연구논문

표본조사에서 일반회귀추정량의 활용*

General Regression Estimators in Survey Sampling

김규성**
Kyu-Seong Kim

표본조사에서 사용 가능한 보조변수가 있는 경우에 추정의 효율을 높이기 위하여 보조변수를 활용하는 방법이 다각적으로 개발되어 왔다. 이 논문은 보조변수를 효과적으로 이용하는 방법 중의 하나인 일반회귀추정량에 대한 개괄적인 고찰이다. 일반회귀추정량의 출현부터 분산추정법의 제안까지 이론전개 과정을 살펴보았으며, 보정추정량 및 QR추정량과의 관련성을 통하여 일반회귀추정량의 성질을 알아보았다. 특히 분산추정에서 통상적인 설계기반 분산추정량이 가지는 조건부 성질의 약점을 보완하기 위하여 가중잔차기법을 사용하는 과정을 살펴보았다. 층화표집이나 집락표집과 같은 복합설계에서 활용할 수 있는 일반회귀추정량의 형태를 소개하였고, 마지막으로 일반회귀추정량의 장단점, 그리고 향후 이론적인 발전방향 및 실용적인 발전방향을 언급하였다.

주제어 : 가중잔차기법, 모형보조추론, 보정추정량, 분산추정, g-가중값, QR추정량

This paper is a broad review about general regression estimators, which are very useful when auxiliary variables are available in survey sampling. We investigate the process of development of general regression estimators from birth to suggestion of variance estimation method and examine some properties of general regression estimators by comparing with calibration and QR estimators. We also present some forms of general regression estimators available under complex sampling designs such as stratified sampling and cluster sampling. Finally, we comment some advantages as well as disadvantages of general regression estimators and theoretical and practical development in the future.

* 본 논문은 2002년도 서울시립대학교 해외연구교수 연구비 지원에 의해 연구되었음.

** 교신저자(corresponding author) : 서울시립대학교 통계학과 부교수 김규성.
E-mail : kskim@uos.ac.kr

key words : calibration estimator, g-weight, model-assisted inference, QR estimator, variance estimation, weighted residual technique

Ⅰ. 서론

전통적인 표집이론(sampling theory)에서는 보조정보를 적절히 이용하면 추론의 효율을 높일 수 있기 때문에 보조정보를 추론에 이용하는 방법이 다양하게 개발되어 왔다. 그 중에 설계단계에서 보조정보를 추출확률에 반영하거나 혹은 추정단계에서 추정량에 반영하는 방법이 대표적이다. 집락추출에서 집락의 크기에 비례하는 확률비례추출의 경우는 전자에, 반대로 단순임의표집에서 비추정량이나 회귀추정량의 경우는 후자에 속한다고 볼 수 있다. 설계기반추론(design-based infer-ence)을 대표하는 호르비쯔-톰슨(Horvitz-Thompson, HT) 추정량이나 비추정량, 혹은 회귀추정량은 많은 실제문제에서 사용되고 있긴 하지만 이론적인 최적성과 보조변수의 최대 활용에 관해서는 부족한 면이 있다. 즉, 포함확률을 계산할 수 있는 대부분의 설계에서 HT 추정량은 손쉽게 활용될 수 있으나 HT추정량이 최적인지 여부는 별개의 문제이며, 게다가 보조변수는 제한적으로만 HT추정량에 적용될 수 있다. 또한 비추정량이나 회귀추정량은 암시적인 회귀모형을 전제로 보조변수를 추정에 활용할 수 있으나 보조변수가 2개 이상일 때에는 적용이 그리 쉽지 않다. 따라서 설계기반추론의 패러다임을 유지하면서 보조변수를 최대로 활용하는 동시에 추정량의 최적성을 추구하는 것은 설계기반추론에서 매우 중요한 문제였다.

이 문제에 대하여 일반회귀추정량(general regression estimator, GREG)은 근사한 하나의 답이 될 수 있다(Cassel et al. 1976). 명시적인 회귀모형을 염두에 두고, GREG는 점근적 설계비편향성을 가지며(Robinson & Sarndal 1983), 설계와 모형을 동시에 고려했을 때 점근

적 최적성을 갖는다(Cassel et al. 1976). 따라서 표본의 수가 충분히 클 때에 GREG는 보조정보를 최대한 활용하면서, 설계비편향성을 유지하는 좋은 추정량이라고 볼 수 있다. 추정량의 형태의 관점에서 볼 때 GREG는 보정추정량(calibration estimation, Deville & Sarndal 1992)의 일종이며 또한 QR추정량(QR estimator, Wright 1983)의 일종이다. 여기서 주목해야 할 점은 표본의 수가 클 때 보정추정량들은 거리함수의 형태에 관계없이 GREG와 점근적으로 동일하고, 또한 점근적 설계비편향성을 갖는 QR추정량은 GREG와 동일하게 될 수 있다는 점이다. 바꿔 말하면, 표본의 수가 클 때 보정추정량과 점근적 설계비편향성을 만족시키는 QR추정량은 GREG의 추론의 방식을 빌려서 추론을 할 수 있는 것이다. GREG의 활용 가능성이 그만큼 넓음을 의미한다.

GREG는 보조변수가 있을 때(1개 혹은 2개 이상) 보조정보를 충분히 활용하면서 추론의 타당성을 확보하고 있는 유용한 추정량임에 분명하다. 그러나 여기서 우리는 GREG의 추론의 기반이 표본설계에 의해서 생성된 확률분포임에 주목할 필요가 있다. 즉, 보조정보를 이용하는 수단으로 회귀모형을 염두에 두고 회귀모형에서 회귀계수를 추정한 후 모평균 혹은 모총계를 추정하지만, 추정량의 비편향성이나 분산추정량의 일치성, 그리고 신뢰구간의 포함 범위 확률 등은 모두 회귀모형이 아닌, 표본설계에서 유도된 확률분포에서 구한다는 사실이다. GREG를 이용한 추론은 GREG를 구하는 과정에서 모형을 이용하기는 하지만 추론의 확률계산에는 이용하지 않는다는 점에서 모형기반추론(model-based inference)과 다르며, 또한 설계기반추론과도 다르다. 이러한 추론을 모형보조추론(model-assisted inference)이라고 하는데 넓게 보면 설계기반추론의 기초 위에서 보조변수를 활용하는 방식으로 볼 수 있다.

GREG에 대한 이론 전개는 GREG에 대한 분산추정법에 소개되면

서(Sarndal et al. 1989) 거의 완성된 것으로 보인다. 그 이후에는 GREG를 여러 경우에 적용한 사례들이 발표되고 있으며 적용의 범위가 점점 넓어지는 추세에 있다. GREG는 통상적으로 표본조사에서 이용되는 다수의 추정량을 포함하고 있다. 예를 들면, 보조변수가 연속형 변수일 때 사용되는 비추정량이나 회귀추정량은 비모형과 단순회귀모형을 염두에 두고 만들어진 GREG의 형태이다. 또한 사후층의 크기를 추정량에 반영하는 사후층화 추정량이나 혹은 래킹비(raking ratio) 추정량은, 보조변수가 이산형 변수일 때 만들어지는 GREG의 일종으로 볼 수 있다. 이와 같이 GREG를 이용하면 기존의 여러 추정량에 일관성 있게 적용할 수 있고, 또한 상당히 넓은 다양한 문제에 효과적으로 활용할 수 있는 장점이 있다.

이 논문의 2절에서는 GREG이 출현부터 GREG의 분산추정법의 완성까지 이론적 발전과정을 살펴보았으며, 3절에서는 보정추정량과 QR추정량 등 GREG와 연관된 추정법과의 관계를 통하여 GREG의 성질을 검토하였다. 또한 4절에서는 복합설계에서 GREG를 활용하는 예를 보였으며 마지막으로 5절에는 GREG에 관한 토의가 있다.

II. 일반회귀추정량의 출현과 발전

1. 일반회귀추정량의 출현

네이만(Neyman 1934)에 의해 설계기반추론의 타당성이 입증된 이후, 설계기반추론은 표집이론에서 정통적인 추론방식으로 자리를 잡는다. 그러나 고담베(Godambe 1955)에 의해서 최량의 설계기반 추정량이 존재하지 않는다는 사실이 증명됨으로써 설계기반추론은 이론적인 위기를 맞는다[1]. 그러나 그는 같은 논문에서 자신이 보인 최량 추정량의

1) 이론적인 위기는 분명하지만 실제문제를 다루는 조사통계학자들에게 준 타격은

무존재성의 위기를 돌파할 방법을 제시하는데, 그것은 조사변수에 대한 모형을 가정하고 모형설계분산(model-design variance)을 도입하여 그것을 추정량의 평가기준[2]으로 사용하는 것이었다. 조사변수에 대하여 독립성과 평균 그리고 분산을 가정하면 모형설계분산을 최소로 하는 설계비편향추정량을 유도할 수 있다. 이로써 설계기반추론이 당면한 최량추정량의 문제는 극복할 돌파구가 마련되지만, 보조변수를 추정량에 활용하는 문제는 더 많은 시간을 필요로 하였다. 보조변수를 이용하는 방법으로 회귀모형을 고려하는 것은 이전의 예에서도 찾아볼 수 있다(예를 들면, Cochran 1939). 비추정량이나 회귀추정량이 암시적인 회귀모형을 염두에 두고 만들어진 추정량이기 때문이다. 단지 차이는 회귀모형을 암시적으로 이용할 뿐 명시적으로 이용하지는 않는다는 점이다. 그런데 고담베(1955) 이후에는 초모집단 모형(superpopulation model)의 명시적인 도입과 추정량 평가기준으로서 모형설계분산이 자연스럽게 받아들여진다.

그러나 초모집단 모형을 도입하고 모형설계분산을 받아들였을 때 다음의 세 가지 문제에 대한 고려가 요구되었다. 첫째는 추정량이 갖추어야 하는 성질이 무엇이어야 하는 점이고, 둘째는 모형설계분산을 수리적으로 계산하는 문제, 마지막으로 설계와 모형의 연관성 문제였다. 모형설계분산을 구하기 이전에 추정량이 취해야 하는 기본 성질로는 설계비편향성(design unbiasedness)과 모형비편향성(model unbi-

---

크지 않은 것 같다(Hansen et al. 1983에 이에 대한 토론이 있다). 그러나 고담베(1955) 이후 많은 수리통계학자들이 표집이론에 관심을 갖는 계기가 된 점은 분명하다.

2) 추정량의 평가기준으로 모형설계분산은 여러 연구자들에 의하여 애용되었다. 그러나 추정량을 구하는 과정 혹은 비교를 위한 수단으로 이용될 뿐, 정작 추정량의 확률 계산에는 이용되지 않는다. 설계기반추론(모형보조추론 포함)은 설계에 기반한 확률을 이용하며, 모형기반추론에서는 모형에 근거한 확률분포를 이용한다. 마찬가지로 GREG를 유도하는 과정에서는 모형설계분산이 중요한 역할을 하지만, GREG에 대한 분산 및 분산추정에는 설계에 기초한 확률만 이용된다.

asedness)이 있는데, 물론 양자를 모두 만족하는 경우도 있다. 어느 조건을 전제로 하느냐에 따라 결과는 판이하게 달라진다. 예를 들어 널리 쓰이는 비모형을 가정했을 때, 모형비편향성을 가지며 모형설계분산을 최소로 하는 추정량은 비추정량이며 최적의 설계는 보조변수의 값이 가장 큰 조사단위를 취하는 유의설계(purposive design)이다(Royall 1970). 반면에 설계비편향성을 전제로 하면 최적의 추정량은 HT추정량이며, 이때 포함확률은 조사변수의 표준편차에 비례하는 값이다. 이처럼 어떤 비편향성을 추정량에 부여하느냐에 따라 결과는 다르게 나온다. 다음으로, 단순임의추출을 제외한 대부분의 경우, 모형설계분산을 직접 계산하기는 쉽지 않다. 그 이유는 회귀계수 추정량은 모형의 관점에서는 선형이기 때문에 계산이 쉬우나 설계의 관점에서는 비선형이기 때문에 복잡한 계산을 필요로 하기 때문이다. 마지막으로 전통적인 설계기반추론에서는 조사변수의 값에 관계없이 타당한 추론을 할 수 있다고 주장을 하는데, 보조변수가 등장하게 되면 이러한 주장이 모호해진다. 왜냐하면 보조변수를 이용하는 이유는 조사변수와 연관성 있는 정보를 설계에 활용하여 추론의 효율을 높이기 위함이므로, 보조변수를 추론에 이용하는 것은 암묵적으로 조사변수에 의존하는 추론을 하는 것을 의미하기 때문이다. 이에 대한 해결방안으로 보조변수가 주어졌을 때, 설계와 조사변수가 서로 무관하다고 하는 조건부 분포를 가정하는 것이 일반적이다. 현실적으로 이러한 가정을 확인하는 것은 쉽지 않은 일인데, 여하튼 이러한 가정에서는 수리적으로 모형-기대값과 설계-기대값의 계산을 서로 바꾸어 할 수 있기 때문에 모형설계분산 계산이 쉬워진다. 따라서 대부분의 경우 보조변수가 주어졌을 때 설계와 조사변수는 독립이라고 하는 조건부 분포가 가정된다.

GREG는 일반편차추정량(general difference estimator, GDE)을 통하여 출현하게 되는데(Cassel et al. 1976), 그 이유는 기술적으로

GREG에 대한 모형설계분산의 계산이 너무 복잡하여 GREG를 직접 유도하기 어렵기 때문이다. 카셀 등(Cassel et al. 1976)은 조사변수에 대한 전이모형(transformation model)을 가정하고, 설계비편향성을 만족시키는 선형추정량 중에서 모형설계분산을 최소로 하는 추정량을 구하였다. 이때 대상이 되는 설계는 고정표본크기 설계이다. 이렇게 구해진 추정량을 GDE라고 부른다. 전이모형에서 위치상수에 회귀모형에서의 조사변수에 대한 기대값을 대입하여 구한 GDE에서, 미지의 회귀계수 대신 회귀계수추정량을 대입한 추정량이 우리가 고찰하고자 하는 GREG이다.

$$\bar{y}_{GREG} = \frac{1}{N}\left(\sum_s \frac{Y_k}{\pi_k} + [\sum_U x_k - \sum_s \frac{x_k}{\pi_k}]'\hat{\beta}\right) \qquad (1)$$

여기서 N은 모집단 크기, Y_k는 k번째 모집단 단위의 조사변수, π_k는 k번째 모집단 단위가 표본에 포함될 포함확률, x_k는 q차원의 벡터로서 k모집단 단위에서의 보조변수값, 그리고 s는 표본, 그리고 $\hat{\beta}$는 회귀모형에서 구한 회귀계수추정량이다. GREG는 HT추정량에 보조변수를 이용하여 구한 항이 추가된 형태인데, 만일 회귀모형이 잘 맞으면 HT추정량과 보조변수를 이용하여 만든 HT추정량의 잔차는 서로 음의 상관계수가 있을 것이므로 GREG의 분산은 보조변수를 이용하지 않는 HT추정량보다 작아짐을 보일 수 있다.

GREG의 출현에 즈음하여 GREG에 관한 관심은 두 방향에서 나타난다. 하나는 회귀계수의 형태에 관한 것이다. 일반통계학에서 회귀계수추정량으로 최량선형비편향추정량(best linear unbiased estimator, BLUE)을 사용하는 것에 대해서는 별다른 이의가 없을 것이다. 그러나 복합조사에서는 조사단위가 서로 다른 포함확률을 가지고 있는 것이 보통이기 때문에 포함확률을 회귀계수에 반영하는 문제는 진지하게 생각해 봐야 하는 문제였다. 다른 하나는, GREG의 성질에 관한 것이

다. GDE는 전이모형의 전제 아래 일정기준을 만족시키는 최량추정량이지만, GREG는 그렇지 못하다. 따라서 GREG를 GDE의 직접적인 확장으로서 직관적으로 받아들일 수는 있지만, 이론적으로 어떤 성질을 가지고 있는지 사후적으로 규명될 필요가 있었다.

2. 회귀계수의 형태

GREG에 포함되는 회귀계수에 관하여 샴달(Samdal, C.E.)을 비롯한 GREG 연구자들은 꽤나 고심을 한 듯 하다. 앞서 언급했듯이 모형만을 고려하면 BLUE를 사용하는 것이 간편한 일이지만, 복합조사를 염두에 두고 GREG의 활용성을 넓히기 위해서는 가능하면 일반적인 형태의 회귀계수를 정의해야 했기 때문이다. GREG 관련 주요 논문에서 서로 다른 형태의 회귀계수를 사용하여 서로 다른 GREG를 정의하는 것은 그런 이유 때문이다. 연구 초기에는 회귀계수로 BLUE를 사용하고 있다(Cassel et al. 1976). 그러나 위에서 지적한 문제점, 즉 설계기반추론에서 포함확률을 무시하고 직접적으로 BLUE를 사용하기는 쉽지 않은 일이므로 포함확률과 BLUE를 회귀계수에 반영하는 방법에 대한 심층적인 비교가 이루어진다(Samdal 1980). 양자의 효율이 유사하다는 결과를 근거로 회귀계수에 포함확률을 포함시키고 또한 분산항을 일반화한 상수항을 포함시키는 형태로 발전하게 된다(Samdal et al. 1989).

$$\hat{\beta}=\left(\sum_{s}\frac{x_k x'_k}{c_k\pi_k}\right)^{-1}\sum_{s}\frac{x_k y_k}{c_k\pi_k} \qquad (2)$$

여기서 $c_k > 0$은 상수인데, 분산항을 대입하려면 $c_k=\sigma_k^2$으로 하면 된다. 상수항 c_k를 임의로 정할 수 있으므로 GREG가 포함하는 추정량의 형태는 상당히 넓어지게 된다. 나중에 설명할 QR추정량이나 보정

추정량에서도 위의 회귀계수를 이용한 GREG를 정의하고 있다. 그러나 그 간편성과 유용성 때문에 Samdal et al. 1992에서는 $c_k = \sigma_k^2$을 고정하고 있다. c_k의 형태는 GREG 추론의 결과에 큰 영향을 미치지 못하므로 GREG의 범위를 넓히기 보다는 직관적으로 이해하기 쉬운 $c_k = \sigma_k^2$의 형태로 되돌아 간 것 같다.

3. 설계일치성

가장 바람직하게는 유한 모집단에서 GREG에 대한 편향과 분산을 구하여 GREG의 성질을 규명하는 것이다. 초기에 이러한 시도가 보인다(Cassel et al. 1976; Samdal 1980). 그러나 예상했던 대로 회귀계수를 포함한 상태에서 GREG의 편향이나 평균제곱오차를 구하는 일은 만만한 일이 아니었다. 따라서 유한 모집단에서 성질 규명이 어려우므로 자연스럽게 점근적인 성질 규명으로 옮겨가게 된다.

유한 모집단을 대상으로 하는 표집이론에서 추정량의 성질을 점근적으로 규명하는 것은 일견 역설적이다. 이에 대한 근거는 네이만(1934)으로 거슬러 올라가는데, 네이만은 층화추출에서 층화평균이 최량선형비편향추정량임을 보임과 동시에 점근적으로 정규분포를 따르기 때문에 모평균에 대한 신뢰구간을 구할 때 신뢰계수로써 정규분포의 분위수를 이용할 수 있다고 하였다. 더 나아가 한센 등(Hansen et al. 1983)은 추정량이 갖추어야 할 성질로서 점근적 설계일치성과 극한분포에서의 추론이 가정된 모형에 의존하지 않아야 한다고 하였다. 따라서 GREG를 이용한 추론이 타당성을 확보하려면, 점근적 설계일치성 혹은 점근적 설계비편향성과 GREG의 극한 분포에 대한 언급이 있어야 한다.

GREG에 대한 점근적 성질을 살펴보기 전에 표집이론에서 나타나는 점근성(asymptotics)에 대하여 짚고 넘어갈 필요가 있다. 단순임의

추출에서는 표본크기 및 모집단 크기가 증가하더라도 단순임의설계의 속성을 유지할 수 있다. 그러나 대부분의 다른 설계에서는 그렇지 못하다. 모집단 크기와 표본의 크기가 증가함에 따라 층화, 집락화, 불균등 추출 확률 등이 변하게 되기 때문에 모집단 및 표본크기의 증가는 현실적으로 발생하지 않음은 물론, 이론적으로도 다소 분명하지 않을 가능성이 많다(Sarndal et al. 1992, p. 167). 한센 등(1983)에서 결합비 추정량과 분리비 추정량이 점근적으로 서로 상이한 결과로 나타나는 것은 그 예이다. GREG의 점근적 성질을 규명하기 위해서는 모집단의 크기가 증가함에 따라 조사변수, 보조변수 그리고 포함확률이 어떤 양태로 변하는지에 대한 고찰이 필요한데, 로빈슨과 샴달(Robinson & Sarndal 1983)은 GREG가 점근적으로 설계비편향이고 설계일치이기 위한 충분조건을 구하였다(정리 1, p. 244). 그 가정은 보조변수 및 조사변수의 2차 적률, 그리고 회귀계수추정량의 제곱의 합에 대한 모형 기대값이 모두 모집단의 크기가 증가하더라도 유한하다고 하는 것이며, 포함확률에 관한 조건은 다음과 같다(Robinson & Sarndal 1983, p.243) :

$$\lim_{t\to\infty} N_t \min_{1\le k\le N_t} \pi_{kt} = \infty \tag{3}$$

$$\lim_{t\to\infty} \max_{1\le k\ne l\le N_t} \left| \frac{\pi_{klt}}{\pi_{kt}\pi_{lt}} - 1 \right| = 0$$

위의 조건들은 GREG의 점근적 설계일치성 및 설계비편향성을 입증하는 데 필요한 충분조건들인데, 이론적인 규명과는 달리 현실적으로 확인하기는 어렵다. 대신, 위의 가정들은 표본의 크기가 큰 대규모 조사 등에서 변수들의 2차 적률이 너무 크면 안되고, 1차 포함확률(π_{kt})은 모두 양수이며 최소값이 너무 작으면 안되고, 또한 2차 포함확률(π_{klt})의 최대값은 1차 포함확률의 곱에 근접해야 한다는 사실을 암

시하고 있다고 볼 수 있다.

4. 분산 및 분산추정

GREG의 신뢰도를 측정하기 위해서는 GREG에 대한 분산 및 분산추정량을 구해야 한다. 원래 GREG는 설계 및 모형을 이용했기 때문에 모형설계분산을 구하는 것이 당연한 것처럼 보이고 초기에는 모형설계분산을 구하려고 한 시도가 있었다. 그러나 회귀계수추정량으로 인하여 분산의 식이 깔끔하게 정리가 되지 않기 때문에 추가적으로 점근적인 모형설계분산을 구하려는 시도가 있었다(Robinson & Sarndal 1983). 그러나 결과는 만족스럽지 못했으며 분산추정량까지는 이르지 못한 것으로 보인다. 결과적으로 GREG는 모형설계분산을 최소로 하여 구해졌지만, GREG의 평가를 위한 측도로는 설계분산과 설계분산추정량이 제안되었다(Sarndal 1989; Sarndal et al. 1992). 샹달 등(Sarndal et al. 1989)이 언급한 것처럼 GREG의 분산추정량이 갖추어야 할 성질로는 (i) 타당한 설계기반 신뢰구간을 만드는 데 제공되어야 하고, (ii) 모형을 가정했을 때 근사적으로 비편향이며, 마지막으로 (iii) 조건부 추론에 적합해야 하는 점 등이다.

샹달 등(1989)은 GREG의 분산을 구하는 방법으로 이전에 겪었던 계산의 복잡성에서 벗어나기 위하여 가중잔차기법(weighted residual technique)을 제안한다. 이 기법은 기존의 설계기반추론이나 모형기반추론에서 사용하던 기법과는 다른 새로운 기법인데 간략하게 소개하면 다음과 같다. 우선 GREG를 테일러 전개를 하여 근사시킨 후 유한모집단 모수값을 대입하여 근사 GREG를 구한다.

$$\tilde{y}_{GREG}=\frac{1}{N}\left(\sum_{s}\frac{Y_k}{\pi_k}+[\sum_{U}x_k-\sum_{s}\frac{x_k}{\pi_k}]'B\right) \qquad (4)$$

여기서 B는 유한모집단에서의 회귀계수이다. 표본에서 구한 추정량 $\hat{\beta}$ 대신에 유한모집단에서의 구한 회귀계수 B를 이용하면 설계기대값 및 분산을 구하는 과정이 훨씬 간단해지는 효과가 있다. 따라서 $\tilde{y}_{GREG}$를 이용하여 GREG의 근사 기대값과 근사 설계분산을 어렵지 않게 구할 수 있다(Sarndal et al. 1992). 또한 분산추정량은 위에서 구한 근사 설계분산에 대응하는 설계비편향추정량을 구한 후, B 대신에 $\hat{\beta}$을 대입하여 얻는다.

$$v_1(\bar{y}_{GREG}) = \frac{1}{N^2}\sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{e_{ks}}{\pi_k}\cdot\frac{e_{ls}}{\pi_l}\right) \tag{5}$$

여기서 $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$, 그리고 $e_{ks} = y_k - x'_k\hat{\beta}$이다. 이 분산추정량은 쉽게 구할 수 있고, 설계기반추론에서 필요로 하는 설계일치성을 가지고 있어서 GREG에 대한 분산추정량으로 사용할 수도 있으나, 근사 GREG에서 유도되었기 때문에 모형의 효과를 둔감하게 반영하는 약점이 있다. 따라서 모형의 특성을 분산추정량에 반영할 필요성이 있었다. 상달 등(1989)은 이러한 문제를 해결하기 위하여 g-가중값(g-weight)을 이용한 분산추정방법을 소개하였다. 즉, GREG는 조사변수에 대한 선형추정량이기 때문에 GREG를 조사변수를 중심으로 표현하면 다음과 같이 된다.

$$\bar{y}_{GREG} = \frac{1}{N}\sum_s g_{ks}\frac{Y_k}{\pi_k} \tag{6}$$

여기서

$$g_{ks} = 1 + [\sum_U x_k - \sum_s \frac{x_k}{\pi_k}]'\left(\sum_s \frac{x_k x'_k}{\sigma_k^2\pi_k}\right)^{-1}\frac{x_k}{\sigma_k^2} \tag{7}$$

이며, 모형의 특성을 GREG에 반영하는 역할을 한다. 따라서 g-가중

값을 반영한 분산추정량은 분산추정량 v_1에서 잔차 e_{ks}대신에 g-가중값과 잔차의 곱을 대입하여 얻을 수 있다.

$$v_g(\bar{y}_{GREG}) = \frac{1}{N^2}\sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{g_{ks}e_{ks}}{\pi_k} \cdot \frac{g_{ls}e_{ls}}{\pi_l}\right) \qquad (8)$$

표본의 크기가 증가하면 g-가중값은 1로 수렴하기 때문에 v_1이 가지는 설계기반 특성을 v_g도 모두 가지며, 더불어 모형의 특성을 g-가중값을 통하여 반영하기 때문에 모형의 특성에 민감한 장점이 있다.

Ⅲ. 일반회귀추정량과 관련된 추정량

1. QR추정량

1980년대 초반에 즈음하여 설계기반추정량과 모형기반추정량의 성질을 규명하고, 두 추정량을 통합하는 합성추정량을 만들려는 시도가 나타나는데, 그 중 QR추정량이 대표적이다. QR추정량은 다음과 같이 정의된다(Wright 1983).

$$\bar{y}_{QR} = \bar{x}'\hat{\beta} + \frac{1}{N}\sum_s r_k e_k \qquad (9)$$

여기서 $\hat{\beta} = (\sum_s q_k x_k x'_k)^{-1}\sum_s q_k x_k Y_k$, $e_k = Y_k - x_k'\hat{\beta}$, 그리고 $r_k \geq 0$, $q_k > 0$는 상수이다. QR추정량은 기존의 많은 추정량들, 예를 들어 HT비추정량, BLUE, 브르워의 추정량 (Brewer 1979) 등을 포함하는데, GREG도 예외는 아니다. 즉, QR추정량에서 $q_k = 1/(c_k\pi_k)$, $r_k = 1/\pi_k$를 대입하면 GREG가 된다.

QR추정량을 고려하게 된 이유는 모형기반추정량과 설계기반추정량

이 제각기 각자의 기준에 의한 성질을 가지고 있기 때문에 추정량의 형태를 합성하여 공통의 성질을 규명해 보려고 했기 때문이다. 재미있는 결과는 점근설계비편향성(asymptotic design-unbiasedness, AUD)을 만족하는 QR추정량은 GREG가 된다는 사실이다(Wright 1983, Theorem 2). 즉, 상당히 넓은 범위의 추정량의 집합인 QR추정량 중에서 ADU를 만족하는 추정량은 GREG가 되므로, QR추정량의 성질을 규명할 때 GREG의 이론적 성질들을 그대로 이용할 수 있다는 뜻이 된다.

2. 보정추정량

보정추정량(calibration estimator)은 GREG와는 다른 출발점에서 다른 방식으로 만들어졌다(Deville & Samdal 1992). 주어진 상황은 GREG와 유사하다. 설계기반추론의 HT추정량이 있고 더불어 활용가능한 보조변수가 주어져 있다. HT추정량은 보조변수를 충분히 활용하지 않기 때문에 보조변수를 충분히 활용하는 설계기반추정량을 만들고 싶다. 보정추정량은 거리함수를 도입한 후 거리함수를 최소로 하는 추정량을 구하는데, 이때 구한 추정량에 보조변수를 대입하면 추정하고자 하는 보조변수의 모수값과 정확하게 일치하여야 한다. 즉, 기존의 HT추정량을 보조변수로 보정한 추정량이 보정추정량이라고 볼 수 있다. 거리함수에 따라 보정추정량의 형태는 달라지는데 제곱거리를 가정하고 구한 추정량이 GREG이다. 즉, $\bar{y}_{HT}=\sum_s d_k y_k/N$ 를 HT추정량이라고 하고, $d_k=1/\pi_k$, 보조변수에 의하여 보정된 보정추정량을 $\bar{y}_{cal}=\sum_s w_k y_k/N$이라고 하자. 보정된 가중치 w_k는 다음의 성질을 만족시킨다.

$$\sum_{k\in s} w_k x_k = \sum_{k=1}^{N} x_k \qquad (10)$$

보정된 가중치는 거리함수 $G(w_k/d_k)$에 따라 다르게 나타나는데, 만일 거리함수를 제곱거리로 가정하고, $G(x)=(x-1)^2/2$, 제곱거리 함수를 최소로 하면 보정추정량은 GREG와 동일한 추정량이 된다. 따라서 GREG는 보정추정량의 일종으로 간주할 수 있다.

보정추정량은 HT추정량이 갖는 포함확률의 임의성[3]에 대한 약점을 보조변수를 활용하여 보완해주기 때문에 매우 실제적인 추정량이라고 할 수 있다. 이러한 실제적인 장점과 더불어 보정추정량을 이용한 추정법이 추론의 방법으로 자리를 잡으려면 보정추정량에 대한 점근적인 성질들이 규명되어야 하는데, 본질적으로 보정추정량은 거리함수에 따라 다른 결과를 도출하므로 각 보정추정량마다 점근적인 성질을 규명하는 것이 쉬운 일은 아니다. 거리함수가 제곱거리가 아니라 하더라도 일정조건을 만족하는 거리함수를 이용하여 구한 보정추정량은 점근적으로 설계일치성을 갖는다는 결과가 있다(Deville & Samdal 1992, Result 4). GREG와 연관하여 흥미로운 사실은 보정추정량은 점근적으로 GREG와 동일하게 된다는 점이다(Deville & Samdal 1992, Result 5). 즉, 표본의 수가 크면 보정추정량에 대한 추론을 GREG의 추론 방식을 이용해 할 수 있다는 뜻이다.

Ⅳ. 복합설계에서 GREG의 활용

앞의 식(1)에 주어진 GREG는 매우 일반적인 형태로서 (i) 양수의 포함확률을 가지는 설계와 (ii) 일반적인 회귀모형을 전제로 하고 있다 :

$$Y_1,\ldots,Y_N \sim E(Y_k)=\sum_{j=1}^{J}\beta_j x_{jk},\quad V(Y_k)=\sigma_k^2,\ \text{서로 독립} \qquad (11)$$

3) 논란의 여지는 있지만 HT추정량의 포함확률의 임의성에 대한 대표적인 예로는 바수(Basu 1971)의 코끼리 예제를 들 수 있다.

따라서 사용된 설계와 회귀모형에 따라 GREG의 형태 및 식 (8)에 주어진 분산추정량은 그 형태를 달리한다. 본 절에서는 표본조사에서 널리 이용되는 층화추출, 집락추출에서 활용할 수 있는 GREG의 형태를 간략하게 알아보기로 한다.

1. 층화추출

층화추출에 쉽게 고려할 수 있는 회귀모형은 일원분산분석모형(one-way ANOVA model)이다. (Samdal 1992, p.261; Lohr 1999, p.113):

$$Y_{hj} = \beta_h + \varepsilon_{hj}, \quad E(\varepsilon_{hj}) = 0, \quad Var(\varepsilon_{hj}) = \sigma_h^2,$$
$$h = 1, \ldots, L, \quad j = 1, \ldots, N_h, \tag{12}$$

여기서 h는 층을 나타내며, j는 조사단위를 나타낸다. 위의 모형에 상응하는 GREG는

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \left(\frac{\sum_{s_h} y_{hj}/\pi_{hj}}{\sum_{s_h} 1/\pi_{hj}} \right), \quad W_h = \frac{N_h}{N} \tag{13}$$

이다. 그리고 위의 GREG에 대한 분산추정량은 다음과 같다.

$$v(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \sum_{k \in s_h} \sum_{l \in s_h} \left(\frac{\pi_{hk}\pi_{hl} - \pi_{hkl}}{\pi_{hkl}} \right) \left(g_{ks_h} \frac{e_{ks_h}}{\pi_{hk}} \right) \left(g_{ls_h} \frac{e_{ls_h}}{\pi_{hl}} \right) \tag{14}$$

여기서 $e_{ks_h} = y_{hk} - (\sum_{s_h} y_{hk}/\pi_{hk})/(\sum_{s_h} 1/\pi_{hk})$, $g_{ks_h} = N_h/(\sum_{s_h} 1/\pi_{hk})$.

위의 경우는 층별로 조사변수를 모형화하는 경우이다. 다른 경우로 층과는 무관하게 조사변수에 대하여 모형화를 고려할 수 있다. 예컨

대, 사후층화의 경우 사후층에 대하여 회귀모형을 도입하면 회귀모형은 기존의 사전층과는 무관하게 설정되며, GREG 형태도 사전층과 사후층의 효과를 고려하는 모양을 갖게 된다.

2. 집락추출

일차추출단위(PSU)가 N_I개 있다고 하고, j번째 PSU의 총계는 보조변수 u_j에 비례한다고 하자. 그러면 다음과 같은 비모형을 고려할 수 있다 :

$$T_j = \beta u_j + \varepsilon_j,\ E(\varepsilon_j) = 0,\ Var(\varepsilon_j) = w_j\sigma^2, j = 1, \ldots, N_I \quad (15)$$

여기서 T_j는 j번째 PSU의 총계이며, u_j와 w_j는 알려진 양수라고 하자. 그러면 위의 모형에 대응하는 GREG는 다음과 같다.

$$\bar{y}_{cl} = \frac{1}{N}\left(\sum_s \frac{T_j}{\pi_j} + (\sum_{j=1}^{N_I} u_j - \sum_s \frac{u_j}{\pi_j})\hat{\beta}\right) \quad (16)$$

여기서

$$\hat{\beta} = \frac{\sum_s u_j T_j / w_j \pi_j}{\sum_s u_j^2 / w_j \pi_j} \quad (17)$$

이다. 또한 분산추정량은

$$v(\bar{y}_{cl}) = \frac{1}{N^2}\sum_s\sum_s\left(\frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}\right)(g_{ks}\frac{e_{ks}}{\pi_k})(g_{ls}\frac{e_{ls}}{\pi_l}) \quad (18)$$

이며,

$e_{js} = y_j - \hat{\beta}u_j, \quad g_{js} = 1 + (\sum_{j=1}^{N_I} u_j - \sum_s u_j/\pi_j)(\sum_s u_j^2/w_j\pi_j)^{-1}u_j/w_j$

이다.

이단추출을 하면 위의 식 (16)에 있는 PSU 총계를 추정해야 하기

때문에, 이단추출에 의한 변동이 GREG 및 분산추정량에 추가되어야 한다(Samdal et al. 1992, p314, Result 8.6.1 참조).

앞에서 소개한 층화추출 및 집락추출 이외에 이중추출(double sampling)에서도 GREG를 활용할 수 있다. 이중추출에서 GREG를 활용하려면 1상 표본(first phase sample)과 2상 표본(second phase sample)에 대하여 각각 모형화를 해야 한다. 그리고 각각에서 구한 회귀계수 추정량을 이용하여 GREG를 만들 수 있다. 또한 각각에서 구한 g-가중치를 이용하여 분산추정량을 만들 수 있다(Samdal et. al 1992, p.362, Result 9.7.1 참조).

V. 토의

이 논문에서는 1976년 GREG가 만들어진 후 1989년 분산추정량의 제안까지 GREG의 이론적 발전과정을 고찰하였다. 가중잔차기법을 이용한 분산추정법이 제안됨으로써 GREG에 관한 독자적인 추론 체계가 완성된 듯 하다. 설계기반추론에 보조변수를 활용하는 유용한 추론체계를 GREG가 제공하는 것이다.

GREG를 주로 이용하는 모형보조추론(model-assisted inference)은 GREG 생성과정에서 명시적인 회귀모형을 이용하므로, 설계기반추론보다는 모형기반추론에 더 가까워 보일 수 있다. 그러나 GREG를 이용한 추론은 표본추출이 만들어 내는 확률분포에서 하므로 실제적으로는 설계기반추론에 더 가깝다. 명시적인 회귀모형은 GREG를 만들어 내는 도구로써 활용될 뿐이기 때문이다. 따라서 가정된 모형이 다소 틀리더라도 모형보조추론은 큰 추론의 오류를 범하지 않는다. 왜냐하면, GREG는 점근적 설계일치성을 가지고 있기 때문에 모형 가정이 적절하지 않더라고 설계에서 보완이 되기 때문이다.

앞의 4절에서도 언급한 바와 같이 다양한 복합설계에서 GREG의 활용은 가능하다. 추정량의 형태에서 사후층화 추정량이나 래킹비 추정량은 GREG로 표현이 가능하며, 보조변수의 주변합을 일치시키는 보정추정량도 GREG로 나타낼 수 있다. 통상적으로 사용하는 비추정량이나 회귀추정량은 비모형과 회귀모형을 가정하고 유도한 GREG이다. 이러한 GREG의 넓은 활용성은 실제 조사에 응용된 사례로 나타나고 있다(예를들면, Amstrong et al. 1994; Estevao, V., Hidiroglou, M.A. & Sarmdal, C.E. 1995 등). 캐나다의 경우는 GREG를 이용하여 일반적인 추정 시스템을 구축한다는 보고가 있다(Estevao et al. 1995). 실제조사에 GREG를 활용할 때 고려해야 하는 점은 적절한 회귀모형을 찾는 일일 것이다. 보조변수의 효과적인 활용은 회귀모형을 통해서 GREG에 반영되기 때문이다. 이러한 면에서 GREG는 설계기반추론에서 흔히 사용하는 비추정량이나 회귀추정량보다 더 효율적일 수 있다. 비추정량이나 회귀추정량은 비모형이나 단순회귀모형을 암시적으로 염두에 두고 만들어진 것이기 때문에 추정량의 형태로만 본다면 비추정량이나 회귀추정량은 GREG의 특수한 형태에 지나지 않기 때문이다.

설계기반추론이 그렇듯이 GREG도 점근적인 성질을 기초로 구축되었기 때문에 표본의 크기가 일정 규모 이상일 때 잘 작동한다는 사실을 기억할 필요가 있다. 또한 모형기반추론의 옹호자들이 늘 지적하듯이 추론의 바탕이 설계이기 때문에 조사변수에 부여된 모형 가정이 틀리지 않을 때에는 모형기반추론보다 효율성이 다소 떨어질 수 있다는 점도 기억할 필요가 있다. 그렇지만 대부분의 복합조사에서 타당한 모형을 부여하는 것이 쉽지 않기 때문에 GREG에 의한 추론은 매우 현실적인 것으로 생각되며, 표본의 수가 크면 타당성까지 확보할 수 있어 보조변수가 있는 추론에 매우 적합하다고 할 수 있다.

이론적인 측면에서 GREG는 오차항이 서로 독립인 경우를 가정하

고 있다. 반면에 BLUE 등은 조사변수가 서로 독립이 아닌 경우도 가능하다. 따라서 오차항에 독립성 가정을 제외하면 더 일반적인 GREG를 만들 수 있을 것이다. 이에 대한 논의는 몬타나리 등(Montanari et al. 2002)에서 찾아볼 수 있다. 실용적인 면에서는 구체적인 조사에서 상황에 맞는 회귀모형을 설정하는 일이 중요하다. 복합설계의 경우 표본추출과정에 복합적이기 때문에 이에 상응하는 모형을 설정하고, 적절한 GREG를 만들어 낸다면 보조변수를 최대로 활용하는 추론을 할 수 있을 것이다.

참고문헌

Amstrong, J. and St-Jean, H. 1994. "Generalized regression estimation for a two-phase sample of tax records." *Survey Methodology* 20: 97-105.

Basu, D. 1971. "An essay on logical foundation of survey sampling." Part I. in *Foundation of Statistical Inference*, eds. Godambe, V.P. and Sprott, D.R., Toronto, Holt, Rinehart and Winston: 203-242.

Brewer, K.R.W. 1979. "A class of robust sampling designs for large-scale surveys." *Journal of the American Statistical Association* 74: 911-915.

Cassel, C.M., Samdal, C.E. and Wretman, J.H. 1976. "Some results on generalized difference estimation and generalized regression estimation for finite populations." *Biometrika* 63: 615-620.

Cochran, W.G. 1939. "The use of the analysis of variance in enumeration by sampling." *Journal of the American Statistical Association* 34: 349-361.

Devile, J.C. and Samdal, C.E. 1992. "Calibration estimators in survey

sampling." *Journal of the American Statistical Association* 77: 376-382.

Estevao, V., Hidiroglou, M.A. and Sarmdal, C.E. 1995. "Methodological principles for a generalized estimation system at statistics Canada." *Journal of Official Statistics* 11: 181-204.

Godambe, V.P. 1955. "A unified theory of sampling from finite populations." *Journal of the Royal Statistical Society* Ser. B 17: 269-278.

Hansen, M.H., Madow, W.G. and Tepping, B.J. 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association* 78: 776-807.

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

Lohr, S.L. 1999. *Sampling : Design and Analysis*. Duxbury press.

Montanari, G.E. and Ranalli, M.G. 2002. "Asymptotically efficient generalized regression estimator." *Journal of Official Statistics* 18: 577-589.

Neyman, J. 1934. "On the different aspects of the representative method : the method of stratified sampling and the method of purposive selection." *Journal of the Royal Statistical Society* Ser A. 97: 558-625.

Robinson, P.M. and Sarndal, C.E. 1983. "Asymptotic properties of the generalized regression estimator in probability sampling." *Sankhya* Ser. B 45: 240-248.

Royall, R.M. 1970. "On finite population sampling theory under certain linear regression models." *Biometrika* 57: 377-387.

Sarndal, C.E. 1980. "On π-inverse weighting versus best linear unbiased weighting in probability sampling." *Biometrika* 67: 630-650.

Samdal, C.E. 1982. "Implications of survey design for generalized regression estimation of linear functions." *Journal of Statistical Planning and Inference* 7: 155-170.

Samdal, C.E., Swensson, B. and Wretman, J.H. 1989. "The weighted residual technique for estimating the variance of the general regression estimator of the finite population total." *Biometrika* 76: 527-537.

Samdal, C.E. and Wright, R.L. 1984. "Cosmetic form of estimators in survey sampling." *Scandinavian Journal of Statistics* 11: 146-156.

Samdal, C.E., Swensson, B. and Wretman, J. 1989. "The weighted residual technique for estimating the variance of the general regression estimator of the finite population total." *Biometrika* 76: 527-537.

Samdal, C.E., Swensson, B. and Wretman, J. 1992. *Model assisted survey sampling*. Springer-Verlag.

Wright, R. L. 1983. "Finite population sampling with multivariate auxiliary information." *Journal of the American Statistical Association* 78: 879-884.