불균등확률표본에서 붓스트랩

정 주 경[1], 김 규 성[2]

요 약

분산 추정 및 신뢰구간 추정의 한 방법으로 널리 쓰이고 있는 붓스트랩 방법을 복합표본에 적용하는 방법에 대해 알아보았다. 복합 표본은 유한 모집단에서 추출되고 추출확률이 다르기 때문에 i.i.d. 표본에 기초하여 개발된 전통적인 붓스트랩 방법을 직접 적용하면 추론의 오류가 발생할 수 있다. 본 연구에서는 복원 확률비례표본과 랜덤그룹표본에 붓스트랩을 적용하는 방법을 알아보았다.

주요용어 : 랜덤그룹표본, 복원 확률비례표본, 분산추정, 붓스트랩, 신뢰구간 추정.

1. 서론

복합표본에서 이용할 수 있는 분산추정 및 신뢰구간 추정 방법의 개발은 많은 연구자들의 관심의 대상이었다. 대표적인 방법으로 선형화 방법, 잭나이프 방법, 균형이분표본 방법, 그리고 붓스트랩 방법 등이 알려져 있다. 전통적으로 널리 이용된 선형화 방법은 분산의 일치추정량을 제공하며, 분산 추정량의 식을 구체적으로 제공한다는 장점이 있는 반면, 추정량이 바뀌면 분산 추정량의 형태도 바뀌고 또한 중간 과정에서 미분을 해야 하는 등 계산이 많은 단점이 있다. 잭나이프 방법은 비모수적인 분산추정방법으로 널리 이용되고 있으며 추정량의 형태에 관계없이 분산추정량의 형태가 동일하여 이용이 간단하다는 점과 일치 분산추정량을 제공한다는 점이 장점이다. 반면, 중앙값 같은 함수는 추정을 적절하게 하지 못하는 단점이 있다. 균형이분표본 방법은 일치추정량을 제공하고 잭나이프 방법과는 달리 중앙값 같은 함수도 추정할 수 있는 장점이 있는 반면, 층화 표본에 대해서만 적용이 가능하며 층의 수가 많아지면 이분표본을 구하는 과정이 급속도로 많아진다는 단점이 있다. 유한 표본에서 위의 세 가지 방법의 효율은 서로 다르며, 일정 조건에서 점근적으로는 세 방법의 효율이 동일하다는 결과가 알려져 있다. (Krewski, D. 와 Rao, J.N.K., 1981).

위의 세 가지 방법은 모두 추정량의 분산추정을 위하여 개발된 방법이며, 추정량의 신뢰구간을 구하기 위해서는 공통적으로 정규이론을 이용해야 한다. 즉, 위의 방법들은 모두 100(1-α)% 신뢰구간을 구할 때, 적당한 조건에서 표준화된 추정량이 정규분포로 수렴한다는 중심극한 정리를 이용하여

$$[\hat{\theta} - z_{\alpha/2}\sqrt{v(\hat{\theta})},\quad \hat{\theta} + z_{\alpha/2}\sqrt{v(\hat{\theta})}] \tag{1}$$

와 같은 대칭의 신뢰구간을 이용한다. 소표본 문제에서 신뢰구간을 보다 효율적으로 추정하는 방법으로 붓스트랩 방법을 고려해 볼 수 있다. 소표본에서 붓스트랩 방법을 이용하면 비선형

1) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 전산통계학과, 대학원.
2) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 전산통계학과, 조교수.
E-mail : kskim@uoscc.uos.ac.kr

추정량의 분산 및 신뢰구간을 보다 효율적으로 추정할 수 있음이 알려져 있다. 그리고 유한 모집단에서 이용할 수 있는 붓스트랩 방법이 다각적으로 연구되어 많은 결과가 발표되었다.

본 연구에서는 불균등확률표본에 대한 붓스트랩 방법 적용을 고찰한다. 형태가 일정한 붓스트랩 분산추정량은 재표본을 뽑는 단계나 추정량을 만드는 단계에서 불균등확률추출 효과를 반영해야 하므로 구체적인 추출방법에 따라 붓스트랩의 적용도 달라져야 한다. 본 논문에서는 복원 확률비례표본과 랜덤그룹(Rao, Hartley and Cochran, 1962) 방법을 이용한 비복원 확률비례표본에 붓스트랩을 적용해 보았다. 복원 확률비례표본에서 모평균 추정량과 분산추정량이 비편향 추정량이 되도록 재표본 추출방법과 추정량의 형태를 살펴보았고, 랜덤그룹표본에 재척도 방법과 반사매치 방법을 이용한 붓스트랩 적용을 살펴보았다.

2. 복원 확률비례표본에 붓스트랩 방법의 적용

각 단위를 그 크기에 비례하도록 뽑는 복원 확률비례표본은 뽑는 과정과 비편향 분산추정량의 형태가 간단하기 때문에 많이 이용되고 있다. 조사단위 i의 추출확률이 p_i $(\sum p_i = 1)$일 때, 확률비례추출된 표본을 이용한 모평균 추정량은 $\hat{\theta} = n^{-1}\sum_{i=1}^{n}(y_i/Np_i)$이다. 확률비례표본에 붓스트랩 방법을 적용하기 위해서는 재표본추출계획과 붓스트랩 추정량의 형태를 정해야 한다. 재표본추출계획은 단순임의추출과 확률비례추출을 고려해 볼 수 있으며, 추정량의 형태는 단순표본평균과 추출 확률의 역수에 의한 가중평균을 고려해 볼 수 있다.

재표본추출계획	추정량의 형태	
	단순표본평균	가중평균
단순임의추출	$\hat{\theta}_1^* = \frac{1}{n^*}\sum_{i=1}^{n^*} y_i^*$	$\hat{\theta}_2^* = \frac{1}{n^*}\sum_{i=1}^{n^*}(\frac{y_i}{Np_i})^*$
확률비례추출	$\hat{\theta}_3^* = \frac{1}{n^*}\sum_{i=1}^{n^*} y_i^*$	$\hat{\theta}_4^* = \frac{1}{n^*}\sum_{i=1}^{n^*}(\frac{y_i}{Np_i})^*$

위에서 고려한 4개의 추정량 중 두 번째 추정량 $\hat{\theta}_2^*$은 모평균에 대한 비편향 추정량임을 보일 수 있다. 그리고 $\hat{\theta}_2^*$의 분산추정량을 구하면

$$v_*(\hat{\theta}_2^*) = \frac{1}{n^* n}\sum_{i=1}^{n}(\frac{y_i}{Np_i} - \bar{y}_{pps})^2 \qquad (2)$$

이 되어, 재표본의 크기가 $n^* = n-1$ 이면 $\hat{\theta}_2^*$에 대한 붓스트랩 분산추정량은 분산을 비편향 추정하게 된다.

위에서 고려한 4개의 추정량 중 원표본추출방법을 그대로 재표본추출에 이용하고, 추정량의 형태도 그대로 동일하게 유지한 방법은 4번째 방법이다. 따라서 붓스트랩 방법을 복원 확률비례표본에 직접 적용하면 4번째 추정량, $\hat{\theta}_4^*$,을 이용하게 될 것이다. 그런데 네 번째 추정량

$\hat{\theta}_4^*$은 비편향 추정량이 아니며 분산추정량 또한 비편향추정량이 아니다. 모의실험 결과 붓스트랩 추정량 $\hat{\theta}_4^*$의 편의는 붓스트랩 추정량 $\hat{\theta}_2^*$에 비하여 다소 큰 반면, 분산추정량의 상대편의는 더 작게 나타났다.

3. 랜덤그룹표본에 붓스트랩 방법의 적용

랜덤그룹방법은 비복원 확률비례추출방법 중 비교적 간단한 경우로 표본의 크기가 n일 때 모집단 N개의 조사단위를 그룹의 크기가 $N_1, N_2, \cdots, N_n$인 n개의 랜덤그룹 $G_1, G_2, \cdots, G_n$으로 분할한 다음, 각 그룹에서 1개의 단위를 확률 p_g에 비례하여 추출한다. 여기서, $p_g = m_g/M_g$이고, $m_j = z_j/\sum_{j=1}^{N} z_j$, $M_g = \sum_{j \in G_g} m_j$ $(\sum M_g = 1)$이다. 모평균 추정량은

$$\bar{y}_{RHC} = \frac{1}{N}\sum_{g=1}^{n} M_g \frac{y_g}{m_g} = \frac{1}{N}\sum_{g=1}^{n} \frac{y_g}{p_g} \tag{3}$$

이며, 추정량의 분산 $V(\bar{y}_{RHC})$은 다음과 같다.

$$V(\bar{y}_{RHC}) = \frac{\sum N_g^2 - N}{N(N-1)} \sum_{i=1}^{N} \left(\frac{y_i}{Np_i} - \bar{Y}\right)^2 p_i \tag{4}$$

또한 분산에 대한 비편향 분산추정량은

$$v(\bar{y}_{RHC}) = \lambda^2 \sum_{g=1}^{n} M_g \left(\frac{y_{i_g}}{Nm_g} - \bar{y}_{RHC}\right)^2 \tag{5}$$

이고, 여기서 $\lambda^2 = (\sum N_g^2 - N)/(N^2 - \sum N_g^2)$이다.

랜덤그룹표본에 기초한 붓스트랩 방법으로 Rao & Wu (1988)가 제안한 재척도 방법이 있다. 재척도 방법의 붓스트랩 절차는 다음과 같다.

단계 1. $\{y_g, m_g\}_{g=1}^{n}$으로부터 $\{y_g^*, m_g^*\}_{g=1}^{n^*}$를 확률 M_g에 비례하여 복원 추출하고, $z_g^* = y_g^*/Nm_g^*$ 라 하자. 그리고 붓스트랩 추정량

$$\hat{\theta}^* = g(\tilde{\bar{y}}^*) \tag{6}$$

을 만든다. 여기서, $\tilde{\bar{y}}^* = \bar{y}_{RHC} + \lambda\sqrt{n^*}(\bar{z}^* - \bar{y}_{RHC})$, $\bar{z}^* = \sum z_g^*/n^*$ 이다.

단계 2. 단계 1을 B번 반복하여, $\hat{\theta}_1^*$, $\hat{\theta}_2^*, \cdots$, $\hat{\theta}_B^*$을 구하고 추정량의 분산은

$$v_* = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta}^*_{(\cdot)})^2 \tag{7}$$

로 구한다.

이렇게 구한 붓스트랩 추정량 $\hat{\bar{y}}^*$는 모평균의 비편향 추정량이 되고, 붓스트랩 분산추정량은 비편향 분산추정량이 된다.

Sitter(1992)는 앞서 설명한 재척도 방법과 달리 랜덤그룹표본의 추출효과를 붓스트랩 표본에 반영하는 반사매치 방법을 제안하였다. 반사매치 방법은 다음과 같다.

단계 1. $v_i = (nM_i/Nm_i)y_i$, $i = 1, 2, \cdots, n$ 라 하자.

단계 2. n개의 v_i들을 그룹의 크기가 $n_1, n_2, \cdots, n_{n^*}$인 n^* ($1 \le n^* < n$)개의 랜덤 그룹 $\Gamma_1^*, \Gamma_2^*, \cdots, \Gamma_{n^*}^*$ 으로 분할한다.

단계 3. 각 그룹에서 1개의 v_i를 확률 m_j^*/M_g^* ($M_g^* = \sum_{j \in \Gamma_g^*} m_j$)에 비례하여 추출하고, $v_1^*, v_2^*, \cdots, v_{n^*}^*$ 를 얻는다.

단계 4. 단계 2-3을 $k = (\sum_{g=1}^{n^*} n_g^2 - n)(N^2 - \sum_{g=1}^{n} N_g^2)/n(n-1)(\sum_{g=1}^{n} N_g^2 - N)$번 독립적으로 반복하여 붓스트랩 추정량

$$\hat{\theta}^* = \hat{\theta}(v_1^*, v_2^*, \cdots, v_{kn}^*)$$

을 만든다. 여기서, k가 정수가 아니면 확률화에 의해 값을 바꾸어 준다.

단계 5. 단계 2-4를 B번 반복하여 $\hat{\theta}_1^*$, $\hat{\theta}_2^*$, $\cdots$, $\hat{\theta}_B^*$를 구하고 추정량의 분산을

$$v_* = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta}^*_{(\cdot)})^2 \tag{8}$$

로 구한다.

반사매치 방법도 재척도 방법과 마찬가지로 모평균의 비편향 추정량이 되며, 분산을 비편향 추정한다.

4. 토의 및 결론

전통적인 붓스트랩 방법은 i.i.d. 표본에 기초하기 때문에 모집단의 분포와 표본추출방법을 구분할 필요가 없었다. 그런데 유한 모집단에서 복합표본은 유한모집단의 분포와 표본추출확률이 서로 다르기 때문에 전통적인 붓스트랩 방법을 적용하는데 어려움이 있다. 앞에서 살펴본 재척도 방법이나 반사매치 방법은 복원확률비례표본 및 랜덤그룹 표본에 이러한 어려움을 극복하는 방법들이다. 더 일반적인 비복원 확률비례표본에 붓스트랩 방법을 적용하기 위해서는 더 많은 숙고를 필요로 한다. 향 후 연구과제로 남겨둔다.

참 고 문 헌

[1] Bickel, P. J., and Freedman, D. A. (1984). Asymptotic normality and the Bootstrap in Stratified sampling. Annals of Statistics, 12, 470-482

[2] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. Annals of Statistics, 9, 1196-1217.

[3] Booth, J.G. Bulter, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. Journal of the American Statistical Association, 89, 1282-1289.

[4] Chao, M.T. and Lo, S.H. (1994). Maximum likelihood summary and the bootstrap method in structured finite populations. Statistica Sinica, 4, 389-406.

[5] Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1-26.

[6] Efron, B. (1994). Missing data, imputation and the bootstrap. Journal of the American Statistical Association, 89, 463-479.

[7] Kovar, J. G., Rao, J.N.K., and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. Canadian Journal of Statistics, 16, Supplement, 25-45.

[8] Krewski, D., and Rao, J.N.K. (1981). Inference from stratifed samples: Properties of the linearization, jackknife, and balanced repeated replication methods. Annals of Statistics, 9, 1010-1019.

[9] Nigam, A.K. and Rao, J.N.K. (1996). On balanced bootstrap for stratified multidtage samples. Statistica Sinica, 6, 199-214.

[10] Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data. Journal of the American Statistical Association, 83, 231-241.

[11] Rao, J.N.K., Hartley, H. O., and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. Journal of the Royal Statistical Society, B24, 482-491.

[12] Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. Survey Methodology, 18, 209-217.

[13] Shao, J. and Chen, Y. (1998). Bootstrapping sample quantiles based on complex survey data under hot deck imputation. Technical report.

[14] Shao, J. and Sitter, R.R. (1996). Bootsrap for imputed survey data. Journal of the American Statistical Association, 91, 1278-1287.

[15] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. Annals of Statistics, 9, 1187-1195.

[16] Sitter, R.R. (1992). A resampling procedures for complex survey data. Journal of the American Statistical Association, 87, 755-765.

[17] Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. Canadian Journal of Statistics, 20, 135-154.