▨응용논문

무응답 상황하에서 최적추정량에 관한 연구

손창균
동신대학교 컴퓨터과학과
정훈조
한서대학교 컴퓨터과학과

The Study on the Optimal Estimators in the Presences of Nonresponse

Chang-Kyoon Son
Dept. of Computer Science, Dongshin University
Hun-Jo Jung
Dept. of Computer Science, Hanseo University

Abstract

In the survey, it is very hard to get the complete response. Because the respondents tend to refuse to the questionnaire with something like incomes of the individual or may not be at home in the survey time. These nonresponses are classified into two groups as the item nonresponse and the unit nonresponse. When the nonresponse happen to the special item of the questionnaire, it is called item nonresponse. On the other hand the unit nonresponse occurs to the totally missing in questionnaire. In this paper, we only consider to the unit nonresponse situation. We propose that the optimal estimator which is minimized the variance of the estimator under a fixed cost function for the survey and response.

1. 서론

일반적인 표본조사의 목적은 관심모집단으로부터 표본을 추출하여, 모집단의 총계나 비율, 평균인 모집단의 특성치를 추정하는데 있다. 이러한 표본조사에 관한 이론은 표본추출법, 추출한 표본자료를 분석하여 모집단의 특성치를 추정하는 방법, 그리고 이러한 추정치의 정도나 조사의 정확성을 평가하는 방법 등으로 구분할 수 있다. 통계적 관점에서 추정이론의 중요한 과제로는 표본으로부터 편의가 없는 최소분산 불편추정량을 구하는 것이다. 그러나 이러한 추정량은 현실적으로 구하기가 곤란한 것이 사실이다. 왜냐하면, 표본으로부터의 관찰과정이나, 면접과정으로부터 발생하는 오차와 분석과정이나 집계과정에서 발생하는 오차 때문이다. 특별히 실사과정에서 빈번히 발생하는 질의에 대한 무응답으로 인하여 모수를 추정하는 과정에서 오차가 수반된 편향된 추정량을 얻게된다. 따라서 조사에 대한 완전한 응답을 가정한 일반적인 추정방법으로는 이러한 문제를 해결할 수 없으며, 무응답이 발생한 상황하에서 일반적인 추정량을 사용함으로서 모수의 추정에 대한 왜곡을 초래하게 된다.

이러한 관점에서 Cassel, Sarndal 그리고 Wretman(1983)은 통계적인 무응답모형을 제안하였고, Godambe 과 Thompson(1986)은 무응답상황 하에서 추정함수를 제안하였고, 설계분산을 가장 작게 하는 최적의 추정함수를 유도하였다.

따라서, 이러한 무응답 상황을 고려하여 표본으로부터 모수를 추정한다면, 전형적인 추정과정보다 현실성을 가질 것이며, 또한 무응답 편의를 줄임으로서 보다 정도가 높은 추정을 할 수 있을 것이다.

본 논문은 무응답이 존재하는 상황에 대해 Sarndal(1986)이 제안한 회귀적인 접근에 무응답을 처리하는데 소요되는 비용을 함께 고려하여, 고정된 비용하에서 분산을 최소로 하는 최적의 포함확률을 구하고, 이를 이용하여 무응답 편의를 줄일 수 있는 추정량을 제안하고자 한다.

2. 추정량과 응답모형

2.1 무응답 수정 추정량

유한모집단 $U=\{1, 2, cdots, N\}$를 고려하고, 모집단으로부터 k번째 단위가 추출될 확률이 π_k인 추출설계로서 크기가 n인 표본 s를 추출한다고 하자. k번째 단위와 연관된 추출가중치는 π_k^{-1}이며, 이를 표본 가중치라고 하였다. 단위 k와 l이 동시에 표본에 포함될 확률은 π_{kl}이라 하자. 표본 s가 주어진 상태에서 특정한 단위 무응답이 발생한다고 가정하자. s에 대해 응답집합을 r이라 하고, 이 집합의 크기를 m이라 하자. 또한 관심변수 y가 단지 $k\in r$인 단위에 대해 관찰되었다고 하자. 무응답 편의에 대한 효과를 줄이기 위해 Sarndal(1986)은 다음과 같은 수정그룹방법(adjust-

ment group method)를 이용하였다.

표본 s가 H개의 그룹인 $s_1, s_2, \cdots, s_H$으로 나뉘고 각 그룹의 크기는 $n_1, n_2, \cdots, n_H$ 라 하자. 응답집합 r 또한 $r_1, r_2, \cdots, r_H$으로 나뉘고, 이들의 크기를 각각 $m_1, m_2, \cdots, m_H$ 라 하자. h그룹에서의 응답률은 $f_h = m_h/n_h$ 이라 하자. 이 방법은 표본가중치 뿐만 아니라 h그룹으로부터 나온 관찰치를 수정한 가중치인 f_h^{-1}를 부여해야 하며, 이를 무응답 가중치라 한다.

$\tau_y = \sum_U y_k$ 를 추정해야할 미지의 모집단 총계라 하자. 그러면 τ_y의 일반적인 수정계급추정량(adjustment class estimator)은 다음과 같다.

$$\hat{\tau}_y = \sum_{h}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k} \tag{2.1}$$

이와 같은 수정그룹 방법은 Godambe과 Thompson(1986)이 제안한 것으로서 동일한 그룹내에서 단위들의 응답확률이 같다는 가정에 근거한 것이다. 이 방법은 각각의 단위 $k \in s$에 대해 그룹을 구별할 수 있어야하기 때문에 그룹을 결정하기 위한 변수들이 기본적으로 응답모형을 추정하기 위해 사용될 수 있다. Sarndal(1986)은 이러한 변수를 공변수(covariate)라고 하였으며, 이와는 다르게 관심변수를 설명하는 변수로서 보조변수(auxiliary variable)로 구별하였다. Sarndal(1986)은 $k \in s$에 대해 이용 가능한 경우를 공통변수 또는 공변량 이라 하였고, $k \in U$인 단위에 대해 이용 가능한 경우 보조변수라 하였다. 식(2.1)의 보다 일반적인 추정량으로는 추출가중치인 π_k^{-1}을 이용하면, 다음과 같은 형태의 추정량을 얻을 수 있다(Sarndal(1986)).

$$\begin{aligned} \hat{\tau}_{\exp} &= \left(\sum_s \frac{1}{\pi_k}\right) \frac{\sum_{h}^{H} f_h^{-1} \sum_{r_h} y_k/\pi_k}{\sum_{h}^{H} f_h^{-1} \sum_{r_h} 1/\pi_k} \\ &= \hat{N}\, \tilde{y}_r \ . \end{aligned} \tag{2.2}$$

여기서, $\sum_s 1/\pi_k = \hat{N}$, $\tilde{y}_r = \dfrac{\sum_{h}^{H} f_h^{-1} \sum_{r_h} y_k/\pi_k}{\sum_{h}^{H} f_h^{-1} \sum_{r_h} 1/\pi_k}$ 이다.

$\tilde{y}_r$은 응답자들의 평균으로 k번째 단위의 곱 가중을 h번째 수정그룹의 각각의 단위 k에 대해 부여하여 사용하였다. 이 추정량은 무응답 상황하에서 적절한 추정량으로 추출설계를 포함하며, 무응답에 대한 수정에 이용할 수 있다.

만일 단순 공변량 x가 $k \in s$에 대해 관찰되었다고 하면, 전형적인 비 추정량의 관

점에서 무응답 상황에 대한 모집단 총계의 추정량은 다음과 같다(Sarndal(1986)).

$$\hat{\tau}_{RA} = \left(\sum_s \frac{x_k}{\pi_k}\right) \frac{\sum_H f_h^{-1} \sum_{r_h} y_k/\pi_k}{\sum_H f_h^{-1} \sum_{r_h} x_k/\pi_k} = \hat{N}\,\tilde{x}_s \frac{\tilde{y}_r}{\tilde{x}_r} \quad . \tag{2.3}$$

여기서, $\tilde{x}_s = \frac{\sum_s x_k/\pi_k}{\sum_s 1/\pi_k}$ 이고, $\tilde{x}_r = \frac{\sum_{r_h} x_k/\pi_k}{\sum_{r_h} 1/\pi_k}$ 이다.

위와 같은 평균은 x-변수에 대해 계산이 가능하며, 이는 모든 $k \in s$에 대해 관찰되며, y-변수에 대해서는 $k \in r$에 대해서만 관찰되기 때문에 계산이 불가능하다.

다음으로 전통적인 회귀추정량은 다음과 같다(Sarndal(1986)).

$$\hat{\tau}_{reg} = \hat{N}[\,\tilde{y}_r + b(\,\tilde{x}_s - \tilde{x}_r)]. \tag{2.4}$$

여기서 회귀계수 b에는 표본가중치와 무응답 가중치가 포함되어있으며, 이에 대한 식은 다음과 같다.

$$b = \frac{\sum_H f_h^{-1} \sum_{r_h} (y_k - \tilde{y}_r)(x_k - \tilde{x}_r)/\pi_k}{\sum_H f_h^{-1} \sum_{r_h} (x_k - \tilde{x}_r)^2/\pi_k}.$$

위의 추정량의 식으로부터 만일 모집단의 크기 N이 기지이면, 각각의 추정량의 $\hat{N}$를 N으로 대치하면 될 것이다.

2.2 응답모형

앞에서 살펴본 3가지 추정량에서 기본적으로 가정된 응답모형은 무응답 가중치들이 주어진 각각의 그룹내에 있는 단위에 대해 응답확률이 일정하다는 것으로서 이러한 가정을 보다 구체적으로 표현하면 다음과 같다(Sarndal(1986)).

가정1) 모든 단위 $k \in s_h$에 대해 응답확률은 미지의 모수 θ_h로 일정하다.

가정2) 각 단위들의 응답확률은 서로 독립이다.

응답확률 θ_h는 서로 다른 그룹간에 상당히 다양할 것이며, 이러한 이유로 수정그룹을 만들고 가중치를 부여하는 동기를 제공한다.

고정된 표본 s를 고려하면, 각각의 그룹수 $\boldsymbol{n}=(n_1,n_2,\cdots,n_H)$는 고정된다. 따라서 각각의 그룹에 대해 고정된 응답자 수 또한 $\boldsymbol{m}=(m_1,m_2,\cdots,m_H)$을 고려할 수 있다. 고정된 s와 $\boldsymbol{n}$으로 부터 위의 가정된 모형 하에서 응답집합 r_h을 추출하는 것은 n_h로 부터 m_h를 추출하는 단순임의 추출과 같게 된다. 그러므로 h그룹의 k번째 단위에 대한 조건부 응답확률은 $k\in s_h$에 대해 다음과 같다(Sarndal 외(1992)).

$$\pi_{k|s,\boldsymbol{m}}=\frac{m_h}{n_h}=f_h\ . \tag{2.5}$$

이와 유사하게 s와 $\boldsymbol{m}$이 주어진 조건하에서, 위에서 가정된 모형에 대해 단위 k와 l의 동시 응답확률은 다음과 같다(Sarndal 외 (1992)).

$$\pi_{kl|s,\boldsymbol{m}}=\begin{cases} f_h, & k=l \\ \dfrac{f_h(m_h-1)}{n_h-1}, & k\neq l\in s_h \\ f_hf_{h'}, & k\in s_h,\ l\in s_{h'} \end{cases}$$

위의 값들은 결국 h층의 n_h로 부터 m_h개를 층화 추출하는 것과 같다.

사실 위와 같은 가정들이 잘못 설정될 수 있으며, 특히 가정1)에서는 각각의 그룹내의 단위들이 서로 다른 응답확률을 가질 수 있다. 역시 가정2)에서는 각각의 응답확률이 서로 다른 단위에 의존할 수도 있다. 만일 위의 가정이 성립한다면, 이에 따른 추정치의 분산과 분산추정량에는 다음의 2가지 확률이 적용될 것이다.

첫째, π_k와 π_{kl}로서 이 값은 추출단계와 연관된 1차와 2차 포함확률이다. 둘째, $\pi_{k|s,\boldsymbol{m}}$과 $\pi_{kl|s,\boldsymbol{m}}$으로 이는 위에서 가정한 응답모형과 연관된 1차와 2차 조건부응답확률이다.

특히 π_k와 π_{kl}은 일반적인 포함확률이므로 추출단계에서 임의의 추출설계가 이용될 수 있다. 추정량의 식(2.2), 식(2.3) 그리고 식(2.4)로부터 총 분산을 다음과 같은 2가지 요소로 분해할 수 있을 것이다.

$$V(\hat{\tau})=V_1(\hat{\tau})+V_2(\hat{\tau}).$$

여기서 $V_1(\hat{\tau})$는 모집단으로부터 표본 추출에 따른 분산이고, $V_2(\hat{\tau})$는 표본을 조사하는 과정에서 발생하는 무응답에 따른 분산항이다.

따라서 만일 표본으로부터 완전응답을 얻었다면, 위의 분산식으로 부터 두 번째항은 무시될 수 있으며, 순수한 표본추출에 따른 분산항만이 남게된다. 만일 $V_2(\hat{\tau})$가 존재하는 경우에는 관심변수 y와 상관이 높은 공변량을 사용하여 현저히 줄일 수 있을 것이다. 결국, 무응답 편의를 줄이기 위해서는 강한 공변수를 사용하는 것이라고 할 수 있다. 이러한 공변수의 이용은 표본추출에 따른 분산항에는 별 영향을 끼치지 않음을 알 수 있다.

일반적으로 $V_1(\hat{\tau})$과 $V_2(\hat{\tau})$는 임의의 추정량에 대해 다음과 같은 형태를 가진다 (Sarndal 외 (1992)).

$$V_1(\hat{\tau}) = \sum\sum_U (\pi_k\pi_l - \pi_{kl})\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}, \tag{2.6a}$$

$$V_2(\hat{\tau}) = E_\xi E_m\left(\sum_h^H n_h^2 \frac{1-f_h}{m_h} S^2_{e_{r_h}}|s\right). \tag{2.6b}$$

이에 대한 각각의 추정량은 다음과 같다.

$$\hat{V}_1(\hat{\tau}) = \sum\sum_r \left(\frac{1}{\pi_k\pi_l} - \frac{1}{\pi_{kl}}\right)\frac{1}{\pi_{kls,m}}\left(\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\right), \tag{2.7a}$$

$$\hat{V}_2(\hat{\tau}) = \sum_h^H n_h^2 \frac{1-f_h}{m_h} S^2_{e_{r_h}}. \tag{2.7b}$$

여기서 $S^2_{e_{r_h}} = \frac{1}{m_h-1}\sum_{r_h}(e_k - \tilde{e}_{r_h})^2$ 이고 $\tilde{e}_{r_h}$는 다음의 각 경우에 대한 평균이다. 또한 e_k는 추정량의 종류에 따라 다음과 같다(Sarndal(1986))

$$e_k = \frac{(y_k - \tilde{y}_k)}{\pi_k}, \ \hat{\tau}_{\exp} \text{에 대해,} \tag{2.8a}$$

$$e_k = \frac{y_k - (\tilde{y}_k/\tilde{x}_k)x_k}{\pi_k}, \ \hat{\tau}_{RA} \text{에 대해,} \tag{2.8.b}$$

$$e_k = \frac{y_k - \tilde{y}_k - b(x_k - \tilde{x}_k)}{\pi_k}, \ \hat{\tau}_{reg} \text{에 대해 .} \tag{2.8.c}$$

3. 제안된 최적 포함확률

3.1 최적 추정 방정식

Godambe과 Thompson(1986)은 무응답을 고려한 상황에 대해 다음과 같은 일반적인 초모집단 모형을 가정하였다.

$$E_\xi(y_k-\theta x_k)=0 \quad . \tag{3.1}$$

따라서, 모형(3.1)은 $E_\xi(e_k)=0$, $V_\xi(e_k)=\sigma_k^2$, $E_\xi(e_k e_l)=0$을 만족한다.

이들은 모형(3.1) 하에서 $\pi_k>0$ 을 만족하는 임의의 추출설계 p에 대해 다음을 만족하는 최적 추정함수 ψ''^*가 존재함을 보였다.

$$E_\xi E(\psi''^*)^2 \le E_\xi E(\psi'')^2 \quad . \tag{3.2}$$

여기서, $\psi''^* = \sum_{k\in r_k}(y_k-\theta x_k)a_k/\pi_k$ 이다.

또한 이들에 의해 제시된 최적 추정 방정식은

$$\sum_{k\in r_k}(y_k-\theta x_k)a_k/\pi_k=0$$

이며, 초 모집단 모수 θ의 추정치 $\hat{\theta}_s$는 다음과 같다.

$$\hat{\theta}_s=\frac{\sum_{k\in r_k} y_k a_k/\pi_k}{\sum_{k\in r_k} x_k a_k/\pi_k} \quad . \tag{3.3}$$

3.2 최적 포함확률

앞에서도 살펴보았듯이 추정함수의 분산을 가장 적게 하는 최적의 포함확률을 찾을 수 있으며, 또한 무응답이 존재하는 경우에도 최적의 포함확률을 구할 수 있을 것이다.

따라서 2.2절의 분산을 최소로 하는 최적의 포함확률을 구하기 위해 다음과 같은 제한 조건을 가정하자.

$$\sum^N \pi_k = Constant \text{ (일정)}, \tag{3.4a}$$

$$\sum^N \pi_k \pi_{Hs.\,m} = Constant \text{ (일정)}. \tag{3.4b}$$

제한조건 (3.4a)는 앞의 분산식 중에서 표본을 추출하여 발생하는 첫 번째 분산항과 연관것이고, 식(3.4b)는 무응답에 기인한 두 번째 분산항에 연관 것이다.

따라서 위의 제한조건 하에서 분산식 (2.6a)와 (2.6b)를 최소로 하는 각각의 포함확률을 구하면 다음과 같다. (단, $\pi_{hs\,m}$은 응답확률이지만, 표본의 관점에서 보면 최종 표본에 대한 포함확률로 간주해도 무방하다.)

분산식 (2.6a)으로부터 $k=l$인 경우 다음과 같은 식으로 축소된다.

$$V_1(\hat{\tau}) = \sum_U \left(\frac{1}{\pi_k} - 1\right)\sigma_k^2 \ . \tag{3.5}$$

따라서 제한조건 (3.4a) 하에서 분산식 (3.5)를 최소로 하는 최적의 1차 포함확률은 다음과 같다.

$$\pi_k = \frac{n\sigma_k}{\sum_U \sigma_k} \ . \tag{3.6}$$

또한 제한조건 (3.4b) 하에서 표본 s가 주어진 조건하에서 분산식 (2.6b)을 최소로 하는 최적의 응답확률은 다음과 같다.

$$\pi_{hs,\,m} = \left(\frac{n_h S_{e_{rs}}^2}{\pi_k}\right)^{1/2} \ . \tag{3.7}$$

만일 모집단 크기 N이 기지이고, 표본 s가 고정이라면, 이는 1단계에서 단순임의 추출을 이용하고, 2단계에서 층화추출을 적용한 층화에 대한 이중추출의 경우로 축소될 수 있을 것이다.

3.3 비용을 고려한 최적 포함확률

다음으로 비용을 고려한 경우 무응답 상황하에서의 최적포함확률을 구하기 위해 다음과 같은 비용함수를 고려하자.

$$C_1 = c_0 + \sum_N^N \pi_k c_1 \ , \tag{3.8a}$$

$$C_2 = c_0 + \sum_N^N \pi_k c_1 + \sum_{r_h} \pi_k \pi_{hs,\,m} c_{2h} = C_1 + \sum_{r_h} \pi_k \pi_{hs,\,m} c_{2h} \ . \tag{3.8b}$$

여기서 C_1을 표본을 추출하는데 소요되는 총 비용이라 하고, C_2를 표본을 추출하여

응답을 얻는데 소요되는 총 비용이라 하자. 또한 c_1을 각 단위를 추출하는데 소요되는 추출비용이라 하고, c_{2h}를 추출된 표본에 대해 각 층별 응답을 얻는데 소요되는 응답비용이라 하자.

그러면 고정된 비용함수 하에서 분산을 최소로 하는 각각의 포함확률 중에서 π_k는 식(3.6)와 같고, $\pi_{Hs,m}$는 다음과 같다.

$$\pi_{Hs,m} = \frac{n\sqrt{c_1}}{\sum_U \sigma_k}\left(\frac{n_h S_{e_{r_h}}^2}{c_{2h}\pi_k}\right)^{1/2} . \tag{3.9}$$

따라서 비용함수를 최적의 포함확률을 분산식에 대입하면 다음과 같은 최소 분산을 구할 수 있다. 식(3.5)로부터 $V_1(\hat{\tau})$에 π_k를 대입하면,

$$V_{1\min} = \frac{(\sum\sigma_k)^2}{n} - \sum\sigma_k^2 . \tag{3.10a}$$

이고, 또한 π_k를 (3.9)에 대입하고, 표본 s가 고정된 조건하에서 이를 식(2.6b)에 대입하면, 다음과 같이 최소분산을 구할 수 있다.

$$V_{2\min} = n^{1/2}\sum^{H}\left(\frac{c_{2h}n_h\sigma_k}{c_1 S_{e_{r_h}}^2 \sum\sigma_k}\right)^{1/2} - \sum n_h S_{e_{r_h}}^2 . \tag{3.10b}$$

식(3.10a)와 식(3.10b)를 원래의 분산식에 대입함으로서 비용을 고려했을 경우 다음과 같이 최소분산을 구할 수 있다.

$$V_{\min}(\hat{\tau}) = \frac{(\sum\sigma_k)^2}{n} - \sum\sigma_k^2 + n^{1/2}\sum^{H}\left(\frac{c_{2h}n_h\sigma_k}{c_1 S_{e_{r_h}}^2 \sum\sigma_k}\right)^{1/2} - \sum n_h S_{e_{r_h}}^2 . \tag{3.11}$$

식(3.11)에서 각각의 추정량에 대한 분산항을 대입하면, 이들에 대한 최소 분산을 구할 수 있다.

4. 모의실험 및 결론

4.1 모의실험

이 절에서는 모의실험을 통해 고정된 비용함수 하에서 최소 분산을 가지는 1차와 2

차 포함확률을 무응답이 존재하는 경우 회귀모형을 이용하여 도출하고자 한다.

우선 가상적인 유한모집단의 관심변수와 보조변수를 발생시키고, 이들 자료를 보조변수의 크기에 따라 층화한 후 층화된 모집단으로부터 고정된 비용 하에서 일정 수만큼의 1단계 표본을 추출한 후 추출된 1단계 표본에 대해 고정된 비용 하에서 1단계에 대한 최적의 포함확률을 결정한다.

다음으로 추출된 1단계 표본으로부터 2단계 표본을 고정된 비용함수 하에서 분산을 최소로 하도록 표본을 추출하고, 이때 2단계 포함확률이 바로 구하고자 하는 최적의 응답확률이 된다. 모집단내의 관심변수는 다음의 모형으로부터 난수를 생성하였다.

$$y_i = \theta + \varepsilon x_i \ .$$

여기서, $\theta = 2$라 하였으며, 변수 x_i는 [0,1]의 값을 가지며, ε는 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 하자.

이 모형으로부터 모집단의 각 개체의 응답확률은 [0,1]사이의 값을 갖도록 하였고, 1단계와 2단계 조사비용은 포함확률과 응답확률에 대해 각각 반비례하도록 하였다.

모집단 단위의 수를 $N=1,200$로 하여, 보조변수 x의 값이 $0<x\leq 0.2$, $0.2<x\leq 0.4$, $0.4<x\leq 0.6$, $0.6<x\leq 0.8$ 에 따라 각각 4개의 층으로 구분하였다. 각 층별 크기는 각각 300 단위로 하여 응답확률에 따라 표본을 추출하였으며, 고정 비용은 1000 단위로 하였고, 1단계 조사비용은 200단위로 고정시키고, 4개의 층에 따른 2단계 조사비용의 변화에 따라 2단계 포함확률과 분산의 추정량의 변화를 30회 반복하여 살펴보았다.

<표 4.1>에서 2단계 응답 확률은 비용을 고려한 경우에 대해 [경우1]은 층별 비용을 1층은 290단위 2층은 450단위 3층은 400, 4층은 290으로 하였으며, [경우2]는 1층 700단위 2층은 600단위 3층은 450, 4층은 350으로, [경우3]에 대해서는 1층 900단위 2층은 750단위 3층은 500, 4층은 400단위인 경우에 층별 평균응답률을 나타낸 것이다.

< 표 4.1 > 무응답 하에서의 비용을 고려하지 않은 경우와 비용을 고려한 경우의 총합추정량 (단, 관심변수의 모집단 총합: 1197.69)

비 용 / 구 분	비용을 고려하지 않은 경우	비용을 고려한 경우		
		경우1	경우2	경우3
2단계 응답확률	0.253	0.087	0.079	0.071
총합 추정량	1024.54	1024.97	1025.36	1025.63
표준 편차	86.297	45.660	47.338	48.819
최적 표본수	48	18	16	15

비용을 고려하지 않은 경우에 응답확률은 0.253으로 비용을 고려했을 때에는 0.087, 0.079, 0.071로 나타났으며, 이 값을 추정량의 식에 적용한 결과 각각 1024.54 와 1024.97, 1025.36, 1025.63으로 추정량의 변화에는 별 차이를 보이지 않으나, 표준편차를 살펴보면 거의 2배의 차이가 있는 것으로 나타났다. 이러한 결과는 본 논문에서 제안한 결과와 일치하는 것으로 고정된 비용 하에서 분산을 최소로하는 추정량을 임을 보여주는 것이다.

또한 거의 동일한 추정량의 값을 얻는데 필요한 최적 표본수는 비용을 고려하지 않은 경우에 48단위에서 비용을 고려한 경우는 18, 16, 15 단위로 급격히 줄어드는 것을 볼 수 있다.

4.2 결론

본 논문에서는 무응답이 존재하는 상황에서 비용함수를 고려한 최적의 추정량에 대해 살펴보았다. 초 모집단 모형의 접근으로 Godambe과 Thompson(1986)이 제안한 최적추정량 식을 이용하여 무응답이 존재하는 상황에서 무응답 편의를 줄이는 최적의 포함확률을 구했으며, 이때 1단계 포함확률과 응답확률은 다음과 같다.

$$\pi_k = \frac{n\sigma_k}{\sum_U \sigma_k},$$

$$f_h = \pi_{hs,m} = \frac{m\sqrt{c_1}}{\sum_U \sigma_k}\left(\frac{n_h S_{e_r}^2}{c_{2h}\pi_k}\right)^{1/2}.$$

이 값을 2장에서 언급한 각각의 추정량에 대입하여 최적의 추정량을 얻을 수 있게 된다. 위의 식을 추정량의 분산에 대입함으로서 추가적으로 무응답에 기인한 편의를 최소로 하게 된다. 이때의 최소 분산은 식(3.11)과 같다. 분산 식으로부터 2절에서 언급한 추정량에 따라 각각 다른 분산을 얻게 된다.

모의 실험으로부터 본 연구에서는 다음과 같은 결론을 얻을 수 있었다.

첫째, 1차 조사비용이 고정되고, 무응답으로 인한 2차 조사비용을 고려했을 때 분산을 최소로 하는 2단계 포함확률은 기존의 2단계 포함확률에 비해 상당히 작아짐을 알 수 있었다.

둘째, 1단계 포함확률과 2단계 포함확률이 최적인 경우 분산을 최소로 한다는 것이다.

셋째, 최적의 2단계 포함확률은 2차 조사단위에 대한 층별 비용의 변화에 민감하지 않음을 알 수 있었다.

넷째, 앞에서도 언급한 사실로서 두 개의 최적 값을 추정량의 식에 대입함으로서 무응답이 존재하는 경우에서 최적의 추정량을 얻을 수 있으며, 이와 같은 사실은 최적 추정량이 무응답에 대해 로버스트 하다는 Sarndal(1986)의 사실과도 일치하는 결과이다.

참고문헌

[1] Cassel, C.M., Sarndal, C.E., and Wretman, J.H.(1983), *Foundation of Inference in Survey Sampling*, New York, Wiley.

[2] Godambe, V.P., Thompson, M.E.(1986), "Some Optimality Results in the Presence of Nonresponse," *Survey Methodology*, Vol. 12, pp. 29-36.

[3] Oh,H.L, Scheuren, F.J.(1983), *Incomplete Data in Sample Surveys*, Vol. 2, (Madow, W.G., Olkin, I. and Rubin, D.B. Eds), New York, Academic Press.

[4] Sarndal, C.E.(1986), "A Regression Approach to Estimation in the Presences of Nonresponse," *Survey Methodology*, Vol. 12, pp. 306-312.

[5] Sarndal, C.E., Swensson, B. and Wretman, J.H.(1992), *Model Assisted Survey Sampling*, New York, Springer-Verlag.