Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

무응답 보정에서 변수 선택을 이용한 보조정보의 결정에 관한 연구

손 창 균[1] · 홍 기 학[2] · 이 기 성[3]

요 약

조사과정에서 필연적으로 발생하는 무응답을 보정하기 위해 보조정보를 사용한다. 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(Lundström과 Sändral, 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 소 지역(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 많은 양의 보조변수를 사용하는 경우 분산의 과소 추정문제가 발생할 수 있다. 이와 같은 문제 이외에도 보조변수를 이용함에 있어서 보조변수들간의 상호작용에 의해 추정과정에서 불필요한 보조변수를 사용하게 되는 경우가 발생할 수도 있다.

최근에 Silva와 Skinner(1997)는 유한모집단에 대한 회귀추정량을 도출함에 있어서 변수선택방법을 적용하였다. 이들은 보조변수들간의 다중공선성(multicollinearity) 문제를 고려하여 능형회귀모형을 사용하였으며, 변수선택방법으로 조건수 축소(condition number reduction) 과정을 적용하였다. 그러나 이들은 단순히 회귀추정량의 도출에 차원축소를 위한 변수선택방법을 적용

1) (520-714) 전남 나주시 대호동 동신대학교 컴퓨터 응용학부 전임강사
2) (520-714) 전남 나주시 대호동 동신대학교 컴퓨터 응용학부 부교수
3) (565-701) 전북 완주군 삼례읍 후정리 우석대학교 전산통계학과 부교수

하였으며, 조사 무응답과 같은 현실적인 문제는 고려하지 않았다.

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.

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

2. 보정추정량

2.1 단위 무응답하에서 보정추정량의 도출

우선 이론전개를 위해 몇 가지 기호를 정의하자. $U=\{1,2,\cdots,N\}$을 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 $r(\subseteq s)$을 고려하자.

$\boldsymbol{x}_k=(x_{k1},x_{k2},\cdots,x_{kq})'$을 응답 단위 k와 연관된 $q\times 1$인 보조변수 벡터라 하자. 보조변수의 모집단 총합인 $q\times 1$ 벡터 $\boldsymbol{X}_{tot}=\sum_{k\in U}\boldsymbol{x}_k$와 $\boldsymbol{x}_k$의 표본 총합이 기지라고 가정하자. 이 때 표본에 대한 $q\times 1$인 총합 벡터 $\boldsymbol{x}_{tot}=\sum_{k\in s}d_k\boldsymbol{x}_k$라 하자. y_k를 k번째 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 s로부터 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)추정량이다.

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

여기서 $d_k=1/\pi_k$는 단위 k의 추출가중치이다.

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

$$\hat{Y}_r=\sum_r d_k g_k y_k \tag{2.2}$$

여기서 $g_k=1+(\boldsymbol{X}_{tot}-\sum_r d_k\boldsymbol{x}_r)'(\sum_r d_k\boldsymbol{x}_k\boldsymbol{x}_k')^{-1}\boldsymbol{x}_k$이다.

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

$$\sum_{U} \boldsymbol{x}_k = \sum_{r} w_k \boldsymbol{x}_k \tag{2.3a}$$

$$\sum_{s} d_k \boldsymbol{x}_k = \sum_{r} w_k \boldsymbol{x}_k \tag{2.3b}$$

위의 보정방정식 (2.3a)와 (2.3b)를 각각 만족하는 새로운 가중치 w_k는 다음과 같은 최소 제곱 거리함수를 최소로 한다.

$$G(w_k, d_k) = \sum_{r} (d_k - w_k)^2 / d_k \tag{2.4}$$

이러한 조건으로부터 원래의 추출가중치와 가장 근접한 새로운 가중치는 모집단 보조정보를 이용하는 경우 다음과 같다.

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

표본의 보조정보를 이용하는 경우는 가중치 식(2.4)에서 $\boldsymbol{X}_{tot}$ 대신 $\sum_{s} d_k \boldsymbol{x}_k$를 대입하면 된다. 식(2.2)를 다음과 같은 회귀추정량의 형태로 다시 표현할 수 있다.

$$\begin{aligned} \hat{Y}_{reg} &= \hat{Y}_r + (\boldsymbol{X}_{tot} - \boldsymbol{x}_r)' (\boldsymbol{X}_r' \boldsymbol{D} \boldsymbol{X}_r)^{-1} \boldsymbol{X}_r' \boldsymbol{D} \boldsymbol{y}_r \\ &= \hat{Y}_r + (\boldsymbol{X}_{tot} - \boldsymbol{x}_r)' \boldsymbol{b} \end{aligned} \tag{2.6}$$

여기서 $\boldsymbol{x}_r = \sum_{r} d_k \boldsymbol{x}_k$은 응답단위에 대한 보조변수벡터로서 $q \times 1$이며, $\boldsymbol{b} = (\boldsymbol{X}_r' \boldsymbol{D} \boldsymbol{X}_r)^{-1} \boldsymbol{X}_r' \boldsymbol{D} \boldsymbol{y}_r$은 $q \times 1$ 벡터이다. 이 때 $\boldsymbol{X}_r$은 응답단위에 대해 $\boldsymbol{x}_k$로 구성된 $m \times q$인 보조변수 행렬이며, $\boldsymbol{D}$는 가중치 d_k를 대각원소로 갖는 $m \times m$인 대각행렬이다. 또한 $\boldsymbol{y}_r$은 응답단위들의 $m \times 1$인 벡터이다.

2.2 보조변수 차원에 대한 보정추정량의 분산의 종속성

식(2.6)으로부터 모집단 총합의 일반화 회귀추정량은 다음과 같은 기본 선형모형(underlying linear model)으로부터 유추할 수 있다.

$$y_k = \boldsymbol{x}_k' \beta + \varepsilon_k \tag{2.7}$$

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

이러한 선형회귀 모형으로부터 $\boldsymbol{x}_k$의 조건하에서의 추정오차 $\hat{Y}_{reg} - Y$의 분산은 다음과 같다.

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

여기서 $\hat{S}_x = \sum_{k \in r} d_k \boldsymbol{x}_k \boldsymbol{x}_k' = \boldsymbol{X}_r' \boldsymbol{D} \boldsymbol{X}_r$이다.

$c_g^2 = (\boldsymbol{X}_{tot} - \boldsymbol{x}_r)' \hat{S}_x^{-1} (\boldsymbol{X}_{tot} - \boldsymbol{x}_r)$라 하면 분산 식(2.8)은 c_g^2에 의존함을 알 수 있다. 이러한 종속성을 보다 명확히 하기 위해 $\boldsymbol{x}_k$가 독립이고, 동일한 정규분포를 가정하면, $(\boldsymbol{X}_{tot} - \boldsymbol{x}_r)$과 $\hat{S}_x^{-1}$의 독립성과 $E(\hat{Y}_{reg} - Y \mid \boldsymbol{x}_k) = 0$을 이용하여 비 조건부 분산은 다음과 같이 유도된다.

$$V(\hat{Y}_{reg}-Y) = N^2\frac{\sigma^2}{m}\left[(1-f)+tr\{E(\boldsymbol{X}_{tot}-\boldsymbol{x}_r)'(\boldsymbol{X}_{tot}-\boldsymbol{x}_r)\}E(\hat{S}_x^{-1})\right]$$
$$= N^2\frac{\sigma^2}{m}(1-f)[1+q/(m-q-2)] \quad (2.9)$$

결과적으로 식(2.9)는 보조변수의 차원인 q에 의존함이 명확하다.

보조정보의 수준에 따른 보정 추정과정으로부터 일반화 회귀추정량의 근사적인 MSE의 추정량은 다음과 같다.

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

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

3. 변수선택 후 보정 추정

3.1 변수선택

본 논문에서 보조정보의 선택방법은 전진선택법(forward selection)과 조건수 축소(condition number reduction) 방법을 적용하고, 이 때 평균제곱오차(MSE)의 추정치를 최소로 하는 보조변수를 선택하는 방법을 전개하고자 한다. 또한 보조변수들간에 존재하는 다중공선성의 문제를 해결하기 위해 Bardsley와 Chambers(1984)에 의해 제안된 방법을 함께 고려한다.

먼저 전진선택법의 과정을 살펴보면, 우선 가장 상관계수가 큰 보조변수를 선택하여, 보정추정량을 구한 후, 보조변수를 하나씩 추가하여, MSE의 추정량이 증가하는 시점의 보조변수들을 선택한다. 결과적으로 선택된 보조변수들을 이용한 보정추정량과 그에 따르는 최소의 MSE 추정량을 얻게 된다.

다음으로 조건수 축소과정을 살펴보면, 식(2.6)으로부터 회귀추정량이 $CP_w=(\boldsymbol{X}_r'\boldsymbol{D}\boldsymbol{X}_r)$에 의존함을 알 수 있다.

Bankier(1992)에 의해 제시된 조건수 축소과정은 $CP_w=(\boldsymbol{X}_r'\boldsymbol{D}\boldsymbol{X}_r)$에 대해 가장 큰 조건수를 나타내는 보조변수들을 제거해 가는 방법으로 이에 대한 과정을 살펴보면 다음과 같다.

단계 1] 모든 이용 가능한 보조변수에 대한 $CP_w=(\boldsymbol{X}_r'\boldsymbol{D}\boldsymbol{X}_r)$를 계산한다.

단계 2] $CP_w=(\boldsymbol{X}_r'\boldsymbol{D}\boldsymbol{X}_r)$의 Hermite 정준행렬 H를 계산한다. 이 행렬에서 선형종속을 나타내는 0인 각각의 열(columns)을 제거한다.

단계 3] 선형종속인 열을 제거한 후 축소된 CP_w로부터 조건수 $c=\lambda_{\max}/\lambda_{\min}$를 계산하고, 만일 특정한 값 L과 비교하여 $c<L$이면 과정을 종료하고, 남아있는 모든 보조변수를 이용한다.

끝으로 보조변수들간의 상호 연관성에 의해 발생하는 다중공선성의 문제를 해결하기 위해 보정추정량의 식을 다음과 같은 능형회귀추정량으로 대치하고, 그에 따르는 MSE를 추정하여 추정량의 안정성을 살펴본다.

$$\hat{Y}_{reg}=\hat{Y}_r+(\boldsymbol{X}_{tot}-\boldsymbol{x}_r)(\lambda C^{-1}+\boldsymbol{X}_r'\boldsymbol{D}\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r'\boldsymbol{D}\boldsymbol{y}_r \quad (3.1)$$

3.2 변수선택 후 보정 추정과정

3.1절에서 선택된 보조변수를 이용하여 새로 구한 보정가중치는 다음과 같다.

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

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

따라서 최종적으로 무응답을 보정한 모집단 총합추정량은 다음과 같다.

$$\begin{aligned}\hat{Y}_{reg}^* &= \hat{Y}_r+(\ \boldsymbol{X}_{tot}^* - \boldsymbol{x}_r^*)'(\ \boldsymbol{X}_r^{*\prime}\boldsymbol{D}\ \boldsymbol{X}_r^*)^{-1}\ \boldsymbol{X}_r^{*\prime}\boldsymbol{D}\ \boldsymbol{y}_r \\ &= \hat{Y}_r+(\ \boldsymbol{X}_{tot}^* - \boldsymbol{x}_r^*)'\ \boldsymbol{b}^*\end{aligned} \tag{3.3}$$

여기서 $\boldsymbol{x}_r^* = \sum_r d_k\ \boldsymbol{x}_k^*$은 변수선택 후 최종적으로 보조변수로 결정된 응답단위에 대한 보조변수벡터로서 $h\times 1$이며, $\boldsymbol{b}^* = (\ \boldsymbol{X}_r^{*\prime}\boldsymbol{D}\ \boldsymbol{X}_r^*)^{-1}\ \boldsymbol{X}_r^{*\prime}\boldsymbol{D}\ \boldsymbol{y}_r$은 변수선택 후 추정된 $h\times 1$인 회귀계수 벡터이다. 이 때 $\boldsymbol{X}_r^*$은 변수 선택 후 응답단위에 대해 $\boldsymbol{x}_k^*$로 구성된 $m\times h$인 보조변수 행렬이며, $\boldsymbol{D}$는 가중치 d_k를 대각원소로 갖는 $m\times m$인 대각행렬이다. 또한 $\boldsymbol{y}_r$은 응답단위들의 $m\times 1$인 벡터이다.

보정추정량 (3.3)에 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 h는 변수선택 후 최종적으로 보조변수로 결정된 변수의 개수를 나타내며, $g_k^* = 1+(\ \boldsymbol{X}_{tot}^* - \sum_r d_k\ \boldsymbol{x}_k^*)'(\sum_r d_k\ \boldsymbol{x}_k^*\ \boldsymbol{x}_k^{*\prime})^{-1}\ \boldsymbol{x}_k^*$이며, $k\in r$에 대해 $\hat{e}_k^* = \hat{y}_k - \boldsymbol{x}_k^{*\prime}\ \boldsymbol{b}^*$이다.

4. 모의실험

4.1 통계량의 정의

모의실험을 위해 다음과 같이 정의된 Särndal 등(1992)의 "MU284 스웨덴 데이터"를 사용하였다. 목적은 1985년의 스웨덴의 총 납세소득을 추정하고자 하며, 이 때 284개 시도의 납세소득의 총합은 미지라고 가정하자. 관심변수 y와 그에 따르는 7개의 보조변수를 위와 같이 정의한 후 표본 추출설계는 단순임의 추출을 가정하여 약 20% 표본인 60개 시도를 표본으로 추출한다. 다음으로 무응답 가정을 위해 추출된 표본으로부터 90%, 80%, 70%로 응답률을 가정하여 최종적으로 응답단위를 결정한다. 이 때 응답단위의 선택은 전개를 간단히 하기위해 균등분포로 가정한다. 효율성 비교를 위해 먼저 다음과 같은 모집단 총합추정량과 편향추정량을 정의하였다.

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

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

이 때 K는 반복 수이다.

<표 4.1> MU284 데이터 세트의 변수에 대한 정의

변 수	내 용
y	1985년 284개시도의 납세소득(단위 : 백만 Kroner)
x_1	1985년 284개시도의 인구수(단위 : 천명)
x_2	1975년 284개시도의 인구수(단위 : 천명)
x_3	1882년 시위회의 보수당 의석수(단위 : 명)
x_4	1982년 시의회의 사회민주당 의석수 (단위 : 명)
x_5	1982년 시의회의 총 의석 수(단위 : 명)
x_6	1984년 도시근로자의 취업 인구수 (단위 : 명)
x_7	1984년 부동산 가치 (단위 : 백만 Kroner)

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

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

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

마지막으로 추정량의 정도를 살펴보기 위해 근사 정규분포 이론에 기초한 모집단 총합의 95% nominal coverage rate을 계산하였다.

참고문헌

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

[2] Chambers, R. L. (1996). Robust Case-Weighting for Multipurpose Establishment Surveys. *Journal of Official Statistics*, 12, pp. 3-32.

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

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

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

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

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

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

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