표본조사의 가중치 작업에 대한 고찰

- Composite calibration estimator의 소개 -

김 재 광*

1. 서론

표본조사에서 사용되는 대부분의 추정량은 표본 관측치의 가중 합으로 종종 표현되는데 이렇게 관측치의 일차 가중합으로 추정량을 구현하는 방법은 하나의 가중치가 여러 개의 항목에 공통적으로 쓰이게 됨으로써 다목적 조사(multi-purpose survey)의 추론에 편리하다. 또한, 예를 들어 총수입과 같은 항목은 여러 가지 세부 수입 항목들의 합으로 표현되는데 이렇게 세부 수입 항목치에 대한 통계치 합이 총수입 항목의 통계치와 같아 짐으로써 일관성있는 통계치가 구현되기 위해서는 그 통계 추정을 일차 가중합으로 구현해야 할 것이다. 이러한 성질을 통계량의 내적 일관성(internal consistency)이라고도 하는데 이는 누구나 같은 결과를 얻을수 있다는 점에서 특히 자료가 일반에게 공개되는 주요 국가 통계에서 반드시 갖추어야할 중요 성질이라 할 수 있을 것이다.

가중치 작업은 표본 추출을 통해서 얻어지는 표본에 대하여 적절한 가중치를 부여해 줌으로써 추정량의 효율성과 신뢰성을 제고하고자 하는 방법으로 표본 조사에서의 필수적인 요소이다. 가중치 작업은 표본 추출 확률을 이용한 설계 가중치 (design weight)외에도 외부 보조 변수와의 일치성을 위해 조정해 주는 calibration, 단위 무응답 자료의 처리를 위한 무응답 가중치 조정, 그리고 지나치게 가중치 이상값(outlier)를 판별하여 이를 처리하는 가중치 이상값 처리, 이러한 요소들로 구성되어 있다. 이를 정리하면 다음과 같다.

설계 가중치 작성 -> 무응답 가중치 조정 - > calibration -> 가중치 이상값 처리

설계 가중치는 일차 표본 포함 확률(first-order inclusion probability)의 역수를 사용하여 계

*연세대학교 응용통계학과

산되는데 이 설계 가중치를 사용한 추정량은 Horvitz-Thompson 추정량(이하 HT 추정량)이 된다. 무응답 가중치 조정은 응답 자료만을 사용할 때 발생할수 있는 무응답 오차(nonresponse error)를 줄여주기 위해서 가중치를 응답 확률을 추정하여 그 역수로 사용한다. 무응답 가중치 조정에 관하여는 Sarndal and Lundstrom (2004)을 참고하기 바란다. calibration 은 보조변수에 대한 모집단 값을 알 때 그 보조변수의 모수(예를 들면 모평균)에 대해 일치하는 통계량을 구현하도록 최종 가중치를 결정하는 방법론이다. 즉, 보조변수를 x 라고 하고 표본원소 집합을 S, 모집단 원소 집합을 U 라고 하면

$$\sum_{i \in S} w_i x_i = \sum_{i \in U} x_i \quad \cdots\cdots(1)$$

을 만족하도록 최종 가중치(w_i)를 결정해 준다.

본 연구에서는 보조변수 x 의 모집단 총계 $X = \sum_{i \in U} x_i$ 를 알지 못하지만 다른 조사 등을 통하여 그 추정량 $\widehat{X}$ 을 알 경우 어떻게 이를 추정에 반영할 수 있을 것인가를 다루고자 한다. 이를 위하여 2절에서는 모집단 총계를 아는 경우의 calibration 방법론에 대하여 reivew 하고 3절에서는 모집단 총계의 추정량 $\widehat{X}$ 을 사용한 calibration 방법론에 대해 기존의 연구들을 소개한다. 4 절에서는 이 경우 새로운 calibration 방법론을 소개하고 그 성질을 다룬다.

2. 전통적인 calibration 추정량

이해를 돕기 위하여 표본 자료가 단순 임의 추출(SRS)로 추출되었다고 가정하자. 관심 변수를 y_i 라 하고 보조 변수를 x_i 라 하며 관심 변수는 표본에서만 관측 가능하지만 보조 변수는 모집단 전체에 대하여 관측 가능하다고 할때 보조 변수의 모평균 $\overline{X_N} = N^{-1} \sum_{i=1}^{N} x_i$ 은 표본 외에서 얻어지는 보조 정보로써 y의 모평균 추정에 이용되어질 수 있는데 회귀 추정량은 calibration 추정량의 대표적인 형태로 다음과 같이 표현될 수 있다.

$$\widehat{\theta_{reg}} = \overline{y_n} + (\overline{X_N} - \overline{x_n})b \quad \cdots\cdots(2)$$

이 때 $b = \frac{\sum_{i \in S}(x_i - \overline{x})y_i}{\sum_{i \in S}(x_i - \overline{x})^2}$ 으로 계산된다. 이 회귀 추정량은 최종 가중치를

$$w_i = \frac{1}{n} + (\overline{x}_N - \overline{x}_n)\frac{(x_i - \overline{x}_n)}{\sum_{i \in S}(x_i - \overline{x}_n)^2} \quad \cdots\cdots (3)$$

으로 사용하는 선형 통계량으로 구현된다. (3)의 최종가중치는

$$\sum_{i \in S} w_i(1, x_i) = (1, \overline{x}_N) \quad \cdots\cdots (4)$$

을 만족시킴을 보일수 있고 실제로 (3)의 최종가중치는 (4)를 만족시키는 모든 가중치 중에서

$$\sum_{i \in S}\left(w_i - \frac{1}{n}\right)^2 \quad \cdots\cdots (5)$$

를 최소화 하는 최적해로 표현된다. Deville and Sarndal (1992) 은 여러 가지 형태의 목적함수를 사용한 calibration 추정량 구현 방법에 대해 소개하였다.

3. composite calibration 추정량

다음으로는 보조변수 x 의 모집단 평균을 알지 못하지만 다른 조사 등을 통하여 그 추정량을 알 경우 이를 추정에 어떻게 반영할 것인가 하는 방법에 대하여 다루고자 한다. Zieschang (1990)은 미국 노동 통계국(Bureau of Labor Statistics)에서 실시하는 Consumer Expenditure Survey (CES)의 가중치 작업에서 이러한 문제를 다루었다. CES 에서는 똑같은 모집단에 대하여 두 개의 표본 집단을 독립적으로 추출하고 각각의 표본에 다른 조사를 실시하였는데 하나는 가계부를 기입하는 방식으로 조사하였고 다른 하나는 면접 조사를 실시하였다. 이러한 경우 같은 항목이 두 조사에 걸쳐서 나타나는 경우 각각 다른 표본에서 다른 결과가 나타날 수 있을 것이다. 이러한 경우 어떻게 두 개의 다른 통계량을 잘 결합하여 일치된 결과를 얻어낼수 있을 것인가는 현실적으로 매우 중요한 문제일 것이다.

이를 설명하기 위하여 동일한 모집단에서 단순 임의 추출로 두 개의 독립적인 표본 S_1과 S_2를 추출하였다고 하자. 또한 S_1에서는 항목 x 와 z를 관측하고 S_2에서는 항목 x와 y를 관측한다고 하자. 이러한 경우 S_1에서 얻어진 $\overline{X}_N$ 의 불편 추정량을 $\overline{x}_1$이라 하고 S_2에서 얻어진 $\overline{X}_N$의 불편 추정량을 $\overline{x}_2$ 이라 하고 우리의 관심 모수는 y 의 모집단 평균이라고 하자. 이러한 경우 (2)의 회귀 추정량을 사용하고자 한다면 $\overline{X}_N$ 대신에 $\overline{x}_1$ 을 사용하면

$$\widehat{\theta}_{sr} = \overline{y}_2 + (\overline{x}_1 - \overline{x}_2)b \quad \cdots\cdots (6)$$

의 형태로 표현될수 있을 것이다. 이 단순 회귀 추정량은 x 에 대하여 같은 추정량 $\overline{x_1}$ 을 구현해 낸다는 일종의 calibration 성질을 지니는 장점이 있지만 전통적인 회귀 추정량과는 달리 (6)의 추정량의 분산이 y 의 단순 추정량 $\overline{y_2}$의 분산보다 더 작게 된다는 보장이 없게 된다.

이러한 문제점을 해결하기 위하여 Zieschang (1990)과 Renssen and Nieuwenbroek (1997)은 $\overline{x_1}$ 을 control 로 사용하는 대신

$$\overline{x_a} = a\overline{x_1} + (1-a)\overline{x_2} \quad \text{.....(7)}$$

을 control 로 사용하는 regression 추정량을 제안하였다. 즉, (7)을 사용한 회귀 추정량은

$$\widehat{\theta_{reg}} = \overline{y_2} + (\overline{x_a} - \overline{x_2})b \quad \text{.....(8)}$$

의 형태로 표현되며 $a=0$ 인 경우에는 (8)의 회귀 추정량이 단순 추정량 $\overline{y_2}$ 와 동일해 지고 $a=1$ 인 경우에는 (6)의 단순 회귀 추정량이 된다. 최적 계수 a는 (8)의 분산을 최소화 하도록 결정하면 된다.

(8)의 추정량은 또한 다음과 같이 표현된다.

$$\widehat{\theta_{reg}} = a\widehat{\theta_{sr}} + (1-a)\overline{y_2} \quad \text{.....(9)}$$

즉, 식 (9)는 식(6)의 단순 회귀 추정량과 단순 추정량 $\overline{y_2}$와의 가중평균인 일종의 복합 추정량 (composite estimator)의 형태가 된다. 이 복합 추정량은 x 의 추정에 대해 calibration을 유지하므로 최적 계수 a 를 사용한 복합 추정량은 calibration 을 유지하면서 분산을 최소화하는 추정량이 될 것이다. 여기서 a는 $\widehat{\theta_{sr}}$을 $\overline{y_2}$ 방향으로 값을 보정해주는 Shrinkage 계수로도 불리울 수 있다. 이 복합추정량의 분산을 최소화하도록 그 계수를 구하면

$$a^* = \frac{Var(\widehat{\theta}_{sr}) - Cov(\overline{y}_2, \widehat{\theta}_{sr})}{Var(\overline{y}_2) + Var(\widehat{\theta}_{sr}) - 2Cov(\overline{y}_2, \widehat{\theta}_{sr})}$$

으로 표현된다.

4. 복합 추정량의 확장

이 절에서는 3절에서 다룬 복합 추정량처럼 calibration 을 유지하면서 분산을 최소화하는 추정량을 고려해 보기로 한다. 본 절에서는 보다 일반적인 형태의 calibration 조건을 이용한 복합 추정량을 연구하고자 한다.

3절에서처럼 하나의 모수에 대해 두가지 다른 불편 추정량 $\overline{x_1}$ 과 $\overline{x_2}$ 이 존재한다고 하면

이 두 추정량은 다음과 같은 moment condition을 가진다.

$$E(\overline{x_1} - \overline{x_2}) = 0 \quad \cdots\cdots (10)$$

이 조건은 단순 추정량 $\overline{y_2}$에 결합하면 다음과 같은 추정량들은 항상 unbiased가 될 것이다.

$$\widehat{\theta}_\alpha = \overline{y_2} + \alpha(\overline{x_1} - \overline{x_2}) \quad \cdots\cdots (11)$$

(11)은 모든 α 값에 대해서 항상 unbiased 이고 최적 α의 결정은 (11) 추정량의 분산을 최소화하도록 결정하면 된다. 이렇게 해서 얻어지는 최적 추정량은

$$\widehat{\theta}^* = \overline{y_2} - \frac{Cov(\overline{y_2}, \overline{x_1} - \overline{x_2})}{Var(\overline{x_1} - \overline{x_2})}(\overline{x_1} - \overline{x_2}) \quad \cdots\cdots (12)$$

으로 표현된다. 단순 임의 추출의 경우에는 (12)의 최적 추정량이 (9)의 최적 composite 추정량과 동일해 지지만 다른 표본 추출의 경우에는 다르게 계산된다. 물론 (12)에서 분산과 공분산 항은 모수이므로 그의 추정치를 사용해서 구현해야 할 것이다.

만약 두 조사가 서로 독립인 경우에는 (12)의 추정량은

$$\widehat{\theta}^* = \overline{y_2} + \frac{Cov(\overline{y_2}, \overline{x_2})}{Var(\overline{x_1}) + Var(\overline{x_2})}(\overline{x_1} - \overline{x_2})$$

으로 표현되고 그 분산 추정도

$$Var(\widehat{\theta^*}) = Var(\overline{y_2}) - Cov(\overline{y_2}, \overline{x_2})[Var(\overline{x_1}) + Var(\overline{x_2})]^{-1} Cov(\overline{x_2}, \overline{y_2})$$

를 이용하여 구현할 수 있다.

참고 문헌

Deville and Sarndal (1992) "Calibration estimatos in survey sampling", *Journal of the American Statistical Association*, 87, p376-382.

Renssen and Nieuwenbroek (1997) "Aligning estimates for common variables in two or more sample surveys", *Journal of the American Statistical Association*, 92, p368-374.

Sarndal and Lundstrom (2004) "Estimation in surveys with nonresponse", *Wiley*.

Zieschang (1990) "Sample weighting methods and estimation of totals in the consumer expenditure survey", 85, p986-1001.