymbol{\lambda}_0)\,\mathbf{x}_i'\right\}^{-1}\sum_{i\in A} d_i\mathbf{h}_i(\boldsymbol{\lambda}_0)\,y_i$$

으로 정의된다.

식 (2.7)으로부터 가중치 보정 추정량의 분산식을 다음과 같이 유도할 수 있다.

$$V\left(\hat{Y}_w\right) = V\left(\sum_{i\in A} d_i g_i(\boldsymbol{\lambda}_0)\,(y_i-\mathbf{x}_i'B)\right) + o\left(n^{-1}N^2\right). \tag{2.8}$$

만약 $\boldsymbol{\lambda}_0$를 안다면 $g(\mathbf{x}_i;\boldsymbol{\lambda}_0)(y_i-\mathbf{x}_i'B)$은 관측 가능하고 따라서 (2.1)의 분산 추정식에 y_i 대신 $g(\mathbf{x}_i;\boldsymbol{\lambda}_0)(y_i-\mathbf{x}_i'B)$를 대입하여 분산 추정량을 구현할 수 있을 것이다. 그러나 실제로는 $\boldsymbol{\lambda}_0$를 모르므로 그의 추정치인 $\hat{\boldsymbol{\lambda}}$를 사용하여

$$\hat{g}_i\hat{e}_i = g\left(\mathbf{x}_i;\hat{\boldsymbol{\lambda}}\right)\left(y_i-\mathbf{x}_i'\hat{B}\right)$$

을 계산한 후 (2.1)의 분산 추정식에 y_i 대신 $\hat{g}_i\hat{e}_i$를 대입하여 분산 추정량을 구현하면 될 것이다. 여기서

$$\hat{B} = \left\{\sum_{i\in A} d_i\mathbf{h}_i\left(\hat{\boldsymbol{\lambda}}\right)\mathbf{x}_i'\right\}^{-1}\sum_{i\in A} d_i\mathbf{h}_i\left(\hat{\boldsymbol{\lambda}}\right)y_i \tag{2.9}$$

으로 계산된다. 이 경우 분산 추정량

$$\hat{V} = \sum_{i\in A}\sum_{j\in A}\Omega_{ij}\hat{g}_i\hat{g}_j\hat{e}_i\hat{e}_j \tag{2.10}$$

이 (2.8)의 분산값에 대하여 근사적 일치성(asymptotic consistency)을 가지려면 $g_i(\hat{\boldsymbol{\lambda}}) = g(\mathbf{x}_i;\hat{\boldsymbol{\lambda}})$가 $g_i(\boldsymbol{\lambda}_0) = g(\mathbf{x}_i;\boldsymbol{\lambda}_0)$에 대하여 확률적으로 일양 수렴(uniformly converge)하여야 할 것이다. 즉,

$$\hat{V}_0 = \sum_{i\in A}\sum_{j\in A}\Omega_{ij}g_i\left(\hat{\boldsymbol{\lambda}}\right)g_j\left(\hat{\boldsymbol{\lambda}}\right)\hat{e}_i\hat{e}_j \tag{2.11}$$

은 (2.8)의 분산에 대해 근사적 일치성을 가지므로 (2.10)의 분산 추정량이 (2.11)에 대해 근사적 일치성을 가지면 (2.10)의 근사적 일치성이 보여지게 되는데 이에 대한 충분조건으로는

$$\lim_n \max_{i\in A}\partial g_i(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda} < K \tag{2.12}$$

이 됨을 보일 수 있는데 (2.12)은 결국 $g_i(\hat{\boldsymbol{\lambda}})$가 $g_i(\boldsymbol{\lambda}_0)$로 수렴하는 속도가 i와 상관없이 이루어지기 위한 충분조건이다. 대부분의 경우에는 최종 가중치가 지나치게 큰 값을 가지지 않도록 제약 조건을 주게 되므로 (2.12)의 조건은 만족하게 될 것이다. 만약 이러한 일양 수렴 조건이 만족되지 않는 경우에는 (2.10)에서 $g_i(\boldsymbol{\lambda}_0)=1$을 사용한

$$\hat{V} = \sum_{i\in A}\sum_{j\in A}\Omega_{ij}\hat{e}_i\hat{e}_j \tag{2.13}$$

을 사용할 것을 추천한다.

Deville과 Särndal(1992)은 (2.10)의 분산 추정식을 사용하되 $\hat{B}$으로

$$\hat{B} = \left\{\sum_{i\in A} d_ig_i\mathbf{x}_i\mathbf{x}_i'\right\}^{-1}\sum_{i\in A} d_ig_i\mathbf{x}_iy_i$$

을 사용할 것을 제안하였는데

$$\frac{\partial g_i(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}} = g_i(\boldsymbol{\lambda})\mathbf{x}_i \tag{2.14}$$

인 경우에 (2.9)의 $\hat{B}$와 동일해 진다. Deville과 Särndal(1992)에서 소개된 대부분의 가중치 보정치는 (2.14)의 조건을 만족시킨다.

3. 적용

3.1. 모형 기반 가중치 보정추정량에의 적용

가중치 보정 추정량은 유한 모집단의 관심 변수와 보조 변수간에 대한 모형을 바탕으로 종종 구현된다. 이 때 사용되는 모형은 유한 모집단 자체를 특정 무한 모집단에서 얻어지는 것으로 가정하게 되는데 이를 초모집단 모형(superpopulation model)이라고 한다. Wu와 Sitter(2001)은 초모집단 모형이

$$E\left(y_i \mid \mathbf{x}_i\right) \quad = \quad \mu\left(\mathbf{x}, \boldsymbol{\theta}\right)$$

으로 표현되는 경우에 (1.4)의 보정식(calibration equation)을

$$\sum_{i \in A} w_i \left(1, \hat{\mu}_i\right) = \sum_{i=1}^{N} \left(1, \hat{\mu}_i\right) \tag{3.1}$$

으로 (여기서 $\hat{\mu}_i = \mu(\mathbf{x}, \hat{\boldsymbol{\theta}})$이다) 사용한 모형 기반 가중치 보정 추정량을 제안하였다. Wu와 Sitter(2001)은 이러한 경우의 분산 추정량을 고려하였는데 단순한 카이제곱 거리 형태의 목적함수를 이용한 가중치 보정에만 적용하였고 일반적인 형태의 가중치 보정 추정량은 고려하지 못하였다.

본 논문의 2절에서 다룬 내용을 이용하면 보다 일반적인 경우도 그대로 적용 가능할 것이다. 이 경우 장애모수는 g_i의 계산에 사용되는 $\boldsymbol{\lambda}$와 $\hat{\mu}_i$의 계산에 사용되는 $\boldsymbol{\theta}$이다. 따라서 $\boldsymbol{\eta} = \left(\boldsymbol{\lambda}', \boldsymbol{\theta}'\right)'$ 으로 정의하고 이 장애 모수의 추정식(estimating equation)을

$$U\left(\boldsymbol{\eta}\right) = \begin{pmatrix} U_1\left(\boldsymbol{\lambda}, \boldsymbol{\theta}\right) \\ U_2\left(\boldsymbol{\theta}\right) \end{pmatrix} = \mathbf{0}$$

으로 정의하면

$$U_1\left(\boldsymbol{\lambda}, \boldsymbol{\theta}\right) = \sum_{i \in A} d_i g\left(\mathbf{x}_i; \boldsymbol{\lambda}\right) \mu\left(\mathbf{x}, \boldsymbol{\theta}\right) - \sum_{i=1}^{N} \mu\left(\mathbf{x}, \boldsymbol{\theta}\right)$$

이고 $U_2\left(\boldsymbol{\theta}\right)$은 초모집단 모형을 바탕으로한 모수 추정식이 될 것이다. 따라서 식 (2.7)의 유도과정과 동일한 방법을 사용하면

$$\hat{Y}_w \quad = \quad \sum_{i \in A} d_i g_i\left(\boldsymbol{\lambda}_0\right) y_i - \left\{ \frac{\partial \hat{Y}_w\left(\boldsymbol{\lambda}_0\right)}{\partial \boldsymbol{\eta}'} \right\} \left\{ \frac{\partial U\left(\boldsymbol{\eta}_0\right)}{\partial \boldsymbol{\eta}} \right\}^{-1} U\left(\boldsymbol{\lambda}_0\right) + O_p\left(n^{-1} N\right)$$

이 될 것이다. 여기서 $\mathbf{z}_i = \left(1, \hat{\mu}_i\right)'$이라 정의하고

$$\begin{aligned} \frac{\partial \hat{Y}_w\left(\boldsymbol{\lambda}_0\right)}{\partial \boldsymbol{\eta}'} \quad &= \quad \left(\frac{\partial \hat{Y}_w\left(\boldsymbol{\lambda}_0\right)}{\partial \boldsymbol{\lambda}'}, \mathbf{0}' \right) \\ \frac{\partial U\left(\boldsymbol{\eta}\right)}{\partial \boldsymbol{\eta}} \quad &= \quad \begin{pmatrix} \sum_{i \in A} d_i \mathbf{h}_i\left(\boldsymbol{\lambda}\right) \mathbf{z}_i' & \partial U_1 / \partial \boldsymbol{\theta} \\ \mathbf{0} & \partial U_2 / \partial \boldsymbol{\theta} \end{pmatrix} \end{aligned}$$

을 이용하면

$$\hat{Y}_w = \sum_{i=1}^{N} \mathbf{z}_i' B_1 + \sum_{i \in A} d_i g_i(\boldsymbol{\lambda}_0)(y_i - \mathbf{z}_i' B_1) + B_1' B_2' U_2(\boldsymbol{\theta}) + O_p\left(n^{-1}N\right) \tag{3.2}$$

으로 표현되고 여기서

$$B_1 = \left\{\sum_{i \in A} d_i \mathbf{h}_i(\boldsymbol{\lambda}_0)\mathbf{z}_i'\right\}^{-1} \sum_{i \in A} d_i \mathbf{h}_i(\boldsymbol{\lambda}_0) y_i$$

이고

$$B_2 = \left(\partial U_2/\partial \boldsymbol{\theta}\right)^{-1} \partial U_1/\partial \boldsymbol{\theta}$$

으로 표현된다. 이 때, 적절한 조건 하에서 (Wu와 Sitter, 2001, theorem 1 참조) $B_2 = o_p(1)$이 성립하고 따라서

$$\hat{Y}_w = \sum_{i=1}^{N} \mathbf{z}_i' B_1 + \sum_{i \in A} d_i g_i(\boldsymbol{\lambda}_0)(y_i - \mathbf{z}_i' B_1) + o_p\left(n^{-1/2}N\right) \tag{3.3}$$

의 테일러 전개가 유도되며 이 때 얻어지는 모형 기반 가중치 보정 추정량의 분산 추정량은 (2.10)의 형태로 구현하되

$$\hat{e}_i = y_i - \mathbf{z}_i' \hat{B}_1$$

으로 표현될 수 있을 것이고 이 경우 $\mathbf{z}_i = (1, \hat{\mu}_i)'$이고

$$\hat{B}_1 = \left\{\sum_{i \in A} d_i \mathbf{h}_i\left(\hat{\boldsymbol{\lambda}}\right)\mathbf{z}_i'\right\}^{-1} \sum_{i \in A} d_i \mathbf{h}_i\left(\hat{\boldsymbol{\lambda}}\right) y_i$$

가 될 것이다.

3.2. 경험적 우도 함수를 이용한 가중치 보정 추정량에의 적용

경험적 우도 함수(empirical likelihood function)는 관측치가 취할 수 있는 값들의 공간(space)이 실제 관측된 값들에만 존재함을 가정하는 일종의 비모수적 다항 분포에서 얻어지는 우도 함수로써 Chen과 Qin(1993)은 이를 가중치 보정에서의 목적 함수로 이용할 수 있음을 밝혔고 Chen과 Sitter(1999)와 Kim과 Lee(2006)은 이를 비균등 확률 추출(unequal probability sampling)로 확장 시켰다. Chen과 Sitter(1999)의 방법을 통해서 얻어지는 가중치 보정항은

$$g(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{1}{\lambda_1 + \boldsymbol{\lambda}_2(\mathbf{x}_i - \bar{\mathbf{x}}_N)}$$

으로 표현되고 Kim과 Lee(2006)의 방법을 이용하는 경우에는

$$g(\mathbf{x}_i; \boldsymbol{\lambda}) = \frac{1}{\lambda_1 + \boldsymbol{\lambda}_2 d_i(\mathbf{x}_i - \bar{\mathbf{x}}_N)}$$

으로 표현된다. 여기서 λ_1과 $\boldsymbol{\lambda}_2$는

$$N^{-1}\sum_{i\in A} d_i g\left(\mathbf{x}_i;\boldsymbol{\lambda}\right)\left(1,\mathbf{x}_i'\right)' = \left(1,\bar{\mathbf{x}}_N'\right)'$$

을 만족하도록 결정되어지는 장애 모수의 추정치이고 $\bar{\mathbf{x}}_N = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i$이다. 이러한 경우 가중치 보정항은

$$g\left(\mathbf{x}_i,\boldsymbol{\lambda}\right) = \frac{1}{\lambda_1 + \boldsymbol{\lambda}_2\mathbf{u}_i} \tag{3.4}$$

으로 일반적으로 표현될 수 있을 것이다. 분산 추정식 (2.10)에 사용되는 잔차항은

$$\hat{e}_i = y_i - \left(1,\mathbf{x}_i'\right)\hat{B}$$

으로 표현되며 여기서 $\tilde{\mathbf{u}}_i = \left(1,\mathbf{u}_i'\right)'$이라 할 때

$$\hat{B} = \left(\sum_{i\in A} d_i g_i^2 \tilde{\mathbf{u}}_i\tilde{\mathbf{u}}_i'\right)^{-1}\left(\sum_{i\in A} d_i g_i^2 \tilde{\mathbf{u}}_i y_i\right) \tag{3.5}$$

이 될 것이다. Chen과 Sitter(1999)는 위와는 달리 일반화 회귀 추정량의 분산 추정식을 이용한

$$\hat{B} = \left(\sum_{i\in A} d_i \tilde{\mathbf{u}}_i\tilde{\mathbf{u}}_i'\right)^{-1}\left(\sum_{i\in A} d_i \tilde{\mathbf{u}}_i y_i\right) \tag{3.6}$$

을 사용할 것을 제안하였다.

4. 모의 실험

3.2절에서 다룬 경험적 우도 함수를 이용한 가중치 보정 추정량의 분산 추정량에 대한 성질을 모의 실험을 통하여 확인해 보았다. 이를 위하여 크기가 $N = 10{,}000$인 유한 모집단을 아래의 두 가지 경우로 각각 발생시켰다.

[A] $y_i = 1 + \sqrt{2}\left(x_i - 3\right) + e_i,$

[B] $y_i = \left(x_i - 3\right)^2 + e_i.$

여기서 $x_i \sim N\left(3,1\right)$ 과 $e_i \sim N\left(0,1\right)$을 따르도록 독립적으로 발생시켰다. 이 유한 모집단으로부터 각각 크기가 n인 단순 임의 표본을 얻었는데 여기서 사용된 표본은 $n = 50$과 $n = 200$이었다. 이러한 표본 추출을 $B = 5{,}000$회 반복한 몬테 카를로 표본을 얻어내었다.

각각의 표본으로부터 $\bar{X}_N = N^{-1}\sum_{i=1}^{N} x_i$ 값을 보조 정보로 사용한 경험적 우도 함수를 이용한 가중치 보정 추정량을 구현하였는데 이 점추정량에 대한 분산 추정량을 다음과 같이 세가지로 구현해 보았다.

1. $\hat{V}_1 = \sum_{i\in A}\sum_{j\in A}\Omega_{ij}\hat{e}_{i1}\hat{e}_{j1}$: 여기서 $\hat{e}_{i1}$은 (3.6)으로 정의된 회귀 계수 추정식을 이용한 잔차항으로 $\hat{V}_1$은 Chen과 Sitter(1999)이 제안한 분산 추정량을 의미한다.

2. $\hat{V}_2 = \sum_{i\in A}\sum_{j\in A}\Omega_{ij}\hat{e}_i\hat{e}_j$: 여기서 $\hat{e}_i$은 (3.5)으로 정의된 회귀 계수 추정식을 이용한 잔차항이다.

3. $\hat{V}_3 = \sum_{i\in A}\sum_{j\in A}\Omega_{ij}\hat{g}_i\hat{e}_i\hat{g}_j\hat{e}_j$: 여기서 $\hat{e}_i$은 (3.5)으로 정의된 회귀 계수 추정식을 이용한 잔차항이다.

또한 각각의 분산 추정량으로부터 95% 신뢰 구간을 얻어 그 신뢰 구간의 포함도(coverage)도 함께 계산하였다.

표 4.1은 몬테 카를로 표본으로부터 위에서 설명한 세가지 분산 추정량의 상대 편향과 그 크기의 유의성에 대한 t-검정 통계량을 계산한 결과를 보여준다. 이 결과에 따르면 $\hat{V}_2$가 편향의 절대값이 작은 것으로 나타난다. 2절의 유도 결과에 따르면 $\hat{V}_3$가 가장 편향이 작을 것으로 기대되나 2절 마지막 부분에서 지적하였듯이 $\hat{g}_i$가 g_i로 일양수렴하지 않는 경우에는 $\hat{V}_2$가 더 나은 결과를 보여주므로 (3.4)의 형태를 갖는 가중치 조정항은 일양 수렴 조건 (2.12)을 만족하지 못하고 따라서 $\hat{V}_3$의 결과가 별로 좋지 않게 되는 것이다. 이러한 현상은 표본수가 작은 ($n = 50$) 경우에 현저하게 나타나며 표본수가 커지는 경우에는 ($n = 200$) 이러한 차이가 유의하지 않았다.

표 4.1: 세 가지 분산 추정량의 시뮬레이션 비교 실험 결과

n	Population	Variance Estimator	Relative Bias (%)	t-statistic H_0: Bias =0	Coverage (%)
50	A	$\hat{V}_1$	1.11	0.53	94.8
		$\hat{V}_2$	1.48	0.72	94.8
		$\hat{V}_3$	2.75	1.33	94.9
	B	$\hat{V}_1$	−12.66	−6.30	92.1
		$\hat{V}_2$	−4.81	−2.39	93.2
		$\hat{V}_3$	−9.96	−4.99	92.7
200	A	$\hat{V}_1$	−0.01	0.00	95.1
		$\hat{V}_2$	0.01	0.00	95.1
		$\hat{V}_3$	0.47	0.23	95.2
	B	$\hat{V}_1$	−2.00	−1.00	94.4
		$\hat{V}_2$	0.48	0.24	94.7
		$\hat{V}_3$	−0.64	−0.32	94.5

감사의 글

논문의 표현 및 내용 전개 방식에 대해 좋은 의견을 주신 익명의 심사 위원님과 편집 위원님께 감사를 드린다.

참고문헌

Chen, J. H. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information, *Biometrika*, **80**, 107–116.

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys, *Statistica Sinica*, **9**, 385–406.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376–382.

Fuller, W. A. (2002). Regression estimation for sample surveys, *Survey Methodology*, **28**, 5–23.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, **77**, 89–96.

Kim, J. K. and Lee, T. H. (2006). Calibration estimation using empirical likelihood in survey sampling, submitted.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, **96**, 185–193.

[2006년 10월 접수, 2007년 2월 채택]

Variance Estimation for General Weight-Adjusted Estimator*

Jae Kwang Kim[1)]

ABSTRACT

Linear estimator, a weighted sum of the sample observation, is commonly adopted to estimate the finite population parameters such as population totals in survey sampling. The weight for a sampled unit is often constructed by multiplying the base weight, which is the inverse of the first-order inclusion probability, by an adjustment term that takes into account of the auxiliary information obtained throughout the population. The linear estimator using the weight adjustment is often more efficient than the one using only the base weight, but its variance estimation is more complicated.

We discuss variance estimation for a general class of weight-adjusted estimator. By identifying that the weight-adjusted estimator can be viewed as a function of estimated nuisance parameters, where the nuisance parameters were used to incorporate the auxiliary information, we derive a linearization of the weight-adjusted estimator using a Taylor expansion. The method proposed here is quite general and can be applied to wide class of the weight-adjusted estimators. Some examples and results from a simulation study are presented.

Keywords: Calibration, Taylor expansion, regression estimator.

* The research was supported by 2004 Yonsei research grant.

1) Associate Professor, Department of Applied Statistics, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, Korea
E-mail: kimj@yonsei.ac.kr