층화에서 최적경계점 결정에 관한 연구

박진우 1), 김영원2)

요 약

층화 추출법에서 층의 경계점을 정하는 문제는 추정의 효율에 직접적으로 영향을 미치기 때문에 매우 실제적이고 중요한 문제이다. 층화변수가 일변량 연속변수인 경우 널리 알려진 방법으로는 누적도수제곱근법과 Ekman법이 있는데 이 두 방법은 모두 나름의 약점을 지니고 있다. 본 논문에서는 Breiman 등(1984)이 제시한 CART 기법 중 회귀나무(regression tree)모형을 이용하여 층의 경계점을 정하는 방법을 소개한다. 그리고 통계청의 어업총조사 자료를 사용하여 층의 경계점을 정하는 여러 다른 방법들의 효율을 비교한다.

주요용어 : 최적경계점, 누적도수제곱근법, Ekman법, 회귀나무모형

1. 서 론

효율적인 표본설계를 위해 중요한 작업 중 하나로 효과적인 층화 작업을 들 수 있다. 실제 조사하고자 하는 관심변수의 값을 모르는 상태에서 효과적인 층화를 하기 위해서는 관심변수와 큰 상관관계를 갖는 층화변수를 찾아야 한다. 일단 층화변수가 선정되고 나면 H개의 층을 형성하기 위해 $H-1$개의 경계점을 정하여야 한다. 층화변수가 범주형일 때 층의 경계점을 정하는 문제는 비교적 단순하지만 연속형일 때에는 여러 가지 고려할 점들이 생긴다. 연속형 층화변수일 경우 층의 수, 층별 표본수가 미리 결정되었을 때 분산을 최소로 하는 층의 경계점을 최적경계점(optimum point of stratification ; ops)이라고 한다(박홍래, 2000).

층화의 최적 경계점을 찾는 문제는 Dalenius (1950)에 의해 처음 연구된 이래 여러 연구자들에 의해 논의가 되어왔는데 가장 널리 알려진 방법으로 Dalenius와 Hodges (1959)의 누적도수제곱근법(일명 누적 $\sqrt{f}$ 법)과 Ekman (1959)의 방법 등이 있다.

Dalenius-Hodges법은 Sarndal 등 (1992)의 책에 잘 설명되어 있는데 먼저 추출틀의 단위들을 층화변수의 크기순으로 정렬한 후 그것을 J개의 구간으로 나누어 정리한다. 여기서 구간수 J는 충분히 큰 수여야 한다. j 번째 구간에 속하는 모집단 단위들의 수를 f_j라고 하자. 먼저 $\sqrt{f_j}$ $(j=1,2,\cdots,J)$ 값을 계산한 후 인접한 구간들을 적절히 통합하여 전체 H개의 층으로 그

1) Associate Professor, Department of Applied Statistics, The University of Suwon.
E-mail: jwpark@suwon.ac.kr
2) Professor, Department of Statistics, Sookmyung Women's University.
E-mail: ywkim@sookmyung.ac.kr

룹화시키는데 이 때 각 층에 속하는 구간들의 누적 $\sqrt{f}$ 값이 가능한 한 같도록 해주면 된다. 즉, 가능한 한 다음의 식을 만족시키도록 층화를 시키면 최적에 근접하는 층화를 할 수 있다:

$$\sum_{j=1}^{J_1}\sqrt{f_j}=\sum_{j=J_1+1}^{J_2}\sqrt{f_j}=\cdots=\sum_{j=J_{H-1}+1}^{J}\sqrt{f_j} \quad (1)$$

Dalenius-Hodges법은 먼저 J 값을 정한 후 적절한 $J_1<J_2<\cdots<J_{H-1}$을 찾는 것이라고 할 수 있다. Dalenius-Hodges법의 문제점으로 지적되는 사항은 첫째 최적의 J 값을 어떻게 선택할 수 있는가에 대해 아무런 단서가 없고, J값이 달라짐에 따라 결과가 달라진다는 점이다. 다시 말해 어느 정도 임의적으로 하는 것을 허용할 수밖에 없는 방법이다. 둘째는 이 방법을 구현하기 위한 프로그램을 작성하기가 무척 어렵다는 사실이다. Sweet과 Sigman (1995)는 사용자에게 미리 J값을 지정하게 할 때 Dalenius-Hodges 방법을 SAS 프로그램으로 구현하였다.

Ekman 법은 층의 경계값인 $b_1, b_2, \cdots, b_{H-1}$을 정할 때 가능한 한 아래의 식을 만족시킬 수 있도록 해주는 방법이다:

$$N_1(b_1-b_0)=N_2(b_2-b_1)=\cdots=N_H(b_H-b_{H-1}) \quad (2)$$

여기서 N_h는 h $(h=1,2,\cdots,H)$번째 층에 속하는 모집단 단위의 수를 나타낸다. Ekman 법은 Dalenius-Hodges 법과 마찬가지로 정확한 해답을 찾기가 어려운 경우가 대부분이다. 아울러 이 방법도 수치해석적인 알고리즘의 도움 없이는 계산이 어렵다.

Cochran (1961), Sethi (1966), Murthy (1967) 등은 여러 방법들의 효과를 실증적으로 고찰하는 연구를 수행한 바 있는데 대체로 Ekman 법과 Dalenius-Hodges 법의 효율이 높은 것을 알 수 있다. 한편 Hedlin (2000)은 확장된 Ekman 방법을 제안하고 그 계산을 위한 알고리즘을 제시하였다.

본 연구에서는 일변량 연속형 층화변수를 사용하는 경우 최적경계점을 정하는 문제를 해결하기 위해 결정나무(decision tree) 분석기법 중 하나인 CART(Classification and Regression Tree; Breimann 등, 1984)를 사용하는 방법을 소개한다. CART 기법의 경우 SAS E-miner를 비롯한 데이터마이닝 기법을 다루는 여러 소프트웨어에서 지원이 되므로 사용자가 손쉽게 층화작업을 할 수 있다는 점에서 효과적이다. 한편 통계청에서 실시한 2000 어업총조사 자료(통계청, 2001)를 이용하여 SAS E-miner에서 제공하는 회귀나무(regression tree) 방법을 사용한 층의 경계점을 정하는 방법과 기타 다른 방법의 효율성을 실증적으로 비교 분석한다.

2절에서는 회귀나무모형을 사용한 층의 경계점을 정하는 방법에 대해 구체적으로 소개하고, 3절에서는 어업총조사 예를 이용하여 여러 층 경계를 정하는 기법들을 사용할 경우의 효율을 비교한다. 마지막으로 4절에서는 전체적인 연구결과에 대해 결론을 내린다.

2. 회귀나무모형을 이용한 층 경계점 결정

최근 다양한 통계분석분야에서 결정나무모형(decision tree model)이 폭 넓게 활용되고 있다. 이 방법은 판별분류분석, 신경망분석, 로지스틱분석 등과는 달리 연구자가 목표변수(target variable)와 입력변수(input variable)의 구조적인 관계(structural relationship)를 쉽게 이해하고 설명할 수 있다는 장점을 갖고 있다. 또한 기존의 분석기법들은 변수의 형태뿐만 아니라 정규성 및 등분산성 등에 대한 가정에 있어서 상당히 제한적이지만, 결정나무모형은 비모수적인 방법이기 때문에 목표변수 및 입력변수 모두에 있어서 연속형 또는 범주형 변수들이 아무런 제한

없이 사용될 수 있고 동시에 정규성 등 분포에 대한 가정이 필요 없다(강형철 등, 1999).

효율적인 표본설계를 의한 추출단위들의 층화에 있어서 주요 관심대상은 일반적으로 관심변수는 연속형 변수이고, 층화변수는 범주형 또는 연속형 변수인 경우일 것이다. 1절에서도 언급한 것과 같이 특히 층화변수가 연속형인 경우 적절한 층화를 위한 층의 개수의 선택 및 최적층화를 위한 층의 경계점을 결정하는 문제는 여러 연구자들에 위해 계속 연구되고 있다. 일반적인 층화작업에서 관심변수는 연속형 변수라는 점을 고려할 때, CHAID, CART, C4.5 등 다양한 형태의 결정나무모형을 위한 알고리즘 중에서 층화작업에 활용할 수 있는 기법은 Breiman 등(1984)이 제시한 CART (classification and regression trees)기법 중 회귀나무(regression tree) 모형이다.

회귀나무모형은 최소제곱회귀모형에서 최적의 나무구조를 찾아내기 위해 Morgan과 Sonquist(1963)가 개발한 AID(Automatic Interaction Detection) 프로그램에서 출발되었다. 이 방법은 후에 Sonquist 등(1973)의 SEARCH 알고리즘으로 발전하였고, Brieman 등(1984)이 제시한 회귀나무모형은 이런 기법을 일부 수정 보완한 것으로 볼 수 있다. 이런 기법과 관련해서 표본조사분야에서는 Kalton (1983)이 SEARCH 알고리즘을 활용하여 결측값 대체를 위한 대체층(imputation cell)의 구성을 위해 활용하였고, 이를 바탕으로 Ryu 등 (2000)은 2000년 우리나라 인구주택총조사의 결측값 대체를 위한 대체층 구성을 위해 회귀나무모형을 활용하고 있다.

회귀나무모형을 추출단위의 층화를 위해 적용하는 경우, 어떤 측면에서 효율적인지 살펴볼 필요가 있다. 우선 일반적인 결정나무모형에서는 대부분 목표변수가 범주형인 경우를 가정하고 있고, 주요 관심대상이 판별분류분석에 있다. 이 경우 결정나무모형의 분리기준으로 카이제곱 통계량의 p값, 지니 지수, 엔트로피 지수 등을 사용하고 있으며, 이런 기준들은 표본설계를 위한 층화작업에 적용하기에는 적합하지 않다. 하지만 연속형 목표변수를 가정하고 있는 회귀나무모형의 경우, SAS E-miner에서는 마디(node) 분리기준으로 F-통계량 또는 분산감소량(variance reduction)을 사용할 수 있다. 이런 분리기준을 적용하면 결과적으로 각 마디에 속하는 목표변수의 평균제곱오차, 즉 각 마디내 분산의 최소화를 기준으로 순차적인 이지분리(binary split)를 하게 된다. 결과적으로 회귀나무기법을 통해 각 층(마디)내 분산을 가능한 최소화하도록 추출단위들을 동질적인 것끼리 순차적으로 층화(분류)하는 것과 동일한 층화결과를 얻을 수 있다.

둘째, 층화작업에서는 적절한 층의 개수를 결정하는 문제를 해결해야 한다. 회귀나무모형을 층화를 위해 적용하는 경우, 먼저 순차적인 이지분리를 통해 일단 최대 크기의 나무구조를 형성한 후에 적절한 수준의 가치치기(pruning)을 통해 단순화된 최종 회귀나무모형을 구성하게 된다. 이 과정에서 표본설계자의 판단에 따라 층의 크기 및 층내분산의 감소량 등을 감안하여 가치치기 작업을 수행하는 경우 표본설계자의 의도를 반영한 층의 개수 결정 및 구체적인 층화작업이 손쉽게 수행될 수 있다.

셋째, 표본추출에서 각 층은 어떤 의미를 갖고 구성되었는지 설명이 가능한 것이 바람직하다. 회귀나무모형은 이런 관점에서 기존의 다른 판별분류기법에 비해 장점을 지니고 있다. 특히 여러개의 층화변수를 동시에 사용하는 경우 기존의 판별분류분석을 적용하는 경우 분류기준은 층화변수들이 선형결합형태로 나타나게 되는 데, 이런 분류기준은 알기 쉽게 해석하는 데는 많은 어려움이 있다.

넷째, 비록 본 연구에서는 하나의 층화변수를 선택한 상태에서 이 층화변수에 의한 최적 층화 방법에 대해 비교연구를 수행하고 있지만, 실제 층화작업에서는 범주형, 연속형 등 다양한 형태의 많은 변수 중 가장 최적의 층화변수를 선택하는 작업이 선행되고 동시에 이 변수를 기준으로 층화 경계점을 결정하여 층화작업을 수행해야 한다. 회귀나무기법을 사용하게 되면 자

연스럽게 층내 분산이 최소화되도록 층화효과를 가장 극대화 할 수 있는 층화변수를 선택하여 층화를 하게 되고, 순차적으로 분리된 층에서 다시 주요변수를 선택하여 층을 세분화하게 됨으로 각 과정에서 주요 층화변수의 선택 및 이에 따른 층화작업이 동시에 이루어 질 수 있다.

마지막으로 층화를 위한 어떤 알고리즘이 효과적이라고 해도 실제 표본설계에 적용할 수 있는 프로그램을 손쉽게 사용할 수 있는가 하는 것도 매우 중요하다. 앞 절에서 언급한 것과 같이 주요 층화방법으로 널리 알려진 Ekman법과 Dalenius-Hodges법과 같은 경우 현재 이런 방법을 지원하는 소프트웨어를 쉽게 얻을 수 없기 때문에 사용상 한계를 갖게 된다. 반면에 회귀나무모형의 경우 흔히 접할 수 있는 SAS E-Miner를 통해 쉽게 적용이 가능하다는 장점을 갖고 있다.

따라서 만약 회귀나무모형에 의한 층화방법이 다른 층화방법과 비교해 유사한 효율성을 유지할 수 있다면 앞에서 언급한 여러 가지 이유에 의해 실제 표본설계에서 좀더 간편하고 효과적인 층화방법으로 향후 폭넓게 활용될 수 있을 것으로 기대된다.

3. 예 제

어업총조사는 전국의 모든 어가를 대상으로 어업의 경영구조와 어가인구의 취업상황 및 생활상태 등을 파악하여 어업정책 수립 및 국가경제 주요지표를 확보하기 위한 조사로서 통계청에서 2001년 3월에 실시한 조사이다 (통계청, 2001). 이 조사에서는 어가의 특성, 어업판매액, 어선보유톤수 등 다양한 항목에 관한 조사가 이루어져 있다.

층의 수	Dalenius-Hodges법	Ekman 법	CART 법
2	5.13 %	5.13 %	5.13 %
3	4.69 %	4.58 %	4.45 %
4	4.50 %	4.35 %	4.30 %
5	4.40 %	4.25 %	4.38 %
6	4.33 %	4.18 %	4.16 %
7	4.28 %	4.14 %	4.13 %
8	4.23 %	3.89 %	4.29 %
9	4.17 %	4.08 %	4.07 %
10	4.15 %	4.07 %	4.05 %

<표 1 > 층경계점 결정 방법에 따른 변동계수 비교

본 예제에서는 어업판매액을 조사하기 위해 어선보유톤수를 층화변수로 사용하는 경우를 다룬다. 즉, 어선보유톤수를 층화변수, 어업판매액을 조사변수로 생각한다. 우리 나라의 어가 현황은 지역에 따라 많은 차이를 나타내고 있으나 본 예제에서는 문제를 단순하게 하기 위해 지역별로 따로 층화는 하지 않고 전국을 한꺼번에 층화하기로 한다. 류제복, 김영원, 박진우 (2002)

는 어업판매액과 어선보유톤수와의 상관계수가 0.511로 상당히 높은 양의 상관을 지니고 있음을 보여준다. 따라서 어업판매액 조사를 위한 표본설계에서 어선보유톤수를 층화변수로 사용하는 것은 합리적이다.

앞의 <표 1>은 표본의 크기 1,200어가, 최적배분법을 사용하는 것을 전제로 전체 층의 수를 각각 2개에서 10개까지로 변화시켰을 경우 각 방법을 사용했을 때의 변동계수(coefficient of variation: CV)를 계산하여 나타낸 표이다. 이 표를 보면 세 가지 방법의 CV값이 크게 다르지는 않지만 그래도 회귀나무방법을 적용한 경우가 대체로 더 효율적인 것으로 나타났다. 다른 방법에 비해 회귀나무방법이 가지는 가장 큰 장점으로는 널리 알려진 통계소프트웨어를 이용하여 손쉽게 계산할 수 있다는 점인데 그런 장점 외에 이 예에서는 효율성 측면에서도 우수한 것으로 나타났다.

4. 결 론

일변량 연속변수를 층화변수로 사용할 경우 층의 최적경계점을 정하는 방법으로 널리 알려진 방법은 누적도수제곱근법과 Ekman 법이다. 그런데 이 두 방법은 개념적으로는 단순하지만 구체적인 계산과정에서 구현하기가 쉽지 않다는 약점을 지니고 있다. 기존의 어떤 통계 소프트웨어에서도 이 층의 최적경계점을 구하는 문제를 다루지 않고 있기 때문이다.

본 연구에서는 CART 기법 중 회귀나무모형을 사용하여 층의 경계점을 정하는 방법을 소개하였다. 최근 데이터마이닝을 지원하는 모든 통계 소프트웨어에서 회귀나무모형을 처리하고 있기 때문에 이 방법을 사용한다면 계산상 매우 편리하다는 장점이 있다. 또한 한꺼번에 층의 개수를 여러 가지로 고려할 수 있기 때문에 적절한 층의 수를 결정하는데 있어서도 편리하다.

통계청에서 실시한 2000 어업총조사 자료(통계청, 2001)를 이용하여 SAS E-miner에서 제공하는 회귀나무 방법을 사용한 층의 경계점을 정하는 방법과 기타 다른 방법의 효율성을 실증적으로 비교 분석하였는데 이 경우 계산상의 편리함 외에 효율면에서도 다른 방법들보다 대체로 높게 나타났다. 이 결과는 회귀나무를 이용한 최적경계점 결정 방법이 실용적이면서도 효율적인 방법이라는 것을 시사한다고 하겠다.

참 고 문 헌

(1) 강형철, 한상태, 최종후, 김은석, 김미경 (1999). 「데이터마이닝: 방법론 및 활용」(3판), 자유아카데미.

(2) 류제복, 김영원, 박진우 (2002). "어가경제조사를 위한 표본설계 연구", 「통계분석연구」, 제 7권 제 2호,

(3) 박홍래 (2000). 「통계조사론」(2판), 영지문화사.

(4) 통계청 (2001). 「2000 어업총조사 잠정보고서」.

(5) Breiman, L., Friedman, J.H., Olshon, R.A., and Stone, C.J., (1984). *Classification and Regression Trees*, Chapman & Hall.

(6) Cochran, W. G.(1977), *Sampling Techniques (3rd ed.)*, New York: John Wiley.

(7) Dalenius, T. and Hodges, J.L. (1959). "Minimum Variance Stratification", *Journal of the American Statistical Association*, 54, 88-101.

(8) Ekman, G. (1959). "An Approximation Useful in Univariate Stratification", *The Annals of Mathematical Statistics*, 30, 219-229.

(9) Hedlin, D. (2000). "A Procedure for Stratification by an Extended Ekman Rule", *Journal of Official Statistics*, 16, 15-29.

(10) Hess, I. Sethi, V.K. and Balakrishnan, T.R. (1966). "Stratification: A Practical Investigation", *Journal of the American Statistical Association*, 61, 74-90.

(11) Kalton, G. (1983). *Compensating for Missing Survey Data*, Institute for Social Research, University of Michigan.

(12) Morgan, J.N. and Sonquist, J.A. (1963). "Problems in the analysis of survey data, and a proposal", *Journal of the American Statistical Association*, 58, 415-434.

(13) Murthy, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

(14) Ryu, J., Kim, Y., Park J. and Lee, J. (2000). "Imputation Methods for the Population and Housing Census 2000 in Korea", *Bulletin of the 53rd Session of International Statistical Institute (Book 2)*, 421-422.

(15) Sarndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

(16) Sonquist, J.A., Baker, E.L. and Morgan, J.N. (1973). *Searching for Structure (Rev. ed.)*, Institute for Social Research, University of Michigan.

(17) Sweet, E. M. and Sigman, R. (1995). *User Guide for the Generalized SAS Univariate Stratification Program*. ESM Report Series, ESM-9504. U.S. Bureau of the Census.