Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 3, pp. 585 ~ 598

Treatment of Missing Data by Decomposition and Voting with Ordinal Data[1)]

Young M. Chun[2)] · Hong K. Son[3)] · Sung S. Chung[4)]

Abstract

It is so difficult to get complete data when we conduct a questionaire in actuality. And we get inefficient results if we analyze statistical tests with ignoring missing values. Therefore, we use imputation methods which evaluate quality of data. This study proposes a imputation method by decomposition and voting with ordinal data. First, data are sorted by each variable. After that, imputation methods are used by each decomposition level. And the last step is selection of values with voting. The proposed method is evaluated by accuracy and RMSE. In conclusion, missing values are related to each variable, median imputation method using decomposition and voting is powerful.

Keywords : Decomposition, Imputation Method, K-Nearest Neighbor, Voting

1.서론

문명이 발달함에 따라 사람들은 한정된 자료에서 최대의 정보를 얻으려고 노력한다. 이러한 일환으로 양질의 정보를 얻기 위해서 결측값에 관심을 갖게 되었으며 다양한 결측값에 대한 연구가 이루어지고 있다. 이러한 연구는 설문지 작성, 설문조사를 거쳐 자료 가공까지의 과정에서 근본적으로 결측값이 발생하지 않도록 하는 노력이

1) This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST)(No. R01-2005-000-10752-0).

2) PostDoc Researcher, National Rural Resource Development Institute, 88-2, Seodun-Dong, Suwon, GyeongGi-Do, Korea.
E-mail : zzari@rda.go.kr

3) Analyst, E-Banking Strategy Institute, 143-40 Samsung-Dong, Kangnam-Gu, Seoul, Korea.
E-mail : shk03158@hdsbank.co.kr

4) Corresponding Author : Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University(Institute of Applied Statistics), Chonbuk, Korea.
E-mail : sschung@chonbuk.ac.kr

선행되어야 하지만 결측값이 없는 자료를 얻는다는 것은 매우 어려운 것이 현실이다(남궁평, 1992). 결측값은 주로 조사대상으로부터 응답회피, 시간의 제약, 문항의 난해함, 문항의 누락, 잘 교육되지 않은 조사원 등에 의해서 발생 되며, 결측값을 그대로 이용할 경우 조사 결과에 영향을 주어 분석의 신뢰도와 정확성이 떨어지게 된다. 즉, 이는 결측값의 비율이 높거나 낮더라도 비표본 오차를 발생시켜, 결측값이 있는 자료를 그대로 이용했을 때 편의를 일으킬 수 있다(Huisman, 2000).

이런 표본자료의 결과 왜곡과 비표본오차 발생을 방지하여 조사의 신뢰도를 높이기 위해 결측값 대체법이 이용되는데, 이에 앞서 결측값을 사전에 줄이는 노력이 필요하다. 이러한 노력의 일환으로 설문지 문항은 복잡한 서술형에서 간결한 폐쇄형 질문으로 바뀌어 가고 있다. 이 때 서술형 질문은 설문지 문항에서 응답자가 자기의 의견을 구체적으로 기술하는 것을 말하며, 폐쇄형 질문은 사전조사 등을 이용하여 한 문항당 2~5개의 항목으로 구성된 질문을 설정하여 응답자가 손쉽게 선택할 수 있는 범주형 성격을 가진 질문을 말한다(박태성, 이승연 1998). 범주형 성격을 가진 설문지에서 결측값을 새로운 값으로 대체하는 방법도 연속형 자료와 더불어 상당히 다양한 분야에서 개발되어 있다. 이러한 기법에는 랜덤 대체, 핫덱 대체, 단순 최빈값 대체, 단순 중앙값 대체, k-최근접 이웃대체 등이 널리 알려져 있다.

위의 대체법들은 자료의 질을 향상시키는 동시에 여러 가지 단점을 수반한다. 단일대체법에 해당하는 최빈값 대체와 중앙값 대체는 각 문항에 국한하여 하나의 값으로 결측값을 대체한다. 이때 문항 간의 상관관계가 높아서 문항 간에 서로 상관이 있는 경우에도 결측이 포함된 문항에 대한 대체를 실시할 때, 다른 문항의 정보를 이용하지 못하게 된다. 그리고 k-최근접 이웃대체는 결측값이 있는 개체의 경우를 제외하고 결측값이 없는 개체에서 근접이웃을 찾고 근접이웃으로 선택된 개체들의 정보를 이용하여 대체한다. 이 때 결측값이 있는 개체의 정보를 전혀 이용하지 않고 근접이웃을 찾는 단점을 가지고 있다. Little(1998)은 결측값을 포함한 관련변수의 정보를 대체과정에 이용하였다. 또한 Jonsson과 Wohlin(2004, 2006)은 결측값이 포함된 개체의 정보를 이용하여 전통적인 k-최근접 이웃대체를 수정하였다. 이 방법은 k-최근접 이웃대체법보다 이용하는 자료가 많지만 최소조건을 만족해야 하는 제약이 있기 때문에 이용 가능한 자료의 개수는 한정된다는 단점이 있다.

이에 본 연구는 최빈값 대체와 중앙값 대체의 단점을 보완하기 위하여 각 항목에 따라 층을 형성하여 항목별로 특성에 따라 각기 다른 대체값으로 대체한다. 이로써 문항사이의 상관정도를 고려하지 않았던 기존 방법과 달리 문항사이의 관계를 반영하여 분석결과의 신뢰도와 정확성을 향상시켜 보았다. 또한 Jonsson과 Wohlin(2006)의 k-최근접 이웃 방법에서 정보이용 극대화를 시도하였다.

전체적인 구성은 다음과 같다. 1장에서는 본 논문의 연구 배경 및 목적을 소개하고 2장에서는 결측값 발생 형태와 대체법의 종류를 설명하고 Jonsson과 Wohlin이 제안한 결측값이 있는 개체를 이용한 k-최근접 이웃대체에 대해 알아본다. 3장에서는 본 연구에서 제안한 decomposition(Latkowski, 2002)과 voting을 이용한 대체법을 소개하고 4장에서는 기존의 방법인 랜덤 대체, 중앙값 대체, k-최근접 이웃대체와 본 연구에서 제안한 대체법을 비교하기 위하여 모의실험과 실제 자료를 이용하여 실험을 실시하였다. 마지막으로 5장에서는 연구에 대한 결론과 향후 연구 방향을 제시하였다.

2. 결측값 형태와 대체법

2.1 결측값 형태

자료에 포함된 결측값의 존재는 통계분석을 실시할 때 문제를 발생시키는 원인이 된다. 만약 결측값이 관심 있는 결과와 관련이 있다면 그것이 무시되었을 때 통계 검정의 결과에 편의가 발생 할 것이다. 게다가 대부분 통계 프로그램은 자동적으로 분석에 이용된 문항에서 결측값이 포함된 개체를 제외한 나머지 개체를 가지고 분석을 하게 된다(박성률, 2005). 이러한 사실은 표본의 크기를 감소시키고, 검정력을 낮추는 결과를 초래한다. 이러한 결과를 유발시키는 결측값의 3가지 유형은 통계적 특성에 따라 완전 임의결측(MCAR), 조건적 임의결측(MAR) 그리고 비임의결측(NMAR)으로 분류된다(Little and Rubin, 1987; Scheffer, 2002).

완전 임의 결측(MCAR; Missing Completely At Random)이란 어느 항목이 측정 또는 결측 되는지가 완전하게 임의로 정해지는 경우로써 특정 항목의 결측은 개체의 어떤 특성과도 관련 없이 발생한다. 따라서 특정항목이나 특정 문항에서 결측값이 발생하는 것이 아니라 항목과 문항에 관계없이 랜덤하게 발생되는 경우가 이에 해당된다.

조건적 임의결측(MAR; Missing At Random)이란 항목의 결측 확률이 다른 측정 항목의 관측 값에 따라 결정되는 경우이다. 예를 들면 가계조사에서 가구소득 응답률이 중학교 졸업자 90%, 고등학교 졸업자 80%, 대학교 졸업자 70%, 대학원 졸업자 60%가 나왔을 때, 교육수준이 높아짐에 따라 가구소득을 보고하지 않는 경향이 있음을 알 수 있다. 이 상황에서 가구소득이 결측 되어 개별 추정 값이 필요한 경우 교육수준을 활용하여 가구소득을 추정할 수 있다.

비임의결측(NMAR; Not Missing At Random)이란 무응답 확률이 응답문항의 값에 의존하여 서로 다른 문항들의 값과 연관되어 있는 경우라고 하였는데, 이 때 보조문항이 동일한 값을 가지고 있는 응답 값과 무응답 값은 구조적인 차이를 나타내게 된다. 여기에서 구조적인 차이란 결측값의 대체값은 결측값의 후보 값들과 일률적인 차이를 보이는 경향을 의미한다(박성률 2005). 예를 들면, 가계조사에서 교육수준이 높아짐에 따라 가구소득을 보고하지 않는 경향이 있고, 또한 반대로 가구소득이 높을수록 교육수준을 보고하지 않는 경향이 있을 때, 대학교 졸업자의 응답률이 전반적으로 70%수준이지만 그 가운데 저소득층, 중간소득층, 고소득층의 응답률이 각각 90%, 70%, 50%등으로 차이가 있는 경우이다. 따라서 대학교 졸업자의 가구소득이 결측된 경우, 그 가구의 소득을 대학교 졸업자 집단의 중앙값으로 예측한다면 과소 편의가 발생할 것이다.

2.2 대체법과 대체법의 분류

결측값 대체법에는 여러 가지 많은 방법이 있는데 본 연구에서는 범주형 자료에 대한 대체법만 살펴보았다. Huisman(2000)은 결측값 대체법을 정보가 없는 경우와 정보가 있는 경우로 대체법을 나누어 비교하였다.

정보가 없는 대체법(Uninformed Imputation)은 결측값이 포함된 자료에서 결측값을 분석대상에서 제외한 나머지 자료(Complete data set)만을 이용하거나, 결측값이 없는

응답 항목 중 무작위로 선택한 항목의 값으로 결측값을 대체하는 기법이다.

정보를 이용한 대체법(Informed Imputation)은 분석대상이 되는 자료의 결측값을 결측이 없는 자료를 이용해 적절한 값으로 대체하는 방법으로 핫덱 대체, 축차 핫덱 대체, 최빈값 대체, 중앙값 대체 등이 널리 알려져 있다. 이는 분석대상이 되는 통계 자료의 결측된 부분을 결측값이 없는 자료를 이용하여 적절한 값으로 대체하는 방법이다.

먼저 가장 오래된 핫덱 대체법은 현재 해당 문항에 존재하는 실제 값들 중에서 랜덤하게 선택하여 대체하는 방법으로, 특히 동일한 항목에 대해 결측값의 바로 전에 조사된 자료의 문항을 이용하여 대체하는 방법이 축차 핫덱 대체이다. 최빈값 대체법은 한 문항에 대해 결측값이 발생했을 때 해당 문항의 최빈값을 결측값의 대체값으로 채워 넣은 방법으로 주로 명목형(nominal) 자료에 많이 이용된다. 중앙값 대체법은 최빈값으로 대체하는 방법과 비슷하여, 해당 문항을 크기순으로 나열한 후 가장 중앙에 있는 중앙값으로 대체값을 채워 넣는 방법으로 순서형(ordinal) 자료에 주로 이용된다. 연속형 자료에서는 중앙값이 둘인 경우에 두 값의 평균으로 결측값을 대체하지만 순서형 자료에서는 중앙값이 둘인 경우는 평균으로 대체가 불가능하여 본 논문에서는 랜덤하게 한 값을 선택하여 대체하였다. 마지막으로 k-최근접 이웃대체법은 결측값이 포함된 개체를 제외한 자료에서 유사성 척도인 유클리드 거리(euclidean distance)를 이용하여 대체할 결측값이 포함된 개체와 가장 가까운 k개의 개체를 이용하여 대체하는 방법이다.

대부분 결측값이 있는 자료의 분석은 결측값을 다른 값으로 대체함으로써 분석 결과의 신뢰성과 정확성을 향상시킬 수 있다. 이 때 결측값을 대체하는 방법을 크게 두 가지로 구분할 수 있다. 최빈값과 중앙값과 같은 일정한 값을 이용하여 대체하는 단일 대체법(single imputation method)과 다수의 값을 생성하여 대체하는 다중 대체법(multiple imputation method)으로 나눌 수 있다(Rubin 1987). 일반적으로 많이 이용되는 단일 대체법은 일반적 통계방법을 통해 얻어진 특정한 값을 이용해 대체를 실시한다는 장점이 있지만, 관측되지 않은 값이 통계 분석모형에 충분히 반영되지 못한다는 단점이 있다. 반면 Rubin(1978)에 의해 제안된 다중 대체법의 장점을 정리하면 첫째, 하나의 결측값에 대하여 통계모형에 기반하여 생성된 다수의 값을 대체하므로 분석에 필요한 추정량에 대한 분포를 표현하여 모수의 추정에 효율을 높일 수 있다. 둘째로 특정 모형만이 아니라, 다수의 모형에 기반한 결측값에 대한 대체값의 생성이 가능하다. 따라서 다양한 분석방법의 반복적 이용을 통하여 불완전 자료에 대한 통계적 추론의 유효성을 점검할 수 있다.

2.3 k-최근접 이웃대체법

최근접 이웃대체란 결측값이 포함된 문항을 제외한 보조문항을 이용하여 결측값이 포함된 개체가 보조문항에서 갖는 값과 가장 유사한 보조문항 값을 갖는 대체군내에서 대체값을 찾아 대체하는 방법으로 여기서 유사성의 척도는 일반적으로 유클리디안 거리(euclidean distance)를 사용한다. 이 때 k-최근접 이웃대체는 완전한 개체를 이용하여 대체를 실시하므로 유클리디안 거리의 정의역은 모든 문항이 된다. k-최근접 이웃대체에서 개체 a와 b사이의 유클리디안 거리는 식 (1)과 같다.

$$E(a,b)=\sqrt{\sum_{i}(x_{ai}-x_{bi})^2} \quad (1)$$

단, x_{ai}와 x_{bi}는 각각 개체 a와 b의 i번째 문항의 값을 의미한다. 한편 Jonsson과 Wohlin (2006)은 결측값이 있어도 정보의 가치가 있다면 완전하지 않은 개체를 이용하는 방법을 제시하였다. 또한 결측값의 비율이 전체의 30% 이하인 경우는 결측값이 포함된 개체를 이용한 k-최근접 이웃대체의 결과와 결측값이 포함된 개체를 이용하지 않는 k-최근접 이웃대체의 결과가 비슷함을 보였다. 여기서 이용한 개체 a와 b 사이의 유클리디안 거리는 식 (2)와 같다.

$$F(a,b)=\sqrt{\sum_{i\in D}(x_{ai}-x_{bi})^2} \quad (2)$$

단, x_{ai}와 x_{bi}는 문항 i에서의 각 개체의 값을 의미하고, $i\in D$ 는 두 개체 중 문항 i가 공통으로 있는 정의역을 의미한다. 이 때 식 (1)과 식 (2)의 차이점은 항목을 뜻하는 i의 정의역 D가 다르다는 것이다. 두 식에서 사용한 정의역과 k-최근접 이웃대체법의 차이에 대한 이해를 돕기 위해 Jonsson과 Wohlin(2006)이 사용한 <표 1>과 같은 예시 자료를 살펴본다. 자료는 총 5개의 문항과 4개의 개체로 되어 있다.

<표 1> k-최근접 이웃대체법 예시 자료

id	x_1	x_2	x_3	x_4	x_5
1	2	3	4	2	1
2	2	-	4	2	5
3	-	-	**2**	**4**	-
4	2	-	-	-	-

<표 1>에서 유클리디안 거리의 정의역을 설명하기 위해 4개의 개체 중에서 3번째 개체를 살펴보면 첫 번째 문항에 있는 결측값을 대체할 경우 기존의 k-최근접 이웃대체법은 결측값이 없는 완전한 개체만을 이용하여 대체를 실시하므로 첫 번째 개체만이 대체의 정의역을 만족한다. 하지만 Jonsson과 Wohlin(2006)이 제시하는 정의역 D 는 x_3과 x_4뿐이므로 이 정의역을 포함하는 개체는 첫 번째 개체의 경우뿐 아니라 두 번째 개체도 해당된다. 즉, 두 개체와 세 번째 개체 사이의 유클리디안 거리인 $E(1,3)=E(2,3)=\sqrt{2\times(4-2)^2}\approx 2.8$ 을 이용하여 최근접 이웃을 결정한다. 이 때 두 번째 개체가 첫 번째 개체보다 결측값은 많지만 세 번째 개체에서 x_1이나 x_5을 대체 시킬 경우 두 번째 개체는 첫 번째 개체와 동일한 정보를 지닌 개체가 되는 것이다. 이렇게 Jonsson과 Wohlin(2006)은 완전한 개체만이 아니라 결측값이 있는 개체를 포함하여 결측값을 대체하였다.

3. 분해(decomposition)와 투표(voting)를 이용한 대체법

기존의 대체법은 결측값이 포함된 개체의 정보를 무시하여 결측값 대체에 유용한 정보를 사용하지 않았다. 따라서 본 논문은 결측값이 포함된 개체라도 대체에 최대한 이용하여 분석결과의 질을 향상시키고자 순서형 자료에 분해와 투표 방법을 접목시켰다.

분해와 투표에 의한 대체법은 대체층 형성, 대체법 적용, 반복, 투표 방법의 4단계로 구성되어 있다. 대체층 형성과정은 정렬과 분리단계로 나눌 수 있다. 먼저 정렬 과정에서는 한 문항에 대해 정렬을 하게 되는데 순서형 자료이므로 1번 항목부터 정렬하게 되며 이 때 결측값은 가장 하단에 정렬된다. 이렇게 정렬된 자료에서 각각의 항목에 따라 분리하여 대체층을 형성한다. 형성된 대체층은 문항의 특성에 따라 나누어지게 되며, 다음 단계로 이렇게 분리된 대체층에 대하여 중앙값 대체, k-최근접 이웃 대체, 랜덤대체 등을 이용하여 결측값을 대체하게 된다. 이 때 각 대체층에서 결측값은 있지만 대체할 값이 없는 경우에 해당 결측값은 대체하지 않고 다른 대체층으로 넘어간다. 이렇게 각 문항에 대해 대체층을 형성한 후 기존의 대체법을 적용하여 문항 개수보다 하나 적은 데이터 셋을 형성하게 된다. 대체가 모두 끝나면 하나의 결측값에 대하여 투표 방법을 통하여 최종 값을 선택하게 된다. 이 때 하나의 결측값에 대해 가장 많이 나온 값이 두 개 이상일 경우에는 두 개의 값 중에서 임의로 하나의 값을 최종적으로 선택한다. 이러한 일련의 과정을 자세히 살펴보면 다음과 같다.

n개의 개체와 p개의 문항을 가지고 있는 자료들의 집합을 행렬 D라 하자. 그리고 $x_{ij}^{*}(i=1,\cdots,p, j=1,\cdots,n)$를 결측값이라고 하면 행렬 D는 아래와 같다.

$$D=\begin{pmatrix} x_{11} & x_{12}^{*} & x_{13} & \cdots & x_{1p} \\ x_{21}^{*} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33}^{*} & \cdots & x_{3p}^{*} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3}^{*} & \cdots & x_{np} \end{pmatrix}$$

이 때, j번째 문항에 존재하는 결측값을 대체하는 방법의 절차는 다음과 같다.

step 1.1 행렬 D를 첫 번째 문항을 기준으로 정렬한다. 정렬한 행렬을 D_1이라 한다.

step 1.2 행렬 D_1을 각 항목별로 분리한다. 이렇게 정렬과 분리의 과정을 합쳐서 대체층 형성과정이라 한다. 대체층이 형성되면 항목별로 그룹이 만들어지며 결측값은 가장 아래쪽에 위치한다.

step 2 대체층을 형성한 후에 중앙값 대체, k-최근접 대체, 랜덤 대체 등을 이용하여 결측값을 대체한다. 이 때 결측값 대체층은 제외 한다. 단, 중앙값 대체시 선택된 중앙값이 정수가 아닌 경우에는 인접한 두 개의 값 중 하나를 임의로 선택한다.

step 3 step 1.1부터 step 2의 단계를 거치면 $D_1^{'}$이라는 하나의 행렬이 만들어진다. 이런 반복을 자료 정렬부터 대체층 형성, 대체하는 일련의 과정을 p번째 문항까지 반복하여 $D_2^{'},\cdots,D_{j-1}^{'},D_{j+1}^{'},\cdots,D_p^{'}$을 만든다. 이 과정이 끝나면 하나의 원

자료에서 대체가 실시된 새로운 $p-1$개의 자료가 생성된다. 이 때 하나의 결측값에 대해 최대 $p-1$개의 대체값이 나올 수 있다.

step 4 하나의 결측값에 대해 생성된 여러 개의 후보 대체값 중에서 다수 투표(majority voting) 방법을 통하여 최종적인 대체값을 결정한다. 단, 투표 방법에서 두 개의 값이 선택될 경우에는 두 개의 값 중 하나를 임의로 선택한다.

대체층을 형성하여 중앙값을 선택한 후에 투표 방법을 이용한다고 할 때, 두 번째 문항의 결측값에 대한 대체값을 선택하는 전체적인 흐름을 예제와 함께 살펴보면 <그림 1>과 같다.

이 때 대체층 형성 k-NN 대체법의 경우에는 대체하려는 결측값이 포함된 문항과 자료를 정렬하여 대체층을 형성하는 데 사용된 문항을 제외한 다른 문항들의 정보를 이용하여 개체들 사이의 거리를 계산하여 대체를 실시하면 된다.

(1)

id	x_1	x_2	x_3	x_4	x_5
1	3	2	1	4	2
2	*	2	1	2	4
3	2	*	1	3	4
4	1	3	3	*	5
5	4	4	5	*	3
6	3	3	3	1	3
7	3	1	4	5	1
8	5	*	3	1	4
9	4	5	2	3	*
10	2	2	2	4	4
11	5	5	4	3	1

⇒

(2)

id	x_1	x_2	x_3	x_4	x_5
4	1	3	3	*	5
3	2	*2	1	3	4
10	2	2	2	4	4
1	3	2	1	4	2
6	3	3	3	1	3
7	3	1	4	5	1
5	4	4	5	*	3
9	4	5	2	3	*
8	5	*5	3	1	4
11	5	5	4	3	1
2	*	2	1	2	4

⇒

(3)

id	x_1	x_2	x_3	x_4	x_5
1	3	2	1	4	2
2	*	2	1	2	4
3	2	*2	1	3	4
9	4	5	2	3	*
10	2	2	2	4	4
4	1	3	3	*	5
6	3	3	3	1	3
8	5	*3	3	1	4
7	3	1	4	5	1
11	5	5	4	3	1
5	4	4	5	*	3

⇒

(4)

id	x_1	x_2	x_3	x_4	x_5
6	3	3	3	1	3
8	5	*3	3	1	4
2	*	2	1	2	4
3	2	*5	1	3	4
9	4	5	2	3	*
11	5	5	4	3	1
1	3	2	1	4	2
10	2	2	2	4	4
7	3	1	4	5	1
4	1	3	3	*	5
5	4	4	5	*	3

⇒

(5)

id	x_1	x_2	x_3	x_4	x_5
7	3	1	4	5	1
11	5	5	4	3	1
1	3	2	1	4	2
5	4	4	5	*	3
6	3	3	3	1	3
2	*	2	1	2	4
3	2	*2	1	3	4
8	5	*2	3	1	4
10	2	2	2	4	4
4	1	3	3	*	5
9	4	5	2	3	*

⇒

(6)

id	x_1	x_2	x_3	x_4	x_5
1	3	2	1	4	2
2	*	2	1	2	4
3	2	2	1	3	4
4	1	3	3	*	5
5	4	4	5	*	3
6	3	3	3	1	3
7	3	1	4	5	1
8	5	3	3	1	4
9	4	5	2	3	*
10	2	2	2	4	4
11	5	5	4	3	1

<그림 1> 대체층 형성 중앙값 대체법의 적용 과정

4. 모의실험

앞에서 살펴본 정보가 없는 대체법(Uninformed Imputation)과 정보를 이용한 대체법(Informed Imputation)은 각각의 문항들의 정보를 이용하지 않거나 정보를 이용했을 경우에도 부분적인 정보를 가지고 분석을 하기 때문에 문항들 사이의 영향이 상쇄되기도 한다. 결측값이 존재하는 문항들의 정보를 이용하기 위하여 제안한 방법으로 모의실험을 실시하여 정확도와 RMSE를 비교하였다. 이 때 실험은 R 2.4.0을 이용하였다.

한편 본 연구에서 제안한 방법의 적용은 순수한 순서형으로 얻어진 자료가 아닌 연속형으로 얻어진 자료를 순서형으로 변환한 자료에 대하여 적용하였다.

첫 번째 모의실험은 시뮬레이션을 통해 생성된 4가지 유형의 자료를 이용하여 비교분석하였다. 즉, 생성된 자료가 서로 연관이 없을 때, 서로 $\rho = 0.3$, $\rho = 0.6$, $\rho = 0.9$의 관계를 가질 때 등으로 나누어 실험을 실시하였다.

효과적인 대체법을 수행하기 위해서는 일반적으로 대체층의 형태로 표본을 나누는 작업이 필요하다. 대체층이란 결측값과 통계적으로 밀접한 관계가 있는 문항인 보조문항들의 교차분류의 형태로 나타나는 층을 말한다. 따라서 많은 대체법에 있어서 성공적인 대체층의 구성이 대체법의 효율성을 좌우하게 된다. 이처럼 대체층으로써 각 문항별로 나누어 분석하고 정확도와 RMSE를 통하여 비교 분석하였다.

모의실험에 사용된 분석 방법으로는 대체층을 나누지 않고 중앙값 대체법, k-최근접 이웃대체법, 랜덤 대체법을 적용하고 각 문항별로 대체층을 만든 후에 각각의 대체층에서 중앙값 대체법, k-최근접 이웃대체법, 랜덤 대체법을 적용하여 분석하였다. 성능 비교를 위해 정확도와 RMSE를 이용하였다. 첫 번째 성능 비교 방법인 정확도는 대체법의 가장 기초가 되는 부분으로써 모의실험에서 결측값이 발생하기 전에 있는 실제 항목 값과 대체를 한 후 얻어진 대체 항목 값을 비교하게 되며 항목에 있는 값을 정확하게 맞출 경우만을 이용하여 정확도를 계산한다. 만일 정확도가 떨어지면 결측값이 있는 개체를 제거하는 것보다도 더 좋지 않은 결과를 얻을 수 있기 때문이다. 정확성을 $A_i (i = 1,2,\cdots,n)$라 했을 때, 기존값과 대체값이 같으면 $A_i = 1$이고 같지 않으면 $A_i = 0$이라 한다면, 정확도는 식 (3)과 같이 정확성을 계산하여 누적한 것으로 정의할 수 있다.

$$Accuracy = \sum_{i=1}^{n} A_i \tag{3}$$

다음으로 RMSE는 결측의 대체값과 실제값의 차이를 알려주는 척도로 편차가 적어야 잘 대체된 것으로 식 (4)와 같다.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \hat{e_i})^2}, \tag{4}$$

단, e_i는 원래의 값, $\hat{e_i}$는 대체법을 이용하여 추정한 값, n은 결측된 자료의 개수이다. 이 때 정확도와 RMSE는 1000번 반복한 것의 평균 횟수로 측정하여 비교하였다.

모의실험 절차는 다음과 같다. 먼저 5개의 변수로 이루어진 다변량 정규분포에서 각 변수들 사이의 상관계수가 동일하도록 하여 500개의 개체를 생성하였다. 이 때 사용된 상관계수는 각각 $\rho = 0.0$, $\rho = 0.3$, $\rho = 0.6$, $\rho = 0.9$ 등으로 총 네 가지의 서로 다른 실험 자료가 이용되었다. 정규분포에서 생성된 연속형 자료의 값을 5점 척도의 순서형 자료로 변환하여 실험을 실시하여 비교 분석하였다. 이렇게 생성된 자료에서 무작위로 각 문항당 20개씩 총 100개의 결측값을 만든 후 6가지의 대체법에 따라 대체값을 구하였다. 6가지의 대체법을 이용한 경우의 정확도는 <표 2>이고 RMSE를 비교한 것은 <표 3>이다.

<표 2> 모의실험 정확도(Mean±SD)

대체법	$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
대체층 중앙값 대체	20.33±3.93	22.24±4.11	33.23±4.33	59.91±4.59
대체층 k-NN 대체	20.36±4.07	21.42±4.14	27.99±4.51	44.22±4.90
대체층 랜덤 대체	19.45±4.00	19.60±3.89	19.80±4.02	23.51±4.22
중앙값 대체	20.27±3.93	20.52±3.91	19.97±3.90	19.67±3.82
k-NN 대체	20.91±4.02	21.66±4.01	25.60±4.43	33.10±4.81
랜덤 대체	19.53±4.06	19.59±4.01	19.54±3.92	19.62±3.91

<표 2>에서 정확도를 비교해 보면, 랜덤대체와 중앙값 대체의 경우에는 상관계수의 증가와 상관없이 정확도가 낮은 반면에 다른 네 가지 대체법들은 상관계수가 증가함에 따라 정확도가 증가하는 것을 알 수 있다. 특히 대체층 중앙값 대체의 경우에는 상관계수가 증가함에 따라 정확도가 가장 많이 증가하는 것으로 나타났다. 또한 대체층을 형성한 대체법들과 대체층을 형성하지 않은 대체법들을 비교해 보면, 상관계수가 낮은 경우에는 차이가 거의 없지만 상관계수가 높은 쪽에서는 대체층을 형성한 세 가지 대체법들이 대체층을 형성하지 않은 세 가지 대체법보다 정확도가 높은 것으로 나타났다. 한편 대체층을 형성한 세 가지 대체법 중에서는 대체층 중앙값 대체법, 대체층 k-NN 대체법, 대체층 랜덤 대체법의 순으로 정확도가 높은 반면에 대체층을 형성하지 않은 세 가지 대체법 중에서는 대체층 k-NN 대체법, 대체층 중앙값 대체법, 대체층 랜덤 대체법의 순으로 정확도가 높은 것으로 나타났다.

<표 3> 모의실험 RMSE(Mean±SD)

대체법	$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
대체층 중앙값 대체	1.405±0.062	1.340±0.073	1.199±0.082	0.684±0.086
대체층 k-NN 대체	1.974±0.114	1.862±0.113	1.598±0.111	1.173±0.109
대체층 랜덤 대체	1.980±0.119	1.976±0.114	1.971±0.117	1.690±0.108
중앙값 대체	1.404±0.060	1.397±0.059	1.403±0.061	1.407±0.059
k-NN 대체	1.942±0.116	1.857±0.116	1.651±0.117	1.253±0.094
랜덤 대체	1.984±0.118	1.977±0.112	1.979±0.117	1.981±0.115

<표 3>에서 RMSE를 비교해 보면 <표 2>의 정확도와 거의 유사한 결과를 보이는 것으로 나타났다. 따라서 상관계수가 높아질수록 대체층을 형성한 대체법이 대체층을 형성하지 않은 대체법보다 더 좋은 결과를 보여주고 있음을 알 수 있다.

다음으로 각 문항 및 항목별로 결측값 발생건수와 비율 뿐만 아니라 6가지 대체법에 따라 대체를 실시했을 때 정확하게 대체한 대체수 및 대체비율을 구해 보았는데, $\rho = 0.6$일 때만을 정리한 것이 <부록 1>이다. 원자료에 대한 결측비율을 살펴보면 5가지 항목에 대해 거의 20%에 근사한 결측비율을 갖는 것으로 나타났다. 그리고 6가지의 서로 다른 대체법은 5 문항 각각에 대해 거의 비슷한 정확성을 갖고 있는 것으로 나타났다. 이는 5 문항 사이의 상관계수를 동일하게 고정했기 때문인 것으로 생각할 수 있다. 한편 대체층을 나누지 않은 중앙값 대체의 경우에는 대체되는 값이 거의 '3'이기 때문에 항목 3만이 100%에 가까운 정확도를 가질 뿐 나머지 항목은 0%의 정확성을 갖는 것으로 나타나 실제 자료에 적용하기에는 무리가 있을 것으로 생각된다.

앞서 4번의 시뮬레이션을 통해 상관정도에 따른 대체법의 효과를 알아봤다. 상관정도가 높을수록 대체층을 형성하지 않은 대체법보다는 대체층을 형성한 대체법이 결과가 좋았다.

두 번째 모의실험은 "초등학교의 자아존중감과 건강행위"(문수미, 2006)에서 설문조사된 실제 자료에 적용해 보았다. 자료의 항목 및 문항은 <표 4>와 같다.

<표 4> 실제 자료의 항목과 빈도수

문항 \ 항목	1	2	3	4
부 교육정도	초등졸(6)	중졸(5)	고졸(105)	대졸 이상(187)
모 교육정도	초등졸(5)	중졸(8)	고졸(137)	대졸 이상(153)
성적	90이상(174)	89~70(118)	60이하(11)	
평소 여가시간	한시간 미만(34)	2시간 미만(118)	3시간 미만(72)	3시간 이상(79)
가족 여가시간	자주(127)	가끔(133)	거의 없음(43)	

( ):항목당 빈도수

먼저 총 332개 개체에서 결측값이 있는 29개 개체를 제외한 303개의 개체를 사용했으며, 사용된 문항은 부 교육정도(항목:4), 모 교육정도(항목:4), 성적(항목:3), 평소 여가시간(항목:4), 가족 여가시간(항목:3) 등 5개의 문항을 사용하였다. <표 4>에 있는 항목들의 비율을 살펴보면, 문항에 따라 항목들의 비율이 순서에 상관없이 서로 다름을 알 수 있는데, 이는 모의실험에서 모든 항목에 동일한 비율로 구성된 자료와 다름을 알 수 있다. 한편 실험 방법은 모의실험과 동일하게 3가지 대체법과 대체층으로 나눈 후 투표하는 방법으로 6가지의 대체법을 사용하였으며, 1000회 반복하여 분석하였다. 그리고 5개 문항간의 상관관계를 알아보기 위한 스피어만의 상관계수는 <표 5>와 같다.

<표 5> 실제 자료의 상관계수

	부교육정도	모교육정도	성적	평소여가시간	가족여가시간
부교육정도	1.000	.683(.000)	-.209(.000)	-.100(.081)	-.081(.161)
모교육정도		1.000	-.203(.000)	-.098(.090)	-.050(.390)
성적			1.000	.067(.248)	.042(.462)
평소여가시간				1.000	-.088(.125)

<표 5>에 있는 문항들간의 상관계수를 살펴보면, 부 교육정도와 모 교육정도 사이의 상관계수만 0.683으로써 양의 상관관계가 강한 편이고 다른 문항들 사이의 상관관계는 약한 편임을 알 수 있다. 실제 자료를 이용하여 정확도와 RMSE를 계산한 결과는 <표 6>과 같다. <표 4>에서 부 교육정도, 모 교육정도, 성적과 같은 문항의 경우에 일부 항목은 상대적으로 개체의 수가 적기 때문에 대체방법의 적용에 있어서 이런 항목의 포함여부가 문항간의 상관계수 뿐만 아니라 실험결과에 영향을 줄 수도 있다. 하지만 항목수가 적은 정보를 포함한 개체의 제거는 대체법의 적용에 문제를 나타낼 수 있다. 먼저 문항당 10개 정도의 개체가 세 문항에 개별적으로 나타난다면 30개 정도의 개체가 제외되는 문제가 생긴다. 또한 한 문항에 대해서 대체를 실시하고자 할 때, 해당되는 개체수가 적은 항목을 제외하고 대체를 실시한다면, 다른 문항을 기준으로 대체하게 될 때마다 대체에 참여하게 되는 개체들의 정보가 달라지는 문제가 발생할 수 있다. 따라서 본 연구에서는 개체수가 적은 항목도 대체과정에 모두 포함하여 대체를 실시하였다. 실제 자료에서도 전체 문항에 대해 100개의 결측을 발생시킨 후에 6가지의 서로 다른 대체법으로 대체값을 구하여 정확도와 RMSE를 계산하였다. <표 6>에 있는 결과를 살펴보면, 대체층 중앙값 대체법이 정확도와 RMSE 측면에서 가장 좋은 결과를 보여주고 있다. 중앙값 대체법과 k-NN 대체법은 대체층을 형성한 경우가 좋은 결과를 나타내는 것에 반해 랜덤 대체법의 경우에는 대체층을 형성하지 않은 경우가 좀 더 좋은 결과를 나타내지만 차이가 크지 않음을 알 수 있다.

<표 6> 실제자료의 정확도와 RMSE(Mean±SD)

대체법	정확도	RMSE
대체층 중앙값 대체	50.80±4.70	0.79±0.062
대체층 k-NN 대체	43.09±4.73	0.97±0.077
대체층 랜덤 대체	26.11±4.23	1.49±0.085
중앙값 대체	44.43±5.10	0.86±0.064
k-NN 대체	40.17±4.79	1.01±0.079
랜덤 대체	27.67±4.37	1.39±0.082

5. 결론

현재 결측값을 줄이려는 노력으로 사전조사 문항수의 축소, 적절한 응답자의 선택 등을 하고 있음에도 불구하고 결측값의 발생을 완전히 차단할 수는 없다. 따라서 결측값을 줄이려는 노력과 더불어 결측값을 적절한 다른 값으로 대체하려는 연구들도 활발히 진행되고 있다. 본 연구에서는 기존에 무응답 개체를 이용하지 않았던 여러 가지 대체법들과 무응답 개체의 정보를 어느 정도 이용한 k-최근접 이웃대체(Jonsson P. and Wohlin C. 2006)에 대해 알아보았다. 또한 최빈값 대체, 중앙값 대체와 같은 단일 대체법의 한계를 극복하고 한정적 정보를 이용하는 k-최근접 이웃대체를 보완하기 위해 decomposition과 voting을 이용한 결측값 대체를 제안하였다.

모의실험에서 대체층을 형성한 대체법이 상관정도가 높을수록 더 좋은 결과를 보인 이유는 상관관계가 높으면 높을수록 항목간의 관계보다 문항간의 관계가 더 중요해지기 때문인데 대체층 중앙값 대체법의 경우 기존 중앙값 대체와는 달리 문항간의 상관정도가 높을수록 정확도와 RMSE가 모두 상당히 좋은 결과를 나타내었다. 실제 자료를 가지고 실험한 경우에도 대체층 중앙값 대체법이 가장 좋은 결과를 보여주었다. 단일대체법의 단점을 보완하고 결측값이 포함된 개체의 정보를 최대한 이용하기 위해 대체층을 형성하였으며, 대체층 내의 정보들을 이용하여 대체를 실시할 경우에 중앙값을 이용한 대체방법이 대체층을 형성하지 않을 경우에 비해 결과가 많이 우수해짐을 알 수 있었다. 랜덤 대체법의 경우에는 대체층의 형성 유무에 상관없이 무작위로 선택된 대체값이 대체되므로 인해 대체층의 정보가 잘 반영되지 않는 결과를 보여주고 있다.

하지만 제안된 방법은 변수들의 상관관계가 높은 경우에 더욱 효과적인 것으로 판단되므로 변수의 선택 문제에 대한 효과적인 접근을 한 후에 적용한다면 더욱 좋은 결과를 나타낼 것으로 생각된다.

추후에는 지금까지 이용하였던 방법 이외에도 대체층을 이용한 Bayesian 대체, 엔트로피 대체 뿐만 아니라 러프 셀 이론에서 사용되는 결정 규칙을 이용한 대체 등을 적용해 볼 수 있을 것으로 생각한다.

참고문헌

1. 남궁평 (1992). 표본조사에서 발생하는 무응답에 관한 다중대체, *한국경제*, Vol. 19, No. 1.
2. 문수미 (2006). *초등학생의 자아존중감과 건강행위*, 석사학위논문, 전북대학교.
3. 박성률 (2005). *결측 자료가 발생할 경우 조정된 편의를 사용한 비가중 대체*, 석사학위논문, 성균관대학교.
4. 박태성, 이승연 (1998). 무응답을 포함하는 범주형 자료의 분석, *응용통계연구* 제 11권 1호, pp. 83-95.
5. 배현주 (2004). *결측값 대체 방법에 대한 비교 연구*, 석사학위논문, 중앙대학교.
6. Chen G. (2003). How to Deal with Missing Categorical Data: Test of a

Simple Bayesian Method, *Organizational Research Methods*, Vol. 6, No. 3, Month, pp. 309-327.
7. Huisman M. (2000). Imputation of Missing Item Responses: Some Simple Techniques, *Quality & Quantity*, Vol. 34, No. 4, pp. 331-351.
8. Jonsson P. and Wohlin C. (2004). An Evaluation of k-Nearest Neighbour Imputation Using Likert Data, *10th international symposium on software metrics*, Chicago, September.
9. Jonsson P. and Wohlin C. (2006). Benchmarking k-nearest neighbour imputation with homogeneous Likert data, *Empirical Software Engineering* Vol. 11, pp. 463-489.
10. Latkowski, R. (2002). Application of Data Decomposition to Incomplete Information Systems, *Proceedings of the IIS'2002 Symposium on Intelligent Information Systems*, Physica-Verlag, pp. 321-330.
11. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York, Wiley.
12. Little, R. J. A. (1988). Missing data in large surveys. *Journal of Business and Economic Statistics*, Vol. 6, No.3, pp. 287-301.
13. Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*, 63, 581-592.
14. Rubin, D. B. (1978). Multiple Imputations in Sample Surveys, *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1978, pp. 20-34.
15. Rubin, D. B. (1987). *Multiple Imputation for nonresponse in surveys*, New York, Wiley.
16. Scheffer, J. (2002). Dealing with Missing Data, *Research Letters Information Mathematical Sciences*, Vol. 3, pp. 153-160.

[2007년 3월 접수, 2007년 7월 채택]

<부록> $\rho = 0.6$일 때의 정확성과 각 항목별로 정확히 대체한 비율

구 분			건수					비율				
			x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
원 자료		1	3,884	3,706	4,301	3,705	4,364	19.4	18.5	21.5	18.5	21.8
		2	3,743	4,064	3,702	4,352	3,394	18.7	20.3	18.5	21.8	17.0
		3	3,643	3,665	3,801	3,937	3,734	18.2	18.3	19.0	19.7	18.7
		4	4,411	4,066	3,724	3,468	4,269	22.1	20.3	18.6	17.3	21.3
		5	4,319	4,499	4,472	4,538	4,239	21.6	22.5	22.4	22.7	21.2
		전체	20,000	20,000	20,000	20,000	20,000	100.0	100.0	100.0	100.0	100.0
대체층 대체	중앙값 대체	1	1,767	552	2,038	1,260	2,077	45.5	14.9	47.4	34.0	47.6
		2	648	1,385	820	1,506	1,209	17.3	34.1	22.2	34.6	35.6
		3	1,759	1,291	1,247	1,137	1,406	48.3	35.2	32.8	28.9	37.7
		4	2,168	1,429	1,652	1,300	1,809	49.1	35.1	44.4	37.5	42.4
		5	792	776	744	691	715	18.3	17.2	16.6	15.2	16.9
		전체	7,134	5,433	6,501	5,894	7,216	35.7	27.2	32.5	29.5	36.1
	k-NN 대체	1	1,332	1,281	2,087	1,607	2,253	34.3	34.6	48.5	43.4	51.6
		2	620	1,005	588	943	472	16.6	24.7	15.9	21.7	13.9
		3	599	538	616	653	506	16.4	14.7	16.2	16.6	13.6
		4	878	1,028	548	487	1,231	19.9	25.3	14.7	14.0	28.8
		5	1,658	1,641	1,816	1,908	1,803	38.4	36.5	40.6	42.0	42.5
		전체	5,087	5,493	5,655	5,598	6,265	25.4	27.5	28.3	28.0	31.3
	랜덤 대체	1	828	821	871	749	885	21.3	22.2	20.3	20.2	20.3
		2	788	829	702	862	729	21.1	20.4	19.0	19.8	21.5
		3	781	753	732	814	767	21.4	20.5	19.3	20.7	20.5
		4	869	820	714	699	876	19.7	20.2	19.2	20.2	20.5
		5	806	916	888	911	890	18.7	20.4	19.9	20.1	21.0
		전체	4,072	4,139	3,907	4,035	4,147	20.4	20.7	19.5	20.2	20.7
기존 대체	중앙값 대체	1	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
		2	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
		3	3,643	3,665	3,801	3,937	3,734	100.0	100.0	100.0	100.0	100.0
		4	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
		5	0	0	0	0	0	0.0	0.0	0.0	0.0	0.0
		전체	3,643	3,665	3,801	3,937	3,734	18.2	18.3	19.0	19.7	18.7
	k-NN 대체	1	1,199	894	1,609	1,207	1,442	30.9	24.1	37.4	32.6	33.0
		2	643	1,239	769	935	427	17.2	30.5	20.8	21.5	12.6
		3	607	481	454	741	655	16.7	13.1	11.9	18.8	17.5
		4	936	879	694	678	1,199	21.2	21.6	18.6	19.6	28.1
		5	1,549	1,616	1,590	1,634	1,553	35.9	35.9	35.6	36.0	36.6
		전체	4,934	5,109	5,116	5,195	5,276	24.7	25.5	25.6	26.0	26.4
	랜덤 대체	1	781	717	857	720	845	20.1	19.3	19.9	19.4	19.4
		2	741	818	760	906	706	19.8	20.1	20.5	20.8	20.8
		3	725	719	763	803	755	19.9	19.6	20.1	20.4	20.2
		4	814	782	760	714	827	18.5	19.2	20.4	20.6	19.4
		5	866	898	923	901	801	20.1	20.0	20.6	19.9	18.9
		전체	3,927	3,934	4,063	4,044	3,934	19.6	19.7	20.3	20.2	19.7