패널무응답의 가중수정 방법

신민웅*, 윤연옥**

<요 약>

패널 무응답자(panel nonrespondent)란 처음 조사에서는 응답을 하였으나 나중 조사에서는 응답을 하지 않은 사람을 의미한다.패널조사에서는 앞 단계에서의 응답으로부터 뒷 단계의 무응답에 대한 정보를 얻을 수 있다.무응답에 대한 수정 방법은 어떤 보조 변수들을 선택하고, 그 변수들이 수정하는 데 어떻게 사용하는 가를 결정하는 것이다.우리는 가중 수정을 패널 무응답자에 대해서만 생각한다.이러한 가중은 패널 무응답자에 대하여 보상하기 위하여 패널 응답자의 가중값을 수정한다. 종속 변수로서 패널응답 상태(status)는 로지스틱 회귀분석으로 패널 무응답에 대한 모형을 선택하는 방법이다. 로지스틱 회귀분석에서 패널무응답과 상관이 있는 변수들은 패널무응답 편향을 감소시키기 위하여 가중 수정에서 사용하기 위한 변수들이다.

I. 서 론

어떤 조사에서는 많은 보조변수들이 무응답 수정에 이용 가능하다. 패널조사에서는 앞 단계에서의 응답으로부터 뒷 단계의 무응답에 대한 정보를 얻을 수 있다. 패널 무응답자(panel nonrespondent)란 처음 조사에서는 응답을 하였으나 나중 조사에서는 응답을 하지 않은 사람을 의미한다.패널응답자란 모든 조사에서 응답한 사람을 의미한다. 단위 무응답을 보상하기 위하여 무응답 수정 가중을 사용한다.이러한 가중값을 얻기 위해서는 응답자와 무응답자에 대한 정보가 필요하다.또한 패널조사에서는 앞 단계에서의 응답으로부터 뒷 단계의 무응답에 대한 정보를 얻을 수 있다.무응답에 대한 수정 방법은 어떤 보조 변수들을 선택하고, 그 변수들이 수정하는 데 어떻게 사용하는 가를 결정하는 것이다.우리는 가중 수정을 패널 무응답자에 대해서만 생각한다. 이러한 가중은 패널 무응답자에 대하여 보상하기 위하여 패널 응답자의 가중값을 수

* (449-791) 경기도 용인시 모현면, 한국외국어대학교 통계학과, 교수
E-mail: mwshin@stat.hufs.ac.kr

** (302-701) 대전시 서구 둔산동 920번지 정부대전청사 3동 통계청, 사무관
E-mail: yyoon@nso.go.kr

정한다.우리는 때때로 무응답을 보정하기 위하여 수정-칸(adjustment cell)을 사용한다. 수정-칸이란 근사적으로 같은 응답확률이나 같은 값의 단위들의 모임이다. (ex.수입) 그러면 각 수정-칸은 안에서 가중수정이나 단순 핫-덱 대체를 할수가 있다. 각 칸에서 만일 조사 항목과 응답확률사이에 공분산이 근사적으로 영이라면 모평균과 모총계의 수정 추정량의 무응답 편향은 근사적으로 영이 될 것이다.
무응답-수정은 인구학적 또는 지리적 분류 변수들로 수정-칸을 형성하여 이루어졌으나, Little(1986)과 다른 학자들은 응답확률이나 항목(item) 값들에 따라서 칸을 형성하였다.

Eltinge(1997)은 칸 형성의 유용성에 대한 진단(diagnostics)을 논했다. 주요한 관심은 칸의 수에 대한 판정, 수정된 칸과 수정안된 칸의 비교, 추가적으로 더 나눠야 할 칸의 발견, 그리고 응답확률과 항목값들로 칸을 형성했을 때에 두 방법 각각 추정결과 를 비교하는 것이다. Rizzo(1996)등은 패널 무응답의 가중수정 방법을 논하였다.

II. 응답 propensity의 예측

패널 무응답 수정의 첫 단계는 어느 항목들이 수정 과정에서 사용되느냐를 결정하는 것이다. 응답자가 모든 조사에서 응답할 항목을 선택한다.종속 변수로서 패널응답 상태(status)는 로지스틱 회귀분석으로 패널 무응답에 대한 모형을 선택하는 방법이다.독립변수가 될 수 있는 항목을 선택하는 일반적 지침(guideline)은 그 항목에 대한 임의의 2개의 범주들 사이에 응답율이 통계적으로 유의한 것이다.예외적인 변수는 성별로 반드시 독립변수로 포함되어야 한다. 로지스틱 회귀분석에서 패널무응답과 상관이 있는 변수들은 패널무응답 편향을 감소시키기 위하여 가중 수정에서 사용하기 위한 변수들이다.그러나 검열(screening)분석은 제한되어 있다.왜냐하면,변수들 간에 교호작용을 생각하지 않거나,패널 무응답 수정을 만들기에는 실제로 사용하기에 너무 많은 변수들이 남아 있기 때문이다.예컨대,두 항목들이 너무 높게 응답상태와 서로 상관이 되어 있어서 수정하는데 있어서 한 항목만 사용해도 충분하다.

패널무응답의 독립변수들을 선택하는 다음 단계는 패널응답 상태를 예측하는 데 있어서,어떻게 항목들을 조합(combination)하여야 하느냐를 조사하는것이다.로지스틱 회귀분석이 패널응답상태에 여러 항목들의 관계를 조사하는 데 사용된다.
크기 N의 모집단 U에서 조사항목 $Y_i,\ i \in U$ 에 대해서 , 모평균은

$$\overline{Y} = N^{-1} \sum_{i \in U} Y_i \tag{2.1}$$

이다.. 표본 s는 크기 n이고, π_i는 단위 i가 표본에 포함될 확률이다.

무응답은 유사-확률모형을 만족한다고 가정한다.(Oh와 Scheuren, 1983).

그리고 R_i는 지시변수로

$$R_i = \begin{pmatrix} 1 & \text{단위 } i \text{ 가 응답} \\ 0 & \text{그 외의 경우} \end{pmatrix} \tag{2.2}$$

라 하자.

R_i는 서로독립인 베르누이(η_i) 확률변수이거, 고정된 응답확률 η_i 는 단위는 다르다고 가정한다, 조사 가중값은 $\lambda_i = 1/\pi_i$이고, 수정안된 조사-가중 평균응답은

$$\hat{\overline{Y}}_1 = (\sum_{i \in s} \lambda_i R_i)^{-1} \sum_{i \in s} \lambda_i R_i Y_i \tag{2.3}$$

이다.

수정안된 추정량 $\hat{\overline{Y}}_1$의 무응답 편향은 근사적으로

$$N^{-1} \eta^{-1} \sum_{i \in U} \eta_i (Y_i - \overline{Y}) \tag{2.4}$$

이다. 단, $\eta = N^{-1} \sum_{i \in U} \eta_i$이다. 무응답 편향을 줄이기 위하여, k " 수정-칸들" 로 분할한다. U_h는 U가 분할된 k번째 칸이고 표본 s도 s_k 로 분할한다, 그러면, 수정된 추정량은

$$\hat{\overline{Y}}_k = \sum_{k=1}^{k} w_h \overline{Y}_{hk} \tag{2.5}$$

이다. 단, $w_h = (\sum_{i \in s} \lambda_i)^{-1} \sum_{i \in s_h} \lambda_i$이고, $\overline{Y}_{kR} = (\sum_{i \in s_h} \lambda_i R_i)^{-1} \sum_{i \in s_h} \lambda_i R_i Y_i$이다.

수정 추정량 $\hat{\overline{Y}}_k$ 은 근사적으로 나머지 무응답 편향

$$N^{-1} \sum_{k=1}^{k} \eta^{-1} \sum_{i \in U_h} (\eta_i - \eta_h) (Y_i - \overline{Y}) \tag{2.6}$$

을 갖는다. 단, N_h 는 U_h의 단위의 수이고, $\overline{\eta}_i = N_h^{-1} \sum_{i \in U_h} \eta_i$,

$\overline{Y}_h = N_h^{-1} \sum_{i \in U_h} Y_i$이다.

결과적으로 우리는 각 칸에서 η_i 와 Y_i 사이에 모공분산이 근사적으로 영이 되도록 수정칸을 만들려고 한다. 실제로 각 칸에서 응답확률 η_i 나 항목 Y_i 가 동질적으로 되도록 한다.

III. 로지스틱 모형에 기초한 수정

예측된 로지스틱 수정(predicted logistic adjustment)이라 불리우는 패널 무응답 가중수정은 모형에서 예측변수들의 교차-분류된 칸에 대하여 주-효과 로지스틱 회귀모형으로부터 예측된 응답률의 역수를 취하므로서 계산된다. 이 모형에서는 변수간에 교호작용은 없다고 가정한다.

표본크기가 25이상인 칸에서는 무응답 수정은 관찰된 칸 응답율의 역수이고, 25인 이하인 칸에서는 무응답 수정은 그 칸에 대하여 예측된 응답율의 역수를 취할 때, 그 수정을 혼합 로지스틱 수정이라고 부른다. 칸들의 표본크기가 30인이 초과될 때까지 붕괴시키고, 붕괴된 칸안에서 관찰된 응답율의 역수를 무응답 수정으로 사용하다.붕괴된 칸들에 대한 전략은 유사한 예측응답율을 갖는 칸들을 같은 그룹으로 묶는 것이다. 이러한 무응답 수정은 붕괴된 로지스틱 수정이라고 부른다.

X_i 가 응답이나 무응답 표본단위 i 에 대한 보조 변수들의 벡터라 하고, 표본 (R_i, X_i) 값들로 $\eta_i = \eta(X_i)$에 대한 모형을 적합시킨다. 표본 칸들 S_k 는 추정된 응답 확률들 $\hat{\eta}_i$ 에 따라서 표본 단위들을 그룹화하므로서 형성된다. 또는 Y_i 의 보조변수 X_i 에 대한 회귀 추정값 $\hat{Y}_i$ 들을 그룹화하여 표본 칸들 s_k 를 형성한다.

Little(1986)은 칸 분할을 결정하는 법칙을 제안하지 않았다. 그러나 η_i 나 $\hat{Y}_i$ 모집단을 k_j^{-1} (j=1, 2,…, k-1) 분위수(quantile)로 나누어 칸 분할을 할 수 있다.

더우기, 주어진 보조변수 X_i 에 대하여, 작은 수의 칸에 의해서도 편향이 많이 감소될

수도 있다. 그러나 중요한 보조변수가 빠지면, 편향이 감소되지 않는다. 마지막으로, 핫-덱 대체로 주어지 수정칸 안에 결측값을 대치하는 것이다. 그러면, 평균추정량은

$$\hat{\bar{Y}}_{imp} = (\sum_{i \in s} \lambda_i)^{-1} (\sum_{i \in s} \lambda_i Y_i^*) \qquad (3.1)$$

단, Y_i^* 는 관찰치이거나 대체값이다.

IV. 무응답 자료 분석

패널 무응답 편향을 감소시키기 위하여는 사용되는 변수는 패널 무응답과 상관이 있는 변수는 모두된다.너무 많은 변수가 있거나, 교호작용을 생가하지 않으면 스크린(screen) 분석은 제한되어 있다. 적당한 수의 칸을 선택하는 문제는 편향-분산 trade-off 에 의하여 결정된다. 칸의 수가 늘어나면 편향이 감소되지만 부산은 증가할 수있다, 어떤 칸의 편향이 크면 그 칸은 다시 분할하여야 한다.

V. 토 의

무응답 가중은 단위 무응답을 보상하기 의하여 널리 사용된다.표본조사에서 기본적으로 필요한 것은 가중을 정하기 위하여 응답자와 무응자에 대한 보조 변수의 정보이다. 많은 조사에서 이러한 정보는 오직 적은 수의 보조 변수들에서만 얻을 수 있다.그러한 조사에서는,무응답 가중값들을 보조 변수들의 교차분류에 초한 크래스들의 집합에 대한 가중 크래스 수정으로 발전시킬수 있다.

무응답 가중값을 수정하기 위한 많은 수의 보조변수에 대한 정보를 알수 있는 경우도 있다.예를 들어 행정적 기록 시스템이 이용될 때에는 무응답 가중값을 수정하기 위한 많은 정보를 얻을 수 있다.조사를 여러번 계속할 때에,처음에 조사한 자료는 나중에 조사한 무응답을 보상하는 데 이용될 수 있다.패널조사에서도 마찬가지이다.

모든 표본추출된 단위들에 많은 수의 보조변수들이 있을 때에 두 가지 선택이 필요하다.첫째,수정에서 사용해야 할 보조변수들을 선택한다.두째,응용되어야 할 수정 방법을 선택하는 것이다.로지스틱 회귀모형을 이용하여 보조변수를 선택한다. 로지스틱 회귀

모형에서 여러 항목들의 응답율에 대한 결합 관계를 시험할수 있다. 주효과 변수들의 모형에서 교호작용항들을 찾어 낼수 있다. 응답 상태와 관련이 있는 두 항목이 상관이 클 때에는 하나의 항목만 사용해도 충분하다.

<참 고 문 헌>

Eltinge.J.L. and Yansane.I.S.(1997) Diagnostics for formation of nonresponse adjustment cells. Survey methodology.23. 33-40.

Rizzo.L., Kalton.G. and Brick J.M.(1996). A comparison of some weighting adjustment methods for panel nonresponse. Survey Methodology.22.43-53.

Losinger.W.C., Garber.L.P., Wagner.B.A and Hill.G.W.(2000). A cautionary note on adjusting weights for nonresponse. Survey Methodology 26.109-111

Deville.J.C., and Sarndal.C.E.(1992). Calibration estimators in survey sampling. Journai of the American Statistical Association. 87.376-382.