응용통계연구 제 12권 2호, 1999년, pp. 527-535

# 양적 확률화응답을 이용한 회귀추정에 관한 연구*

최경호[1)]

요 약

양적 확률화응답을 이용한 민감사안에 대한 평균이나 분산의 추정시 보조정보를 활용한 회귀추정법에 대해서 언급하고, 유도된 회귀추정량과 Greenberg et al.의 추정량 그리고 비추정량과의 비교를 통하여 회귀추정량이 효율적일 수 있는 조건을 찾았다. 또한 각 질문에 대한 응답의 분포가 포아송 분포인 경우 회귀추정량의 효율이 증대될 수 있는 조건에 대해서도 논하였다.

## 1. 서론

사회조사시 발생되는 문제점 중, 응답자로부터 신뢰할만한 정보를 얻지 못하는 경우를 고려할 수 있다. 이는 조사의 내용이 법적이나 도덕적으로, 즉 사회통념상 인정되기 어려운 민감한 사안(sensitive issue)일 때 주로 발생된다. 그래서 민감한 사안에 대한 조사시 직접 질문을 하게되면 응답거절이나 거짓응답률이 높게 발생되어, 즉 비표본오차의 증대로 인하여 추정의 신뢰도가 떨어진다.

이에대한 해결방안으로, 응답자의 신분보호(privacy protection)를 통하여 신뢰할만한 응답을 얻음으로써, 추정의 신뢰도를 높일 수 있는 간접질문방식인 확률화응답기법(randomized response technique)이 Warner(1965)에 의하여 개발 되었고, Chudhuri와 Mukerjee(1988) 그리고 류제복 등(1993)은 이를 체계적으로 정리하였다. 이후 질적(qualitative) 확률화응답에 대한 개선된 기법이 많이 소개되고 있다. 한편 양적 (quantitative) 확률화응답을 이용하여 관심사안(변수)에 대한 평균과 분산 등의 추정방법에 대해서는 Greenberg et al.(1971)에 의하여 소개되고 있다. 이방법은 질적기법의 한 형태인 무관질문기법을 응용한 것이다.

Greenberg et al.의 방법을 이용함에 있어서의 문제점은 민감사안과는 무관(unrelated)인 질문을 선택해야 함에 있다. 이는 어디까지가 민감사안과 무관인지에 대한 최적의 기준이 없기에 더욱 그러하다. 또한 양적 확률화응답을 이용하여 관심의 대상이 되는 민감사안에 대한 평균이나 분산의 추정시 반드시 무관질문을 이용해야 할 필요도 없다. 오히려 관련이 있는, 그러나 덜 민감하거나 혹은 민감하지 않은 질문을 선택함으로써 이로부터 얻어지는 정보를 보조정보(auxiliary information)로 이용한다면 통계적측면에서 추정의 효율이 증대될 수 있다.

본 논문에서는 양적 확률화응답을 이용하여 민감사안에 대한 평균이나 분산의 추정시 보조정보를 활용한 회귀추정량(regression estimator)을 유도하고, Greenberg et al.의 추정량 그리고 비추정량(ratio estimator)과의 효율비교도 행하고자 한다. 또한 각 질문에 대한 응답의 분포가 포아송분포인 경우 회귀추정량의 효율이 증대될 수 있는 조건에 대해서도 논한다.

* 이 논문은 1999년도 전주대학교 학술연구비 조성에 의하여 수행되었음.

1) (560-759) 전주시 효자동 1200, 전주대학교 정보통계학과, 부교수

## 2. 회귀추정량

회귀추정의 목적은 민감사안($X$)에 대한 평균등의 추정에 있어, 관련사안($Y$)의 정보를 보조정보로 이용하여 추정의 정도(precision)를 높이는데 있다. 이를 위하여 회귀추정에서는 Greenberg et al.의 방법과는 달리 민감질문과 대비되는 질문으로, 관련이 있는 그러나 덜 민감하거나 혹은 민감하지 않은 질문을 이용한다. 예를들어 낙태횟수에 관한 조사시 Greenberg et al.의 방법에서는 확률장치를 통하여 다음과 같이 구성된 질문중 하나에 응답을 하게 한다.

$Q_1$ : 지금까지 귀하의 낙태횟수는 몇번입니까?

$Q_2$ : 생계를 위하여 여성이 직업을 갖기 위해서는 몇 명의 자녀가 적당하다고 생각하십니까?

위에서 $Q_2$는 무관질문이다. 이에 반하여 회귀추정에서는 $Q_2$를 다음의 질문2와 같이 관련질문으로 그러나 민감하지 않은 질문으로 구성한다.

질문1 : 지금까지 귀하의 낙태횟수는 몇번입니까? ($X$)

질문2 : 귀하는 결혼이후 몇 명의 자녀를 출산하셨습니까? ($Y$)

이렇게 구성된 관련질문을 이용하여 민감사안에 대한 평균 $\mu_X$에 대한 회귀추정량을 구하기 위하여 단순임의복원으로 추출된 크기가 각각 $n_1, n_2$인 두 개의 독립표본에 속하는 응답자들에게 확률 $p_i (i = 1, 2)$로 질문1에 그리고 확률 $q_i$ ($q_i = 1 - p_i$)로 질문2에 응답하도록 한다. 이 때 $i (i = 1, 2)$번째 표본내의 응답자로부터 얻어지는 응답을 $Z_1, Z_2, \cdots, Z_{n_i}$라면 $\bar{Z}_i$는 이들의 평균으로 $\bar{Z}_i = \sum_{j=1}^{n_i} Z_j / n_i$이다. 따라서,

$$E(\bar{Z}_i) = p_i \mu_X + q_i \mu_Y \tag{2.1}$$

이며, $p_1 \neq p_2$이면 $\hat{\mu}_X$와 $\hat{\mu}_Y$은 각각 다음과 같다.

$$\hat{\mu}_X = \frac{[(1-p_2)\bar{Z}_1 - (1-p_1)\bar{Z}_2]}{(p_1 - p_2)} \tag{2.2}$$

$$\hat{\mu}_Y = \frac{[p_2 \bar{Z}_1 - p_1 \bar{Z}_2]}{(p_2 - p_1)} \tag{2.3}$$

한편, 식 (2.2)와 (2.3)의 $\hat{\mu}_X$와 $\hat{\mu}_Y$은 식 (2.1)을 이용하면 각각 $\mu_X$와 $\mu_Y$의 불편추정량임을 쉽게 알 수 있다.

이상을 토대로 민감사안에 대한 평균 $\mu_X$의 회귀추정량을 고려하면 이는 다음과 같다.

$$\hat{\mu}_{reg} = \hat{\mu}_X + b(\mu_Y - \hat{\mu}_Y) \tag{2.4}$$

### 2.1. 회귀추정량의 성질

식 (2.4)의 회귀추정량에서 $b$는 민감변수(사안) $X$와 관련변수 $Y$간의 회귀계수의 추정량으로, 먼저 $b$가 $b_0$로 미리 정해진 경우 식 (2.4)의 회귀추정량은 식 (2.2)와 (2.3)의 $\hat{\mu}_X$와

$\hat{\mu}_Y$이 각각 $\mu_X$와 $\mu_Y$의 불편추정량임을 이용하면 관심의 대상이 되는 민감변수의 모평균인 $\mu_X$의 불편추정량이 됨을 쉽게 알 수 있다.

$$\begin{aligned} E(\hat{\mu}_{reg}) &= E(\hat{\mu}_X) + b_0 E(\mu_Y - \hat{\mu}_Y) \\ &= \mu_X \end{aligned} \tag{2.5}$$

나아가 회귀추정량 $\hat{\mu}_{reg}$의 분산은 $p_1 \neq p_2$에 대하여 다음과 같다.

$$\begin{aligned} V(\hat{\mu}_{reg}) &= V(\hat{\mu}_X) + b_0^2 V(\hat{\mu}_Y) - 2b_0 Cov(\hat{\mu}_X, \hat{\mu}_Y) \\ &= V(\hat{\mu}_X) + b_0^2 V(\hat{\mu}_Y) - 2b_0 \rho \sqrt{V(\hat{\mu}_X) V(\hat{\mu}_Y)} \end{aligned} \tag{2.6}$$

여기서, $\rho$는 민감변수 $X$와 관련변수 $Y$간의 모상관계수이며

$$V(\hat{\mu}_X) = \frac{[(1-p_2)^2 V(\bar{Z}_1) + (1-p_1)^2 V(\bar{Z}_2)]}{(p_1 - p_2)^2} \tag{2.7}$$

$$V(\hat{\mu}_Y) = \frac{[p_2^2 V(\bar{Z}_1) + p_1^2 V(\bar{Z}_2)]}{(p_2 - p_1)^2} \tag{2.8}$$

$$V(\bar{Z}_i) = \frac{1}{n_i}[\sigma_X^2 + p_i(\sigma_X^2 - \sigma_Y^2) + p_i(1-p_i)(\mu_X - \mu_Y)^2] \tag{2.9}$$

이다. 식 (2.6)으로부터 알 수 있는 사실은 다음과 같다.

첫째, 회귀추정량의 분산 $V(\hat{\mu}_{reg})$은 전체 표본크기 $n$ $(n = n_1 + n_2)$이 증가할수록 감소한다.

둘째, 회귀추정량의 분산 $V(\hat{\mu}_{reg})$은 $\rho = \frac{b_0}{2}\sqrt{\frac{V(\hat{\mu}_Y)}{V(\hat{\mu}_X)}}$일 때 Greenberg et al.방법의 분산과 같아진다.

셋째, 회귀추정량의 분산 $V(\hat{\mu}_{reg})$은 $\rho$가 1에 가까울수록 감소한다.

넷째, 회귀추정량의 분산 $V(\hat{\mu}_{reg})$은 $\mu_X = \mu_Y$, $\sigma_X^2 = \sigma_Y^2$ 그리고 $p_1 \neq p_2 \neq 1/2$이면 감소한다.

한편 회귀추정량 식 (2.4)에서 $b$가 $b_0$로 미리 정해지는 경우 최적의 $b_0$는 식 (2.6)을 $b_0$에 관하여 미분함으로써 구할 수 있다.

$$b_0 = B = \rho\sqrt{\frac{V(\hat{\mu}_X)}{V(\hat{\mu}_Y)}} \tag{2.10}$$

따라서 식 (2.10)을 이용하면 회귀추정량의 최소분산은 다음과 같다.

$$\begin{aligned} V_{min}(\hat{\mu}_{reg}) &= V(\hat{\mu}_X) + \rho^2 \frac{V(\hat{\mu}_X)}{V(\hat{\mu}_Y)} V(\hat{\mu}_Y) - 2\rho^2 V(\hat{\mu}_X) \\ &= V(\hat{\mu}_X)(1-\rho^2) \end{aligned} \tag{2.11}$$

실질적인 측면에서 볼 때 계속조사를 수행할 때에는 식 (2.4)에서의 $b$값을 이전조사의 자료를 이용하여 정할 수 있다. 이에 반하여 $b$값을 미리 정할 수 없는 경우라면 식 (2.4)의 회귀추정량은 더 이상 불편추정량이 되지 못한다. 그러나 통계적인 관점에서 전체표본의 크기가 증가하거나 $\rho$가 $\pm 1$에 접근하면 편의가 감소하여 이 경우에도 식 (2.4)의 회귀추정량은 불편추정량으로 간주할 수 있으며, 역시 식 (2.11)을 이의 최소분산으로 이용할 수 있다.

### 2.2. 층화추출로의 확장

사용상의 복잡성 때문에 실용적인 측면은 적지만, 크기 $N$인 모집단이 $L$개의 층으로 형성되고 층의 크기가 $N_h$ ($\sum_{h=1}^{L} N_h = N$)인 $h$번째 층에서 크기 $n_h(n_h = n_{h1} + n_{h2})$의 독립표본을 단순임의 복원추출한 경우 $b$가 일정할 때 결합회귀추정량(combined regression estimator)은 다음과 같다.

$$\hat{\mu}_{reg/c} = \hat{\mu}_{st/X} + b(\mu_Y - \mu_{st/Y}) \tag{2.12}$$

여기서, $\hat{\mu}_{st/X} = \sum_{h=1}^{L} W_h \hat{\mu}_{Xh}$, $\hat{\mu}_{st/Y} = \sum_{h=1}^{L} W_h \hat{\mu}_{Yh}$, $W_h = N_h/N$이며, $\hat{\mu}_{Xh}$와 $\hat{\mu}_{Yh}$은 식 (2.2)와 (2.3)을 $h$번째 층에 응용한 결과이다.

나아가 식 (2.12)의 분산과 최적 $b$는 다음과 같다.

$$V(\hat{\mu}_{reg/c}) = \sum_{h=1}^{L} {W_h}^2 V(\hat{\mu}_{Xh}) + b^2 V(\hat{\mu}_{Yh}) - 2b\rho_h \sqrt{V(\hat{\mu}_{Xh}) V(\hat{\mu}_{Yh})} \tag{2.13}$$

$$b_{opt} = \frac{\sum_{h=0}^{L} {W_h}^2 Cov(\hat{\mu}_{Xh}, \hat{\mu}_{Yh})}{\sum_{h=0}^{L} {W_h}^2 V(\hat{\mu}_{Xh})} \tag{2.14}$$

단, $\rho_h$는 $h$번째 층에서의 민감변수($X_h$)와 관련변수간($Y_h$)간의 상관계수이다.

## 3. 효율비교

표본의 크기가 충분히 크다고 가정할 때, 본 논문에서 제시한 회귀추정량과 Abul-Ela et al.(1985)의 비추정량(ratio estimator), 그리고 Greenberg et al.의 추정량과의 효율비교를 행해보자. 먼저 이들 추정량들의 분산은 다음과 같다.

$$\text{회귀추정} V(\hat{\mu}_{reg}) = V(\hat{\mu}_X)(1 - \rho^2) \tag{3.1}$$

$$\text{비추정} V(\hat{\mu}_{ratio}) = V(\hat{\mu}_X) + R^2 V(\hat{\mu}_Y) - 2R\rho\sqrt{V(\hat{\mu}_X) V(\hat{\mu}_Y)}\ \text{, 단, } R = \mu_X/\mu_Y \tag{3.2}$$

$$\text{Greenberg et al.} V(\hat{\mu}_G) = V(\hat{\mu}_X) \tag{3.3}$$

식 (3.1)과 (3.2)로부터 회귀추정량이 비추정량보다 효율적인 조건은 다음과 같다.

$$-\rho^2 V(\hat{\mu}_X) < R^2 V(\hat{\mu}_Y) - 2R\rho\sqrt{V(\hat{\mu}_X) V(\hat{\mu}_Y)}$$

따라서,

$$\left(R\sqrt{V(\hat{\mu}_Y)} - \rho\sqrt{V(\hat{\mu}_X)}\right)^2 > 0$$

인데 $\rho = B\sqrt{\frac{V(\hat{\mu}_Y)}{V(\hat{\mu}_X}}$이므로 $B = \frac{Cov(\hat{\mu}_X, \hat{\mu}_Y)}{V(\hat{\mu}_Y)}$에 대하여

$$(R - B)^2 > 0 \tag{3.4}$$

이다.

식 (3.4)로부터 $R = B$, 즉 민감변수 $X$에 대한 관련변수 $Y$의 회귀직선이 원점을 통과하지 않는 한 회귀추정량이 비추정량보다 항시 효율적임을 알 수 있는 바, 이는 조사자료가 확률화응답인 경우에도 일반적인 표본조사에서 사용되는 직접질문에 의한 경우에서와 동일한 결과임을 알 수 있다.

한편 회귀추정량과 Greenberg et al.의 추정량의 효율을 비교해 보면 $\rho = 0$이 아닌 이상 회귀추정량이 항상 효율적이다. 즉, 민감사안인 질문1에 대비되는 질문2가 무관질문이 아닌 경우에는 회귀추정량의 사용이 바람직하다. 실질적인 측면에서 무관질문의 선택이 용이하지 아니한 점을 고려한다면 관련질문이지만 덜 민감하거나 혹은 민감하지 않은 질문을 질문2로 선택하고 회귀추정량을 이용하는 것이 바람직 하다고 할 수 있다.

## 4. 예제

식 (2.11)에 주어진 회귀추정량의 분산에 대한 성질을 Greenberg et al.에 제시된 자료를 이용하여 알아보자. 그 자료에 의하면 인종별로 여성의 낙태횟수에 대한 평균과 분산은 다음과 같다.

표 4.1: 1968년 미국 North Caroline 낙태조사에서 인종별 낙태횟수에 대한 평균과 분산

| 인 종 | $\hat{\mu}_X$ | $\sqrt{V(\hat{\mu}_X)}$ |
|---|---|---|
| 백 인 | 0.415 | 0.107 |
| 흑 인 | 0.645 | 0.177 |

이로부터 백인여성의 경우 $\rho$의 변화에 따른 회귀추정량의 분산은 다음과 같다.

| $\rho$ | -1 | -0.5 | 0 | 0.5 | 1 |
|---|---|---|---|---|---|
| $V_{min}(\hat{\mu}_{reg})$ | 0 | -0.009 | 0.011 | 0.009 | 0 |

앞의 이론적 고찰에서와 동일하게 $|\rho|$가 1에 가까울수록 회귀추정량의 효율이 증가하며, $V_{min}(\hat{\mu}_{reg})$은 $\rho = 0$에 대하여 대칭임을 알 수 있다.

한편, 본 논문에서 제시된 질문1과 질문2의 분포를 구체적인 분포-예를들어 포아송분포-로 가정하고 회귀추정량의 최소분산인 $V_{min}(\hat{\mu}_{reg})$의 최적조건을 수치적으로 알아보자.

응답을 얻는 과정을 Greenberg et al.의 방법을 응용하면, 서로 독립인 표본 1과 표본 2에 대한 확률장치를 통하여 얻어지는 각 응답자의 응답에 대한 확률함수는 다음과 같다.

$$\text{표본1} \ : \ h_1(Z_1) = p_1 f(Z_1) + (1-p_1) g(Z_1) \tag{4.1}$$

$$\text{표본2} \ : \ h_2(Z_2) = p_2 f(Z_2) + (1-p_2) g(Z_2) \tag{4.2}$$

단, $f(Z)$와 $g(Z)$는 각각 민감질문(질문1)과 관련질문이나 덜 민감하거나 혹은 민감하지 않은 질문(질문2)에 대한 확률함수이다. 질문1과 질문2의 응답분포가 포아송분포를 한다고 가정하면

$$f(Z) = \frac{e^{-\mu_X}\mu_X^Z}{Z!}, \ Z = 0, 1, 2, \cdots \tag{4.3}$$

$$g(Z) = \frac{e^{-\mu_Y}\mu_Y^Z}{Z!}, \ Z = 0, 1, 2, \cdots \tag{4.4}$$

이다. 그러면 포아송분포의 성질로부터 $\mu_X = \sigma_X^2$, $\mu_Y = \sigma_Y^2$ 이다. 이로부터 $i(i = 1, 2)$번째 표본에서의 응답평균에 대한 분산은 다음과 같다.

$$V(\bar{Z}_i) = \frac{1}{n_i}[\mu_Y + p_i(\mu_X - \mu_Y) + p_i(1-p_i)(\mu_X - \mu_Y)^2] \tag{4.5}$$

한편 식 (2.11)의 $V_{min}(\hat{\mu}_{reg})$에서 $V(\hat{\mu}_X)$은 식 (2.7)인데 이를 구하기 위한 $V(\bar{Z}_i)$는 식 (4.5)를 이용하면 된다.

이제 모수 $\mu_X$, $\mu_Y$, $p_1$, $p_2$ 그리고 $\rho$의 변화에 따른 $V_{min}(\hat{\mu}_{reg})$의 최적조건을 수치적으로 찾기위해 $n_1 = n_2 = 1000$, $p_1 = 1 - p_2 = 0.6, 0.9$, $\mu_X = 1, 2, 3, 4$, $\mu_Y = 2, 3, 5$ 그리고 $\rho = 0.3, 0.5, 0.9$일 때의 회귀추정량의 분산을 구해보자.

표 4.2로부터 알 수 있는 $V_{min}(\hat{\mu}_{reg})$의 최적조건, 즉 최적모수의 선택은 다음과 같다.

첫째, $\mu_X$, $\mu_Y$ 그리고 $\rho$의 고정된 값에 대해서 $V_{min}(\hat{\mu}_{reg})$는 $p_1$이 0.6에서 0.9로 증가함에 따라 감소한다.

둘째, 고정된 $\mu_X$, $\mu_Y$ 그리고 $p_1$에 대해서는 $\rho$가 0.3에서 0.9로 증가할수록 $V_{min}(\hat{\mu}_{reg})$은 감소한다.

셋째, 고정된 $\rho$에 대하여 $|\mu_X - \mu_Y|$가 증가할수록 $V_{min}(\hat{\mu}_{reg})$은 증가한다.

이는 앞에서 살펴본 이론적 결과와 동일한 결과이며, 나아가 표 4.2를 근거로 이상의 내용을 종합해 볼 때 양적 확률화응답을 이용한 회귀추정량의 정도(precision)를 높이기 위 한 방안으로 다음 사항을 고려하는 것이 바람직 하겠다.

1. 관련변수 $Y$는 가급적 민감변수 $X$와 상관의 정도가 높은 사안을 선택하도록 한다.

2. $p_1 + p_2 = 1$ ($p_1! = p_2! = 0.5$)이 되는 범위 내에서 $p_1$ 또는 $p_2$를 0.5에서 벗어나도록 선택한다. $p_1$이나 $p_2$가 0.5에 가까운 경우에는 응답자의 의심을 완화시킬 수 있는 방안이 요구된다.

3. 민감변수 $X$와 관련변수 $Y$의 측도(예, Kg, cm, 명 등)는 동일하게 한다.

이상에서 1번 항목을 제외하고는 Greenberg et al.의 권고사항과 같다.

표 4.2: 모수의 변화에 따른 $V_{min}(\hat{\mu}_{reg})$

| $\mu_X$ | $\rho$ | $\mu_Y$ | $p_1 = 0.6$ | $p_1 = 0.9$ |
|---|---|---|---|---|
| 1 | 0.3 | 2 | 0.02013 | 0.00140 |
| | | 3 | 0.03411 | 0.00184 |
| | | 5 | 0.07910 | 0.00366 |
| | | 7 | 0.14680 | 0.00571 |
| | 0.5 | 2 | 0.01659 | 0.00115 |
| | | 3 | 0.02811 | 0.00152 |
| | | 5 | 0.06519 | 0.00277 |
| | | 7 | 0.12099 | 0.00471 |
| | 0.9 | 2 | 0.00420 | 0.00029 |
| | | 3 | 0.00712 | 0.00038 |
| | | 5 | 0.01651 | 0.00070 |
| | | 7 | 0.03065 | 0.00119 |
| 2 | 0.3 | 2 | 0.02366 | 0.00233 |
| | | 3 | 0.03196 | 0.00256 |
| | | 5 | 0.06559 | 0.00366 |
| | | 7 | 0.12194 | 0.00560 |
| | 0.5 | 2 | 0.01950 | 0.00192 |
| | | 3 | 0.02634 | 0.00211 |
| | | 5 | 0.05406 | 0.00302 |
| | | 7 | 0.10050 | 0.00461 |
| | 0.9 | 2 | 0.00494 | 0.00049 |
| | | 3 | 0.00667 | 0.00054 |
| | | 5 | 0.01370 | 0.00076 |
| | | 7 | 0.02546 | 0.00117 |
| 3 | 0.3 | 2 | 0.03287 | 0.00347 |
| | | 3 | 0.03549 | 0.00350 |
| | | 5 | 0.05777 | 0.00417 |
| | | 7 | 0.10276 | 0.00569 |
| | 0.5 | 2 | 0.02709 | 0.00286 |
| | | 3 | 0.02925 | 0.00288 |
| | | 5 | 0.04761 | 0.00344 |
| | | 7 | 0.08469 | 0.00469 |
| | 0.9 | 2 | 0.00686 | 0.00073 |
| | | 3 | 0.00741 | 0.00073 |
| | | 5 | 0.01206 | 0.00087 |
| | | 7 | 0.02145 | 0.00119 |
| 4 | 0.3 | 2 | 0.04776 | 0.00483 |
| | | 3 | 0.04470 | 0.00464 |
| | | 5 | 0.05562 | 0.00490 |
| | | 7 | 0.0825 | 0.00599 |
| | 0.5 | 2 | 0.03936 | 0.00398 |
| | | 3 | 0.03684 | 0.00382 |
| | | 5 | 0.04584 | 0.00404 |
| | | 7 | 0.07356 | 0.00494 |
| | 0.9 | 2 | 0.00997 | 0.00101 |
| | | 3 | 0.00933 | 0.00097 |
| | | 5 | 0.01161 | 0.00102 |
| | | 7 | 0.01864 | 0.00125 |

## 5. 결론

민감사안에 대한 조사시 직접질문에 의한 조사는 응답자의 거짓응답이나 응답거절로 인하여 신뢰할만한 응답을 얻기가 어려운 경우가 종종있다. 이에대한 해결방안으로 확률장치를 통하여 응답자의 신분을 보호하여 줌으로써, 조사의 신뢰도를 높일 수 있는 간접조사방법으로 확률화응답기법이 있다.

본 논문에서는 양적 확률화응답을 이용하여 민감사안에 대한 평균이나 분산의 추정시 보조정보를 활용한 회귀추정법에 대해서 언급하였다. 그리고 이에따른 회귀추정량과 이의 최적분산을 유도하였다. 또한 유도된 $V_{min}(\hat{\mu}_{reg})$를 비추정량 그리고 Greenberg et al.의 추정량의 분산과 비교하여 본 논문에서 제시된 회귀추정량이 효율적일 수 있는 조건을 이론적으로 살펴봄과 동시에, 각 질문에 대한 응답분포가 포아송분포인 경우 수치적 최적해에 대해서도 알아보았다.

본 논문에서 고찰된 사항을 종합하여 볼 때, 양적 확률화응답기법을 이용한 조사수행시 무관질문의 선택에 어려움이 있는 Greenberg et al.의 방법보다는 본 논문에서 제시한 관련질문 그러나 덜 민감하거나 혹은 민감하지 않은 질문을 사용하여 회귀추정량을 구함으로써 추정의 정도도 높일 수 있고 Greenberg et al.의 방법이 갖는 문제점도 해소 될 수 있음을 알 수 있었다.

## 참고문헌

[1] 류제복, 홍기학, 이기성(1993). 〈확률화응답모형〉, 자유아카데미, 서울.

[2] Chaudhuri, A., and mukerjee, R.(1988). *Randomized Response - Theory and Technique*, Marcel Dekker, Inc., New York.

[3] Abul-Ela, A.L., and Abdel-Hamied, S.M.(1985). A randomized response ratio estimate from quantitative data. *Proceedings of the Social Statistics Section, American Statistical Association.* 300-305.

[4] Greenberg, B.G., Kuebler, R., Abernathy, J.R., and Horvitz, D.G.(1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association.* Vol. 66, 243-250.

[5] Warner, S.L.(1965). Randomized response : A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association.* Vol. 60, 63-69.

[ 1998년 12월 접수, 1999년 4월 최종수정 ]

# A Study on Randomized Response Regression Estimate from Quantitative Data*

Kyung-Ho Choi[1)]

ABSTRACT

In this paper, a new estimate of the mean and variance of the sensitive characteristic based on the regression procedure is proposed. A comparison between the regression estimate and Greenberg's estimate, ratio estimate is introduced. The properties of the regression estimate are studied. Conditions underwhich the efficiency of the regression estimate, assuming the Poission distribution, may be increased are explored.

* This research was supported by Jeonju University Research Fund, 1999.

1) Department of information statistics, Jeonju University, Wansan-Gu, 560-759, Korea.