무응답 대체 후 발생하는 문제점과 해결 방안

김 규 성*

요 약

대부분의 통계조사에서 흔히 발생하는 무응답을 처리하기 위한 방법으로 최근에는 표본 대체 방법이 널리 이용되고 있다. 본 논문에서는 여러 가지 표본 대체 방법을 소개하고 각 방법의 장·단점을 비교·설명한다. 그리고 대체된 데이터를 응답 데이터인 것처럼 활용했을 때 발생하는 문제점들을 지적하고 모의실험을 통하여 그 정도를 살펴본다. 이와 더불어 제기된 문제점을 해결하는 몇 가지 해결 방안을 소개한다.

Ⅰ. 서 론

대부분의 통계 조사는 무응답을 포함한다. 무응답은 여러 가지 이유로 발생하며, 현실적으로 무응답이 발생하는 것을 피하기는 어려운 일이다. 이러한 무응답은 조사 결과에 상당한 영향을 주기 때문에 조사의 신뢰도를 높이기 위해서는 무응답률을 낮추는 일은 매우 중요하다. 예를 들어 국회의원 선거 여론조사에서는 상당 비율의 무응답이 발생하며, 무응답 처리 결과에 따라 조사의 결과가 다르게 나타남은 어렵지 않게 알 수 있다. 무응답을 줄이기 위해서는 정비된 조사 체계, 숙련된 조사원, 그리고 검증을 거친 설문지 등 여러 요소를 유기적으로 갖출 필요가 있으며 가능하면 무응답률이 낮은 조사를 하는 것이 바람직하지만, 시간적·경제적 여건 때문에 불가피하게 발생하는 무응답에 대해서는 사후적으로 처리하는 방법을 고려해야 한다.

가장 손쉽게 생각해 볼 수 있는 방법은 무응답은 제외하고 응답만을 조사 결과로 이용하는 것이다. 과거에 주로 이용됐던 이 방법은 조사 변수들 간의 관계를 고려하지 않고 무응답을 제외시킴으로써 자료를 효과적으로 이용하지 못하는 단점이 있어 바람직

* (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 전산통계학과, 조교수.
e-mail : kskim@uoscc.uos.ac.kr

하지 않다. 대신에 무응답 대체를 통하여 완전자료를 만든 후 대체 후 자료를 조사 분석에 이용하는 방법이 효과적인 방법이다. 적절한 방법으로 무응답을 대체하면 응답자료의 손실을 줄일 수 있을 뿐 아니라 기존의 분석방법을 그대로 이용할 수 있는 장점이 있어서 무응답 자료를 제거하는 것보다 훨씬 바람직한 방법으로 평가되고 있다.

무응답 대체 방법은 여러 가지가 있으며, 개별 문제에 따라 적절한 대체 방법을 채택하여 이용하는 것이 통상적인 방법이다. 본 연구에서는 무응답 대체 방법을 개괄적으로 고찰한 후, 대체 후 자료를 이용하여 추론을 하는 경우 제기되는 문제점에 초점을 맞추었다. 무응답 대체를 하여 얻은 완전자료는 여러 가지 바람직한 성질을 갖고 있는 반면, 사후적으로 대체된 값이기 때문에 대체값을 응답값인 것처럼 취급하는 경우에 대체에 따른 새로운 문제점이 제기된다. 크게 두 가지 문제가 중요하게 여겨지는데 하나는 대체 후 추정량이 구조적 편향을 가질 수 있다는 점이고 다른 하나는 대체 후 추정량의 분산이 과소추정될 수 있다는 점이다. 이 두 가지 문제는 무응답 대체로 인한 심각한 오류로 평가되는데, 전자는 추정량이 대체로 인한 구조적 편향성을 갖을 수 있어 추정치의 정확성이 감소됨을 의미하며, 후자는 추정량의 분산이 실제보다 작게 추정됨으로써 조사 결과의 신뢰도가 실제보다 높은 것으로 평가됨을 의미한다. 본 논문에서는 후자 쪽에 초점을 맞추어 대체후 추정량의 분산의 과소추정 문제를 고찰하기로 한다.

Ⅱ. 무응답 대체 방법

무응답 대체 방법은 크게 결정적인 대체 방법과 확률적인 대체 방법으로 구분할 수 있다(Kalton & Kasprzyk, 1986, 김영원 & 조선경, 1996). 결정적인 대체 방법은 무응답 항목에 대한 대체값이 유일하게 결정되는 대체방법이며, 확률적인 대체 방법은 대체값이 확률적으로 정해지는 방법이다. 결정적인 대체 방법으로는 연역적 대체, 시기적 대체, 평균대체, 축차 핫덱 대체, 비 대체, 회귀 대체, 최근방 대체 등이 있으며, 확률적 대체 방법에는 핫덱 대체, 가중 핫덱 대체, 랜덤 비 대체, 랜덤 회귀 대체 등이 있다.

대체 방법을 전체 데이터에 직접 적용하기보다는 대체 군을 만들어 대체 군내에서 대체 방법을 적용하는 것이 합리적이다. 대체 군은 제어 변수를 이용하여 대체 군내의 변수 값들이 가능한 서로 비슷해지도록 만드는 것이 바람직하다. 대체군을 이용하면 무응답 편향의 효과를 감소시키고 대체의 정확도를 높이는 효과가 있다.

연역적 대체 방법은 논리적인 제약조건이나 다른 기록에 의하여 확실하게 대체값을 지정하여 무응답을 대체하는 방법을 말한다. 시기적 방법은 반복조사에서 유용한데, 만일 반복조사에서 동일 항목의 응답값이 조사 시점에 따라 안정된 값을 보이고 전회

조사값과 금회 조사값의 상관관계가 높으면 금회 조사의 무응답을 전회 조사값으로 대체하는 방법이다. 평균대체는 무응답 항목에 대체군 내의 응답값 평균을 구하여 무응답 항목에 대체하는 방법이다. 평균 대체 방법은 간단하여 이용되기 쉬운 장점이 있으며, 항목 변수가 양적 변수이고 구하고자 하는 통계량이 평균일 때 유용하다. 그러나 대체 후 값들은 평균값의 빈도수가 지나치게 많아져 응답값들의 분포가 왜곡되고, 평균이 아닌 통계량, 예를 들면 백분위수 같은 통계량을 구할 때는 효율이 저하되는 단점이 있다. 축차 핫덱 대체 방법은 데이터 파일을 만들 때, 직전에 응답한 단위의 항목값으로 대체하는 방법으로, 사회·인구통계조사에 유용하다. 즉 표본은 사회·인구통계적 지표에 의해 자연스럽게 대체 군으로 구분되며, 대체 군내의 항목값은 서로 유사할 가능성이 높다. 또한 조사는 지리적인 연속성을 가지고 이루어지므로 무응답이 발생하면 직전의 응답값으로 대체하는 것이 타당성이 있다.

비 대체와 회귀 대체는 보조변수를 이용하여 비 예측치나 회귀 예측치로 무응답을 대체하는 방법으로, 항목 변수와 상관관계가 높으며, 전체 조사 단위에서 이용가능한 보조변수가 있을 때 유용한 방법이다. 이 방법이 효과적이기 위해서는 항목 변수가 양적인 변수이며 항목 변수와 보조변수의 관계를 비 모형이나 회귀모형이 잘 설명할 수 있어야 한다. 최근방 대체는 보조변수를 이용하여 무응답 조사 단위와 가장 유사한 응답 조사 단위를 찾아, 대응되는 항목값을 대체하는 방법으로 축차 핫덱 방법과 유사성이 있다. 이때 가장 유사한 단위는 거리 함수를 이용하여 찾는다.

결정론적 대체 방법은 대체 값을 유일하게 지정하여 대체를 하므로 항목 변수의 변동을 줄이는 경향이 있다. 이러한 단점을 보완하기 위하여 확률적 대체가 제안되었다. 핫덱 대체는 대체 군내의 응답값 중에서 하나를 랜덤하게 선정하여 무응답 항목에 대체하는 방법으로, 핫덱 대체를 이용하면 무응답 대체 후에도 표본의 분포가 그대로 유지되는 장점이 있다. 평균 대체나 비 대체 등과는 달리 표본 분포가 유지되므로 통계량의 형태에 무관하게 이용될 수 있다. 가중 핫덱 대체는 응답값 중 하나를 선정할 때 가중값을 두어 선정하는 방법이다. 이 방법은 층화가 되어 있거나 집락화가 되어 있어 대체 군내의 응답값들이 서로 다른 추출확률을 갖을 때 유용하게 이용될 수 있다. 랜덤 비대체 방법이나 랜덤 회귀대체 방법은 은 비대체 값이나 회귀대체 값에 확률오차를 포함시켜 대체하는 방법으로 포함되는 확률오차는 대체 전 자료의 변동과 대체 후 자료의 변동이 같아지도록 변동의 폭을 계산하여 포함시킨다. 확률대체 방법을 이용하면 결정적 방법에 비하여 대체하는 방법은 다소 복잡하나 자료의 변동을 유지할 수 있다는 장점을 가지고 있다.

Ⅲ. 무응답 대체 효과

크기 N인 모집단에서 단순확률추출한 표본을 A_s로 나타내고 이 중 응답 표본을 A_r로 나타내며 크기는 각각 n과 r이라 하자. 무응답에 대체 방법을 사용하여 대체한 자료를 $\hat{y}_k$로 표현하고, 대체 후 자료를 y_k^*로 나타내자.

$$y_k^* = \begin{cases} y_k, & k \in A_r, \quad \text{응답값} \\ \hat{y}_k, & k \in A_{s-r}, \quad \text{대체값} \end{cases} \tag{3.1}$$

그러면 대체 후 자료로 만든 표본평균은 다음과 같다.

$$\bar{y}_I = \frac{1}{n}\left(\sum_{k \in A_r} y_k + \sum_{k \in A_{s-r}} y_k^* \right) \tag{3.2}$$

대체 후 표본평균 $\bar{y}_I$는 대체 값이 포함된 추정량이므로 표본평균이 갖는 좋은 통계적 성질을 그대로 유지하지는 못한다. 그러나 적절한 대체 방법을 선택하여 이용하면, 대체 후 표본평균 $\bar{y}_I$는 모평균의 비편향추정량이 되는 바람직한 성질을 가지고 있으므로 모평균 추정에 쉽게 이용될 수 있으며, 또한 널리 이용되고 있다. 그러나 대체값으로 인해 대체 후 추정량 $\bar{y}_I$는 표본평균보다 더 큰 분산을 갖는데 주목할 필요가 있다. 더 큰 분산은 추정량의 효율을 나타내는데 중요한 역할을 하며, 이를 간과할 경우 조사의 결과가 과대 평가될 수 있는 위험이 있다. 반면, 완전 자료를 응답자료인 것처럼 간주하여 만든 통상적인 분산 추정량은 도리어 작은 값을 갖는 경향이 있다. 이러한 사실을 확인해 보기 위하여 모의실험을 실시하였다.

자료는 관심변수 y와 보조변수 x로 구성되어 있으며, 관심변수와 보조변수는 다음의 관계가 있다고 하자.

$$y_i = \beta\sqrt{x_i} + \varepsilon_i, \quad i = 1, \ldots, n \tag{3.3}$$

보조변수 x는 감마분포에서 생성하며, 오차 ε는 보조변수 x가 주어졌을 때 정규분포에서 생성한다. 표본의 크기는 $n = 100$으로 하고 무응답은 관심변수와 무관하게 일어나는 것을 가정하며, 무응답률은 10%, 20%, 30%, 40%, 50%인 경우를 고려한다. 대

체방법은 평균대체, 비대체, 핫덱 대체 등 세가지를 고려한다. 각각의 대체 방법에 따라 모의실험을 10,000번 반복하여 대체에 따른 두 가지 대체효과를 알아본다. 대체로 인한 분산의 증가분은 RV로 수량화하고,

<표 1> 표본대체 방법에 따른 대체 효과

무응답률		무응답 대체방법		
		평균대체	비대체	핫덱 대체
10%	대체 후 분산	0.065	0.065	0.071
	RV	12.0 %	12.0 %	22.4 %
	대체후 분산추정값	0.052	0.059	0.058
	RB	-19.2 %	-8.6 %	-17.9 %
20%	대체 후 분산	0.073	0.072	0.084
	RV	25.8 %	24.1 %	44.8 %
	대체후 분산추정값	0.046	0.061	0.058
	RB	-36.2 %	-15.5 %	-31.0 %
30%	대체 후 분산	0.084	0.082	0.102
	RV	44.8 %	41.3 %	75.8 %
	대체후 분산추정값	0.040	0.062	0.058
	RB	-51.2 %	-24.5 %	-43.0 %
40%	대체 후 분산	0.097	0.096	0.120
	RV	67.2 %	65.5 %	106.8 %
	대체후 분산추정값	0.035	0.063	0.058
	RB	-64.0 %	-34.0 %	-51.6 %
50%	대체 후 분산	0.115	0.115	0.145
	RV	98.2 %	98.2 %	150.0 %
	대체후 분산추정값	0.029	0.065	0.057
	RB	-74.9 %	-43.6 %	-60.1 %

$$RV(\%) = \left(\frac{V(\bar{y}_I)}{V(\bar{y}_s)} - 1 \right) \times 100 \tag{3.4}$$

대체 후 분산의 과소 추정은 RB로 수량화한다.

$$RB(\%) = \sum_{k=1}^{K} \frac{(v_k^*(\bar{y}_I) - V(\bar{y}_I))/K}{V(\bar{y}_I)} \times 100 \tag{3.5}$$

모의실험의 결과가 <표 1>에 실려있다.

대체적으로 무응답률이 증가하면 대체 후 분산은 증가하고, 반면에 대체 후 분산 추정값은 감소하는 경향을 볼 수 있다. 평균대체, 비 대체, 핫덱 대체 모두 무응답률이 증가하면 이같은 현상은 동일하게 나타나는데, 세 가지 방법을 비교하면 평균대체가 가장 심하게 분산을 과소 추정하며 그 다음이 핫덱 대체, 그리고 비 대체 순이다.

이 결과는 자료가 생성된 모형에서 설명되는 결과이므로 모형이 바뀌면 결과는 다소 달라질 수 있다. 그러나 위에서 설명한 두 가지 대체로 인한 효과, 즉 대체 후 분산은 증가하고 통상적인 분산추정량은 과소추정하는 효과는 정도의 차이는 있지만 여전히 나타날 것으로 짐작할 수 있다.

Ⅳ. 무응답 대체 효과의 보정

무응답 대체로 인하여 생기는 대체 후 추정량의 분산의 과소 추정문제를 해결하기 위해서는 대체분산까지 추정할 수 있는 분산추정 방법을 고려할 필요가 있다. 즉 대체 효과를 분산 추정 방법에 반영할 수 있어야 한다. 이에 대한 연구 결과는 다각적으로 발표되고 있는데, 대표적인 방법으로는 대체모형을 이용한 분산 추정 방법, 수정된 잭나이프 방법을 이용한 방법, 다중대체를 이용한 분산추정 방법, 그리고 붓스트랩을 이용한 방법 그리고 균형이분표본을 이용한 방법 등이 있다.

Sarndal(1992)는 대체모형을 이용한 분산추정 방법을 제안하였다. 이 방법은 관심변수와 연관이 높은 보조변수가 있을 경우, 보조변수를 이용하여 다음과 같은 비 대체 모형을 설정한 후, 표본추출분산과 대체분산을 각각 추정하여 합한 추정량으로 전체 분산을 추정하는 방법이다. 대체 모형을 이용하면 대체 분산을 표본추출분산과 분리해서 구할 수 있고, 대체분산으로 인한 대체 효과를 명료하게 설명할 수 있으며, 결과적으로 분산의 과소 추정을 방지할 수 있다.

잭나이프 방법은 대표적인 비모수적 분산 추정 방법으로, 전통적인 잭나이프 방법은 독립이며 동일한 분포를 갖는 표본을 대상으로 하여 고안되었다. 그러나 표본조사에서 얻어지는 표본은 대부분 복합표본이어서 추출확률이 다르거나 독립이 아닌 경우가 많다. 따라서 잭나이프 방법을 복합표본에 적용하기 위해서는 약간의 수정이 필요하다(Yung & Rao, 1996). Rao & Shao(1992)는 무응답이 대체된 완전자료에 잭나이프 방법을 이용하여 분산을 추정하는 방법을 제안하였는데, 그들이 제안한 방법은 대체효과를 잭나이프 표본평균을 반영하는 것이다. 수정된 잭나이프 방법은 표본평균 외에 표본평균으로 만들어지는 대부분의 추정량에 적용할 수 있어 적용범위가 넓은 장점이 있다(Rao & Sitter, 1995; Sitter, 1997; Sitter & Rao, 1997).

V. 결 론

대부분의 통계조사에서 무응답은 현실적으로 존재한다. 무응답을 최대한 줄이는 것이 바람직하지만 시간적·경제적 제약으로 불가피하게 발생하는 무응답은 사후적으로 적절하게 처리하여 무응답의 효과를 최대한 보정하는 것이 조사의 신뢰도를 높이는 방법이 될 것이다. 무응답 대체 방법은 개별적인 문제에 따라 취사 선택되어야 하며, 각 방법에 따라 대체 효과가 다르므로 보정 방법 또한 달라져야 한다. 무응답을 원천적으로 없앨 수는 없지만 사후적으로 현명하게 대처하면 무응답의 효과를 상당부분 상쇄시킬 수는 있을 것이다. 본 논문에서 소개한 무응답 대체 방법과 수정된 분산추정 방법들을 실제 문제에 현명하게 적용하면 무응답으로 인한 조사결과의 오류를 줄일 수 있을 것으로 기대한다.

<참 고 문 헌>

[1] 김영원, 조선경 (1996). 표본조사에서 항목 무응답 대체방법. 한국통계학회논문집, 3, 145-159.

[2] Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. Survey Methodology, 12, 1-16.

[3] Kovar, J.G. and Chen, E.J. (1994). Jackknife variance estimation of imputed survey data. Survey Methodology, 20, 45-52.

[4] Kovar, J.G. and Whitridge, P.J. (1995). Imputation of business survey data. Business survey methods, 403-423.

[5] Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. Biometrika, 79, 811-822.

[6] Rao, J.N.K. and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. Biometrika, 82, 453-460.

[7] Rubin, D.B. (1987). Multiple imputation for nonresponse in survey. Wiley, New York.

[8] Sarndal, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. Survey Methodology, 18, 241-252.

[9] Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. Journal of American Statistical Association, 91, 1278-1288.

[10] Shao, J. Chen, Y., and Chen, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. Journal of American Statistical Association, 93, 819-831.

[11] Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. Journal of the American Statistical Association, 92, 780-787.

[12] Sitter, R.R. and Rao, J.N.K. (1997). Imputation for missing values. The Canadian Journal of Statistics, 25, 61-73.

[13] Yung, W. and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. Survey Methodology, 22, 23-44.