응용통계연구 제10권 2호
1997년 9월 pp. 253-260

모형에 기초한 표본추출방법의 알고리듬

강 명 욱[1], 김 영 일[2]

요 약

D-최적과 최소평균제곱오차를 기준으로 모형에 기초한 표본추출에 대한 여러 가지 알고리듬을 연구하였다. 이 두 기준은 서로 다른 관점에서 출발하였지만 기본적으로 같은 취지를 가지고 있어 거의 유사한 표본을 제공한다. 표본대상 개체의 추출비용이 서로 다른 경우를 포함한 간단한 예를 통해 이를 살펴보았으며 향후 연구과제에 대해 언급하였다.

1. 서 론

Royall(1970)이 유한모집단에서 적절한 규모의 표본을 추출하여 효율적인 추정량을 구하는 방법을 제시한 이후로 모형에 기초한 표본조사론(model based sampling)에 실험계획법의 최적화이론(optimal design theory)을 접목시키는 노력이 Wynn(1977)에 의해 시도되었다. 그리고 이를 오차 함수에 대한 정확한 정보가 주어지지 않는 회계감사의 예에 응용시킨 논문을 김영일(1993)이 발표하였다.

본 논문에서는 Royall이 발표한 평균제곱오차(mean square error)를 최소화하여 주는 균형표본(balanced sample)과 실험계획법에서 가장 많이 인용되는 일반화분산(generalized variance)을 줄이는 D-최적이론의 유사성을 알고리듬의 차원에서 살펴본다. 더 나아가 유한모집단 내에 있는 개개의 표본대상 개체의 표본추출비용이 다를 때의 문제점을 짚어 보고, 기존의 알고리듬에 Fedorov(1992a, b)가 발표한 제한-최적화이론(optimization theory with constraint)을 결부시켜 제한된 비용 하에서 주어진 목적함수를 달성하는 표본을 구하는 알고리듬을 개발한다. 다음절에서는 이를 위한 간단한 내용 전개 및 표현방법을 소개한다.

2. 표현방법

k개의 측정값으로 구성되어 있는 N개의 개체로 이루어진 유한모집단 S를 가정하자. 이러한 모집단은 $N\times k$의 행렬 X로 구성된다. 그리고 각각의 개체에는 y_i, $i=1,2,\cdots,N$라는 표본이 수집되기 전까지는 분석가에게 값이 알려지지 않는 반응변수가 있다. 물론 표본이

1) (140-742) 서울특별시 용산구 청파동 2가 53-12 숙명여자대학교 통계학과 부교수.
2) (456-756) 경기도 안성군 대덕면 내리 40-1 중앙대학교 산업정보학과 부교수.

수집되면 y_i 값은 분석가에게 알려진다. 한 예로 회계감사에서 회계장부 참값(true book value)이 y_i 라면 이는 회계감사 전에는 알려져 있지 않을 것이다. 본 논문에서는 y_i 값은 X 와 다음과 같이 선형모형으로 이루어져 있다고 가정한다.

$$E(y) = X\beta$$

그리고 $Var(y_i) = \sigma^2$, $Cov(y_i, y_j) = 0$ 이라고 가정한다. 여기서 분석가의 목적은 주어진 예산 범위 내에서 혹은 주어진 표본크기 n에서, 주어진 목적을 최대한 만족시키는 S의 부분집합 $s \subset S$을 구하는 것이다.

y_i의 선형결합인 모수 $\tau = \sum_{i=1}^{N} c_i y_i$의 불편추정량 T를 고려하여 보자. T의 형태는 다음과 같다.

$$T = \sum_{i \in s} c_i y_i + \sum_{i \in s^c} c_i \hat{y}_i \qquad (1)$$

여기서 s^c는 비표본집단을 의미한다. $s \subset S$을 구하는 기준은 여러 표본 중에서도 식 (1)에 의해서 구한 추정량 T의 평균제곱오차 $E(T-\tau)^2$의 값이 가장 작은 표본을 구하는 것이다. 만약 모든 i에 대하여 $c_i = 1$이면 $E(T-\tau)^2$은 다음과 같이 전개된다.

$$E(T-\tau)^2 = \sum_{s^c} \sigma^2 + \sigma^2 \sum_{s^c} x_i{}^T (X_{s_n}^T X_{s_n})^{-1} x_i \qquad (2)$$

여기서 s_n은 크기가 n인 표본을 의미하고, $X_{s_n}^T X_{s_n}$은 표본추출된 개체로 이루어진 정보행렬(information matrix)이다. 식 (2)의 우변중 우측에 있는 것은 비표본개체 y_i의 예측값의 분산의 합이다. 따라서 등분산 경우는 우측에 있는 요소의 값만 고려하여 최적의 표본을 추출할 수 있다. 물론 y_i의 분산이 각각 다른 경우는 두 항을 모두 고려하여 표본을 추출하여야 한다(김영일, 1993). Royall은 식 (2)를 최소화시켜 주는 표본을 균형표본이라 불렀다. 이러한 균형표본의 특징은 측정값의 표본평균 k개의 값이 모집단의 해당하는 평균과 같다는 점을 들 수 있다. 이러한 표본의 특징에 대해서는 Royall과 Herson(1973)의 논문을 참조 바란다.

Wynn(1977)은 최적실험계획의 이론을 표본조사론에 적용시켜 보았는데 그는 모든 비표본집단의 개체에 대해서 $x_i{}^T (X_{s_n}^T X_{s_n})^{-1} x_i$의 최대값이 최소가 되는 표본추출을 고려하여 보았다. 즉, 최적실험계획법의 G-최적이론을 본 뜬 기준을 제시하고, 이에 대한 알고리듬을 제시하였다. 즉, $X_{s_n}^T X_{s_n}$의 역행렬이 존재하도록 크기가 n인 표본을 임의로 추출한 다음 $X_{s_n}^T X_{s_n}$의 행렬식의 값이 가장 크게 되게끔 매 단계에서 표본 내의 개체와 비표본된 개체의 교환을 통한 알고리듬(exchange algorithm)을 발표하였다. 이러한 배경으로는 연속최적화이론(continuous optimal design theory)에서 나온 Kiefer와 Wolfowitz(1959)의 D-최적과 G-최적의 동격이론(equivalence theory)을 들 수 있다. Wynn에 의하면 비록 $x_i{}^T (X_{s_n}^T X_{s_n})^{-1} x_i$의 최대값을 최소화하는 기준이나 $X_{s_n}^T X_{s_n}$의 행렬식의 값을 최대화하는 기준은 표본추출문제에서는 동격일 수는 없으나 이러한 알고리듬에 의하여 준 최적(near optimal)의 표본을 얻을 수 있다고 하였

다. 다음절에서는 기존 알고리듬과는 다른 관점에서 $X_{s_n}^T X_{s_n}$의 행렬식을 크게 하는 표본추출의 기준과 Royall이 제시한 기준의 유사성을 밝혀본다.

3. 알고리듬

모집단의 크기 N이 유한이기 때문에 목적함수를 최소화하거나 최대화하는 문제는 열거(enumeration)법을 통하면 해결할 수 있다. 그러나 Wynn은 계산상의 문제로 다음과 같은 역행렬 및 행렬식의 성질을 규정하는 공식에 의해 개체의 교환(추가 및 제거)을 이루었다. 이러한 Wynn의 알고리듬은 Mitchell과 Miller(1970)의 교환법칙에 의해서 좀 더 단순화 될 수 있다. 식 (3)과 (4)를 참조하면, 교환은 비표본집단에서 $x^T D_n x$의 최대값에 해당되는 개체를 추가하고, $x^T D_{n+1} x$의 최소값에 해당되는 개체를 제거함으로서 이루어진다. 여기서 D_n은 $X_{s_n}^T X_{s_n}$의 역행렬을 의미하고 $X_{s_n}^T X_{s_n}$은 M_n, 그리고 이의 행렬식은 $|M_n|$으로 표기한다.

$$|M_{n+1}| = |M_n|(1 + x^T D_n x) \tag{3}$$

$$|M_n| = |M_{n+1}|(1 - x^T D_{n+1} x) \tag{4}$$

참고로 D_{n+1}은 다음과 같은 공식에 의해 구할 수 있다.

$$D_{n+1} = D_n - \frac{D_n x x^T D_n}{1 + x^T D_n x}$$

여기서 D_{n+1}은 각각 새로운 개체인 크기 $1 \times k$인 x^T가 표본에 추가되었을 때의 $X_{s_{n+1}}^T X_{s_{n+1}}$의 역행렬을 의미한다. 이러한 알고리듬은 한 번에 여러 개체의 교환이 이루어지도록 하는 Fedorov(1972)나 Johnson과 Nachtsheim (1983)의 알고리듬을 통해 향상될 수 있을 것이다.

그러나 일반적인 실험계획법과는 달리 표본추출에서는 알고리듬의 첫 단계에서 주어진 표본의 크기 n을 고집할 필요는 없다. $X_{s_n}^T X_{s_n}$의 역행렬이 존재하는 한 n보다 작은 크기가 n_0인 표본에서 출발하여도 원하는 표본을 추출할 수 있다. 이러한 방법을 전진선택(forward selection)방법이라 하자. 다만 이러한 단계에서는 최초의 크기가 n_0인 표본을 정하는 문제가 대두되는데 경험적으로 D-최적의 성질 즉, 주어진 실험영역 혹은 모집단의 극단에 위치한 개체들을 중심으로 표본이 구성되는 성질을 이용하여 정할 수 있을 것이다. 한편 알고리듬의 첫 단계에서 표본크기를 모집단의 크기인 N으로 정하여 제거를 통한 방법으로 크기 n인 표본을 추출할 수 있다. 이러한 방법을 후진제거법(backward elimination)이라 하자. 또한 이 방법은 N에 비하여 n이 그렇게 작지 않다면 오히려 전진선택방법보다는 계산 부담이 줄 것이다. 이 방법에 의해 행렬식의 값이 제일 큰 표본을 추출하는 방법은 다음과 같다.

단계 1) 모집단에 있는 모든 개체를 표본에 포함시켜 $D_N=(X_{s_N}^T X_{s_N})^{-1}$을 구한다.

단계 2) 식 (5)에 의거 $x^T D_N x$이 최소가 되는 개체를 표본에서 제거한다.

단계 3) 이와 같은 방법으로 계속 제거하여 표본크기 n이 달성될 수 있도록 한다.

참고로 단계 3)에서 역행렬의 계산은 식 (5)에 의해 수행되어 진다.

$$D_{n-1}=D_n+\frac{D_n x x^T D_n}{1-x^T D_n x} \tag{5}$$

문제가 가지고 있는 이산(discrete)적인 성질상 제시된 세 가지 방법에 의한 표본추출방법은 동일한 표본을 보장하지는 않는다. 모형에 의거한 표본추출방법에서 발생하는 국지-최적(local optimum)을 피하는 방법으로는 Wynn이 제시한 방법과 전진선택법 그리고 후진제거법 등을 혼합 구성하여 최적의 표본을 추출할 수 있으나 경험상 큰 장점은 부각되지 않는다.

이러한 후진제거법을 균형표본을 구하는 방법에 적용시켜 보면 균형표본과 D-최적과의 유사성이 부각된다. 만약 모든 유한모집단의 개체가 전부 표본에 포함될 경우는 식 (2)의 값은 0이다. 이를 출발점으로 한 번에 하나씩 개체를 제거하는 방법을 고려하여 보자. 후진제거법에서 다음 단계인 크기가 $N-1$인 표본은 당연히 $\min_{i\in S} x_i^T(X_{s_{N(i)}}^T X_{s_{N(i)}})^{-1}x_i$에 해당되는 개체가 제거되어야 평균제곱오차가 제일 작을 것이다. 여기서 $D_{N(i)}=(X_{s_{N(i)}}^T X_{s_{N(i)}})^{-1}$은 i번째 개체가 제거된 정보행렬의 역행렬이다. 물론 다음 단계에서도 이와 같은 방법으로 해당되는 개체를 제거하며 표본의 크기가 n이 될 때까지 반복한다. 후진제거법의 단계 2)에서 나온 $x^T D_N x$와 비교하면 이는 단지 이차형식의 가운데 들어간 정보행렬이 i번째 개체를 포함하고 있느냐 아니면 포함하고 있지 않느냐의 차이이지 그 근본적인 형태는 같다. 이러한 맥락에서 표본정보행렬의 행렬식을 최대화하는 기준과 균형표본을 이루는 기준은 적어도 y_i의 등분산 가정 하에서는 비슷한 성질을 가지고 있다고 보아야 한다. 물론 두 기준은 유한모집단의 특성에 따라 각기 다른 표본을 구성할 수 있으나 경험적으로 이로 인한 차이는 거의 없다고 보아진다. 그러나 분산이 다를 경우는 식 (2)의 우변의 첫째항의 값에 따라서 이를 보장할 수는 없다.

참고로 $D_{N(i)}$과 D_N의 차이는 식 (5)에서 찾아 볼 수 있다. 균형표본인 경우에 발생하는 $\min_{i\in S} x_i^T(X_{s_{N(i)}}^T X_{s_{N(i)}})^{-1}x_i$은 $\min_{i\in S}(x_i^T D_N x_i + x_i^T B x_i)$로 표현된다. 여기서 B는 $(D_N x_i x_i^T D_N)/(1-x_i^T D_N x_i)$이다. 개략적으로 설명하면 행렬 B의 분모 중 $x_i^T D_N x_i$의 값이 작은 값일수록 분자가 일정하다는 가정 하에서는 $x_i^T D_N x_i + x_i^T B x_i$의 값은 작아질 것이다. 이러한 의미에서 두 기준은 거의 유사한 표본을 구성할 수 있다고 보여진다. 다음절에서는 모집단의 개체를 추출하는 비용이 각각 다른 경우 제한된 자원인 예산범위 내에서 표본을 구성하는 방법을 고려하여 본다. 이는 Fedorov(1992a)가 제시한 제한-최적실험계획법에 의거하여 구성한 것이다.

4. 추출비용이 다른 경우의 알고리듬

지금까지는 묵시적으로 개개의 개체의 추출비용이 같은 경우를 가정하였다. 그러나 실제적인 표본추출작업에 이러한 가정이 반영되기에는 문제점을 가지고 있다고 보여진다. 한 예로 1절에서 언급한 회계감사의 경우만 하더라도 회사의 규모가 큰 경우와 작은 경우는 표본추출비용이 같다는 가정을 할 수 없을 것이다. 개개의 자료점에서 추출비용을 포함한 손실함수를 일반화하여 Fedorov(1992a)는 $\zeta(x_i)$라 가정하고, 최적실험계획 ξ^*를 구하는 이론적인 틀을 다음과 같이 제시하였다.

$$\xi^* = \text{Arg} \ \min_{\xi} \Psi[M(\xi)]$$

여기서

$$\int \xi(dx) = 1 \tag{6}$$

$$C(\xi) = \int \phi(x)\xi(dx) \leq 0 \tag{7}$$

이다. 이산형태로 말하면 식 (6)은 각 자료점에 반복되는 관측값의 수 r_i의 합, $\sum r_i$가 N임을 의미하고 식 (7)은 C를 표본추출에 배정된 총비용이라고 할 때 $\sum r_i\zeta(x_i) \leq C$의 정규화된 표현이다. 또한 $\phi(x) = \zeta(x) - C/N$을 의미한다. 주어진 목적함수상의 $\Psi[M(\xi)]$은 D-최적인 경우에는 $-\log|M(\xi)|$로 변하고, $M(\xi) = N^{-1}\sum p_i f(x_i) f(x_i)^T$, $p_i = r_i/N$이다. Fedorov(1992a)는 식 (6)을 무시할 수 있고 $\xi'(dx) = \phi(x)\xi(dx)$로 치환한다면 주어진 모형 $y = f(x)^T\beta + \varepsilon$의 $f(x)$를 $f'(x) = \phi^{-1/2}(x)f(x)$로 치환하여 종래의 알려진 알고리듬을 구현할 수 있다고 하였다.

일반적으로 실험계획법에서는 주어진 제약조건에서 (x_i, p_i)을 결정하는 실험계획 ξ을 구하는 것은 어려운 문제이다. 왜냐하면 x_i은 사전에 알려져 있지 않고 실험영역에서 찾아야 하기 때문이다. 제약조건에 따른 실험계획 ξ을 구하는 문제는 Fedorov(1992b)의 논문을 참조하기 바란다. 그러나 본 논문에서 논한 표본추출에서는 보조변수역할을 하는 x_i의 값은 이미 알려져 있다고 가정하였기 때문에 이에 대한 해결이 비교적 용이하다. Fedorov(1992a)가 언급한 치환의 이점을 살리면 주어진 비용 하에서 유한모집단에서 주어진 목적함수를 달성시키는 표본을 추출할 수 있다.

각각의 자료점의 표본추출비용을 $\zeta(x_i)$으로 가정한다면 3절에서 제시된 알고리듬을 쉽게 변경할 수 있다. 즉, 식 (8)과 같은 단위비용당(per unit cost) 예측분산의 값의 감소를 최소화하는 자료점을 파악하여 제거할 수 있다.

$$\zeta^{-1}(x_i)\, x_i^T D_N x_i \qquad \text{혹은} \qquad \zeta^{-1}(x_i)\, x_i^T D_{M(i)} x_i \tag{8}$$

여기서는 후진제거법의 최초 출발점을 기준으로 표기하였다. 이러한 기준을 설정함으로서 자료점에서 발생하는 비용을 감안, 주어진 예산범위 내에서 최적의 표본을 추출할 수 있다. 즉,

식 (8)을 최소화하는 자료점을 하나씩 제거하면서 총비용을 초과하기 바로 전에 표본추출작업을 멈추면 된다. 다음절에서는 이에 대한 간단한 예를 보고 결론에서는 본 연구결과에 따르는 문제점 및 향후 연구과제에 대한 논의한다.

5. 예 제

단순선형회귀모형을 기준으로 먼저 표본추출비용이 같은 경우를 고려하여 본다. 이 모형은 $f(x)^T=(1, x)$을 의미한다. 가상적인 자료로 x_i가 1부터 15까지의 정수값으로 이루어 졌다고 가정하자. 등분산을 가정한다면 균형표본은 2절에서 언급하였듯이 표본평균값이 전체 모집단의 평균값 8과 같은 개체로 구성될 것이다. 물론 이러한 균형표본은 유일하게 결정되지 않는다. 표본크기 $n=4$일 경우, 알고리듬에 의하면 다음과 같은 개체가 D-최적의 표본으로 결정되었다.

$$s=\{1,2,14,15\}$$

이 표본의 평균값은 8이므로 당연히 균형표본의 하나로 보아야 한다.

각 자료점의 표본추출비용이 다른 경우를 고려하여 본다. 정수 1부터 10까지의 자료점에 해당되는 추출비용은 자료값과 같은 값을 가지며 11부터는 동일한 10의 비용을 가지고 있다고 하자. 그리고 표본추출에 배정된 총 예산은 30으로 가정하면, 다음과 같은 표본이 알고리듬에 의해 추출된다.

$$s=\{1,2,3,4,14,15\}$$

이 경우 역시 D-최적과 최소평균제곱오차의 기준은 같은 결과를 가져다주었다. 물론 예산상의 제약조건상 이러한 표본의 표본평균값은 모집단의 평균값인 8이 아님을 알 수 있다. 3절에서 언급하였듯이, 항상 D-최적과 균형표본이 같은 결과를 초래한다고 볼 수는 없다. 경험적으로 3절에서 제시된 3가지 알고리듬을 혼합하여 쓰는 것이 최적은 아니라 하더라도 준 최적에 해당하는 표본을 추출한다고 볼 수 있다.

위의 경우는 표본의 총 추출비용이 정해져 있는 경우였으나 때에 따라서는 원하는 목적함수의 값의 상한선 혹은 하한선을 정한 후에 이를 달성하기 위한 최소비용의 계산문제도 고려하여 볼 수 있다. 그러나 이 문제에 대한 해결은 용이하다. 알고리듬을 전개하는 동안 변하는 목적함수의 값을 추적하여 알고리듬을 수정할 수 있을 것이다. 그리고 또한 본 논문에서 밝힌 D-최적이나 균형표본의 기준만이 아니라 실험계획법에서 나온 여러 가지 최적이론의 기준들(A-최적, D_s-최적 등)을 가지고 목적함수를 구성할 수 있을 것이다.

6. 결 론

본 논문에서는 실험계획이론에서 나온 기준의 하나인 D-최적을 중심으로 균형표본과의 관계를 알고리듬의 차원에서 언급하였고 더 나아가 추출비용이 다른 경우, 이를 표본추출과정에 감안하는 방안을 살펴보았다. 물론 이러한 표본추출방법은 모형과 가정이 맞다는 전제에서 이루어져야 한다. 만약 이러한 가정의 일부가 불확실한 경우에는 특정한 표본이 가지는 로버스트한 문제점들을 감안하여 표본추출을 하여야 할 것이다. 위에서 예제로 제시된 단순선형회귀모형에서 보았듯이 양극점에 위치하고 있는 자료점만을 추출한다는 것은 최적실험계획의 이론의 문제점과 같은 논의를 불러올 수 있다. 그러나 이러한 표본추출은 좀 더 정확한 추론의 근거를 마련하여 주는 이론적인 틀이 된다고 보여진다.

참 고 문 헌

[1] 김영일(1993), Error-Robust Model Based Sampling in Accounting, 응용통계연구, 6권 제1호, 29-40.

[2] Fedorov, V. (1972), *Theory of Optimal Experiments,* Academic Press, New York.

[3] Fedorov, V. (1992a), Optimal Design Construction With Constraints I, Technical Report No. 573, School of Statistics, University of Minnesota.

[4] Fedorov, V. (1992b), Optimal Design Construction With Constraints II, Technical Report No. 576, School of Statistics, University of Minnesota.

[5] Kiefer, J. and Wolfowitz, J. (1959), Optimum Designs in Regression Problems, *Annals of Mathematics*, 30, 271-294.

[6] Johnson, M. E. and Nachtsheim, C. J. (1983), Some guidelines for Constructing Exact D-optimal Designs on Convex Spaces, *Technometrics*, 25, 271-277.

[7] Mitchell, T. J. and Miller, F. L. (1970), Use of Design Repair to Construct for Special Linear Models, Math. Div. Ann. Progr. Report (ORNL-4661), Oak Ridge National Laboratory, Oak Ridge, TN, 130-131.

[8] Royall, R. M. (1970), On Finite Population Sampling Theory Under Certain Linear Regression Models, *Biometrika*, 57, 377-387.

[9] Royall, R. M. and Herson, J. (1973), Robust Estimation in Finite Populations I, II, *Journal of the American Statistical Association*, 68, 880-893.

[10] Wynn, H. P. (1977), Minimax Purposive Survey Sampling Design, *Journal of the American Statistical Association*, 72, 655-657.

On the Algorithm of Constructing the Model-Based Optimal Sample

Myung-Wook Kahng[1], Young-Il Kim[2]

Abstract

Various algorithms are investigated with respect to finding the best model-based samples according to criteria such as D-optimality and minimum mean square error. These two criteria are slightly different, but related to each other. Therefore, it is not surprising that these two are producing the almost identical samples. Some simple examples follow and critiques are provided along with directions for further research.

1) Associate Professor, Department of Statistics, Sookmyung Women's University, Yongsan-Ku, Seoul, 140-742, Korea.

2) Associate Professor, Department of Industrial Information, ChungAng University, Ahnsung-Kun, Kyunggi-Do, 456-756, Korea.