무응답 수정 칸의 형성 진단

신민웅[1], 윤연옥[2], 이주영[3], 이상은[4]

1. 서론

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-칸 안에서 가중수정이나 단순 핫-덱 대체를 할수가 있다. 각 칸에서 만일 조사 항목과 응답확률사이에 공분산이 근사적으로 영(zero)이라면 모평균과 모총계의 수정 추정량의 무응답 편향은 근사적으로 영이 될 것이다.

무응답-수정은 인구학적 또는 지리적 분류 변수들로 수정-칸을 형성하여 이루어 졌으나, Little(1986)과 다른 학자들은 응답확률이나 항목(item) 값들에 따라서 칸을 형성하였다.

Eltinge(1997)은 칸 형성의 유용성에 대한 진단(diagnostics)을 논하였다. 주요한 관심은 칸의 수에 대해 판정, 수정된 칸과 수정 안된 칸의 비교, 추가적으로 더 나눠야 할 칸의 발견, 그리고 응답확률과 항목값들로 칸을 형성했을 때에 두 방법 각각 추정결과를 비교하는 것이다.

2. 무응답 편향

크기가 N인 모집단 U에서 조사항목 $Y_i,\ i \in U$ 에 대해서 , 모평균은

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

이다. 표본 s는 크기 n이고, π_i는 단위 i가 표본에 포함될 확률이다. 무응답은 유사-확률모형을 만족한다고 가정한다(Oh와 Scheuren, 1983).

그리고 R_i는 지시변수로

1) (449-791) 경기도 용인시 모현면, 한국외국어대학교 통계학과, 교수
E-mail: mwshin@stat.hufs.ac.kr
2) (302-701) 대전시 서구 둔산동 920번지 정부대전청사 3동 통계청, 사무관
E-mail: yyoon@nso.go.kr
3) (122-701) 서울특별시 은평구 녹번동 5번지, 국립보건원 유전체연구소, 선임연구원
E-mail; jylee_monte@hanmail.net
4) (442-760) 경기도 수원시 팔달구 이의동, 경기대학교 응용정보통계학과, 조교수
E-mail: sanglee@stat.kyonggi.ac.kr

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

라 하자.

R_i는 서로 독립인 베르누이(η_i) 확률변수이고, 고정된 응답확률 η_i 는 단위에 따라서 다르다고 가정한다. 조사 가중값은 $\lambda_i = 1/\pi_i$ 이고, 수정안된 조사-가중 평균무응답은

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

이다.

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

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

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

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

이다. 단, $w_h = (\sum_{i\in s} \lambda_i)^{-1} \sum_{i\in s_h} \lambda_i$ 이고, $\bar{Y}_{kR} = (\sum_{i\in s_h} \lambda_i R_i)^{-1} \sum_{i\in s_h} \lambda_i R_i Y_i$ 이다. 수정 추정량 $\hat{\bar{Y}}_k$ 은 근사적으로 나머지 무응답 편향

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

을 갖는다. 단, N_h 는 U_h안의 단위들의 수이고, $\bar{\eta}_i = N_h^{-1} \sum_{i\in U_h} \eta_i$, $\bar{Y}_h = N_h^{-1} \sum_{i\in U_h} Y_i$ 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.

3.응답률 또는 예측항목에 기초한 수정 칸

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 그룹화함으로서 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

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

단, Y_i^* 는 관찰치이거나 대체값이다.

4. 무응답 자료 분석

적당한 수의 칸을 선택하는 문제는 편향-분산 trade-off 에 의하여 결정된다. 칸의 수가 늘어나면 편향이 감소되지만 분산은 증가할 수 있다. 어떤 칸의 편향이 크면 그 칸은 다시 분할하여야 한다.

5.토의

무응답 가중은 단위 무응답을 보상하기 의하여 널리 사용된다. 표본조사에서 기본적으로 필요한 것은 가중을 정하기 위하여 응답자와 무응답자에 대한 보조 변수의 정보이다. 로지스틱 회귀 모형에서 여러 항목들의 응답율에 대한 결합 관계를 시험할수 있다. 주효과 변수들의 교호작용항들을 모형에서 찾아 낼 수 있다. 응답 상태와 관련이 있는 두 항목이 상관이 클 때에는 하나의 항목만 사용해도 충분하다.

참고문헌

1. Eltinge.J.L. and Yansane.I.S.(1997) Diagnostics for formation of nonresponse adjustment cells. Survey methodology.23. 33-40.
2. Rizzo.L., Kalton.G. and Brick J.M.(1996). A comparison of some weighting adjustment methods for panel nonresponse. Survey Methodology.22.43-53.
3. 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
4. Deville.J.C., and Sarndal.C.E.(1992). Calibration estimators in survey sampling. Journal of the American Statistical Association. 87.376-382.