한국통계학회논문집
2011, 18권, 1호, 47–55

DOI: 10.5351/CKSS.2011.18.1.047

보조 정보에 의한 이중적 로버스트 대체법

박현아[a], 전종우[a], 나성룡[1,b]

[a]서울대학교 통계학과, [b]연세대학교 정보통계학과

요 약

비대체와 회귀대체는 조사변수의 모형과 조사변수와 보조변수의 관계에 의존하며 모형이 성립되지 않는 경우 이들 대체법을 이용한 추정량의 불편성은 보장되지 않는다. 본 연구에서는 모형이 성립되지 않는 경우에도 추정량의 근사적 불편성이 성립되는 로버스트 대체법을 개발한다. 대체법 개발시 보조변수의 모수 정보를 이용하여 추정량의 효율 증대를 가져오게 한다. 모의실험을 실시하여 본 연구에 대한 이론적 결과의 타당성을 보인다.

주요용어: 대체법, 이중적 로버스트, 비대체법, 보조변수.

1. 서론

실제 조사에서 발생되는 무응답의 형태는 두가지로 분류할 수 있다. 전체 응답의 거절 또는 부재 등으로 나타나는 단위 무응답과 설문조사 항목안에서 발생하는 항목무응답으로 분류된다. 단위무응답이 발생하면 가중치 조정을 통하여 보정이 되며, 항목무응답이 발생하면 대체법을 사용하여 결측자료를 처리하게 된다. 대체법의 종류로는 평균대체법, 핫덱대체법 등 여러가지가 있으며 그중 보조변수를 사용하는 것으로 비대체법과 회귀대체법이 있다. Kalton (1983)과 Groves 등 (2002)은 표본조사에서 발생할 수 있는 여러가지 대체법에 관한 것을 제시하고 있다.

대체법 중 비대체와 회귀대체는 조사변수(target variable)와 보조변수(auxiliary variable)의 모형에 의존하는데 조사변수의 모형이 옳지 않으면 이들 대체를 사용한 추정량의 성질이 좋다는 것을 보장할 수 없게 된다. 즉, 대체 후 추정량의 불편성이 성립되지 않으며 추정량의 효율도 떨어지게 된다. 대체 후의 추정량의 불편성과 효율을 계산하기 위한 방법에는 두 가지의 접근 방법이 있다. 즉 조사변수의 응답한 부분의 모형과 무응답한 부분의 모형을 가정하여 접근하는 모형기반접근법(model-based approach)과 응답변수와 응답확률을 이용하는 이단계접근법(two-phase approach)이 있다. 이단계 접근법을 다른 말로 응답모형접근법(response model approach)이라 한다.

기존의 비대체법과 회귀대체법은 조사변수의 회귀모형을 가정하며 모형기반접근법에 의해 대체 후 추정량의 성질을 규명하는데 회귀모형의 가정이 성립되지 않으면 추정량의 불편성을 만족할 수 없게 된다. 이 문제는 조사변수의 모형이 성립되지 않아도 불편성이 만족되도록 두 가지 접근 방법에서 추정량의 불편성이 만족되는 이중적 로버스트 대체법(doubly robust imputation)을 고려해서 해결할 수 있다. 다시말하면 이중적 로버스트 대체법이란 모형기반접근법과 응답모형접근법에서 이 대체를 사용한 추정량의 불편성이 만족되는 것을 말하며 만약 관심변수의 모형에 관한 가정이 성립하지 않아

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2009-0063964).

[1] 교신저자: (220-710) 강원도 원주시 흥업면 매지리 234, 연세대학교 정보통계학과, 부교수.
E-mail: nasr@yonsei.ac.kr

도 추정량의 불편성이 성립될 수 있는 것을 말한다. 이중적 로버스트 대체법은 Carpenter와 Kenward (2006), Qin 등 (2008), Cao 등 (2009) 등에서 연구되어 있고 Kim과 Park (2006)은 보조변수와 응답확률을 적용하는 대체법에 대하여 모형기반접근법에서도 불편성이 만족되고 이단계접근법에서도 근사적 불편성이 만족되는 것을 보였다.

본 연구에서는 보조변수가 취할 수 있는 값을 표본의 정보에서 모수의 정보로 확장시켜 두 접근 방법에서의 근사적 불편성도 확보하면서 더 많은 정보의 유입으로 추정량의 효율성도 향상시키는 이중적 로버스트 대체법을 개발한다. 이 새로운 대체법은 응답확률을 사용하는데 일반적으로 이 값들은 추정되어져야 한다. Rosenbaum (1987), Robins 등 (1994), Kim과 Park (2006)은 로지스틱 모형을 사용하여 응답확률을 추정하고 있으며 본 연구에서도 응답확률에 대하여 로지스틱모형을 가정하여 모수를 추정한다. 이 논문에서는 새로운 대체법과 응답확률을 사용한 추정량의 근사불편성을 증명하며 효율을 계산한다. 또한 응답확률의 추정을 포함하여 추정량의 근사적 불편성을 계산한다. 이론적 결과들의 타당성을 모의실험을 통하여 입증하며 기존 대체법을 이용한 추정량과의 비교를 통하여 새로운 대체법의 우월성을 알아 본다.

2. 이중적 로버스트 대체법

i번째 개체의 조사변수가 y_i이고 크기가 N인 모집단을 $U = \{y_1, y_2, \ldots, y_N\}$로 정의한다. 표본조사에서 추계되어야 할 대상은 모평균 $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$이며 추정량은 표본설계에 의해 결정되는 가중치 w_i를 사용하여 $\hat{Y}_n = \sum_{i=1}^{n} w_i y_i$이다. 이 모평균의 추정량에 사용된 가중치 w_i는 임의의 표본설계에서

$$E_D\left(\hat{Y}_n\right) = \bar{Y} \tag{2.1}$$

의 불편성을 만족함을 가정한다. 단, $E_D(\hat{Y}_n)$은 표본설계에 의해 계산되는 평균이다. 표본설계에 의해 추출된 표본은 조사과정에서 무응답이 발생 할 수 있다. 응답 여부를 추정과정에 표시하기 위해 추출된 표본하에서의 응답 확률변수

$$R_i = \begin{cases} 1, & \text{표본의 } i\text{번째 개체가 응답한 경우,} \\ 0, & \text{표본의 } i\text{번째 개체가 응답하지 않은 경우} \end{cases}$$

와 응답확률 $\pi_i = P(R_i = 1)$을 $i = 1, \ldots, n$에 대하여 정의한다.

일반적인 비대체(ratio imputation)를 사용하여 모평균을 추정하면

$$\hat{Y}_{RI} = \sum_{i=1}^{n} w_i R_i y_i + \sum_{i=1}^{n} w_i (1 - R_i) x_i \hat{r} \tag{2.2}$$

와 같고 $\hat{r} = (\sum_{i=1}^{n} w_i R_i x_i)^{-1} \sum_{i=1}^{n} w_i R_i y_i$이다. 비대체에 대하여 Rao (1996)와 Rao와 Sitter (1995)를 참고할 수 있다. 기존의 비대체를 사용하여 모평균 등을 추정하면 뽑힌 표본과 x_i 및 응답여부의 정보에 따라 정의되는 모형

$$E_M(y_i) = x_i r \tag{2.3}$$

에 의존하여 추정량의 불편성이 성립된다. 단, $E_M(y_i)$는 관심변수의 모형에 의해 계산되는 평균이다. 즉 모형기반접근법에 의해 불편성이 증명되며 모형의 가정이 성립되지 않으면 기존의 비대체를 사용한 추정량의 불편성은 만족되지 않게 된다.

본 논문에서는 모형의 가정이 성립되지 않아도 근사적 불편성이 만족되는 대체법을 이용한 추정량을 생각하기 위해 모형기반접근법과 응답변수를 이용한 이단계 접근법에서 근사적 불편성이 만족되는

대체법과 그 대체법을 이용한 추정량을 제시한다. 또한 추정량의 효율을 높이기 위해 보조변수의 모수 정보를 사용한 대체법을 생각한다. 먼저 $\bar{X} = N^{-1}\sum_{i=1}^{N} x_i$와 응답확률을 사용하여 이단계 접근법, 즉 응답모형에 의해 근사적으로 불편성이 만족되는 추정량은

$$\tilde{Y}_{ARI} = \sum_{i=1}^{n} w_i x_i \tilde{r}_A + \left(\sum_{i=1}^{n} w_i x_i\right)^{-1} \left(\sum_{i=1}^{n} w_i \pi_i^{-1} R_i y_i\right) \bar{X} - \sum_{i=1}^{n} w_i \pi_i^{-1} R_i x_i \tilde{r}_A$$

식으로 정의되고 이때 식 (2.2)와 같이

$$\tilde{Y}_{ARI} = \sum_{i=1}^{n} w_i R_i y_i + \sum_{i=1}^{n} w_i (1 - R_i) x_i \tilde{r}_A \tag{2.4}$$

식이 성립하도록 $\tilde{r}_A$를 정하면

$$\tilde{r}_A = \left\{\sum_{i=1}^{n} w_i R_i x_i \left(\pi_i^{-1} - 1\right)\right\}^{-1} \sum_{i=1}^{n} w_i R_i y_i \left(\frac{\bar{X}\pi_i^{-1}}{\sum_{i=1}^{n} w_i x_i} - 1\right)$$

의 식을 얻는다. 따라서 이 새로운 대체법을 사용하는 추정량의 최종적인 형태는 식 (2.4)가 된다. 새로운 대체법을 사용한 추정량은 두 접근 방향에 대하여 근사적 불편성이 성립하며, 따라서 모형의 가정이 성립되지 않아도 근사적 불편성이 만족되게 된다. Kim과 Park (2006)도 모형기반접근법에서도 불편성이 만족되고 이단계접근법에서도 불편성이 만족되는 대체법을 소개하고 있다.

제시된 대체법을 사용한 추정량의 근사적 불편성을 증명하기 위해 다음을 가정한다.

(T1) Isaki와 Fuller (1982)가 제시한 모집단과 표본의 연속적인 집합을 가정하고, 관심변수와 보조변수의 적률에 관하여

$$N^{-1}\sum_{i=1}^{N} \theta_i^{2+\tau} = O(1)$$

의 성질이 성립하는데, $\tau > 0$이고 θ는 y, x, z를 나타낸다.

(T2) 음이 아닌 상수 D_1, D_2, D_3, D_4에 대하여

$$D_1 < \max_{1\le i\le N}\{nw_i\} < D_2$$
$$D_3 < n\mathrm{Var}_D(\hat{Y}_n) < D_4$$

이 성립한다. 단, $\mathrm{Var}_D(\hat{Y}_n)$은 표본설계에 의해 계산되는 분산이다.

(T3) 응답확률변수 $R_1, \ldots, R_n$은 서로 독립이고 응답확률에 대하여

$$\pi_i > C_1$$

이 음이 아닌 상수 C_1과 $i = 1, \ldots, n$에 대하여 성립한다.

한편 Fay (1991)가 제시한 R_i의 확장된 정의를 사용한다.

이중적 대체법을 사용한 추정량 $\tilde{Y}_{ARI}$가 두 가지 접근방향에서 근사적 불편성을 만족하는 것을 다음의 정리 1에서 증명한다.

정리 1. 조건 (T1), (T2), (T3)를 가정하자. 이때 $\tilde{Y}_{ARI}$에 대하여

$$E\left(\tilde{Y}_{ARI}\right) = \bar{Y} + o\left(n^{-\frac{1}{2}}\right) \tag{2.5}$$

의 근사적 불편성이 성립한다. 이단계접근법에서

$$\mathrm{Var}\left(\tilde{Y}_{ARI}\right) = \mathrm{Var}_D\left[\sum_{i=1}^{n} w_i(y_i - Rx_i)\right] + E_D\left[\sum_{i=1}^{n} w_i^2\left(\pi_i^{-1} - 1\right)(y_i - r_A x_i)^2\right] + o\left(n^{-1}\right) \tag{2.6}$$

이 성립하는데, 여기에서 $R = \bar{X}^{-1}\bar{Y}$, $\bar{Y}_\pi = N^{-1}\sum_{i=1}^{N}\pi_i y_i$, $\bar{X}_\pi = N^{-1}\sum_{i=1}^{N}\pi_i x_i$, $r_A = (\bar{X} - \bar{X}_\pi)^{-1}(\bar{Y} - \bar{Y}_\pi)$이다. 한편

$$\mathrm{Cov}_M(y_i, y_j) = \begin{cases} \sigma_i^2, & i = j, \\ 0, & i \neq j \end{cases} \tag{2.7}$$

와 (2.3)을 가정하는 모형기반접근법에서

$$\mathrm{Var}\left(\tilde{Y}_{ARI}\right) = \mathrm{Var}_D\left[r\sum_{i=1}^{n} w_i\left(1 - R_X R_\pi\right)x_i\right] + E_D\left[\sum_{i=1}^{n} w_i^2 R_i^2\left\{1 + R_X\left(\pi_i^{-1} - 1\right)\right\}^2\sigma_i^2\right] + o\left(n^{-1}\right) \tag{2.8}$$

식이 성립한다. 단 $\bar{X}_{R\pi} = N^{-1}\sum_i^N R_i\pi_i^{-1}x_i$, $R_\pi = \bar{X}^{-1}\bar{X}_{R\pi}$, $\bar{X}_R = N^{-1}\sum_{i=1}^{N}R_i x_i$, $R_X = (\bar{X}_{R\pi} - \bar{X}_R)^{-1}(\bar{X} - \bar{X}_R)$, $\mathrm{Cov}_M(y_i, y_j)$는 관심변수의 모형에 의해 계산되는 공분산이다.

증명: 먼저 이단계 접근법을 사용하여 추정량의 근사적 불편성을 보이고 분산을 계산한다. 식 (2.1)과 (T1)–(T3)를 사용하여 $\tilde{r}_A$에 대하여 테일러전개를 한다.

$$\tilde{r}_A - r_A = \left(\bar{X} - \bar{X}_\pi\right)^{-1}\left[\left(\tilde{Y}_\pi - \bar{Y}\right) - R\left(\hat{X}_n - \bar{X}\right) - \left(\tilde{Y}_R - \bar{Y}_\pi\right) - r_A\left(\tilde{X}_\pi - \bar{X}\right) - \left(\tilde{X}_R - \bar{X}_\pi\right)\right] + o_p\left(n^{-\frac{1}{2}}\right),$$

여기서 $\tilde{Y}_\pi = \sum_{i=1}^{n} w_i\pi_i^{-1}R_i y_i$, $\tilde{X}_\pi = \sum_{i=1}^{n} w_i\pi_i^{-1}R_i x_i$, $\tilde{Y}_R = \sum_{i=1}^{n} w_i R_i y_i$, $\tilde{X}_R = \sum_{i=1}^{n} w_i R_i x_i$, $\hat{X}_n = \sum_{i=1}^{n} w_i x_i$이다. 위 식을 이용하면

$$\begin{aligned}\tilde{Y}_{ARI} &= \tilde{Y}_R + \left(\hat{X}_n - \tilde{X}_R\right)(\tilde{r}_A - r_A) + r_A\left(\hat{X}_n - \tilde{X}_R\right) \\ &= \bar{Y} + \sum_{i=1}^{n} w_i\left[\left(r_A - R\right)x_i + \pi_i^{-1}R_i\left(y_i - r_A x_i\right)\right] + o_p\left(n^{-\frac{1}{2}}\right)\end{aligned}$$

식을 얻게되고 식 (2.1)에 의하여 식 (2.5)가 증명된다. 또

$$\mathrm{Var}_D\left[E\left(\tilde{Y}_{ARI} \mid (x_1, y_1), \ldots, (x_n, y_n)\right)\right] = \mathrm{Var}_D\left[\sum_{i=1}^{n} w_i(y_i - Rx_i)\right] + o\left(n^{-1}\right)$$

와

$$E_D\left[\mathrm{Var}\left(\tilde{Y}_{ARI} \mid (x_1, y_1), \ldots, (x_n, y_n)\right)\right] = E_D\left[\sum_{i=1}^{n} w_i^2\left(\pi_i^{-1} - 1\right)(y_i - r_A x_i)^2\right] + o\left(n^{-1}\right)$$

의 두 식을 얻을 수 있고 이에 의해 식 (2.6)이 증명된다.

다음으로 모형기반접근법에 의해 추정량의 불편성을 증명한다. 조건 (2.1), (2.3), (2.7), (T1)–(T3)을 사용하여 $\tilde{r}_A$의 테일러전개를 구하면

$$\tilde{r}_A - r = \left(\bar{X}_{R\pi} - \bar{X}_R\right)^{-1}\left[\left(\tilde{Y}_\pi - r\bar{X}_{R\pi}\right) - rR_\pi\left(\hat{X}_n - \bar{X}\right) - \left(\tilde{Y}_R - r\bar{X}_R\right) - r\left\{\left(\tilde{X}_\pi - \bar{X}_{R\pi}\right) - \left(\tilde{X}_R - \bar{X}_R\right)\right\}\right] + o_p\left(n^{-\frac{1}{2}}\right)$$

이고 이를 대입해서

$$\begin{aligned}\tilde{Y}_{ARI} &= \tilde{Y}_R + \left[\hat{X}_n - \tilde{X}_R - \left(\bar{X} - \bar{X}_R\right)\right](\tilde{r}_A - r) + r\left(\hat{X}_n - \tilde{X}_R\right) + \left(\bar{X} - \bar{X}_R\right)(\tilde{r}_A - r)\\ &= \tilde{Y}_R + r\left(\hat{X}_n - \tilde{X}_R\right) + R_X\left[\tilde{Y}_\pi - rR_\pi\hat{X}_n - r\left\{\left(\tilde{X}_\pi - \bar{X}_{R\pi}\right) - \left(\tilde{X}_R - \bar{X}_R\right)\right\}\right] + o_p\left(n^{-\frac{1}{2}}\right)\end{aligned}$$

식을 얻고 이에 의하여 식 (2.5)가 증명된다. 한편 식 (2.3)과 (2.7)을 이용하여

$$\mathrm{Var}_D\left[E\left(\tilde{Y}_{ARI} \mid (x_1, R_1), \ldots, (x_n, R_n)\right)\right] = \mathrm{Var}_D\left[r\sum_{i=1}^{n} w_i\left(1 - R_X R_\pi\right)x_i\right] + o\left(n^{-1}\right)$$

과

$$E_D\left[\mathrm{Var}\left(\tilde{Y}_{ARI} \mid (x_1, R_1), \ldots, (x_n, R_n)\right)\right] = E_D\left[\sum_{i=1}^{n} w_i^2 R_i^2\left\{1 + R_X\left(\pi_i^{-1} - 1\right)\right\}^2\sigma_i^2\right] + o\left(n^{-1}\right)$$

을 구하며 또한 식 (2.8)이 증명된다. □

참고 1. 이중적 대체법에 관한 논문인 Kim과 Park (2006)에서 제시된 추정량 $\hat{Y}_{Id}$을 모평균에 관한 추정으로 바꾸어 분산을 계산하면 다음과 같다.

$$\mathrm{Var}\left(\hat{Y}_{Id}\right) = \mathrm{Var}_D\left(\hat{Y}_n\right) + E_D\left[\sum_{i=1}^{n} w_i^2\left(\pi_i^{-1} - 1\right)(y_i - r_A x_i)^2\right] + o\left(n^{-1}\right),$$

이 때, $\mathrm{Var}(\hat{Y}_{Id})$이 $\mathrm{Var}(\tilde{Y}_{ARI})$ 보다 크게 되는 조건은

$$\rho > \left[2\mathrm{CV}\left(\hat{Y}_n\right)\right]^{-1}\mathrm{CV}\left(\hat{X}_n\right)$$

이다. 단, $\mathrm{CV}(\hat{X}_n) = \bar{X}^{-1}[\mathrm{Var}_D(\hat{X}_n)]^{-1/2}$, $\mathrm{CV}(\hat{Y}_n) = \bar{Y}^{-1}[\mathrm{Var}_D(\hat{Y}_n)]^{-1/2}$이다. 또한 $\rho = [\mathrm{Var}_D(\hat{X}_n)\,\mathrm{Var}_D(\hat{Y}_n)]^{-1/2}\,\mathrm{Cov}_D(\hat{X}_n, \hat{Y}_n)$, $\mathrm{Cov}_D(\hat{X}_n, \hat{Y}_n)$는 표본설계에 의해 계산되는 공분산이다. 따라서 $\rho > 1/2$이고 $[\mathrm{CV}(\hat{Y}_n)]^{-1}\mathrm{CV}(\hat{X}_n) < 1$이면 $\tilde{Y}_{ARI}$의 효율이 $\hat{Y}_{Id}$보다 좋음을 알 수 있다.

일반적으로 응답확률은 알 수 없으므로 추정해서 사용한다. 응답확률의 추정방법으로 모수적 추정법을 이용하는데 응답확률의 모형으로

$$\pi_i = \pi(z_i; \alpha) = \left(1 + \exp\left(-z_i^T\alpha\right)\right)^{-1}$$

의 로지스틱모형을 가정한다. 여기서 z_i는 보조변수들을 나타내고 $\alpha = (\alpha_1, \ldots, \alpha_p)^T$이다. 응답확률의 추정은 α를 추정함으로써 이루어지며 추정량 $\hat{\alpha}$에 대하여

$$n^{\frac{1}{2}}(\hat{\alpha} - \alpha) = n^{-\frac{1}{2}}\sum_{i=1}^{n} H(z_i, R_i; \alpha) + o_P(1) \tag{2.9}$$

표 1: 추정량들의 평균, 분산, 표준화된 MSE

모형	추정량	추정량의 평균	추정량의 분산	표준화된 MSE
선형모형	$\hat{Y}_n$	4.2936	0.0617	100
	$\hat{Y}_{RI}$	4.2926	0.0623	101
	$\hat{Y}_{Id}$	4.2939	0.0626	102
	$\hat{Y}_{Ie}$	4.2938	0.0626	101
	$\tilde{Y}_{ARI}$	4.2878	0.0166	27
	$\hat{Y}_{ARI}$	4.2886	0.0135	22
비선형모형	$\hat{Y}_n$	2.0473	0.0612	100
	$\hat{Y}_{RI}$	2.2047	0.0660	148
	$\hat{Y}_{Id}$	2.0498	0.0653	107
	$\hat{Y}_{Ie}$	2.0507	0.0642	105
	$\tilde{Y}_{ARI}$	2.0418	0.0342	56
	$\hat{Y}_{ARI}$	2.0432	0.0326	53

을 성립한다. 여기서 $E[H(z_i, R_i; \alpha)] = 0$이고 $E[H(z_i, R_i; \alpha)H(z_i, R_i; \alpha)^T]$은 양정치(positive definite) 조건을 만족한다 (Kim과 Park, 2006).

추정된 응답확률을 사용한 대체법에 의한 추정량은 다음과 같이 정의 할 수 있다.

$$\hat{Y}_{ARI} = \sum_{i=1}^{n} w_i R_i y_i + \sum_{i=1}^{n} w_i (1 - R_i) x_i \hat{r}_A,$$

여기서 $\hat{r}_A$는

$$\hat{r}_A = \left\{ \sum_{i=1}^{n} w_i R_i x_i \left(\hat{\pi}_i^{-1} - 1 \right) \right\}^{-1} \sum_{i=1}^{n} w_i R_i y_i \left(\frac{\bar{X} \hat{\pi}_i^{-1}}{\sum_{i=1}^{n} w_i x_i} - 1 \right)$$

식으로 정의된다. $\hat{Y}_{ARI}$에 대하여 테일러 전개를 적용하면 다음의 결과를 얻게 된다.

$$\hat{Y}_{ARI} = \tilde{Y}_{ARI} + (\hat{\alpha} - \alpha)^T \left\{ N^{-1} \sum_{i=1}^{N} \pi_i \left(\frac{\partial \pi_i^{-1}}{\partial \alpha} \right) (y_i - r_A z_i) \right\} + o_p \left(n^{-\frac{1}{2}} \right).$$

그리고 식 (2.5)와 (2.9)에 의하여 $\hat{Y}_{ARI}$의 근사적 불편성이 성립한다.

참고 2. 이중적 로버스트(doubly robust) 대체법이란 관심변수의 모형이 맞거나 응답모형이 맞을 때 이 대체법을 사용한 추정량의 불편성이 성립되는 것을 말한다. 예를 들어 응답모형의 가정이 성립되지 않아도 추정량의 불편성이 성립된다는 것을 뜻한다. 즉, 이중적 로버스트한 성질은 로버스트하다는 것과는 다른 의미를 지닌다. 조사변수 모형의 가정이 성립되지 않을 때도 제안된 대체를 사용한 추정량의 불편성이 만족된다는 것은 응답확률의 모형의 가정이 성립될 때만 가능하다. 응답확률의 모형이 맞지 않는 경우에는 제안된 추정량의 불편성이 성립되지 않는다. 다시 말하면 조사변수의 모형이 맞을 경우에는 응답확률의 모형에 상관없이 제안된 대체를 사용한 추정량의 불편성을 말할 수 있게 된다. 이와 같은 성질을 만족하는 대체 및 추정량을 이중적 로버스트 대체법 및 추정량이라 한다.

3. 모의 실험

새로운 이중적 대체법을 사용한 추정량 $\tilde{Y}_{ARI}$와 $\hat{Y}_{ARI}$에 관한 성질이 이론적으로 증명되었다. 이제 모의실험을 통하여 이론적 증명들을 뒷바침한다. 이중적 대체법을 사용한 추정량의 불편성과 효율의

증대를 보이기 위하여 모집단 모형으로 선형모형과 비선형모형을 생각한다. 선형모형은 아래와 같으며 이 모형에서 모의 자료를 생성한다.

$$y_i = 3.9x_i + \sqrt{x_i}\,\epsilon_i,$$

여기서 $x_i \sim \text{Uniform}(0.1,\ 2.1)$, $\epsilon_i \sim N(0,\ 1)$이고 x_i와 ϵ_i는 서로 독립이다. 비선형모형은 다음과 같으며 이 모형에서 비선형자료를 생성한다.

$$y_i = (1.8x_i - 1)^2 + \sqrt{x_i}\,\epsilon_i$$

응답확률모형으로는 보조변수가 1개인 로지스틱회귀모형을 가정한다.

$$\pi_i = [1 + \exp(-1 + 2.3z_i)]^{-1} \exp(-1 + 2.3z_i)$$

본 모형에 의한 평균적 응답비율은 0.76이다. 응답확률의 추정은 로지스틱모형의 회귀계수들을 추정함으로써 얻어진다. 회귀계수들의 추정방법은 Newton-Raphson방법에 의해 계산된 최우추정법을 사용한다. 모의 실험의 횟수는 1000번이며 각 횟수마다 추출되는 방법은 단순임의복원추출을 사용하며 표본의 크기는 100개이다. 각 횟수마다 $(x_i,\ z_i,\ \pi_i,\ R_i,\ y_i)$의 값들이 100개씩 생성되며 100개의 값으로 완전한 응답에서 계산되는 $\hat{Y}_n$값과 기존의 비대체를 사용하는 추정량 $\hat{Y}_{RI}$의 값과 Kim과 Park (2006)에서 제시되는 추정량 $\hat{Y}_{Id}$, $\hat{Y}_{Ie}$의 값과 본 논문에서 제시되는 추정량 $\tilde{Y}_{ARI}$, $\hat{Y}_{ARI}$의 값이 각각 1개씩 생성된다. 여기서 $\hat{Y}_{Ie}$는 추정된 응답확률을 가지고 계산된 것이다. 추정량의 평균은 각 횟수마다 생성된 1000개의 추정량의 값의 평균값이며 추정량의 분산은 각 횟수마다 생성된 1000개의 추정량의 값의 분산이다. MSE는

$$\text{MSE} = \text{추정량의 분산} + (\text{추정량의 평균} - \text{모평균})^2$$

으로 계산되며 표준화된 MSE는 각 추정량들의 MSE값을 $\hat{Y}_n$의 MSE의 값으로 나눈 것을 백분율로 나타낸 것이다.

표 1의 결과를 보면 제안된 이중적 대체법을 사용한 추정량의 불편성은 선형모형과 비선형모형에서 근사적으로 만족됨을 알 수 있다. 이것은 선형모형의 가정이 성립하지 않아도 제안된 추정량의 불편성이 성립된다는 이론적 사실을 뒷바침하는 결과라 할 수 있다. 또한 Kim과 Park (2006)에서 제안된 추정량도 이중적 로버스트한 성질을 지님을 알 수 있다. 그리고 추정량의 분산을 보면 $\tilde{Y}_{ARI}, \hat{Y}_{ARI}$의 분산이 다른 추정량의 분산보다 작음을 볼 수 있다. 이것은 효율면에서 $\hat{Y}_{Id}$ 과 $\hat{Y}_{Ie}$보다 좋은 성질을 지님을 알 수 있다.

참고 문헌

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data, *Biometrika*, **96**, 723–734.

Carpenter, J. R. and Kenward, M. G. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data, *Journal of the Royal Statistical Society (Series A)*, **169**, 571–584.

Fay, R. E. (1991). A design-based perspective on missing data variance, *The ASA Proceedings of Bureau of the Census Annual Research Conference*, US Bureau of the Census, Washington, D. C., 429–440.

Groves, R., Dillman, D., Eltinge, J. and Little, R. J. A. (2002). *Survey Nonresponse*, Wiley, New York.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, **77**, 89–96.

Kalton, G. (1983). *Compensating for Missing Survey Data*, Institute for Social Research, University of Michigan, Ann Arbor.

Kim, J. K. and Park, H. (2006). Imputation using response probability, *Canadian Journal of Statistics*, **34**, 171–182.

Qin, J., Shao, J. and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses, *Journal of the American Statistical Association*, **103**, 797–810.

Rao, J. N. K. (1996). On variance estimation with imputed survey data, *Journal of the American Statistical Association*, **91**, 499–506.

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.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **89**, 846–66.

Rosenbaum, P. R. (1987). Model-based direct adjustment, *Journal of the American Statistical Association*, **82**, 387–394.

2010년 6월 접수; 2010년 8월 채택

Doubly Robust Imputation Using Auxiliary Information

Hyeonah Park[a], Jongwoo Jeon[a], Seongryong Na[1,b]

[a]Department of Statistics, Seoul National University
[b]Department of Information and Statistics, Yonsei University

Abstract

Ratio and regression imputations depend on the model of a survey variable and the relation between the survey variable and auxiliary variables. If the model is not true, the unbiasedness of the estimator using the ratio or regression imputation cannot be guaranteed. In this paper, we develop the doubly robust imputation, which satisfies the approximate unbiasedness of the estimator, whether the model assumption is valid or not. The proposed imputation increases the efficiency of estimation by using the population information of the auxiliary variables. The simulation study establishes the theoretical results of this paper.

Keywords: Imputation, doubly robust, ratio imputation, auxiliary variable.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. 2009-0063964).

[1] Corresponding author: Associate Professor, Department of Information and Statistics, Yonsei University, 234 Maeji, Heungup, Wonju, Gangwon 220-710, Korea. E-mail: nasr@yonsei.ac.kr