응용통계연구 제 16권 1호, 2003년, pp. 33-44

단위 무응답 보정에서 보조변수의 선택에 관한 연구

손창균[1] 홍기학[2] 이기성[3]

요 약

조사과정에서 필연적으로 발생하는 무응답을 보정하기 위해 보조정보를 사용해야 한다. 이 때, 이용 가능한 보조정보의 차원이 크면, 계산과정에서 많은 시간이 소요되며 데이터를 다루기가 매우 어렵다. 또한 추정량의 분산이 보조정보의 차원에 의존하기 때문에 과소추정의 문제가 발생한다. 이러한 문제를 해결하기 위해 무응답 보정에서 적절한 보조정보의 선택 방법을 제안하였고, 이에 대한 효율성을 모의실험을 통해 살펴보았다.

주요용어: 단위무응답, 보조정보, 보정추정량, 회귀추정량, 변수선택법.

1. 서론

일반적으로 보조정보는 표본조사에서 종종 사용된다. 특히 회귀추정량이나 래킹추정량 또는 사후추정량을 도출하기 위해 추정과정이나 표본설계과정에서 보조정보를 사용한다. 그러나 이러한 보조변수의 선택은 단순히 관심변수와 강한 상관이 있는 보조변수만을 선택하거나, 그렇지 않으면 연구자의 주관에 의해 결정된다(Lundstörm 과 Särndal,1999). 전통적으로 보조정보를 사용함으로서 다음과 같은 장점에 의해 조사의 질을 개선할 수 있다. 첫째, 관심변수와 강한 상관관계가 있는 보조정보를 사용함으로서 표본분산을 줄일 수 있다. 둘째, 특별히 무응답 (nonresponse)이나 비포괄성(noncoverage)에 의한 편향을 감소시킬 수 있다. 셋째, 보조정보를 이용함으로서 다른 데이터로부터 얻은 결과들과 일치성을 가진다. 하지만, 이러한 장점들에도 불구하고 관심변수와 강한 상관이 있는 모든 보조변수를 이용하는 것은 이용 가능한 모든 보조변수의 수가 매우 많을 때, 보조변수 행렬의 차원이 상당히 커지기 때문에 계산과정이 매우 힘든 측면이 있다. 이에 대한 예로서 특별히 소지역(small area) 통계에서 이용 가능한 보조변수는 단순히 상관계수만을 고려할 경우 기존의 전수조사로부터 얻어진 자료 전체가 될 수도 있다. 또한 추정량의 분산이 보조변수의 차원에 의존하기 때문에 많 은 양의 보조변수를 사용하는 경우 분산의 과소추정 문제가 발생할 수 있다. 이와 같은 문제 이외에도 보조변수를 이용함에 있어서 보조변수들 간의 상호작용에 의해 추정과정에서 불필요한 보조변수를 사용하게 되는 경우가 발생할 수도 있다.

1) (520-714) 전남 나주시 대호동 동신대학교 컴퓨터학과, 전임강사
E-mail :ckson85@blue.dongshinu.ac.kr
2) (520-714) 전남 나주시 대호동 동신대학교 컴퓨터학과, 부교수
E-mail :khhong@blue.dongshinu.ac.kr
3) (565-701) 전북 완주군 삼례읍 후정리 우석대학교 전산통계학과, 부교수
E-mail :gisung@core.woosuk.ac.kr

Bankier(1990)는 변수선택과 관련하여 2단계 일반화 회귀추정과정에서 보조변수의 차원을 축소하기 위해 조건수 축소과정을 이용하였고, Bradsley와 Chambers(1984)는 능형회귀에 관련하여 단순히 추정량 개선을 목적으로 보조변수에 내재된 다중공선성의 문제를 다룬바 있지만, 변수선택과정은 고려하지 않았다.

Silva와 Skinner(1997)는 유한모집단에 대한 회귀추정량을 도출함에 있어서 변수선택방법을 적용하였다. 이들은 보조변수들 간의 다중공선성(multicollinearity) 문제를 고려하여 능형회귀모형을 사용하였으며, 변수선택방법으로 조건수 축소(condition number reduction) 과정을 적용하였다. 그러나 이들은 단순히 회귀추정량의 도출에 차원축소를 위한 변수선택방법을 적용하였으며, 조사 무응답과 같은 현실적인 문제는 고려하지 않았다.

최근에 Son, Hong 그리고 Lee(2001)은 무응답을 고려한 분산의 보정 추정과정을 소개하였으며, 이때 이용 가능한 보조정보를 보조변수의 모집단 총합과 분산으로 정의하여 사용하였다. 그러나 이들은 이용 가능한 보조변수의 차원이 클 때 분산의 과소추정의 문제는 다루지 않았다.

이러한 연구들을 기초로 본 논문에서는 먼저 조사과정에서 필연적으로 발생하는 무응답 중에서 단위 무응답의 경우를 고려하여 보정추정과정에서 관심변수와 관련이 있는 이용 가능한 보조정보의 선택에 대해 살펴보고자 한다. 또한 이용 가능한 보조변수들에 대해 변수선택방법을 적용하여 선택된 보조변수들을 이용한 보정추정과정을 적용하여 관심변수에 대한 추정량을 구해 보고자 한다.

2. 보정추정량

2.1. 단위 무응답하에서 보정추정량의 도출

우선 이론전개를 위해 몇 가지 기호를 정의하자. $U = \{1, 2, \cdots, N\}$ 를 N개의 구별 가능한 유한모집단이라 하자. 또한 $s(\subset U)$를 추출설계 $p(s)$에 의해 모집단으로부터 추출된 크기 n인 표본이라 하자. 또한 무응답을 정의하기 위해 표본 s로부터 응답확률 $q(r|s)$ 로서 응답한 크기 m인 응답집합 $r(\subset s)$을 고려하자. $\mathbf{x}_k = (x_{k1}, x_{k2}, \cdots, x_{kq})'$ 을 응답단위 k와 연관된 $q \times 1$인 보조변수벡터라 하자. 보조변수의 모집단 총합인 $q \times 1$ 벡터 $\mathbf{X}_{tot} = \sum_{k \in U} \mathbf{x}_k$와 $\mathbf{x}_k$의 표본총합이 기지라고 가정하자. 이 때 표본총합은 $q \times 1$인 벡터로 $\mathbf{x}_{tot} = \sum_{k \in s} d_k \mathbf{x}_k$라 하자. y_k 를 k 번째 응답 원소에 대한 조사변수 y의 값이라 하고, y_k는 단지 $k \in r$ 에 대해 관찰된다. 목적은 모집단 총합 $Y = \sum_{k \in U} y_k$를 추정하는 것이다.

완전응답의 경우 크기 n인 표본 s로 부터 모집단 총합 Y를 추정고자 할 때, 표본의 각 원소가 포함확률 π_k를 가진다면 Y에 대한 설계 비편향(design unbiased)추정량은 다음과 같은 Horvitz-Thompson(HT)추정량이 된다.

$$\hat{Y}_{HT} = \sum_{s} d_k y_k \tag{2.1}$$

여기서, $d_k = 1/\pi_k$인 단위 k의 추출가중치이다.

보다 실제적인 상황으로 응답확률 $q(r|s) = \theta_k$을 고려하면, 보조변수의 모집단 총합 벡터인 $\mathbf{X}_{tot} = \sum_{k\in U} \mathbf{x}_k$ 가 기지라는 가정 하에서 Y에 대한 의사 설계 비편향추정량(quasi-design- unbiased estimator)은 다음과 같다.

$$\begin{aligned} \hat{Y}_w &= \sum_r d_k g_k y_k \\ &= \sum_r w_k y_k \end{aligned} \tag{2.2}$$

여기서, $g_k = 1 + (\mathbf{X}_{tot} - \sum_r d_k \mathbf{x}_r)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$ 로서 무응답을 보정한 가중인자 이다.

무응답 단위에 대한 가중치 조정과 관련하여 이용 가능한 보조정보의 수준-모집단- 에 따라 다음과 같은 보정방정식을 만들 수 있다.

$$\sum_U \mathbf{x}_k = \sum_r w_k \mathbf{x}_k \tag{2.3}$$

위의 보정방정식을 만족하는 새로운 가중치 w_k는 다음과 같은 거리함수를 최소로 한다.

$$G(w_k, d_k) = \sum_r \frac{(d_k - w_k)^2}{d_k} \tag{2.4}$$

이러한 조건으로부터 $q_k = 1$ 로 놓고, 원래의 추출가중치와 가장 근접한 새로운 가중치는 모집단 보조정보를 이용하는 경우 다음과 같다.

$$w_k = d_k \left[1 + (\mathbf{X}_{tot} - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k \right] \tag{2.5}$$

새로운 가중치 식(2.5)를 다음과 같은 회귀추정량의 형태로 다시 표현할 수 있다.

$$\begin{aligned} \hat{Y}_w &= \hat{Y}_r + (\mathbf{X}_{tot} - \mathbf{x}_r)'\mathbf{b} \\ &= \hat{Y}_{reg} \end{aligned} \tag{2.6}$$

여기서, $\mathbf{x}_r = \sum_r d_k \mathbf{x}_k$ 는 응답단위에 대한 보조변수벡터이며, $\mathbf{b} = (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_r d_k \mathbf{x}_k' y_k$ 인 $q \times 1$ 벡터이다.

2.2. 보조변수 차원에 대한 보정추정량의 분산의 종속성

식(2.6)으로부터 모집단 총합의 일반화 회귀추정량은 다음과 같은 기본 선형모형으로부터 유추할 수 있다.

$$y_k = \mathbf{x}_k'\beta + \epsilon_k \tag{2.7}$$

이 때, ϵ_k 는 평균이 0 이고, 분산이 σ^2 인 독립인 분포를 따른다.

이러한 선형회귀모형하에서 $\mathbf{x}_k$ 의 조건하에서 추정오차 $\hat{Y}_{reg} - Y$ 의 분산은 다음과 같다.

$$Var(\hat{Y}_{reg} - Y|\mathbf{x}_k) = N^2 \frac{\sigma^2}{m} \left[(1-f) + (\mathbf{X}_{tot} - \mathbf{x}_r)' \hat{S}_x^{-1} (\mathbf{X}_{tot} - \mathbf{x}_r)\right] \tag{2.8}$$

여기서, $\hat{S}_x = \sum_{k \in r} d_k \mathbf{x}_k \mathbf{x}_k'$ 이다.

$c_g^2 = (\mathbf{X}_{tot} - \mathbf{x}_r)' \hat{S}_x^{-1} (\mathbf{X}_{tot} - \mathbf{x}_r)$ 이라 하면 분산 식(2.8)은 c_g^2 에 의존함을 알 수 있다.

이러한 종속성을 보다 명확히 하기 위해 $\mathbf{x}_k$ 가 독립이고, 동일한 정규분포를 가정하면, $(\mathbf{X}_{tot} - \mathbf{x}_r)$ 과 $\hat{S}_x^{-1}$ 의 독립성과 $E(\hat{Y}_{reg} - Y|\mathbf{x}_k) = 0$ 을 이용하여 비 조건부 분산은 다음과 같이 유도된다.

$$\begin{aligned} Var(\hat{Y}_{reg} - Y) &= N^2 \frac{\sigma^2}{m} \left[(1-f) + tr\left[E(\mathbf{X}_{tot} - \mathbf{x}_r)'(\mathbf{X}_{tot} - \mathbf{x}_r)\right] E(\hat{S}_x^{-1})\right] \\ &= N^2 \frac{\sigma^2}{m} (1-f)\left[1 + q/(m-q-2)\right] \end{aligned} \tag{2.9}$$

결과적으로 식(2.9)는 보조변수의 차원인 q 에 의존함이 명확하다. 보조정보의 수준에 따른 보정추정과정으로부터 일반화 회귀추정량의 근사적인 MSE의 추정량은 다음과 같다.

$$MSE = \frac{1-f}{m(m-q-2)} \sum_{k \in r} {g_k}^2 \hat{e}_k^2 \tag{2.10}$$

여기서 $g_k = 1 + (\mathbf{X}_{tot} - \sum_r d_k \mathbf{x}_r)' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k)^{-1} \mathbf{x}_k$ 이며, $k \in r$ 에 대해 $\hat{e}_k = y_k - \mathbf{x}_k' \mathbf{b}$ 이다.

3. 변수선택 후 보정추정과정

3.1. 변수선택

본 논문에서 이용한 변수선택방법으로는 전진선택법(forward selection)과 조건수 축소(condition number reduction) 방법이다. 또한, 보조변수들 간에 존재하는 다중공선성의 문제를 해결하기 위해 Bardsley와 Chambers(1984)에 의해 제안된 방법을 함께 고려한다. 먼저 전진선택법의 과정을 살펴보면, 우선 상관계수의 크기에 따라 순서대로 보조변수 를 하나씩 추가하여 보정추정량과 MSE의 추정량을 구하고, 이때 MSE가 증가하는 시점의 보 조변수들을 선택한다. 이와 같이 선택된 보조변수들을 이용한 보정 추정량은 가장 작은 MSE 추정량을 얻게 된다. 다음으로 조건수 축소과정을 살펴보면, 식(2.6)으로부터 회귀추정량이 $CP_w = (\sum_r d_k \mathbf{x}_k' \mathbf{x}_k)'$ 에 의존함을 알 수 있다. Bankier(1992)에 의해 제시된 조건수 축소과정은 $CP_w = (\sum_r d_k \mathbf{x}_k' \mathbf{x}_k)$ 에 대해 가장 큰 조건수를 나타내는 보조변수들을 제거해 가는 방법으로 이에 대한 과정을 살펴보면 다음과 같다.

단계 1] 모든 이용 가능한 보조변수에 대한 $CP_w = (\sum_r d_k \mathbf{x}_k' \mathbf{x}_k)$ 를 계산한다.

단계 2] $CP_w = (\sum_r d_k \mathbf{x}_k' \mathbf{x}_k)$ 의 Hermite 정준행렬 H 를 계산한다. 이 행렬에서 선형종속을 나타내는 0 인 각각의 열(columns)을 제거한다.

단계 3] 선형종속인 열을 제거한 후 축소된 CP_w 로부터 조건수 $c = \lambda_{max}/\lambda_{min}$ 를 계산하고, 만일 특정한 값 L 과 비교하여 $c < L$ 이면 과정을 종료하고, 남아있는 모든 보조변수를 이용한다. 여기서 λ_{max} 와 λ_{min} 은 각각 CP_w의 가장 큰 고유치와 가장 작은 고유치를 나타낸다.

끝으로 보조변수들간의 상호 연관성에 의해 발생하는 다중공선성의 문제를 해결하기 위해 보 정추정량의 식을 다음과 같은 능형 회귀추정량으로 대치하고, 그에 따르는 MSE를 추정하여 추정량의 안정성을 살펴본다.

$$\hat{Y}_{reg} = \hat{Y}_r + (\mathbf{X}_{tot} - \mathbf{x}_r)(\lambda C^{-1} + \sum_r \mathbf{x}_k \mathbf{x}_k)^{-1} \sum_r \mathbf{x}_k' y_k \tag{3.1}$$

3.2. 변수선택 후 보정추정과정

3.1절에서 선택된 보조변수를 이용하여 새로 구한 보정가중치는 다음과 같다.

$$w_k^* = d_k \left[1 + (\mathbf{X}_{tot}^* - \mathbf{x}_r^*)' (\sum_r d_k \mathbf{x}_k^* \mathbf{x}_k^{*'})^{-1} \mathbf{x}_k^* \right] \tag{3.2}$$

여기서 $\mathbf{X}_{tot}^*$ 는 변수선택 후 최종적으로 결정된 보조변수들의 총합벡터이며, $\mathbf{x}_r^*$ 는 변수선택 후 $k \in r$ 인 단위들의 보조변수 벡터이다.

따라서 최종적으로 무응답을 보정한 모집단 총합추정량은 다음과 같다.

$$\begin{aligned} \hat{Y}_{reg}^* &= \hat{Y}_r + (\mathbf{X}_{tot}^* - \mathbf{x}_r^*)' (\sum_r d_k \mathbf{x}_k^* \mathbf{x}_k^{*'})^{-1} \sum_r d_k \mathbf{x}_k^{*'} y_k \\ &= \hat{Y}_r + (\mathbf{X}_{tot}^* - \mathbf{x}_r^*)' \mathbf{b}^* \end{aligned} \tag{3.3}$$

여기서 $\mathbf{x}_r^* = \sum_r d_k \mathbf{x}_k^*$ 은 변수선택 후 최종적으로 보조변수로 결정된 응답단위에 대한 $h \times 1$ 인 보조변수 벡터이며, $\mathbf{b}^* = (\sum_r d_k \mathbf{x}_k^* \mathbf{x}_k^{*'})^{-1} \sum_r d_k \mathbf{x}_k^{*'} y_k$ 는 변수선택 후 추정된 $h \times 1$ 인 회귀계수벡터이다.

따라서 변수선택 후 보정추정량에 대한 근사적인 MSE의 추정량은 다음과 같다.

$$MSE^* = \frac{1-f}{m(m-h-2)} \sum_{k \in r} g_k^{*2} \hat{e}_k^{*2} \tag{3.4}$$

여기서 $g_k^* = 1 + (\mathbf{X}_{tot}^* - \sum_r d_k \mathbf{x}_k^*)' (\sum_r d_k \mathbf{x}_k^* \mathbf{x}_k^*)^{-1} \mathbf{x}_k^*$ 이며, $k \in r$ 에 대해 $\hat{e}_k^* = y_k - \mathbf{x}_k^{*'} \mathbf{b}^*$ 이다.

4. 모의실험

4.1. 통계량의 정의

모의실험을 위해 다음과 같이 정의된 Särndal 등(1992)의 *"MU284 스웨덴 데이터"*를 사용하 였다. 목적은 1985년도 스웨덴의 총 납세소득을 추정하고자 하며, 이 때 284개 시도의 납세소득 의 총합은 미지라고 가정하자. 관심변수 y 와 그에 따르는 7개의 보조변수를 아래와 같이 정의한 후 표본 추출설계는 전개를 간단히 하기 위해 단순임의 비복원 추출로 약 30% 표본인 100개 시 도를 표본으로 추출한다. 다음으로 무응답 가정을 위해 추출된 표본으로부터 90%, 80%, 70% 으로 응답률을 가정하여 최종적으로 응답단위를 결정한다.

표 4.1 : *MU284 데이터 세트* 의 변수에 대한 정의

변 수	내 용
y	1985년 284개 시 도의 납세소득 (단위 : 백만 Kroner)
x_1	1985년 284개 시 도의 인구 수 (단위 : 천명)
x_2	1975년 284개 시 도의 인구 수 (단위 : 천명)
x_3	1982년 시의회의 보수당 의석 수 (단위 : 명)
x_4	1982년 시의회의 사회민주당 의석 수 (단위 : 명)
x_5	1982년 시의회의 총 의석 수 (단위 : 명)
x_6	1984년 도시근로자의 취업 인구 수 (단위 : 명)
x_7	1984년 부동산 가치 (단위 : 백만 Kroner)

효율성 비교를 위해 먼저 다음과 같은 모집단 총합에 대한 추정량과 그에 따른 편향추정량을 정의하였다.

$$E(\hat{Y}_w) = \frac{1}{K}\sum^{K}\hat{Y}_w \tag{4.1}$$

$$Bias(\hat{Y}_w) = \frac{1}{K}\sum^{K}\left[\hat{Y}_w - E(\hat{Y}_w)\right] \tag{4.2}$$

이 때, K 는 반복 수이다.

이와 더불어 총합추정량에 대한 MSE 추정량과 모의실험을 통한 분산추정량은 다음과 같다.

$$MSE(\hat{Y}_w) = \frac{1}{K}\sum^{K}\left[\hat{Y}_w - E(\hat{Y}_w)\right]^2 \tag{4.3}$$

$$E(\hat{V}(\hat{Y}_w)) = \frac{1}{K}\sum^{K}\hat{V}(\hat{Y}_w) \tag{4.4}$$

마지막으로 추정량의 정도를 살펴보기 위해 근사 정규분포 이론에 기초한 모집단 총합의 95% 명목포함률(nominal coverage rate)을 다음과 같이 계산하였다.

$$Coverage(Y) = \hat{Y}_w \pm 1.96\sqrt{MSE} \tag{4.5}$$

4.2. 모의실험 결과

우선 가정된 모집단으로부터 응답한 단위들의 관심변수와 보조변수들 간의 상관계수를 구한 결과가 다음과 같다.

표 4.2 : 응답단위들에 대한 관심변수와 보조변수들 간의 상관계수(r)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y	0.97	0.98	0.43	0.38	0.54	0.99	0.94

표 4.2로부터 단순히 관심변수와 보조변수들 간의 상관관계만을 고려할 경우 7개의 보조변수가 단위 무응답 보정에 사용될 수 있으며, 보다 보수적으로 보조정보를 선택한다면, x_3, x_4, x_5 인 3개의 보조변수를 제외한 나머지 보조변수를 보정추정과정에서 이용할 수 있을 것이다. 그러나 이러한 주관적인 보조정보의 선택방법을 보다 객관적으로 적용하고, 또한 보조변수들 간의 내부적인 종속성을 고려하여 3절에서 제안한 변수선택 절차를 적용하여 다음과 같은 결과들을 얻었다.

표 4.3 : 전진선택법을 이용한 보조변수의 선택에 따른 통계량들의 비교($K = 1,000$)

선택 방법	선택변수	응답률 (%)	$\hat{Y}_w$ ($Y = 68,990$)	Bias ($\times 10^{-11}$)	MSE ($\times 10^6$)	$E(\hat{V})$ ($\times 10^6$)	95% Coverage
	모든 변수	90	63,342.8	6.41	0.788	0.009	0.926
		80	63,356.0	-0.17	1.654	0.028	0.940
		70	63,359.6	1.23	2.712	0.068	0.955
전	x_6	90	68,506.8	4.14	0.070	0.012	0.945
		80	68,511.8	4.22	0.163	0.036	0.944
		70	68,529.2	0.74	0.279	0.089	0.950
진	x_2, x_6	90	68,505.1	0.21	0.070	0.012	0.945
		80	68,510.1	9.30	0.162	0.037	0.944
		70	68,527.5	-1.79	0.279	0.090	0.950
선	x_1, x_2, x_6	90	68,503.0	3.00	0.070	0.012	0.943
		80	68,508.0	-2.69	0.162	0.037	0.944
		70	68,525.4	-7.25	0.279	0.092	0.950
택	x_1, x_2, x_6, x_7	90	63,345.8	0.97	0.788	0.009	0.926
		80	63,359.1	4.46	1.655	0.026	0.940
		70	63,362.8	0.94	2.712	0.065	0.955
법	x_1, x_2, x_5, x_6, x_7	90	63,343.5	-0.63	0.788	0.009	0.926
		80	63,356.7	-1.01	1.655	0.027	0.940
		70	63,360.3	-2.76	2.712	0.066	0.955
	$x_1, x_2, x_3, x_5, x_6, x_7$	90	63,342.4	-2.51	0.788	0.009	0.926
		80	63,356.6	1.66	1.655	0.027	0.940
		70	63,360.2	5.69	2.712	0.068	0.955

전진선택법의 경우 표 4.3으로부터 응답단위들에 대해 관심변수와 가장 상관관계가 큰 보조 변수 x_6 을 보정추정과정에 적용하여 총합 추정치를 구하고, 이 때의 MSE를 구하였다. 다음으로 상관계수가 큰 x_2 를 함께 고려한 x_2, x_6 일때의 총합 추정치와 MSE를 구하였다. 이와 같은 방법을 순차적으로 진행하여 MSE가 증가하는 시점을 찾으면, 보조변수가 3개인 x_1, x_2, x_6 경우이며, 이 때의 MSE가 가장 작음을 알 수 있다. 따라서 전진선택법을 이용했을 때 이용 가능한 보조변수로 x_1, x_2, x_6 를 선택함으로서 최적의 보조변수 집합이 됨을 알 수 있다.

다음으로 표 4.4로부터 조건수 축소방법을 이용하여 선택된 보조변수는 전진선택법과 마찬가지로 x_1, x_2, x_6 이며, 이 변수들의 선택하는 과정은 3.1절의 변수선택 과정에서 살펴본 바와 같이 우선 모든 이용 가능한 모든 변수들에 대한 Hermite 정준행렬을 만들고, 이 때 선형종속인 0인 열을 하나씩 제거해 나감으로서 결과적으로 세 변수를 선택하게 된다. 변수선택기준으로 $L = 900$으로 하여 앞의 3.1절에서 언급한 고유치의 비인 c 가 $L = 900$ 보다 작으로면 모든 변수를 이용하고, 그렇지 않으면, 해당변수를 제거하는 방법으로 적용하였다.

이 과정으로부터 맨 처음에 제거된 변수는 x_7이며, 다음으로 x_5, x_4, x_3 의 순서로 제거되었으며, 최종적으로 남은 변수는 x_1, x_2, x_6 가 되었으며, 이에 따른 추정치를 구하였다.

표 4.4 : 조건수 축소방법을 이용한 보조변수의 선택에 따른 통계량들의 비교($K = 1,000$)

선택 방법	선택변수	응답률 (%)	$\hat{Y}_w$ ($Y = 68,990$)	Bias ($\times 10^{-11}$)	MSE ($\times 10^6$)	$E(\hat{V})$ ($\times 10^6$)	95% Coverage
	모 든 변 수	90	67,129.6	-7.39	4.666	0.016	0.905
		80	66,908.5	3.99	10.420	0.045	0.904
		70	66,888.2	0.62	18.161	0.110	0.988
조 건 수 축 소	$x_1, x_2, x_3, x_4, x_5, x_6$	90	108,621.9	-5.89	0.172	0.132	0.946
		80	108,597.7	-10.10	0.413	0.377	0.951
		70	108,661.8	-4.46	0.746	0.860	0.954
	x_1, x_2, x_3, x_4, x_6	90	103,481.8	-1.57	0.149	0.105	0.942
		80	103,452.9	-7.35	0.352	0.301	0.947
		70	103,502.3	8.45	0.630	0.687	0.950
	x_1, x_2, x_3, x_6	90	97,082.8	-10.20	0.133	0.078	0.941
		80	97,051.2	9.23	0.310	0.224	0.946
		70	97,092.6	-4.65	0.554	0.513	0.951
	x_1, x_2, x_6	90	69,559.9	5.14	0.067	0.017	0.942
		80	69,538.2	2.75	0.157	0.049	0.945
		70	69,567.2	1.24	0.281	0.115	0.950

마지막으로 보조변수들 간의 다중공선성을 고려한 능형회귀법의 경우, x_1, x_2, x_6 으로 선택되었다. 이 때 능형상수 λ 는 0에서 0.05까지 0.002씩 변화를 주면서 최적의 값을 구했으며, 여기서는 $\lambda = 0.002$로 나타났고, VIF 값이 급격히 감소하였다. 단, 모의실험과정에서 비용상수 C 의 값은 1 로 고정했다.

이러한 능형회귀 과정으로부터 응답 집합에서 사용할 최적의 보조변수로는 앞의 두 가지 방법과 마찬가지로 x_1, x_2, x_6 으로 선택되었으며, 이 때 MSE 추정치가 가장 작게 나타났으며, 이 변수들을 이용한 추정치들이 표 4.5에 구해져 있다.

표 4.5 : 능형회귀법을 이용한 보조변수의 선택에 따른 통계량들의 비교($K = 1,000$)

선택 방법	선택변수	응답률 (%)	$\hat{Y}_w$ ($Y = 68,990$)	Bias ($\times 10^{-11}$)	MSE ($\times 10^6$)	$E(\hat{V})$ ($\times 10^6$)	95% Coverage
	모 든 변 수	90	73,853.3	9.02	5.752	0.016	0.914
		80	73,469.1	2.86	15.586	0.052	0.957
		70	73,527.6	-8.34	22.414	0.133	0.933
능	x_6	90	69,208.9	3.33	0.117	0.011	0.915
		80	69,170.3	-2.81	0.303	0.036	0.945
		70	69,172.4	0.28	0.499	0.090	0.948
형	x_2, x_6	90	69,157.0	-13.40	0.102	0.011	0.919
		80	69,124.6	-5.38	0.264	0.037	0.940
		70	69,125.2	-2.06	0.439	0.091	0.949
회	x_1, x_2, x_6	90	69,148.6	-6.66	0.101	0.011	0.918
		80	69,116.7	3.03	0.262	0.037	0.939
		70	69,117.6	-8.65	0.437	0.092	0.948
귀	x_1, x_2, x_6, x_7	90	96,456.5	-0.99	0.207	0.061	0.912
		80	96,409.3	14.10	0.532	0.189	0.940
		70	96,413.8	-11.60	0.890	0.451	0.942
	x_1, x_2, x_5, x_6, x_7	90	102,540.9	-6.19	0.207	0.083	0.911
		80	102,499.6	3.78	0.529	0.256	0.937
		70	102,506.6	5.13	0.903	0.605	0.945
	$x_1, x_2, x_3, x_5, x_6, x_7$	90	107,178.8	-8.37	0.197	0.104	0.922
		80	107,152.4	-6.87	0.497	0.319	0.939
		70	107,162.5	-3.30	0.872	0.752	0.943

위와 같은 변수선택 절차로부터 다음과 같은 결론을 도출할 수 있다.

첫째, 전진선택법을 적용한 경우 MSE의 증가를 살펴보면 선택변수로서 x_1, x_2, x_6 를 선택했을 때 가장 작아지며, 이 후 변수를 하나씩 추가할 때 MSE의 값은 급격히 증가함을 알 수 있다. 따라서 보조정보로서 x_1, x_2, x_6 를 선택하는 것이 바람직함을 알 수 있다.

둘째, 조건수 축소방법을 적용했을 때의 선택된 변수는 역시 전진선택법과 마찬가지로 x_1, x_2, x_6 이며, 이 때의 MSE가 가장 작아짐을 알 수 있다.

셋째, 보조변수들 간의 다중공선성을 고려한 능형회귀법의 경우 역시 보조변수로서 x_1, x_2, x_6 를 선택하는 것이 바람직하며, 이 때의 능형상수는 0.002로 결정되었다.

결과적으로 모의실험을 통해 세 가지 전통적인 회귀분석에서 사용되는 변수선택방법을 적용하여 보조변수를 선택함으로서 무응답 보정에서 사용되는 보조변수의 차원을 축소할 수 있었으며, 또한 보조변수의 차원에 의존하는 추정량의 분산의 안정화에 기여할 수 있음

을 알 수 있었다. 이 때, 총합 추정량들의 응답률에 따른 변동에서 응답률이 큰 경우가 작은 경우에 비해 모집단 총합과 차이가 나는 것은 표본단위들로 부터 응답률 90%, 80%, 70%로 응답집합을 선택하는 과정에서 생긴 변동에 기인한 것으로 사료된다. 결과적으로 응답률에 따른 총합 추정량의 변동 보다는 이용 가능한 보조변수에 대한 MSE 추정량이 최소가 될 때, 이 변수들을 보조변수로 선택하여, 보정 추정에 이용하는 것이며, 이 경우 응답률이 90%, 80%, 70% 에 대해 적절히 모집단 총합을 추정하고 있음을 볼 수 있다.

5. 결론

단위 무응답 보정에서 이용 가능한 모든 보조정보를 이용할 경우 추정량의 분산이 보조정보 의 차원에 의존하기 때문에 과대 또는 과소 추정의 문제가 발생한다. 또한, 보조변수들 간의 내부적인 종속성에 기인한 정보의 왜곡으로 인하여 추정과정에서 잘못된 보조정보를 사용하는 오류를 범할 수 있다.

본 논문에서는 이러한 문제점들을 극복하고, 보다 실제적인 문제로 조사과정에서 발생하는 무응답 문제를 전통적인 회귀분석 방법을 함께 고려하여 가장 적절한 보조정보를 선택하여 이 용함으로서 추정량의 과소추정의 문제와 계산의 복잡성을 해결하고자 하였다.

그 결과 모의실험을 통해 이용 가능한 7개의 보조변수들 중에서 응답집합에 있는 모든 보조정보를 통 계 이용자나 작성자의 주관에 따라 보조변수를 선택하기보다는 객관적인 통계절차에 근거하여 적절한 보조변수를 선택함으로서 관심변수에 대한 편향추정의 문제와 분산의 과소추정 문제를 해결하였다.

마지막으로 전통적인 회귀분석 방법에서 변수선택의 과정에서 발생하는 문제점들에 대해서는 향후 연구 과제로 남겨놓는다.

참고문헌

[1] 손창균, 홍기학, 이기성. (2001). 무응답 보정에서 변수선택법을 이용한 보조정보의 결정에 관한 연구, < 한국 통계학회 추계발표 논문집 >, pp. 63-68.

[2] Bankier, M. D. (1990). Two Step Generalized Least Squares Estimation. Ottawa : Statistics Canada, Social Survey Methods Division, internal report.

[3] Bardsley, P., and Chambers, R. L. (1984). Multipurpose Estimation from Unbalanced Samples, *Applied Statistics*, 33, pp. 290-299.

[4] Cochran, W. G. (1977). *Sampling Techniques (3rd ed.)*. New York : John Wiley & Sons.

[5] Deng, L. Y., and Wu, C. F. J. (1987). Estimation of Variance of Regression Estimator. *Journal of the American Statistical Association*, 82, pp. 568-576.

[6] Deville, J. C., and Särndal, C. E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, pp. 376-382.

[7] Jayasuriya, B. R., and Valliant, R. (1996). An Application of Restricted Regression Estimation in a Household Survey. *Survey Methodology*, 22, pp. 127-137.

[8] Kalton, G., and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, pp. 1-16.

[9] Kott, P. S. (1994). A Note on Handling Nonresponse in Sampling Survey. *Journal of the American Statistical Association*, 89, pp. 693-696.

[10] Lundström, S., and Särndal, C. E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15, pp. 305-327.

[11] Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York : Springer-Verleg.

[12] Silva, P. L. D., and Skinner, C. J. (1997). Variable Selection for Regression Estimation in Finite Population. *Survey Methodology*, 23, pp. 23-32.

[13] Son, C. K., Hong, K. H., and Lee G. S.(2001). The Calibrated Variance Estimator under the Unit Nonresponse. *Korean Journal of Computational & Applied Mathematics*, 8, pp. 975-987.

[14] Theberge, A. (1999). Extensions of Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 94, pp. 635-644.

[2001년 12월 접수, 2002년 5월 채택]

A Study on Auxiliary Variable Selection in Unit Nonresponse Calibration

Chang-Kyoon Son[1] Ki-Hak Hong[2] Gi-Sung Lee[3]

ABSTRACT

Typically, it should be use auxiliary variable for calibrating the survey nonreponse in census or sampling survey. Where, if the dimension of auxiliary information is large, then it may be spend a lot of computing time, and difficult to handle data set. Also because the variance estimator depends on the dimension of auxiliary variables, the variance estimator becomes underestimator. To deal with this problem, we propose the variable selection methods for calibration estimation procedure in unit nonreponse situation and we compare the efficiency by simulation study.

Keywords: Unit Nonresponse; Auxiliary Information; Calibration Estimator; Regression Estimator; Variable Selection

1) Full-time Lecture, Dept.of Computer Science, Dongshin University.
E-mail :ckson85@blue.dongshinu.ac.kr
2) Associate Professor, Department of Computer Science, Dongshin University.
E-mail :khhong@blue.dongshinu.ac.kr
3) Associate Professor, Department of Computer Science & Statistics, Woosuk University.
E-mail :gisung@core.woosuk.ac.kr