회귀나무를 이용한 무응답 가중치 조정

Unit Nonresponse Weighting Adjustment Using Regression Tree

김 세 미* 이 석 훈**

< 요 약 >

가중치 조정(weighting adjustment)으로 단위 무응답(unit nonresponse)을 처리하는 문제에서 성향점수를 추정하는 모형을 만들기 위해 응답변수와 관심변수를 동시에 고려하는 다변량 회귀나무(multivariate regression tree)기법을 제안하였다. 효과적인 무응답 조정층 구축을 위해 응답한 개체들만 사용하는 경우와 모든 개체들을 사용하는 경우를 제시하고 이 두방법을 편향의 관점으로 비교한다.

<Abstract>

This paper considers formation of nonresponse weighting adjustment cell for handling unit nonresponse in sample surveys. We propose a multivariate regression tree mehtod for segmentation using the variable of interest and the estimated response probability simultaneously to construct effective nonresponse adjustment cell. One is using only response data and the other is using response and nonresponse data. These two cases are compared in terms of bias.

I. 서 론

무응답은 단위 무응답(unit nonresponse)과 항목 무응답(item nonresponse)으로 나눌 수 있다. 단위 무응답은 응답 거부, 피조사자 부재 등의 사유로 통계조사에 대해 전혀 응답하지 않은

*충남대학교 통계학과, semi@cnu.ac.kr
**충남대학교 통계학과 교수, shlee@stat.cnu.ac.kr

경우이고, 조사에는 협조 하였으나 전체 조사항목 중 몇 가지에는 응답하고 나머지는 응답하지 않은 항목 무응답이 있다. 무응답이 존재하는 원자료를 완전자료로 만들기 위한 두 가지 방법으로 가중치 조정(weighting adjustment)과 대체(imputation)가 있다. 가중치 조정은 결측값을 갖거나 불완전한 개체들은 무시되고 표본의 하위 층(subclass)의 응답률의 역수를 곱하여서 응답한 개체들의 표본 가중치를 키우는 방법이다.

본 논문에서는 응답변수를 이용하여 층을 형성하면서 관심변수를 보조적으로 이용하는 기존의 방법이 무응답층 설정과정이 복잡하고 분석자의 주관이 많이 들어가는 단점을 보완하고자 하여 응답변수와 관심변수를 동시에 이용하여 무응답층을 형성하는 방법을 제시하고 기존의 방법과 비교 분석하였다. 이때 사용되는 관심변수는 응답한 개체만 활용할 것인지 무응답한 개체를 대체하여 응답한 개체와 무응답한 개체 모두를 활용할 것인지 검토하였다.

2장에서는 무응답 가중치 조정의 전반적인 내용을 소개한다. 3장에서는 무응답층의 형성 방법 중 응답 변수와 관심 변수를 동시에 고려한 다변량 의사결정나무를 이용한 방법을 소개한다. 또한 다변량 의사결정나무의 이용시 고려되는 입력변수 중 관심변수의 무응답 개체도 층 형성에 직접 사용하기 위한 방법을 탐색한다. 4장에서는 2003년 6월 통계청의 가계조사 자료를 이용하여 기존의 방법과 제시된 방법을 비교한다. 관심변수의 응답개체만 사용할 경우와 모든 개체를 사용할 경우의 무응답 조정층 추정치의 편향에 대해서 살펴본다.

Ⅱ. 무응답 가중치 조정

1. 자료의 형태

자료의 형태는 크게 관심 변수, 보조 변수 그리고 응답 변수로 나뉜다. 관심 변수란 표본 조사에서 무응답 발생시 대체 또는 조정하여야 하는 변수이고 보조 변수란 응답여부에 관계없이 수집할 수 있는 인구통계학적 혹은 지리적 변수를 말한다. 응답 변수란 이진 형태(binary type)로서 응답 여부를 가리키는 변수이다. 설명을 위한 기호로써 Y를 관심변수, X를 보조변수, R을 응답 변수라 하자. 이와 같은 자료의 형태는 X는 모든 개체에서 값을 갖고 Y는 모든 항목에 응답한 경우와 모든 항목에 응답하지 않은 경우가 있다. 이때 모든 항목에 응답하지 않은 경우 R의 값을 0으로 하여 무응답 개체를 구분한다.

예를들어 가계 조사 자료에서 Y는 소득과 지출 등에 관계된 연속형 변수이고, X는 표본관리명부에서 얻을 수 있는 가계 조사의 응답 여부에 관계없이 주어지는 변수로써 순서형과 범주형 변수이다. R은 응답여부이므로 응답일때 1을 갖고, 무응답일 때 0의 값을 갖는 이진 변수이다.

2. 단위무응답의 처리

단위 무응답의 가중치 조정 처리의 주요 아이디어는 유사한 응답확률을 갖는다고 믿어지거나 특정 관심 변수가 거의 같은 값을 갖는다고 생각되는 표본 단위들의 층을 정의하는 것이다. 각 층 내에서 무응답한 표본 단위들의 가중치를 응답한 표본 단위들의 가중치에 분배하여 줌으로써 무응답 개체의 가중치도 모수 추정에 사용한다. 무응답 조정 층 형성 과정은 단순히 인구통계학적 범주형 변수의 조합으로 형성할 수도 있으나 Little(1986)과 그 외에 다른 연구자들은 응답 변수를 이용하여 추정된 응답확률 혹은 추정된 관심 변수 값에 따른 표본 단위의 그룹핑에 의한 층의 형성을 고려하였다. Little(1986)의 연구에서 응답확률을 사용한 가중치 조정은 대표본 편의(large sample bias, LSB)는 제어할 수 있지만 분산을 제어하기 어렵고 추정된 관심변수를 이용한 층에서의 가중치 조정은 편의와 분산을 모두 제어한다는 것이 제시되었다. 그러나 추정된 관심변수를 이용하는 방법은 모든 관심변수에 대해 각각의 분리된 모형과 그에 따른 무응답 가중치 조정이 필요하다는 단점이 있다. 따라서 Little(1986)에서는 분산을 제어하기 위해 응답률을 이용한 수정된 응답 성향 가중치 조정법이 제안되었다.

응답률에 대한 모형을 성향 점수(Propensity score)라고도 하는데 대표본 편의의 제어는 관심변수와 응답변수가 서로 독립이라는 가정하에 성립된다. 따라서 응답개체와 무응답 개체의 관심변수 Y의 분포가 같아지도록 조정층을 형성한다. 이렇게 만들어진 조정층 내에서 응답한 개체를 층 개체수로 나누어서 추정한 응답률의 역수, 즉 성향점수를 표본단위들이 갖고 있는 기본 가중치에 곱하여 준다. 이는 무응답이 발생하였을 때 무응답이 발생한 만큼 응답자들의 가중치를 높여줌으로써 무응답의 대체 효과를 고려하는 것이다. 이때 성향 점수의 참값을 알 수 없기 때문에 자료로부터 응답률의 역수를 추정하여 사용한다. 이를 기호로 표현하면 $\hat{\theta} = \sum_{i \in s} w_i y_i$ 을 무응답이 없었을 때의 추정량이라고 하자. 성향 점수의 추정량을 이용하여 무응답 보정을 사용한 추정량은 다음과 같이 표현된다. 이때 $\hat{P}[R_i = 1|X_i]$는 $P[R_i = 1|X_i]$ 의 추정량이다.

$$\hat{\theta}_R = \sum_{i \in s} \frac{w_i R_i y_i}{\hat{P}[R_i = 1|X_i]}$$

응답률을 추정할 때에는 전체 응답률을 사용할 수도 있으나 적절히 층을 나누어 층 내에서 응답률을 추정한다면 좀 더 정교한 추정 값을 얻을 수 있다. 적절한 무응답 조정 층을 만들기 위해서는 응답자와 무응답자 모두에게서 관찰되는 변수가 필요하다. 일반적인 조사에서 이런 보조 정보를 항상 활용할 수 있는 것은 아니지만 패널 조사인 경우 처음 조사에서 모든 가능한 정보를 얻은 후 시간이 갈수록 무응답이 증가하는 경향이 있으므로 초기 조사의 자료를 응답률

에 대한 모형을 만들기 위한 자료로 사용할 수 있다. 패널조사가 아니더라도 조사에 응하지 않은 표본들에 대해 최대한 자료를 수집하여 가능한 정보를 활용할 수도 있다. 이때 활용 가능한 보조변수 X들의 상호 교차로 만들어진 모든 층을 무응답 가중치 조정 층으로 사용하게 되면 표본수가 전혀없는 층이 발생하거나 무응답 혹은 응답 표본 어느 한 쪽만 존재하는 층이 발생할 수도 있기 때문에 통계적 기법을 이용한 다양한 무응답 가중치 조정 층 설정 방법들이 개발되어 있다.

무응답 조정 층을 이용한 무응답 가중치 조정법은 보조 변수 X가 무응답층을 나타내는 변수가 되는 경우이다. 모집단(U)이 C 개의 무응답 조정층 U_c 로 나뉘어 진다고 한다면 표본(S)도 C 개의 무응답 조정층 S_c 로 나뉘어 지고 성향 점수는 각 무응답 조정 층 내에서 일정하다고 가정한다.

$$U=\bigcup_{c=1}^{C} U_c\,, \qquad S=\bigcup_{c=1}^{C} S_c$$

이에 따라 그 층 내의 모든 원소들의 응답률과 그 추정치는 아래 식으로 표현된다.

$$P[R_i=1|X_i]=P[R_i=1|i\in S_c] \quad =\frac{\text{층 } c \text{ 의 총 응답 개체수}}{\text{층 } c \text{ 의 총 개체수}}$$

$$\hat{P}[R_i=1|X_i]=\frac{\sum_{i\in s_c} w_i R_i}{\sum_{i\in s_c} w_i} \quad ,(w_i\text{는 가중치})$$

따라서 이 경우 무응답층을 이용한 가중치 조정 추정량은 다음과 같이 표현된다.

$$\widehat{\theta_R}=\sum_{c=1}^{C}\left(\sum_{i\in s_c} w_i\right)\frac{\sum_{i\in s_c} w_i R_i y_i}{\sum_{i\in s_c} w_i R_i}$$

각 무응답층 내에서 무응답률이 일정하다고 할 때 이 무응답층을 이용한 가중치 조정 추정량은 다음과 같은 통계적 성질이 알려져 있다. 분모에 확률변수가 포함되어 있으므로 $\widehat{\theta_R}$ 는 근사적 불편 추정량이 된다. 분산의 형태는 다음과 같다. (한근식·김재광, 2003)

$$V[\widehat{\theta_R}]\doteq V[\widehat{\theta_n}]+E[\sum_{c=1}^{C} n_g^2(\frac{1}{r_c}-\frac{1}{n_c})\hat{\sigma}^2_{weg}\,]$$

$$\text{where} \qquad \hat{\sigma}^2_{weg}=\frac{1}{n_c-1}\sum_{i\in s_c} w_i^2(y_i-\overline{Y_c})^2$$

n_c : 층 c 의 총 개체 수

r_c : 층 c 의 총 응답 개체 수

$\widehat{\theta_n}$: 무응답이 없는 상황에서의 추정량의 분산

분산의 형태를 보면 등식 오른쪽 첫 번째 항은 무응답과 관계없는 항으로 점추정치의 표본설계로 인한 분산이며 두 번째 항은 무응답으로 인해 증가되는 분산값으로 일반적으로 층의 개수가 많아질수록 그리고 층 내 관심변수 Y의 분산이 작을수록 작아진다는 것을 알 수 있다. 위의 두 가지 통계적 성질을 이용하여 무응답층 내에서 응답률이 비슷하도록 함과 동시에 관심변수가 동질적이 되도록 결정하는 것이 추정치의 편의를 줄이고 효율을 높이는 것으로 유추되었다. 3장에서는 $P[R_i = 1|X_i]$를 추정하는 방법에 대해 논의 한다.

Ⅲ. 무응답 가중치 조정층 형성 방법

1. 다변량 회귀나무의 이용

의사결정나무는 비모수적인 기법으로 분류 혹은 예측에 사용되는 데이터마이닝 방법 중의 하나이다. 종속변수가 하나인 일변량 회귀나무 방법을 Beiman et al.(1984)의 CART (Classification and Regression Tree)에 근거하여 간단히 소개한다. 회귀 나무 모형의 생성은 자식 노드의 개체들의 연속형 종속변수 (y_i) 값이 가능한 동질적이 되도록 하는 독립 변수들 (x_{ip}) 과 그 독립변수의 분리 지점을 선택하는 과정의 반복으로 이루어진다. 이때 다음의 $D(N)$으로 정의된 부모노드의 편차를 가장 많이 감소시키는 변수의 분리 지점을 선택하게 된다. 이와 같은 분리 과정은 모든 개체로 이루어진 뿌리 노드로부터 시작하는데 이로부터 갈라진 노드를 자식노드라하고 그 자식 노드를 부모노드로 하여 다시 자식노드가 형성된다. 즉 자식노드의 $D(N_{left})$, $D(N_{right})$[1]을 각각의 개체 수로 가중 평균을 하여 부모노드의 $D(N)$과 차이를 가장 크게하는 분리 변수로 회귀나무를 생성해 나간다. 노드 N의 편차는 아래 식(1)로 정의한다. 여기서 $\bar{y}(N)$ 은 노드 N에 속한 개체들의 종속변수의 평균이다.

$$D(N) = \sum_{i \in N} y_i - \bar{y}(N)^2 \qquad (1)$$

이와 같은 회귀나무 모형의 형성과정을 단계별로 정리하면 다음과 같다.

(1) 뿌리 노드의 편차를 계산한다.

(2) 입력된 모든 종속변수의 모든 분리지점에서 고려되는 모든 자식노드들의 편차를 계산하여 왼쪽과 오른쪽 자식노드의 편차를 가중평균 한다.

(3) (1)에서 계산된 값과 (2)에서 계산된 값의 차이를 구하여 가장 큰 값을 가지는 변수의

---

1) N_{left}, N_{right} 는 부모노드 N에서 왼쪽 또는 오른쪽으로 분리된 가지에 속한 개체들의 집합

분리 지점을 첫 번째 분리되는 자식노드로 한다.

(4) (3)에서 결정된 자식노드를 부모 노드로 하여 (2), (3)을 반복하여 더 이상 회귀나무가 커지지 않을 때까지 반복한다.

(5) 마지막으로 가지치기 과정을 통하여 적당한 회귀 나무를 선택한다.

또한 독립변수의 유형에 따라 분리 지점을 고려하는 것이 달라지게 된다. 독립변수가 L 개의 범주를 가진 범주형 변수인 경우 가질 수 있는 모든 부분 집합은 2^L 개이나 실제로 고려될 분리 지점은 $2^{L-1}-1$ 개 이다. 순서형 변수인 경우는 $L-1$ 개의 모든 분리 가능한 지점이 있고 연속형 변수인 경우 그 변수의 모든 범위의 각 지점에서 분리를 고려한다. 회귀나무 모형에서 미리 종료노드의 최소 개체 수를 정하거나 편차의 최소 임계값을 정해놓는다면 쓸데없이 큰 나무가 형성되는 것을 막을 수 있다. 그러나 이런 법칙을 미리 정해 놓아도 좋은 모형이라고 할 수 없는 과대 추정된 나무가 생성되기 때문에 가지치기(prunning)를 실행하여 적당한 크기의 나무을 생성한다.

다변량 회귀나무는 두 개 이상의 연속형 종속변수를 가지는 자료의 회귀나무 분석에 쓰이는 기법이다. 두 개 이상의 연속형 종속변수의 회귀나무 분석에서 각 노드의 편차 측정은 아래의 식(2)으로 계산한다. (Larsen, D.R., Speckman, P.L., 2004)

$$D(N) = \sum_{i \in N} y_i - \bar{y}(N)\,' V^{-1}\, y_i - \bar{y}(N) \qquad (2)$$

여기서 y_i 는 종속변수 행렬이고 $\bar{y}(N)$ 은 노드 N에 속한 개체들의 종속변수의 평균 벡터이다. V 는 전체 자료의 분산 공분산 행렬이다. 일변량 회귀나무처럼 아래 식으로 표현된 불순도의 감소량이 가장 큰 분리 변수를 이용하여 회귀나무를 생성해 나간다.

$$Diff. = D(N) - \left(\frac{n_{left}}{n} D(N_{left}) + \frac{n_{right}}{n} D(N_{right})\right)$$

2. 가계조사 무응답 조정층 형성 방법

무응답 조정층을 형성하기 위해 다변량 회귀나무를 이용할 때 종속변수는 관심변수 중 총수입(총지출) 변수와 로지스틱 회귀분석을 이용하여 추정된 각 개체의 응답확률을 연속적인 값으로 변환시킨 변수를 사용한다. 중요한 아이디어는 응답확률을 활용하는 것인데 먼저 응답확률을 응답변수 R과 보조변수 X로부터 로지스틱 회귀모형을 가정하여 추정한다. 그런데 이 값이 0과 1사이의 값이기 때문에 관심변수가 되는 소득, 지출관련 변수값과 함께 다변량 회귀나무에서 목표변수 벡터로 사용하기 위하여 다음과 같이 변환한다.

$$y_{1i} = \ln\left(\frac{\hat{P}(R_i = 1|X)}{1 - \hat{P}(R_i = 1|X)}\right)$$

관심변수에서 지출은 물론 수입 부분에 관련된 항목인 총수입(총지출) 변수가 여러 관심 변수들을 대표할 수 있다고 판단하여 다변량 회귀나무의 종속변수로 사용하였다. 정리하면 로지스틱 회귀모형에서 추정된 응답확률의 오즈비의 로그변환으로 얻어진 변수 y_{1i} 와 총수입(총지출) 변수 y_{2i} 를 다변량 회귀나무의 종속 변수로하여 무응답 조정층을 구축하였다.

Ⅳ. 모의실험

1. 층형성과 방법론 비교

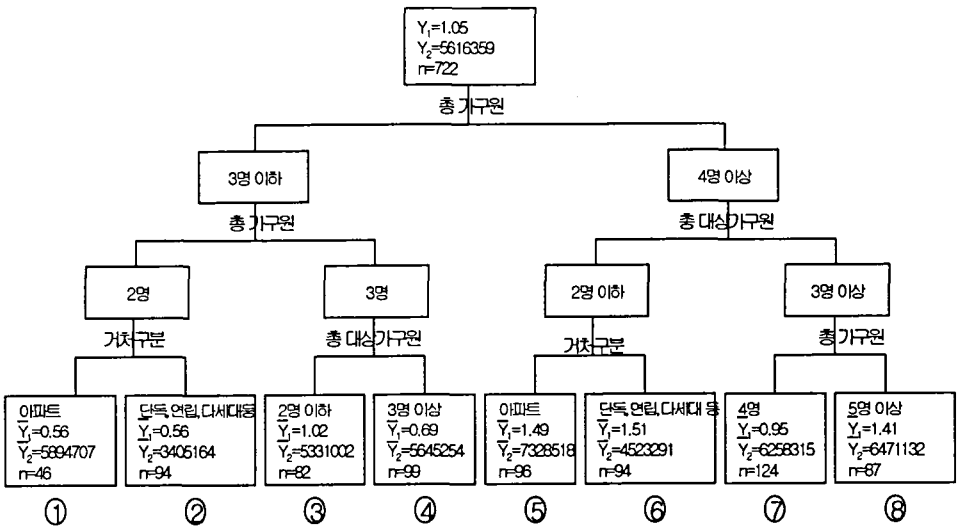
3장에서 소개한 방법으로 2003년 6월의 가계조사 자료 중 서울 자료를 이용하여 무응답 조정층을 만들고 결과를 비교하였다.

첫 번째 방법은 관심변수 중 총수입(총지출)변수와 응답여부의 로지스틱 회귀모형에서 추정된 사후확률 오즈비의 로그변환을 시킨 변수를 종속변수로 하는 이변량 회귀나무를 형성한다. 이때 종속변수 중 총수입(총지출)변수의 이상값을 탐지하여 자료값 중 상위 3배의 IQR을 벗어나는 가구의 자료를 제거한 뒤 무응답 조정층을 구축하였고 관심변수를 종속변수로 사용하기 때문에 원자료에서 응답한 개체만을 무응답 조정층 구축에 사용하였다.

두 번째 방법은 두 번째 방법에서 응답한 자료만 사용하게 되는 점을 보완하기 위해 총수입(총지출)변수를 보조변수 X를 이용하여 대체한 후에 총수입(총지출) 항목에 대해 완전자료를 만들고 응답여부의 사후확률 오즈비의 로그변환 변수와 함께 이변량 회귀나무를 형성한다. 이때 관심변수를 대체하는 방법으로 회귀대체를 사용하였다. 응답 자료에서 관심변수 총수입(총지출)을 종속변수로 보조변수들을 독립변수로하여 회귀모형을 추정한 뒤에 무응답 개체들을 모형에 대입시켜 총수입(총지출)의 추정값을 계산하여 대체하는 방법이다. 무응답한 개체들의 관심변수 값과 응답한 개체들의 관심변수 값의 분포가 서로 다르지 않을 것이라는 가정하에서 응답한 개체들의 분포를 왜곡시키지 않으면서 무응답을 대체하였다.

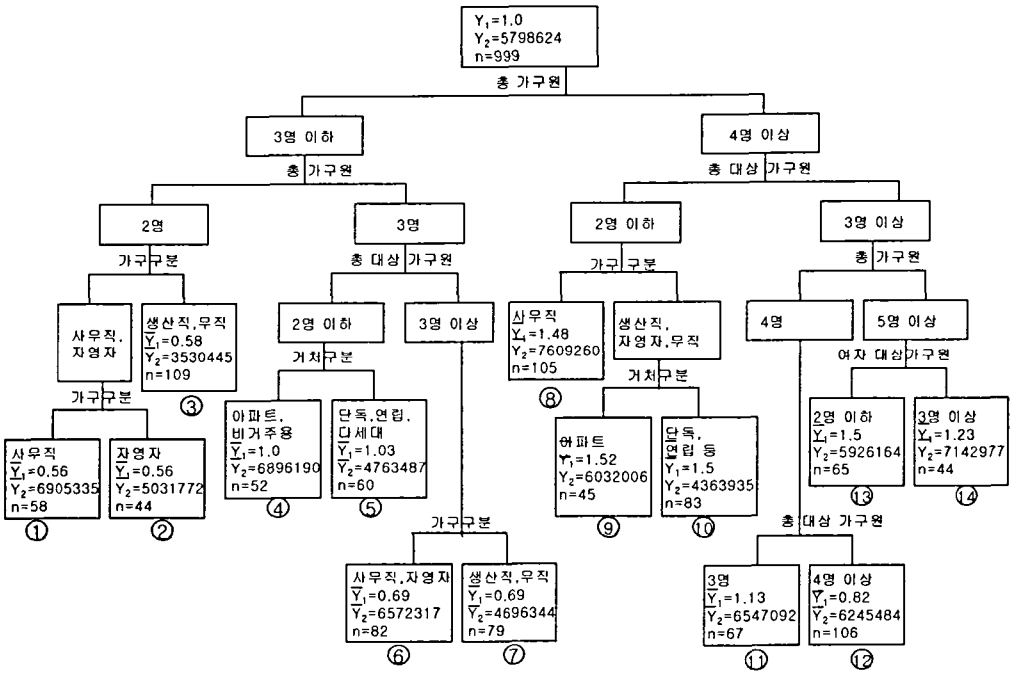
응답변수와 관심변수를 이용하여 만들어진 무응답 조정층은 <그림 1>과 같다. 8개의 층이 최종적으로 형성되었고 해당 노드의 종속변수들의 평균과 개체 수를 나타내었다. 응답변수와 대체된 관심변수를 이용하여 만들어진 무응답 조정층은 <그림 2>와 같다. 14개의 층이 최종적으로 형성되었고 해당 노드의 종속변수들의 평균과 개체 수를 나타내었다. 종료노드의 아래

쪽에 원문자로 표시된 숫자는 최종적으로 결정된 층번호이다. 종료노드에서 $\overline{Y_1}$ 은 로지스틱 회귀모형에서 추정된 응답확률의 오즈비의 로그변환으로 얻어진 변수의 층내 평균이고, $\overline{Y_2}$ 는 총수입(총지출)변수의 층내 평균이다. 두 가지 무응답 조정층을 모형의 구축과 해석의 측면에서 비교하고 통계량을 추정하여 비교해 보았다.



<그림 1> 응답변수와 관심변수를 이용한 무응답 조정층

<그림 1>에서 층 1과 2를 보면 총가구원이 2명인 가구의 Y_1 의 평균값은 낮은 편이지만 거처구분이 아파트인지 아닌지에 따라 Y_2 의 평균값이 크게 다름을 알 수 있다. 응답률은 비슷하지만 관심변수의 분포가 서로 다른 집단으로 분리되어 가중치 조정이 따로 이루어지므로 더욱 정교한 추정값을 기대할 수 있다. 또한 층 3과 4는 Y_2 의 평균값이 비슷한 반면에 Y_1 의 평균이 다르므로 총 가구원이 3명인 가구 중 총 대상가구원이 2명 이하인 가구가 응답해 줄 가능성이 크다고 볼 수 있다. 층 5는 Y_2 의 평균값이 가장 높고 응답 가능성도 높을 것으로 기대되는 층이다. 즉 총가구원이 4명 이상이고 총 대상 가구원이 2명 이하일 때 거처 구분 변수가 응답률이 비슷하면서 관심변수의 분포는 다른 두개의 층으로 나누는데 결정적인 역할을 하고 있다. 이는 층 1과 2가 분리된 것처럼 거처구분이 아파트인 경우 관심변수의 값이 높은 것을 확인할 수 있다. 마지막으로 층 7과 8에서도 관심변수의 표본 평균값은 전체 표본에 비해 크면서 응답률의 분포가 서로 다르다. 다변량 회귀나무 모형으로 구축한 무응답 조정층은 응답변수와 관심변수를 뿌리 노드부터 같이 고려하여 분리해 나갔기 때문에 종료노드에서 이러한 해석이 가능하다.



<그림 2> 응답변수와 대체한 관심변수를 이용한 무응답 조정층

<그림 2>에서 <그림 1>에 비해 종료노드의 개수가 늘어난 것은 무응답 개체들도 회귀나무 모형의 형성에 사용되어서 뿌리 노드의 표본수가 늘어났기 때문으로 판단된다. 층 1, 2, 3에서 Y_1 의 평균은 비슷하지만 Y_2 의 차이로 인해 분리되었고 총 가구원이 2명이라는 조건하에 응답을 안해주는 경향의 사무직과 자영자, 생산직/무직의 차이는 총수입(총지출)에서 나타남을 알 수 있다. 또한 나머지 다른 층에 비해 Y_1 의 평균값이 가장 낮은 걸로 미루어보아 2인 가구가 가계조사에 응답할 가능성이 가장 작은 것으로 예상된다. 층 4, 5에서 총 가구원 3명 중 대상 가구원이 2명 이하인 경우는 거처구분이 아파트/비거주용인 경우 Y_2 의 평균값이 더 크다. 주로 아파트에 거주하는 가구들의 Y_2 값이 더 클것으로 예상된다. 층 6, 7에서 총 가구원 3명 모두가 대상 가구원인 경우 가구구분이 층의 분리에 사용되었고, 층 4, 5, 6, 7에서 총 가구원 3명 중 대상가구원이 2명이하인지 아닌지가 응답에 영향을 주고 있음을 알 수 있다. 층 8과 9, 10은 다른 층에 비해 Y_1 의 평균값이 크고 가구구분이 사무직인 층 8의 Y_2 의 평균값이 가장 크다. 층 9, 10은 가구구분이 생산직/자영자/무직으로 거처구분에 의해 분리가 되었고 Y_2 의 평균은 차이가 있다. 이제까지의 내용을 볼 때 비슷한 응답률을 기대할 수 있는 가구에서 관심변수 Y_2 의 차이는 사무직일수록 아파트에 거주할수록 더 큰값을 갖는 것으로 보인다.

층 11, 12는 Y_1 과 Y_2 의 평균값이 크게 다르지 않아 보이나 다른 층의 표본수에 비해 비교적 큰 표본수를 갖고있기 때문에 가지치기하지 않았다. 층 13, 14는 Y_1 의 평균값이 어느 정도 크면서 여자 대상가구원이 2명 이하인지 아닌지에 따라 Y_2 의 차이에 의해 분리되었다. 여자 대상가구원이 더 많은 가구의 총수입(총지출)이 더 크다. 대상가구원이란 만 15세 이상의 경제활동이 가능한 가구원을 말하는데 대상가구원이 많다는 것은 그만큼 총수입(총지출)에 영향을 주게 된다. 정리하면 뿌리노드에서 처음 2~3번의 분리가 이루어질 때는 주로 Y_1 의 역할이 큰 것으로 보이고 마지막 종료노드의 부모노드에서의 분리에는 대체적으로 Y_2 의 역할이 큰 것으로 보인다.

2003년 6월의 자료를 이용해 만든 무응답 조정층으로 가계조사의 2003년 4, 5, 6월 자료에 가중치 조정을 적용한 뒤에 한근식·김재광(2003)에서 제시한 잭나이프 방법을 이용한 분산 추정식을 이용하여 주요 관심변수의 변동계수를 구하였다. 잭나이프 방법을 이용한 분산 추정은 한 개의 조사구를 제거하고 제거된 조사구의 기본 가중치를 0으로 하는 대신 층내 다른 개체의 가중치를 올려준 후에 무응답 가중치 조정을 통해 최종 가중치를 계산한다. 모든 조사구를 한 번씩 제거해서 구한 최종 가중치로 총 표본 조사구 수만큼의 평균 점추정치를 계산하여 그 값들의 변동을 이용하여 분산 추정치를 구하게 된다.

무응답 조정 가중치를 이용한 모수의 추정은 먼저 각 무응답 조정 층 내에서 기본 가중치에 성향점수를 곱하여 무응답 조정 승수를 아래와 같이 계산한다. 이때 무응답 조정 승수(Nonresponse Adjustment Factor ; NAF)는 항상 1이상의 값을 가지게 된다. 해당 조정 층에 무응답 개체가 많을수록 1보다 커지게 된다.

$$NAF_c = \frac{\sum_{(hij)\in S_c} BW_{hij}}{\sum_{(hij)\in S_c} BW_{hij} \times R_{hij}}$$

h : 층(stratum)을 나타내는 첨자
i : 집락(cluster: 여기서는 조사구)을 나타내는 첨자
j : 개체를 나타내는 첨자
S_g : c 번째 무응답층에 속하는 표본 개체들의 집합
BW_{hij} : h 층 내 i 조사구내 j 개체의 기본 가중치
R_{hij} : h 층 내 i 조사구내 j 개체가 응답하였을 경우에는 1을 갖고 무응답인 경우 0을 가지는 지시변수

또한 최종 가중치(Final Weight ; FW)는 기본 가중치(Base Weight ; BW)에 무응답 조정 승수를 곱하여 구한다.

$$FW_{hij} = BW_{hij} \times NAF_g \times R_{hij} \quad \text{if} \quad (hij) \in S_g$$

이렇게 최종적으로 만들어진 최종 가중치를 이용하여 우리가 원하는 관심변수의 평균을 추정할 수 있다. 관심 변수 y에 대한 평균의 추정치는 다음과 같이 계산할 수 있다.

$$\hat{\theta}_D = \frac{\sum_{(hij)\in S} FW_{hij} \times y_{hij}}{\sum_{(hij)\in S} FW_{hij}}$$

여기서 최종 가중치는 이미 무응답 개체에 0 값을 취하기 때문에 위의 계산은 응답개체 만으로 계산될 수 있음을 알 수 있다.

<표1> 무응답 조정층을 이용한 주요 관심변수의 변동계수(%)

관심변수	2003년4월		2003년5월		2003년6월	
	방법1	방법2	방법1	방법2	방법1	방법2
소득	3.36	3.34	3.64	3.69	12.43	12.88
경상소득	3.49	3.46	3.64	3.69	3.58	3.57
근로소득	3.55	3.50	3.46	3.47	3.73	3.71
가구주소득	4.31	4.29	4.19	4.22	4.36	4.36
배우자소득	9.75	9.6	8.70	8.73	9.91	9.88
기타가구원소득	11.95	11.94	12.19	12.10	12.19	12.22
사업소득	25.36	25.62	18.33	18.53	21.85	22.20
재산소득	30.40	31.12	54.08	56.20	26.65	25.96
이전소득	28.33	28.58	28.04	29.08	32.49	33.25
가계지출	4.25	4.26	4.45	4.52	4.92	4.89
소비지출	3.93	3.93	4.30	4.39	5.07	5.02
식료품	2.46	2.48	2.45	2.49	2.49	2.51
주거	11.08	10.98	12.84	12.58	16.86	16.62
광열수도	3.33	3.35	2.99	3.01	2.84	2.84
가구집기가사	12.61	12.56	15.52	16.35	17.14	17.04
피복및신발	10.00	10.06	5.72	5.71	7.75	7.88
보건의료	13.12	12.95	30.11	31.15	15.19	15.64
교육	11.69	11.67	10.00	9.63	11.42	11.30
교양오락	9.95	9.85	13.67	13.47	12.52	12.53
교통통신	11.99	11.95	11.66	11.69	16.00	15.63
기타소비지출	4.65	4.67	5.34	5.61	8.40	8.39
비소비지출	8.30	8.31	8.13	8.09	8.56	8.53

주요 관심변수의 추정된 통계량을 <표 1>에 정리하였다. <표 1>에서 방법1은 응답변수와 관심변수를 이용한 다변량 회귀나무 무응답 조정층 방법이고 방법2는 응답변수와 대체된 관심변수를 이용한 다변량 회귀나무 무응답 조정층 방법을 의미한다. 표를 보면 두 방법으로 만든 무응답 가중치 조정층에 의해 계산된 변동계수의 3개월 추정치가 비슷하다. 즉 어떤 식으로 무응답 조정층을 구하던지 성능은 비슷하다는 것을 확인할 수 있다. 통계량의 추정으로 방법간의 차이가 발견되지 않았다. 따라서 다음절에서 방법1과 방법2의 편의를 계산하여 비교해 봄으로써 좀 더 정확한 추정방법이 있는지 알아본다.

2. 편향의 비교

방법1과 방법2로 계산한 추정량의 편향을 다음의 실험으로 비교해 본다. 모수의 참값을 알아낼 수 없으므로 가계조사의 전국 자료 약 7000개 중에서 관심변수의 값을 알고 있는 응답한 개체 약 5000여 가구의 자료를 모집단으로 가정하고 응답여부(R_i) 를 0, 1로 랜덤하게 부여한다. 이때 응답여부가 0인 무응답은 30%정도 발생하도록 처리한다. 이렇게 만들어진 모집단(N= 약 5000)에서 크기 1500개인 표본을 비복원 추출한다.

먼저 표본에서 로지스틱 회귀모형으로 $\hat{P}[R_i=1|X_i]$ 를 추정하고 추정된 응답률의 오즈비의 로그변환을 한다. 응답 변수가 1인 약 1000여개의 개체로 방법1의 무응답 가중치 조정층을 형성하여 표본 평균을 계산한다. 이때 이변량 회귀나무에 사용된 관심변수는 총수입(총지출) 변수이다. 계속해서 방법2를 적용하기 위해 응답변수가 0인 개체들의 총수입(총지출) 변수값을 삭제하고 삭제된 총수입(총지출) 변수를 회귀대체로 대체값을 생성하여 준다. 이렇게 준비된 1500개의 개체로 방법3의 무응답 가중치 조정층을 형성하여 표본평균을 계산한다. 이러한 과정을 10번 반복하여 두 방법당 10개의 표본 평균 ($\bar{y}_j$) 을 계산한다. 알고 있다고 가정한 모집단에서 모수($\bar{Y}$) 를 계산한다.

다음의 식으로 편향의 평균값을 계산한다.

$$bias=\frac{\sum_{j=1}^{J}(\hat{y}_j-\bar{Y})}{J} \quad , \quad J=10$$

<표 2>에 방법1로 구한 bias1과 방법2으로 구한 bias2가 있다. 총수입(총지출)항목에서 bias1이 더 작은 것으로 나타났다. 그러나 그 차이가 2263원으로 총수입(총지출)의 추정값의 크기에 비해 미미하므로 방법1이 방법2보다 더 나은 추정값을 제공한다고 볼 수 없다. 다른 항목에서는 소득부터 이전 소득에 이르는 소득 관련 항목에서 사업소득, 재산소득, 이전소득만 빼고 bias2가

작은 것으로 나타났고 가계지출에서 비소비지출에 이르는 항목에서는 가계지출, 소비지출, 식료품, 피복 및 신발, 교육, 교통통신, 비소비지출 항목에서 bias2가 작고, 나머지 6개 지출관련 항목은 bias1이 크다. 그러나 그 차이 역시 거의 없다고 볼 수 있을 정도로 작기 때문에 방법1과 방법2의 편향을 비교했을 때 큰 차이를 발견하지 못했다.

< 표 2 > 방법1과 방법2로 구한 편향의 비교

항목	bias1	bias2
총수입	213147	215410
소득	-83175	-78510
경상소득	30481	29710
근로소득	14312	12899
가구주소득	11607	11011
배우자소득	-2880	-2789
기타가구원소득	5585	4678
사업소득	10247	10573
재산소득	2000	2100
이전소득	3923	4137
가계지출	29498	28108
소비지출	25221	24069
식료품	4933	4722
주거	636	702
광열수도	-197	-303
가구집기가사용품	308	384
피복및신발	-1353	-1345
보건의료	-2102	-2286
교육	5735	5435
교양오락	848	875
교통통신	3947	3395
기타소비지출	12466	12491
비소비지출	4277	4039

V. 결 론

단위 무응답이 발생한 자료에서 비편향된 추정치를 구하기 위해 가중치를 조정하여 무응답을 대체하는 방법에 대해 논의했다. 무응답 조정층을 구축할 때 중요한 관점은 층내 응답률이 얼마나 동질적인지와 관심변수의 층내 분산이 크지 않아야함에 있다. 다변량 회귀나무 알고리즘을 이용한 방법은 응답변수와 관심변수를 처음부터 같이 고려하므로 최종 층을 결정할 때 층내 관심변수의 분산을 살펴보지 않아도 된다는 장점이 있다. 또한 층의 해석상의 관점에서도 응답률과 관심변수의 동시 해석이 가능하다. 본 논문에서 제시된 다변량 회귀나무 알고리즘은 종속변수의 개수를 두 개 이상으로 확장할 수 있기 때문에 층의 형성에 고려하고 싶은 관심변수의 개입이 자유롭다.

다변량 회귀나무 구축에 응답한 개체만을 이용할 때에는 각 개체의 추정된 응답확률 관련 변수인 y_1 이 응답한 가구의 자료만 쓰이게 된다. 따라서 최종 층내에서 모든 가구들의 y_1 의 평균이 응답 가구들의 y_1 의 평균보다 작게 된다. 즉 y_1 변수에 대해서 선택편향(selection bias)을 야기할 가능성이 있다는 한계가 있다. 따라서 무응답 조정층 형성에 다변량 회귀나무를 이용하면서도 위에서 언급한 선택편향에 대한 위험을 제거할 방법으로 무응답하여 관심변수값이 없는 개체에 대해 해당 관심변수를 대체하여 완전자료를 만들어 다변량 회귀나무를 적합시키는 방법을 제시하였다.

각 방법의 효율에 대한 비교는 변동계수를 추정하여 살펴보았으나 방법간에 큰 차이가 없었다. 관심변수와 응답변수를 동시에 사용하여 이변량 회귀나무로 조정층을 만드는 방법에서 응답개체만 사용하는 경우와 모든 개체를 사용하는 경우에 선택 편향이 발생했는지 알아보기 위한 실험에서 두 경우에서 계산된 편향의 차이가 해당 항목의 추정치의 크기에 비해 크지 않았다. 선택 편향의 문제는 무응답자들의 특이성으로 인해 야기되는 편향에 대해서 추적조사하지 않는 이상 그 편향을 측정한다는 것이 어렵다는 한계가 있다. 즉 응답자와 무응답자들이 비슷한 보조변수 X의 값을 갖고있다는 조건하에 관심변수의 분포가 크게 다르다면 선택편향이 커질 것이다. 그렇지 않고 분포가 같다면 조정층 형성과정에서 응답개체만 쓰거나 모든 개체를 쓰는 방법에 있어서의 선택편향의 문제는 없다고 봐야할 것이다.

향후 과제로는 관심변수가 무응답 조정층의 구축에 사용되는 경우 본 논문에서 고려한 이차원 회귀나무에서 종속변수로 선택된 총수입(총지출)변수 이외에 다른 변수들도 포함시킨 다차원 회귀나무의 성능에 관한 검토가 필요하다.

참고문헌

한근식·김재광, (2003), 통계청 학술용역 보고서, 가계조사 무응답 처리기법연구

Breiman, L., Friedman, J.H., Olshen, R.A., stone, R.A., (1984), "Classification and Regression Trees", Wadsworth, Belmont.

Dufour,J., Gagnon, F., Morin, Y., Renaud,M., Samdal, C.-E. (2001) "A better understanding of weight transformation through a measure of change", Survey Methodology, 27, 97-108.

Eltinge & Yansaneh (1997) "Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey", Survey Methodology, 23, 33-40.

Larsen, D.R., Speckman, P.L., (2004), "Multivariate Regression Trees for Analysis of Abundance Data", Biometrics 60, 543-549.

Little R.J.A. (1986), Survey Nonresponse Adjustment for Estimates of Means, International statistical review, 54, 2, 139-157.

Rizzo, L., Kalton,G., and Brick,M. (1996) "A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse", Survey Methodology, 22, 43-53.