범주형 자료의 결측치 처리 방법*

Solutions for Missing Values in Categorical Data

김 덕 준 (호서대학교 행정학과 교수)

Abstract

Duck-Joon Kim

The analytical focus of this study centers on how to treat the missing values in public administration research. Examining the mechanism of missing values in data gathering, the paper constructs a regression model for ordinal categorical data of the survey by the KOSSDA numbered A1-205-0040. Applying four solutions for missing values such as complete case analysis method, EM technique, regression methodology, and multiple imputation, this paper tries to approach the focus of the research.

Results confirm that the complete case analysis method can not be an appropriate solution for missing values in reality. The higher the percentage of missing values in data set, the riskier the application of this solution. On the contrary, the other three solutions turn out to be the appropriate remedies for missing values in general. Especially, the usefulness of EM technique and regression methodology can be highlighted. In short, future research should accommodate the necessity/accessibility of scientific techniques as well as investigate the mechanism of missing values in a given data set.

주제어: 범주형 자료, 결측치, EM 기법, 회귀분석기법, 다중대체기법, 완전사례 분석

Key Words: missing data, categorical data. EM technique, regression, multiple imputation

* 이 논문은 2007년도 호서대학교의 재원으로 학술연구비 지원을 받아 수행된 연구임(2007-0126).

I. 서 론

우리나라 행정연구에 있어서 과학적 방법론으로서의 계량분석기법의 활용이 증대되어 왔다. 2000년대 중반에 있어서는 당시 주요 학술지에 게재된 논문의 약 과반수 정도가 계량분석기법에 의거한 실증분석논문이라는 사실이 밝혀진 바 있다.[1] 지금까지도 이러한 연구경향이 계속되어 최근 2010년을 기준으로 볼 때, 동일한 학술지에 게재된 논문 중 약 67% 정도가 계량기법을 활용한 연구임을 발견할 수 있다.[2]

이러한 계량적 분석기법이 연구방법론의 주류를 형성하고 있는 상황에서 제기될 수 있는 가장 기본적이고 중요한 문제는 바로 연구결과의 신뢰성 문제이다. 즉, 계량분석기법을 활용한 연구결과가 과연 어느 정도로 이론적으로나 현실적으로 타당하며 효용성을 보유하는 것인가가 중요문제로 제기된다. 기본적으로 이 문제에 대한 답은 활용되는 계량분석기법 자체의 방법론적 논리 및 적용되는 대상으로서의 자료의 질과 양이 과연 어느 정도 수준인가에 따라서 좌우될 수 있음에 주목할 필요가 있다. 환언하자면, 분석자료의 질과 양은 이에 적용되는 분석기법의 과학성 및 정치성(精緻性)과 함께 분석결과의 효용성과 타당성을 좌우하는 영향인자로서 기능한다는 사실이 강조되어야 한다는 것이다.

보다 구체적으로, 대부분의 실증적 연구가 한정된 시간, 예산 및 조사과정의 정확성을 확보하기 위한 방편으로서 전수조사 대신 모집단의 일부 표본을 분석대상으로 설정하는 표본조사의 방식을 채택하고 있음이 사실이다. 하지만 문제는 이러한 이유로 인하여 실증분석을 위한 대상으로서 추출된 표본으로부터 모집단의 특성·본질·속성을 정확하게 추론하지 못한다면 연구결과의 신뢰성이 크게 훼손 받을 수밖에 없다는 사실이다.

이러한 신뢰성의 훼손은 곧 표본오차와 비표본오차에 기인하는 것으로서, 전자는 표본이 적합하게 추출되지 못함으로서 발생되는 오류[3]를 의미하며 후자는 추출된 표본으로부터 분석주제에 관한 정보를 얻지 못함으로써 발생되는 오류를 지칭한다. 따라서 연구자의 입장에서는 연구결과의 신뢰성을 확보하기 위해서는, 즉, 자료의 질과

---

1) 강민아·김경아(2006)는 2000년부터 2004년까지 5년간 행정학분야 3대 주요학술지인 한국행정학보, 한국정책학회보, 정책분석평가학회보에 게재된 전체 논문 중 약 53% 정도가 실증분석논문인 것을 밝힌 바 있다.

2) 주석 1)의 3대 주요학술지를 대상으로 했을 때, 2010년에 게재된 전체 일반논문 133편 중 89편이 계량분석기법을 활용하고 있는 것으로 나타난다.

3) 표본오차가 분석결과에 미치는 오류를 심각하게 인식하였던 최초의 학문분야는 생물통계학 분야로 알려져 있으며, 이러한 문제점이 사회과학분야에 인지되기 시작한 것은 Tobin이 계량경제학분야에서 explicit selection의 문제를 취급하기 시작하면서부터이다(Tobin, 1958). 표본오차가 초래하는 오류를 시각적으로 표시함으로써 보다 용이하게 문제의 본질을 설명하고 있는 학자로서는 Berk를 들 수 있겠다(Berk, 1983).

양을 만족할 만한 수준으로 확보하기 위해서는 무엇보다도 먼저 표본오차 및 비표본오차의 축소를 중요목표로서 설정하여야 한다.

하지만, 분석결과의 신뢰성에 직접적으로 영향을 미칠 수밖에 없는 영향인자로서의 표본오차 및 비표본오차의 중요성에도 불구하고, 기존연구들에 있어서는 이러한 분석주제들이 여타의 주제들에 비하여 상대적으로 경시되어 온 것이 사실이다. 특히, 비표본오차는 무응답이나 부정확한 측정도구의 사용과 이에 의거한 응답, 그리고 자료입력시의 오류 등에 기인하는 오차로서 이해될 수 있는 바, 아직까지는 행정학분야는 물론 더 포괄적 관점에서의 사회과학분야에서 이론적 조망의 중요대상으로서 선택받지 못하고 있는 것으로 파악된다. 나아가 비표본오차 중에서도 무응답 혹은 응답거부는 대부분의 조사연구에 있어서 하나의 보편적 현상으로까지 간주되고 있으며 이로 인한 결측치의 존재에 별다른 중요성을 부여하지 않고 있는 것이 현실의 한 단면이라고 평가할 수 있겠다.[4)]

기본적으로 다수의 결측치가 발생된다는 사실은 곧 연구설계에서 계획된 분석대상의 수가 감소됨을 의미한다. 덧붙여 결측치가 통계적인 무작위방식으로 발생되지 않고 모집단을 구성하는 어떤 특별한 하위집단을 중심으로 체계적으로 발생되는 것이라면 관측치는 모집단의 일부집단을 배제하고 있는 것과 마찬가지이기에 결국 모집단을 적절히 대표하는 것으로서의 표본의 대표성이 심각히 위협받을 수밖에 없게 된다. 따라서 결측치의 문제를 심각하게 인식하고 그로부터 연원되는 위험성을 보완할 수 있는 방법론적 기법의 개발과 활용이 필요함은 자명한 사실이다.

특히, 대부분의 행정조사연구에서는 조사대상에게 주어진 항목에 대한 답을 몇 가지 보기 중에서 선택토록 하는 자료수집방법을 활용하고 있음이 일반적이다. 즉, 분석대상으로서 수집되어 활용되는 자료는 범주형 자료(categorical data)의 성격을 보유하는 경우가 대부분이다. 이러한 범주형 자료를 분석함에 있어서는 자료의 특성상 연속형 자료의 분석에 활용되는 기법이 그대로 적용될 수 없다. 그렇다면 범주형 자료

4) King 등의 논문(King et al., 2001)은 1993년부터 1997년까지 미국, 영국의 대표적 3대 정치학 학술지인 American Political Science Review, Ameican Journal of Political Science, British Journal of Political Science에 게재된 설문조사를 활용한 계량분석논문들을 대상으로 하고 있는 바, 논문의 설문조사 응답자들의 약 절반 정도가 최소 한 두 개의 질문에 응답치 않음으로써 결과적으로 수많은 결측치가 발생되었음에 반하여, 실제 논문의 20% 미만만이 결측치의 처리방법을 밝히고 있음을 지적하고 있다. 우리나라의 경우에 있어서는 그러한 정도가 더욱 심각한 것으로 파악되는 바, 김태일·서주현(1998)은 1991년부터 1997년까지 한국행정학보에 게재된 설문조사를 활용한 실증분석논문을 분석대상으로 설정하였을 때, 우선 설문지의 회수율을 밝히지 않았던 논문이 전체의 약 40%에 이르고 있을 정도로 결측치 문제에 관한 인식이 미흡한 것을 밝히고 있다. 또한 강민아·김경아(2006) 역시 2000년부터 2004년까지 한국행정학보, 한국정책학회보, 정책분석평가학회보에 게재된 실증분석논문을 대상으로 할 때 결측치를 언급한 논문들은 매년 한 두편에 불과하다는 것을 보고하고 있다.

의 결측치를 분석하고 처리함에 있어서도 당연히 연속형 자료의 그것과는 차별성을 보유하는 관점이 적용되어야 마땅할 것이다.

바로 이러한 배경에서 본 논문은 실제 행정조사연구의 분석자료 중 대부분을 차지하는 범주형 자료의 수집에 있어서 결측치가 발생되는 경우 그것이 어떻게 발생되는 것인지를 살펴보고 이러한 결측치들을 보완할 수 있는 있는 방법들을 점검키로 한다. 나아가 실제 행정조사연구에 활용되고 있는 자료에 제반 보완기법들을 적용해 봄으로써 각기 보완 방법들의 결과 및 방법론적 특성을 논의하는 기회를 가져보기로 한다.

II. 결측치의 발생기제와 유형

결측치 문제에 관한 논의를 전개해 나가기 위해서는 먼저 결측치가 어떻게 발생된 것인지의 발생기제를 점검해야만 한다. 즉, 결측치의 발생기제에 따라서 상이한 처리방법이 적용되어야 한다는 사실에 모든 연구자들이 의견일치를 보이고 있다. 결측치가 발생되는 기제에 따라서 결측치의 유형은 크게 "완전무작위 결측치"(MCAR: Missing Completely At Random), "무작위적 결측치"(MAR: Missing At Random), "무시할 수 없는 결측치"(NI: Non-ignorable)의 3가지 유형으로 구분되고 있다 (Rubin, 1987; King et al., 2001; Little and Rubin, 2002; Gibson and Olejnik, 2003).

1. 완전무작위 결측치(MCAR)

완전무작위 결측치는 X와 Y의 두 가지 변수를 상정할 때 Y 변수의 결측치가 발생될 가능성이 X, Y 두 변수 어느 것의 영향으로부터도 독립적인 방식으로, 즉 통계학적 의미로서 무작위적으로 발생된 것을 말한다. 쉬운 예를 들자면 설문조사지를 응답자들로부터 무작위로 수거하여 책상 위에 놓아두었는데 갑자기 창문으로 불어 닥친 회오리 바람에 의해서 설문지의 상당 부분이 바닥으로 떨어져서 오염되고 더러는 창문 밖으로 날아가 버려서 전체설문지를 수거할 수 없게 된 경우를 생각해 볼 수 있겠다. 이러한 경우에 결과적으로 획득된 관측치들은 조사대상자집단의 무작위 표본으로 간주될 수 있을 것이며, 반면 발생된 결측치들은 바로 완전무작위 결측치에 해당되는 것으로서 이해될 수 있겠다.

결국, 완전무작위 결측치에 있어서는 특정 설문지를 대상으로 판단해 볼 때, 주어진 설문지 내용으로부터 도출될 수 있는 어떠한 정보를 가지고도 결측치가 발생될 확

률을 예측할 수 없는 상황이라고 할 수 있다. 이러한 맥락에서 어떤 응답자가 주어진 질문에 응답할 가능성은 주사위 던지기와 같은 무작위 확률에 의해서 발생한다고 가정할 수 있겠다.

결측치의 발생기제가 이와 같은 완전무작위적인 것이라면 연구자는 결측치가 없는 자료만을 대상으로 분석을 수행하고 그 결과로 모집단의 속성을 추론하는 것이 가능할 것이다. 왜냐하면, 이 때 관측치들은 본래 모집단의 무작위 표본으로서 간주될 수 있기 때문이다. 하지만 문제는 결측치의 발생기제가 완전무작위적이라는 논거의 현실성이 몹시 희박한 것으로서 판단될 수 있다는 것이다. 즉, 실제 발생되는 결측치들의 어느 정도가 이에 해당될 수 있는가는 상당한 논란이 제기될 수밖에 없는 문제인 것이다.

2. 무작위적 결측치(MAR)

무작위적 결측은 어떤 결측치가 발생할 가능성을 관측된 자료에 의해서 예측할 수 있으나 관측된 자료를 통제한다면 결측자료와는 통계적으로 독립적인 경우를 의미한다. 즉 관측된 변수에서 결측치가 발생할 가능성은 관측된 변수(들)값의 영향을 받는 것이기에 주어진 관측치를 활용하여 결측치가 발생될 가능성을 예측할 수 있게 된다.

이러한 무작위적 결측치와 완전무작위 결측치의 경우 그 발생기제를 “무시할 수 있는(ignorable)” 것으로서 간주하여 바로 다음에 살펴 볼 “무시할 수 없는” 결측치와 구별지울 수 있다. 기본적으로 결측치의 발생기제에 따라서 그 발생기제에 합당한 결측치의 보완기법이 적용되기에 발생기제를 구분하고 정리함이 요구되는 것을 이해할 수 있겠다. 즉, 결측치의 발생기제가 무시할 수 있는 것이라면 그 보완기법으로서 본 연구에서 제시할 여러 기법들이 활용될 수 있을 것이다. 물론, 보다 근원적 문제로서 과연 결측치의 발생기제가 무시할 수 있는 것인가에 관한 구체적이고 설득력 있는 입증이 제시될 수 있는가에 관한 논란이 제기될 수 있음도 지적할 필요가 있겠다.

3. 무시할 수 없는 결측(NI)

결측치가 발생될 가능성을 관측된 자료에 의해서는 예측할 수 없는 경우의 결측을 “무시할 수 없는” 결측이라 한다. 환언하자면, 결측치의 발생 원인이 결측된 변수의 값과 직접적으로 관련이 있다면, 이러한 경우의 결측치 발생기제가 “무시할 수 없는” 것이다. 예로서, 교육수준이 낮은 저학력 계층에서 학력에 관한 설문조사에 응답치 않을 가능성이 높을 때, 입수한 관측치들을 가지고는 어떤 조사대상자가 저학력 계층인

지를 직접적으로 파악할 수 없는 경우가 이에 해당된다.

이러한 "무시할 수 없는" 결측을 보완하기 위해서는 "무시할 수 있는" 결측의 경우보다 더 복잡하고도 난해한 절차와 과정들이 요구된다(King et al, 2001; Little and Rubin, 1987, 2002; Schafer, 1997). 따라서 가능하다면 "무시할 수 없는" 결측을 "무시할 수 있는" 결측의 문제로서 전환시키는 것이 분석의 효율성과 편의성의 제고라는 측면에서 바람직하다. 예로서 학력에 관한 조사에 있어서 학력과 함께 그 대안적 변수로서 고용직종이나 소득수준에 관한 조사가 이루어진다면 저학력 계층에서의 학력에 관한 결측치 발생의 문제를 무작위적 결측치의 문제로서 전환시킬 수 있기에 보다 효율적이고 용이한 보완기법의 적용이 가능할 수도 있다는 것이다.

III. 범주형 자료의 결측치 보완기법

범주형 자료의 결측치를 보완하기 위한 몇 가지 방법론이 제시되었던 바, 그 대표적인 방법론의 기본 논리와 특성은 다음과 같다.

1. 완전사례분석기법(Complete Case Anaysis Method)

결측치가 발생된 경우에 있어서 가장 보편적이며 용이한 해결방법은 완전한 관측치를 보유하는 사례들만을 대상으로 분석을 수행하는 완전사례분석기법 혹은 완전제거기법이다.[5] 기본적으로 이같은 기법은 결측치의 발생기제가 앞서 설명한 바와 같은 완전무작위 결측의 경우에 전형적으로 활용될 수 있다. 또한 결측치의 비율이 극소인 경우에 그 실용성이 제고되며 선호되는 보완기법으로 평가된다.

그러나, 완전사례분석기법을 적용함에 있어서는 실제 분석의 대상이 되는 자료가 완전무작위 결측의 가정을 어느 정도로 만족시키는가가 문제로 제기될 수밖에 없게 된다. 나아가, 결측치의 비율이 극소인 경우에 그 활용이 선호된다면, 과연 어느 정도 수의 결측치가 극소에 해당하는 것인지 – 3%인지, 5%인지, 혹은 10%인지 –가 문제로 제기될 수밖에 없다. 예로서, Pigott(2001)에 따르면 154명의 조사대상자 중 88%가 전체 7개 변수 중에서 단지 한 두개의 변수에 대해서만 결측치를 나타내는 경우에 있어서 완전사례분석기법을 활용한다면, 결과적으로 분석대상자의 수는 20명

5) 이러한 방법은 대부분의 통계프로그램에서 선형회귀분석을 수행하는 경우에서 볼 수 있는 디폴트 분석방식으로서 모형을 구성하는 변수값이 하나라도 결측된 경우들을 모두 제외하고 완전한 관측치를 지닌 경우만을 대상으로 분석을 수행하는 방법을 말한다.

미만으로 축소됨을 보고한 바 있다. 따라서, 이러한 완전사례분석기법은 분석대상의 수를 크게 축소시킴[6]으로써 추정치의 정확성이 훼손된다는 문제점을 노출한다.

2. EM기법

이 기법은 모든 변수들의 결합분포(joint distribution)가 다변량 정규분포를 취하고 있다는 가정과 결측치의 발생기제가 "무시할 수 있는" 것이라는 가정 위에서 결측치를 보완하는 방식이다(Dempster, Laird, and Rubin, 1977; Little and Rubin, 1987; Schafer, 1997). 구체적으로, 이 기법은 결측치를 보완하기 위한 두 단계 – E(Expectation) 단계와 M(Maximization) 단계를 거친다. 먼저 E 단계에 있어서는 관측된 자료들과 계수 추정치들이 주어졌을 때의 결측치들의 조건부 기댓값을 구하여 이로써 결측치를 대체한다. 이어서 M 단계에서는 E 단계를 통해서 기댓값으로써 결측치들이 대체되어 완성된 데이터 셋의 loglikelihood를 최대화시킴으로써 다시금 계수 추정치들의 수치를 수정해 나가는 단계이다.

결국, E 단계와 M 단계가 반복된 결과로서 계수 추정치가 일정한 값에 수렴하고 결측치들의 조건부 기댓값 역시 일정한 수치에 수렴한다. 바로 이렇게 결측치의 대체값을 구하는 것이 EM기법의 핵심이나, 과연 이러한 반복적 과정을 거친 결측치의 대체법이 표준분석방법이 될 수 있는가의 방법론상의 원론적 문제가 제기될 수 있겠다.

3. 회귀분석기법(Regression Method)

회귀분석기법의 기본 논리는 결측이 발생된 변수를 종속변수로 설정하고, 반면 관측된 변수들을 독립변수로 설정하여 회귀분석을 실시한 후, 획득한 종속변수의 회귀계수로서 결측치를 대체하는 방법이다. 여기서 만일 회귀계수 그대로 결측치를 대체한다면 이는 회귀계수의 표준편차를 전혀 반영하지 않게 되기에 일반적으로 결측치 대체값의 표준편차 혹은 분산의 수치가 과소측정됨으로써 이후의 분석결과에 있어서 Type I 오류의 가능성을 증대시키는 문제점을 지닌다(Little, 1992; Allison, 2002). 이러한 문제점을 보완키 위하여 회귀계수를 계산함에 있어서 오차항을 반영하는 기법이 제시되었던 바, 이 역시 변수들이 고도의 상관관계를 지니고 있는 경우라면 회귀계수의 일관성이 문제시될 수 있음이 지적된 바 있다.

---

6) 종속변수와 n개의 독립변수로 구성된 분석모형에서 만일 각개 변수의 x%가 결측된 경우, 완전한 관측치들만을 사용하여 분석을 한다면 이론적으로 최대 $1-(1-X/100)^n$ 의 자료가 분석의 대상으로부터 제외될 수 있다. 예를 든다면 4개 변수 각각에 있어서 10%의 결측치가 존재하는 경우에는 최대 34.39%의 자료가 분석대상에서 제외될 수 있다.

만일 자료가 행정조사연구의 대다수 분석자료에서와 같이 범주형 자료인 경우라면 회귀분석기법의 적용에 있어서도 범주형 자료의 회귀분석을 활용하여야 할 것이다. 즉, 결측이 된 변수가 이분형(binary) 혹은 대조형(contrast) 변수인 경우라면 로짓 회귀분석기법의 사용이 권장될 것이며, 리커트 척도(Likert scales)를 활용한 것과 같은 서열적(ordinal) 범주형 변수라면 최적화 척도법에 의거한 범주형 자료의 회귀분석기법(Regression for Ordinal Categorical Data)을 사용함이 바람직하다.

4. 다중대체기법(Multiple Imputation Method)

다중대체기법은 EM기법과 마찬가지로 모든 변수의 결합분포가 다변량 정규분포를 취하고 있음을 가정하지만 EM기법과는 달리 표준분석방식을 사용하여 결측치의 대체값을 구하는 방식이다. 즉, 다중대체기법에서는 결측치들의 조건부 기댓값을 구하는 대신 그 추정치를 계산한다. 여기서 추정치는 관측치들을 주어진 것으로 할 때, 결측된 변수의 분포로부터 무작위로 추출되는 것으로 간주하여 사후확률분포(posterior probability distribution)로부터 구한다.

실제 사후확률분포를 구하는 절차는 데이터증대(DA: Data Augmentation)라는 두 단계의 복잡한 알고리즘으로 구성된다. 즉, 그 첫 단계인 I 단계에서는 모수를 알고 있다고 가정하고 관측치와 현재의 모수의 추정치를 주어진 것으로 간주할 때의 결측치들의 조건부 분포로부터 결측치의 값을 표본추출한다. 이어서 다음 단계인 P 단계에서는 완전한 데이터가 구성된 것으로 가정하고 모수를 다시를 추정한다(Schafer, 1997; Tanner, 1993; Tanner and Wong, 1987). 결국, 수많은 반복 과정을 거친 데이터증대를 통하여 여러 개의 완성된 데이터 셋[7]이 만들어지며 이를 대상으로 결측

7) 여기서 완성된 데이터 셋의 구체적인 개수는 어느 정도가 적당한가가 문제시 될 수 있는 바, 가상 데이터의 수 m을 많이 생성하면 할수록 보다 바람직한 결과를 도출할 수 있겠지만, 분석에 있어서 너무 많은 시간이 소요되기에 그 효율성이 의문시될 수 있을 것이다. 이러한 맥락에서 일반적으로 m은 아래 <표: 결측정보비 r과 가상데이터 셋의 수 m에 기반한 다중대체법의 효용성>에서 볼 수 있듯이 그 수가 3에서 5개 정도가 적절하다고 주장된 바 있다.

<표: 결측정보비 r과 가상데이터 셋의 수 m에 기반한 다중대체법의 효용성>

	r				
	0.1	0.3	0.5	0.7	0.9
3	97%	91%	86%	81%	77%
5	98%	94%	91%	88%	85%
10	99%	97%	95%	93%	92%
20	100%	99%	98%	97%	96%

표에 관련되어, Rubin(1987)은 m개의 가상 데이터 수에 의거한 대체추정량의 효율성은 다음과 같은 공식을 사용하여 구할 수 있다고 설명한다.

즉, 추정량의 효율성 = $[1+(r/m)]^{-1}$ 이다.

치에 대한 최종 추정치를 구하게 된다(Rubin, 1987; Schafer, 1997).

IV. 보완 방법들의 실제 적용

이제 본 연구는 앞서의 여러 결측치 보완기법들 중에서 완전사례분석기법, EM기법, 회귀분석기법, 다중대체기법 각각이 실제로 행정정책분야의 연구에 활용되는 범주형 자료에 적용되었을 때 어떠한 결과를 가져 올 수 있는가를 점검키로 한다. 이러한 보완 기법들의 적용을 위한 분석자료로서 본 연구는 박중훈・서성아, 『공직부패의 실태에 관한 설문조사 2005』를 활용하며, 여기에 SPSS-V18을 적용하기로 한다.

1. 분석모형의 구성

분석자료로 활용된 『공직부패의 실태에 관한 설문조사 2005』는 전국 5개 도시(서울, 부산, 대구, 대전, 광주)의 자영업자를 비롯한 기업인을 대상으로 우리나라 공직사회의 부정부패 실태와 그 해소방안에 관하여 직접면접에 의한 자기기입식 방법으로 조사한 설문자료이다. 자료에서 총 500명의 응답자들은 약 70개에 가까운 문항에 답하는 바, 이들 대부분의 문항은 서열적(ordinary) 범주형 설문에 해당한다. 본 연구는 결측치를 지니고 있지 않은 이들 500개의 관측치를 모집단으로 설정하며, 4개 설문을 변수화하는 바, 이들은 곧 "금품 제공의 필요성," "금품제공이 행해지는 정도 – 즉, 금품제공의 보편성," "금품 제공이 업무처리에 미치는 영향 – 즉, 금품제공의 영향," 그리고 "월평균소득"이다. 따라서 본 논문에서는 대표적인 결측치 보완기법들을 적용시킬 다음과 같은 분석모형을 구성하였다.[8)]

금품제공의 필요성 = 금품제공의 보편성 + 금품제공의 영향 + 월평균소득
(Var1) (Var2) (Var3) (Var4)

여기서 r은 결측정보비(proportion missing information):

$$r = \frac{\gamma + 2/(df+3)}{\gamma+1}, \quad \gamma = \frac{(1+m^{-1})^B}{\overline{U}}$$ 이다.

간략히 요약하자면, 결측정보비가 주어졌을 때 위의 표는 m의 개수에 따른 다중대체법의 효율성을 나타내는 바, 만약 결측정보비가 0.3이하인 경우라면 m이 3개에서 5개 정도만 되어도 m=∞에 비하여 그 효율성이 크게 저하되지 않음을 알 수 있다. 즉, m의 사이즈를 5개 이상으로 증가시켜도 그 때 추가적으로 획득하게 되는 효율성이 극히 미미한 것임이 드러난다.

8) 본 논문의 분석초점이 결측치 보완 기법의 효과성/효율성에 설정되기에 여기에 제시된 분석모형의 적합성이나 그 결과해석은 논외로 삼기로 한다.

모형의 분석기법으로서는 모형을 구성하는 모든 변수가 서열적 범주형 변수이기에 먼저 모집단을 대상으로 최적화 척도법에 의거한 범주형 자료의 회귀분석(regression for ordinal categorical data)을 수행하여 그 계수를 구한다. 그리고, 임의적으로 모집단의 측정치에 일정 비율의 결측치를 부여하고 이를 대상으로 앞서와 같은 결측치 보완기법들을 적용한 후 동일한 회귀분석을 적용하여 회귀계수를 구하기로 한다. 결국, 이렇게 획득한 결측치를 부여한 자료들을 대상으로 하는 회귀계수들을 모집단의 그것과 비교해 봄으로써 각각의 결측치 보완 기법의 효과 또는 효율을 논하기로 한다.

보다 구체적으로 본 연구에서는 분석모형을 구성하는 종속변수로서의 금품제공의 필요성과 독립변수 중 금품제공의 보편성은 모집단 전체에 있어서 결측치 없이 완전하게 관측되는 것으로 처리하기로 한다.[9] 반면 독립변수 금품제공의 영향은 완전무작위 결측치(MCAR)의 발생기제를 지닌 것으로 설정한다.[10] 덧붙여 독립변수 월평균소득은 무작위 결측치(MAR)의 발생기제를 지니는 것으로 설정하여 분석모형을 구성하는 독립변수인 금품제공의 보편성의 값에 따라서, 즉 금품제공의 보편성의 측정치가 평균이하의 관측치를 보유하는 분석대상에서만 결측치가 발생하는 것으로 설정한다.[11] 또한 금품제공의 영향과 월평균소득에서의 결측치의 비율은 각각 변수별로 전체의 10%(=50개), 20%(=100개), 30%(=150개), 40%(200개)로 설정한다. 여기서 분석의 타당성을 확보하기 위한 방법으로서, 각각의 결측치 비율마다 다섯 개의 상이

9) 앞서의 주석 5)에서 지적한 바와 같이 종속변수에서 결측치가 발생되는 경우는 본 논문의 분석주제에 해당하지 않기에 실증분석에서도 이를 취급하지 않기로 한다.

10) 이러한 설정은 다중대체방법이 그 적용의 기본가정으로서 무시할 수 있는 결측치 발생 중 무작위적 결측치(MAR)를 설정하고 있으나 완전무작위 결측치의 경우에도 적용된다면 어떠한 효과를 발생시키는지를 점검할 수 있는 기회를 제공할 것이다.

11) 즉, 금품제공의 보편성의 값을 1, 2, 3으로 답한 연구대상은 전체 연구대상의 약 56%인 바, 이들에서만 월평균소득의 결측치가 발생하는 것으로 설정한다. 아래의 <표: 변수값 설명>은 본 연구에서 활용된 각 변수의 값이 의미하는 바를 나타낸다.

<표: 변수값 설명>

변수	변수값	변수값 설명	변수	변수값	변수값 설명
금품제공 필요성 Var1	1	매우 필요함	금품제공의 보편성 Var2	1	매우 보편적
	2	필요함		2	보편적
	3	필요한 편		3	보편적인 편
	4	필요없는 편		4	예외적인 편
	5	불필요함		5	예외적
	6	전혀 불필요함		6	매우 예외적
금품제공의 영향 Var3	1	매우 긍정적	월평균소득 Var4	1	100만원 이하
	2	긍정적		2	100~200만원
	3	긍정적인 편		3	200~300만원
	4	부정적인 편		4	300~400만원
	5	부정적		5	400~500만원
	6	매우 부정적		6	500만원 이상

한 자료집단을 구성하고 각각의 자료에 최적화 척도법에 의한 범주형 자료의 회귀분석을 실행한다. 즉, 예로서 금품제공의 영향과 월평균소득 각각에 결측치가 10%가 발생하는 다섯 개의 자료집단을 만들어 그 각각에 회귀분석을 실행하고, 결측치 20%, 30%, 40%의 경우에도 마찬가지로 각각 다섯 개의 자료집단을 구성하고 회귀분석을 실행한다.

2. 실증분석의 결과 해석

완전사례분석기법을 포함한 4개의 결측치 보완 기법을 통한 서열적 범주형 자료에 대한 회귀분석의 수행 결과는 <표 1>, <표 2>, <표3>, 그리고 <표 4>에 제시되고 있다. 모든 표의 두 번째 행에는 분석기법의 효과성/효율성 비교를 위한 기준으로서 결측치가 존재하지 않는 경우, 즉 원자료를 대상으로 한 분석결과가 보고되어 있다.

또한, 이들 표를 기반으로 작성된 <표 5>는 동일한 결측치 비율을 지니는 5개의 자료집단을 대상으로 각각의 보완기법을 활용하여 계산된 변수 추정치들 중 몇 개가 모수의 일정 범위에 위치하고 있는가를 보여줌으로써 추정의 불편성과 일관성을 비교평가하는 근거를 제공한다.

<표 1>부터 <표 4>까지 공통적으로 기재되어 있듯, 결측치가 없는 모집단을 분석대상으로 정하여 서열적 범주형 자료들인 Var1을 종속변수로 Var2, Var3, Var4를 독립변수로 설정한 최적화 척도법에 의한 다중회귀분석의 F값은 9 이상의 값을 지닌다. 따라서, 전체적으로 파악할 때 분석모형의 구성에서는 문제가 없는 것으로 해석된다. 독립변수의 표준화된 계수 추정치에 있어서, Var2와 Var3의 추정계수는 각각 0.224와 0.241로서 그 신뢰수준은 95% 훨씬 이상인 것으로 보고되었다. 반면 Var3은 종속변수에 통계학적으로 영향을 미치지 못하는 것으로 해석된다.

<표 1> 결측치 10%의 경우 보완기법의 결과 비교

		R^2	Adj R^2	*Regression F−value*	Var2	Var3	Var4
모집단 분석결과		0.130	0.116	9.189	0.224** (0.062)	0.241** (0.047)	0.054 (0.114)
LW	1	0.126	0.108	7.130	0.211** (0.091)	0.241 ** (0.063)	−0.096 (0.132)
	2	0.186	0.174	15.094	0.262** (0.065)	0.300** (0.058)	0.094 (0.127)
	3	0.110	0.095	7.027	0.210** (0.062)	0.215** (0.111)	0.080 (0.130)
	4	0.130	0.115	8.479	0.216** (0.070)	0.247** (0.058)	0.056 (0.129)
	5	0.134	0.116	7.650	0.210** (0.075)	0.256** (0.070)	0.082 (0.129)
EM	1	0.141	0.129	11.545	0.222** (0.057)	0.255** (0.047)	0.101 (0.119)
	2	0.141	0.129	11.524	0.221** (0.059)	0.264** (0.056)	0.093 (0.082)
	3	0.138	0.128	13.156	0.220** (0.063)	0.251** (0.064)	0.110 (0.096)
	4	0.146	0.132	10.500	0.216** (0.057)	0.264** (0.044)	0.110 (0.120)
	5	0.141	0.127	10.101	0.209** (0.066)	0.261** (0.053)	0.096 (0.116)
REG	1	0.131	0.121	12.405	0.233** (0.056)	0.234** (0.048)	0.096 (0.105)
	2	0.142	0.128	10.167	0.231** (0.056)	0.262** (0.050)	0.080 (0.092)
	3	0.117	0.105	9.325	0.233** (0.062)	0.201** (0.086)	−0.050 (0.113)
	4	0.133	0.119	9.392	0.229** (0.058)	0.236** (0.046)	−0.064 (0.102)
	5	0.135	0.121	9.556	0.220** (0.060)	0.231** (0.055)	−0.088 (0.096)
MI	1	0.114	0.112	9.534	0.228** (0.024)	0.210** (0.019)	0.029 (0.044)
	2	0.148	0.146	10.902	0.228** (0.023)	0.270** (0.018)	0.089** (0.021)
	3	0.106	0.103	9.814	0.233** (0.024)	0.182** (0.022)	0.065** (0.026)
	4	0.126	0.124	9.814	0.227** (0.024)	0.230** (0.019)	0.062 (0.046)
	5	0.129	0.126	10.557	0.224** (0.025)	0.238** (0.019)	0.063 (0.055)

괄호 안의 수치는 계수 추정치에 대한 표준 붓스트랩(1000) 오류를 나타낸다.
** $p<0.05$, * $p<0.10$

<표 2> 결측치 20%의 경우 보완기법의 결과 비교

		R^2	Adj R^2	*Regression F−value*	Var2	Var3	Var4
모집단 분석결과		0.130	0.116	9.189	0.224** (0.062)	0.241** (0.047)	0.054 (0.114)
LW	1	0.108	0.085	4.722	0.218** (0.099)	0.201 (0.163)	0.136** (0.070)
	2	0.135	0.115	6.927	0.247** (0.088)	0.197 ** (0.115)	−0.178** (0.075)
	3	0.125	0.105	6.397	0.236** (0.084)	0.193** (0.115)	0.148** (0.064)
	4	0.167	0.143	7.018	0.304** (0.104)	0.203** (0.088)	0.087 (0.145)
	5	0.151	0.129	6.864	0.256** (0.086)	0.244** (0.073)	0.042 (0.169)
EM	1	0.148	0.136	12.221	0.221** (0.057)	0.256** (0.054)	0.154** (0.048)
	2	0.137	0.124	11.126	0.233** (0.058)	0.224** (0.047)	−0.110 (0.091)
	3	0.149	0.136	10.785	0.210** (0.058)	0.266** (0.049)	0.146** (0.049)
	4	0.142	0.130	11.656	0.200** (0.058)	0.268** (0.048)	0.105 (0.108)
	5	0.135	0.125	12.850	0.217** (0.061)	0.254** (0.049)	0.079 (0.094)
REG	1	0.122	0.108	8.537	0.230** (0.056)	0.204** (0.065)	0.134** (0.053)
	2	0.136	0.121	8.599	0.245** (0.060)	0.224** (0.050)	−0.110 (0.088)
	3	0.123	0.108	8.575	0.238** (0.060)	0.218** (0.055)	0.072 (0.078)
	4	0.126	0.111	8.811	0.225** (0.056)	0.218** (0.048)	−0.069 (0.071)
	5	0.120	0.110	11.252	0.238** (0.059)	0.211** (0.082)	−0.043 (0.092)
MI	1	0.109	0.107	10.338	0.233** (0.024)	0.193** (0.021)	0.039 (0.026)
	2	0.100	0.098	10.821	0.241** (0.024)	0.166** (0.023)	−0.040* (0.025)
	3	0.106	0.104	9.797	0.241** (0.024)	0.186** (0.021)	0.017 (0.060)
	4	0.120	0.118	10.147	0.221** (0.023)	0.215** (0.020)	0.052 (0.044)
	5	0.122	0.120	10.992	0.232** (0.025)	0.214** (0.018)	0.036 (0.060)

괄호 안의 수치는 계수 추정치에 대한 표준 붓스트랩(1000) 오류를 나타낸다.
** $p<0.05$, * $p<0.10$

<표 3> 결측치 30%의 경우 보완기법의 결과 비교

		R^2	Adj R^2	*Regression F−value*	Var2	Var3	Var4
모집단 분석결과		0.130	0.116	9.189	0.224** (0.062)	0.241** (0.047)	0.054 (0.114)
LW	1	0.166	0.142	6.973	0.274** (0.087)	0.205** (0.123)	−0.063 (0.168)
	2	0.170	0.145	6.877	0.283** (0.127)	0.207 (0.196)	−0.181 (0.169)
	3	0.148	0.123	6.085	0.262** (0.117)	0.195 (0.171)	0.097 (0.105)
	4	0.149	0.123	5.886	0.294** (0.094)	0.164 (0.219)	−0.142 (0.213)
	5	0.141	0.116	5.651	0.297** (0.076)	0.147 (0.169)	0.152* (0.097)
EM	1	0.107	0.097	9.886	0.214** (0.062)	0.191** (0.076)	0.028 (0.084)
	2	0.115	0.104	10.626	0.214** (0.068)	0.201** (0.067)	−0.047 (0.088)
	3	0.136	0.124	11.077	0.183** (0.069)	0.253** (0.049)	0.110** (0.047)
	4	0.110	0.098	8.716	0.208** (0.061)	0.195** (0.101)	0.051 (0.089)
	5	0.131	0.120	12.337	0.203** (0.059)	0.229** (0.049)	0.119 (0.074)
REG	1	0.095	0.079	5.726	0.254** (0.062)	0.119* (0.081)	−0.063 (0.071)
	2	0.096	0.085	8.729	0.242** (0.061)	0.140 (0.098)	−0.037 (0.087)
	3	0.110	0.096	7.600	0.236** (0.066)	0.176** (0.062)	0.085* (0.052)
	4	0.096	0.085	8.719	0.246** (0.054)	0.133 (0.129)	−0.056 (0.064)
	5	0.115	0.102	9.095	0.237** (0.059)	0.184** (0.054)	0.098* (0.062)
MI	1	0.114	0.112	9.287	0.253** (0.025)	0.186** (0.021)	−0.060** (0.020)
	2	0.113	0.110	9.599	0.254** (0.028)	0.154** (0.021)	−0.106** (0.018)
	3	0.100	0.099	11.210	0.247** (0.026)	0.153** (0.023)	0.035 (0.048)
	4	0.112	0.109	10.419	0.233** (0.025)	0.172** (0.020)	−0.083** (0.020)
	5	0.097	0.095	8.962	0.255** (0.027)	0.141** (0.023)	0.004 (0.077)

괄호 안의 수치는 계수 추정치에 대한 표준 붓스트랩(1000) 오류를 나타낸다.
** $p<0.05$, * $p<0.10$

<표 4> 결측치 40%의 경우 보완기법의 결과 비교

		R^2	Adj R^2	*Regression F−value*	Var2	Var3	Var4
모집단 분석결과		0.130	0.116	9.189	0.224** (0.062)	0.241** (0.047)	0.054 (0.114)
LW	1	0.269	0.244	10.795	0.404** (0.125)	0.332** (0.132)	−0.050 (0.159)
	2	0.147	0.129	7.861	0.278 (0.231)	0.242 (0.181)	0.095 (0.109)
	3	0.149	0.125	6.288	0.249** (0.113)	0.241 (0.197)	0.178 (0.117)
	4	0.313	0.280	9.716	0.428** (0.119)	0.267** (0.132)	−0.076 (0.102)
	5	0.311	0.283	11.167	0.399** (0.237)	0.275** (0.141)	−0.170 (0.203)
EM	1	0.167	0.155	14.049	0.166* (0.062)	0.313** (0.038)	0.081 (0.068)
	2	0.181	0.169	15.486	0.184** (0.052)	0.340** (0.045)	0.149** (0.040)
	3	0.141	0.127	10.105	0.168** (0.061)	0.291** (0.040)	0.122** (0.038)
	4	0.164	0.149	10.678	0.182** (0.062)	0.308** (0.044)	−0.054 (0.052)
	5	0.147	0.133	10.594	0.179** (0.065)	0.295** (0.049)	0.078 (0.057)
REG	1	0.122	0.112	11.445	0.223** (0.059)	0.223** (0.046)	−0.016 (0.086)
	2	0.136	0.124	11.054	0.226** (0.057)	0.252** (0.045)	−0.141 (0.079)
	3	0.092	0.079	7.120	0.234** (0.055)	0.131* (0.082)	0.051 (0.077)
	4	0.128	0.116	10.328	0.243** (0.055)	0.230** (0.042)	−0.034 (0.070)
	5	0.101	0.088	7.891	0.246** (0.057)	0.169** (0.049)	0.042 (0.072)
MI	1	0.127	0.125	10.531	0.246** (0.028)	0.209** (0.021)	−0.059** (0.028)
	2	0.097	0.094	9.878	0.277** (0.021)	0.127** (0.022)	0.081** (0.021)
	3	0.101	0.099	10.433	0.222** (0.028)	0.065 (0.068)	−0.156** (0.026)
	4	0.148	0.145	9.796	0.240** (0.024)	0.264** (0.019)	−0.085** (0.023)
	5	0.117	0.114	9.196	0.245** (0.024)	0.195** (0.023)	0.081** (0.023)

괄호 안의 수치는 계수 추정치에 대한 표준 붓스트랩(1000) 오류를 나타낸다.
** $p<0.05$, * $p<0.10$

<표 5> 결측치 비율에 따른 보완기법의 계수 추정치 위치

		LW			EM			REG			MI		
결측치 비율	모수 ±%	Var2	Var3	Var4	Var2	Var3	Var4	Var2	Var3	Var4	Var2	Var3	Var4
10%	±10	4	2	.	5	5	.	5	4	.	5	2	2
	±20	1	2		.	.		.	1		.	2	
	±30	.	.		.	.		.	.		.	1	
	기타	.	1		.	.		.	.		.	.	
20%	±10	2	1.	3	4	3	2	5	3	1	5	.	1
	±20	2	3		1	2		.	2		.	3	
	±30	1	.		.	.		.	.		.	1	
	기타	.	1		.	.		.	.		.	1	
30%	±10	.	.	1	2	2	1	4	.	2	1		3
	±20	1	1		2	2		1	.		4		
	±30	2	.		1	1		.	2		.	2	
	기타	2	4		.	.		.	3		.	3	
40%	±10	.	.	.	.	.	2	5	3	.	4	1	5
	±20	1	2		3	.		.	.		.	2	
	±30	1	.		2	4		.	1		1	.	
	기타	3	3		.	1		.	1			2	

각 변수 아래의 수치는 각 변수 추정치들의 Mean, 괄호 안은 표준편차이다.
Var4의 경우는 p=.10 수준에서의 통계적 유의성을 지니는 계수 추정치의 개수를 나타낸다.

<표 1>은 변수 Var3과 변수 Var4의 결측치가 각각 10%인 경우에 있어서 보완기법들을 활용한 회귀분석 결과를 나타낸다. 완전사례분석기법을 사용한 5번의 회귀분석결과가 LW 1부터 5에, EM기법에 의거한 결과는 EM 1부터 5, 회귀분석기법을 통한 분석결과는 REG 1부터 5, 다중대체기법에 의한 결과는 MI 1부터 5에 보고되어 있다. 먼저, <표 1>에서는 생략되고 있으나, 완전사례분석기법에 의한 5번의 회귀분석결과는 이미 앞서 지적한 바와 같이 분석에 사용되는 관측수를 크게 감소시키는 단점을 노출한다. 즉, 완전사례분석기법을 활용한 경우, 두 변수 Var3과 Var4에 있어서 각각 10%의 결측치가 존재하는 경우, 분석대상에서 제외된 관측 수는 전체 관측 수의 10% 정도가 아니라 실제 20% 정도인 100개 정도인 것으로 판명된다.[12] 이를 반영하여 <표 1>에서 나타나듯, 회귀분석의 F값이 LW 2의 경우를 예외로 하고는 모집단을 대상으로 하는 분석결과에 비하여 전반적으로 저하된 것으로 나타난다. 구체

12) 완전제거법에 의한 5번의 회귀분석에서 첫 번째 자료를 대상으로 하는 LW 1의 경우는 분석대상의 수가 406, 두 번째 자료를 대상으로 하는 LW 2의 경우는 403개, 마찬가지로 LW 3은 404개, LW 4와 LW 5는 모두 405개 씩의 관측 수를 나타낸다.

적인 계수추정치에 있어서는 <표 5>에서 나타나듯 Var2의 경우, 모집단의 계수 0.224의 ±10% 이내(0.2016~0.2464)에 해당하는 계수추정치가 4개이며, 0.224의 ±20% 이내(0.1792~0.2688)의 추정치는 LW 2의 1개 경우이다. Var3에 있어서는, 모집단 계수 0.241의 ±10% 이내(0.2169~0.2651)에 해당하는 계수추정치가 2개, ±20% 이내(0.1928~0.2892)의 추정치가 2개이며, LW3의 경우는 ±30%(0.1568~0.2912) 밖에 위치하는 것으로 나타난다. Var4에 관해서는 통계적 유의성을 지니는 추정치가 존재하지 않는 것으로 나타나 모집단을 대상으로 하는 분석과 크게 차이가 없는 것으로 드러난다.

이에 비하여 EM기법을 활용한 분석결과들은 회귀분석 F값들이 비교적 안정적인 것을 발견할 수 있으며, 계수추정치에 있어서도 Var2와 Var3 모두에 있어서 일관성을 발견할 수 있겠다. 즉, EM기법의 활용에 있어서는 Var2, Var3의 모든 추정치들이 모집단 계수의 ±10% 이내의 값을 기록하고 있기에 추정치의 불편성과 일관성이 향상된 것으로 판명될 수 있겠다. 또한 Var4에 있어서도 모집단을 대상으로 하는 분석과 유사한 것으로 나타난다.

회귀분석기법을 활용한 분석결과들에 있어서 Var2의 계수 추정치들의 평균과 모집단의 계수와의 편이는 EM기법의 활용에 따른 추정치들의 평균과 모집단 계수와의 편이와 거의 유사하며, 그 표준편차 역시 거의 차이를 나타내고 있지 않음을 발견할 수 있다. 즉 EM기법의 활용결과와 마찬가지로 회귀분석기법의 활용결과에 있어서도 Var2의 모든 계수 추정치들은 모집단 계수의 ±10% 이내에 해당하는 안정된 수치를 기록하고 있다. Var3의 추정에 있어서는 모집단 계수의 ±20% 이내에 해당하는 REG 3의 경우를 제외한다면 나머지 추정치들 역시 Var2의 경우와 마찬가지로 ±10% 이내에 포함되는 일관성을 보유한다. Var4에 관해서도 통계적 유의성을 지니는 추정치가 존재하지 않기에 모집단 분석과 유사한 것으로 드러난다.

다중대체기법의 활용 결과에 있어서는 회귀분석의 F값이 일관성을 지니는 것을 관찰할 수 있으며, 계수 추정치에 있어서 Var2의 경우, 5개의 추정치가 그 최소값 0.224에서 최대값 0.233에 이를 정도로 앞서의 모든 보완기법의 활용결과를 뛰어 넘는 극히 안정된 분포를 나타내고 있다. 전체적으로 이러한 추정치들은 모집단 계수(0.224)로부터의 오차가 0~5% 이내에 해당하는 수치이다. 더욱이 Var 2의 평균 추정치는 모집단의 계수와 정확히 일치하는 불편성을 보유하고 있다. 따라서 Var2의 경우 MI기법의 활용은 예상한 것 이상으로의 정확하고 일관성을 지니는 계수 추정치를 가져오는 것으로 판단된다. 이에 비하여 MCAR의 결측치 발생기제를 지니는 Var3의 경우에 있어서는 그 추정의 불편성/일관성이 Var2에 비하여 저하되는 것으로 나타나는 바, 5개의 결과 중 모집단 계수의 ±10% 이내에 해당하는 수치가 2개, ±20% 이

내(0.2169~0.2651)에 해당하는 추정치가 2개, ±30% 이내가 1개로 나타난다. 단, MAR의 결측치 발생기제를 보유하도록 설정된 Var4의 경우에 있어서, 모집단을 대상으로 산출된 계수가 그 통계적 유의성을 결여하고 있음에 반하여 다중대체기법의 활용 결과 5개 중 2개의 추정치가 비록 그 효과는 미미하더라도 95% 이상의 신뢰수준에서 양(+)의 효과를 지니는 것으로 해석됨을 지적할 필요가 있겠다. 즉, Var4의 영향력이 과대평가되는 위험성이 있음을 유념하여야 할 것이다.

<표 2>는 변수 Var3과 Var4 각각에 있어서 결측치 20%가 존재하는 경우, 보완기법을 적용시킨 분석 결과이다. 먼저, 완전제거법을 적용한 분석결과들을 살펴본다면 분석에 활용된 관측수가 더욱 감소함[13]을 반영하듯 회귀분석 F값의 저하가 관측된다. Var2의 추정치에 있어서도 그 분포가 <표 1> 결측치 10% 경우에 비하여 분산되어 있음을 발견할 수 있는 바, <표 5>에서 관찰되듯, 모수 ±10% 이내에 2개, 모수 ±20% 이내에 2개, 모수 ±30% 이내에 1개가 위치하고 있다. Var3의 모수 추정치에 있어서 LW 1의 경우에는 그 추정치의 통계적 유의성이 확보되지 않고 있으며, 나머지 4개 추정치 중 1개만 모수 ±10% 이내에 그리고 나머지 3개는 모수 ±20% 이내에 위치하는 것으로 나타난다. Var4에 관해서는 5개의 추정치 중 3개가 통계적 유의성을 지니는 것으로 드러나 모집단의 분석결과와 차이를 보이고 있다.

EM기법을 활용한 분석결과들은 완전사례분석기법의 결과들보다는 비교적 회귀분석의 F값이 일관성 있게 계산되고 있음을 발견할 수 있겠다. Var2의 계수 추정에 있어서도 5개의 추정치 중 4개가 모수 ±10% 이내에 나머지 하나가 모수 ±20%에 해당하는 것으로 나타난다. Var3의 경우에는 3개의 추정치가 모수 ±10% 이내에 해당하며 2개의 추정치가 모수 ±20% 이내에 위치하고 있음이 발견된다. Var4의 계수 추정치에 관해서는 모집단의 분석결과와는 달리 5개의 추정치 중 2개가 통계적 유의성을 지니는 것으로 나타나고 있다.

회귀분석기법을 활용한 결과들에 있어서는 Var2의 추정치 5개 모두가 모수 ±10% 이내에 위치하는 수치로 나타나는 추정의 일관성을 보여주고 있다. Var3의 추정에 있어서도 EM기법의 활용과 유사하게 5개 중 3개가 모수 ±10% 이내에 해당하며 나머지 2개의 추정치가 모수 ±20% 이내에 해당함을 발견할 수 있다. Var4에 관해서는 통계적 유의성을 지니는 추정치가 단 1개로 감소되었음을 알 수 있다.

다중대체기법에 의거한 추정치 분석결과는 Var2의 경우 5개 추정치 모두가 모수 ±10% 이내에 해당하는 일관성을 나타나지만, Var3의 경우 3개가 모수 ±20% 이내에 해당하며 1개의 추정치가 모수 ±30% 이내에, 그리고 나머지 1개는 모수 ±30% 밖에 위치하고 있음이 드러난다. Var4에 관해서도 1개의 추정치가 통계적 유의성을

13) LW 1의 경우 분석대상이 된 관측수는 322, LW 2는 320개, LW 3은 322개, LW 4는 325개, 그리고 LW 5는 318개로서 평균 약 321개의 관측을 회귀분석의 대상으로 삼는다.

지니고 있는 것으로 나타난다.

Var3과 Var4 각각이 결측치 30%를 지니는 경우의 보완기법들의 활용결과는 <표 3>에 기록되어 있다. 결측치 10% 및 20%의 경우를 기반으로 예상할 수 있듯이, 결측치의 비율이 증가될수록 완전사례분석기법의 활용결과는 계수 추정치에 있어서 일관성/불편성의 문제점을 더욱 심화시키는 것을 관찰할 수 있다. 즉, <표 5>가 보여주듯, Var2의 경우는 LW 추정치 5개 중 모수 ±10% 이내에 해당하는 것은 하나도 존재하지 않으며 모수 ±20% 이내에 위치하는 것이 1개, 모수 ±30% 이내에 해당하는 것이 2개, 나머지는 그 밖에 위치함이 드러나고 있다. Var3의 경우는 단 1개의 추정치만이 모수 ±20% 이내에 위치하고 있으며 나머지 4개는 통계적 유의성을 확보하고 있지 않은 것으로 나타난다. Var4에 관해서도 1개 추정치가 p=0.10 수준의 통계적 유의성을 보유하고 있는 것으로 나타난다.

EM기법을 활용한 Var2의 분석결과 역시 <표 5>에서 잘 드러나고 있는 바, 추정치 2개가 모수 ±10% 이내에, 추정치 2개가 모수 ±20% 이내에, 나머지 추정치 1개가 모수 ±30% 이내에 위치함으로써 완전사례분석기법의 활용결과에 비하여 불편성이나 일관성이 향상되었음을 알 수 있다. Var3에 관해서도 유사한 위치 산정이 이루어져서 추정치 2개는 모수 ±10% 이내에, 다른 2개는 모수 ±20% 이내에, 나머지 1개는 모수 ±30% 이내에 해당됨을 알 수 있다. Var4에 있어서는 추정치 1개가 통계적 유의성을 지니고 있는 것으로 계산된다.

회귀분석기법에 의거한 분석결과는 Var2의 추정치 중 4개가 모수 ±10% 이내에 있고 1개가 ±20% 안에 해당하는 것으로 나타난다. Var3에 관해서는 5개 추정치 중 2개가 모수 ±30% 이내에 위치하며 나머지는 통계적 유의성을 보유하지 못하거나 모수 ±30% 바깥에 위치한다. Var4의 추정에 있어서는 2개의 추정치가 p=.10 수준의 유의성을 보유한다.

다중대체방법을 활용한 Var2의 계수 추정치에 있어서는 추정치 1개가 모수 ±10% 이내에 위치하며 4개는 모수 ±20% 범위에 해당하는 수치를 나타낸다. Var3의 추정에 있어서는 추정치 2개가 모수 ±30%v 이내에 해당하며 나머지 3개는 통계적 유의성을 확보하고는 있으나 모두가 모수 ±30% 범위를 조금씩 이탈하고 있다. Var4의 추정에 있어서는 3개의 추정치가 통계적 유의성을 지니는 것으로 나타나지만 실제 그 계수의 수치는 미미하다는 점을 지적할 필요가 있겠다.

마지막으로 <표 4>와 <표 5>를 통해서 Var3과 Var4의 결측치 비율이 각각 40%로 증가한 경우의 분석결과를 살펴볼 수 있다. 예상과 같이 완전사례분석기법의 활용결과는 회귀분석 F값, Var2와 Var3의 계수 추정치 모두에 있어서 상당한 문제점을 노출시키는 것으로 파악된다. 회귀분석의 F값은 최소값과 최대값의 차이가 거의

배에 이를 정도로 산정된 수치들이 안정성을 결여하고 있음이 드러난다. Var2의 계수 추정치에 있어서 1개가 모수 ±20% 이내에 위피하며 나머지들은 모두 모수±30% 밖에 위치하거나 통계적 유의성을 보유하지 못하는 것으로 판명된다. Var3의 경우에 있어서도 추정치 2개가 모수 ±20% 이내에 있으나 나머지 3개가 모수 ±30% 밖에 위치하는 것으로 나타난다.

EM기법의 활용결과는 Var2의 경우 추정치 5개 중 3개가 모수 ±20% 범위내에 위치하며 나머지 2개는 모수 ±30% 이내에 해당함을 알 수 있다. Var3의 계수 추정에 관해서도 4개가 모수 ±30% 이내에 위치하며 1개만이 그 범위 밖에 있음이 드러난다. Var4에 있어서는 2개 추정치가 통계적 유의성을 지니는 것으로 드러난다.

회귀분석기법의 결과를 살펴보면 Var2의 경우 다섯 개 추정치 모두가 모수 ±10% 이내에 들어가는 안정성과 불편성을 확보하는 것으로 드러난다. Var3의 계수 추정치에 있어서도 3개가 모수 ±10% 범위에 해당하며 1개가 모수 ±30% 이내에 그리고 나머지 1개가 그 밖에 위치하는 것으로 나타난다. 전체적으로 볼 때, 회귀분석기법의 활용은 결측치 30%의 경우보다도 결측치 40%의 경우에 있어서 보다 안정되고 정확한 추정치들을 산출하는 것으로 해석될 수 있겠다.

다중대체기법의 활용결과는 Var2의 경우, 추정치 5개 중 4개가 모수 ±10% 범위 이내에 위치한다. 또한 Var3의 경우는 5개 중 1개가 모수 ±10% 이내에, 2개가 ±20% 이내에, 나머지 2개가 ±30% 외곽에 해당하는 것으로 나타난다. Var4의 계수 추정에 있어서는 5개 모두가 통계학적 유의성을 지니는 것으로 판명되었다. 앞서의 회귀분석결과와 마찬가지로 다중대체기법을 결측치 40%의 경우에 활용한 분석결과는 오히려 결측치 30%의 경우에 활용한 것에 비하여 추정치의 불편성이나 일관성의 측면에서 개선된 것으로서 판단될 수 있겠다.

V. 요약과 함의

행정연구에 있어서 계량분석기법의 활용이 점차 증가되고 있음이 사실이다. 하지만 계량분석기법이 적용되는 자료가 어떠한 성격을 지니는지, 즉, 자료에 수반되는 결측치의 본질이 무엇이며, 그 정도가 어떠하며, 또한 이를 어떻게 처리할 것인가에 관해서는 아직까지 몇몇 연구를 제외하고는 별다른 관심을 표방해 오지 않았음 역시 사실이다. 어쩌면, 본 논문의 주석 4)에서 지적했듯이 거의 대부분의 행정조사연구에 있어서는 결측치의 존재가 관심거리조차 되지 못했던 것이 사실이라고 할 수 있겠다. 하지만, 결측치의 존재는 자료의 불완전성을 가져오며 ,이러한 자료로부터의 추정치는

불편성과 일관성에 문제가 있으며, 통계적 검증의 신뢰성이 과대평가될 위험성을 지니며, 분석결과의 신뢰성과 타당성을 저하시키는 문제를 초래할 수 있음이 분명하다.

결국, 결측치를 보완하기 위한 기법의 개발과 적용이 필요하다. 덧붙여, 행정조사 연구의 대상이 되는 자료 중에는 범주형 자료가 다수를 차지하고 있다. 바로 이러한 맥락에서 본 논문은 범주형 자료에 계량분석기법을 적용하는 경우, 결측치의 문제가 어떻게 전개될 것인가에 관한 분석의 초점을 설정한 것이다. 분석의 결과, 우선 결측치의 비율이 증가함에 따라서 완전사례분석기법이 지니는 문제점들이 더욱 심각하게 노출된다는 사실이 부각되었다. 실제 결측치 10%의 경우에 있어서 완전사례분석기법을 적용한 분석결과들은 회귀분석 F값의 저하를 초래하고 변수의 계수 추정치에 있어서 편의(bias)와 일관성(consistency)이 문제될 수 있음이 드러났다. 이러한 문제점들은 결측치 비율이 증가함에 따라서 더욱 더 증대되어 결측치 40%의 경우에는 과반수 이상의 추정치들이 모수 계수의 ±30% 외곽에 위치하거나 통계적 유의성을 상실하기에 이르렀다. 이러한 자료에 과학적 방법론으로서의 계량기법을 적용한다면 누구도 그 결과를 과학적/객관적인 것으로 신뢰할 수 있다고 주장하기 곤란할 것이다.

그러나, 회귀분석기법, EM기법, 다중대체기법 모두가 완전사례분석이 노출하는 문제점을 감소시켜주는 결측치의 적극적인 보완기법이 될 수 있음이 판명되었다. 하지만, 어느 기법이 보다 더 우수한 기법인지, 더 효과적인 기법인지는 결측치의 비율에 따라서, 또한 결측의 발생기제에 따라서 그 판단이 달라질 수 있겠다. 보다 구체적으로, 결측치가 가장 적었던 10%의 경우에 있어서는 EM기법, 회귀분석기법, 다중대체기법 그 모두가 계수 추정치의 편의성/일관성을 기준으로 설정할 때 효과적이며 적극적인 보완기법으로 평가될 수 있겠다. 결측치 20%의 경우에는 결측치의 보완책으로서 EM기법이나 회귀분석기법의 활용이 추천될 수 있겠다. 이 때 다중대체기법은 결측의 발생기제가 MCAR인 변수 Var3의 계수 추정에서 EM기법이나 회귀분석기법에 비하여 상대적으로 비효율적인 보완기법이 되는 것으로 평가될 수 있겠다. 결측치 비율이 30%로 증대된 경우에 있어서도 EM기법과 회귀분석기법의 사용이 효과적인 것으로 드러난다. 그리고, 결측의 비율이 40%인 경우에 있어서는 회귀분석기법의 사용이 다른 보완기법의 분석결과에 비하여 보다 더 효과적인 것으로 판단될 수 있겠다. 또한 회귀분석기법과 다중대체기법은 결측치의 비율이 40%일 때가 결측치 비율 30%인 경우에 비하여 오히려 더 계수 추정의 편의와 일관성의 측면에서 우수한 결과를 가져오는 것으로 나타났다.

결론적으로, 자료에 결측치가 수반된다면 최선의 보완기법이 모색되어야만 한다. 따라서 다음의 사항들이 강조될 수 있겠다. 우선, 보완기법의 효과와 타당성은 활용되는 기법의 과학성과 접근성에 좌우된다는 사실이다. 따라서 연구자들은 과학적 방법론의

수립/적용과 이를 처리하기 위한 통계 소프트웨어의 활용 능력을 증진시켜야 할 것이다. 그리고 과학적 방법론의 적용 혹은 결측치 보완기법의 실제 활용과 동전의 양면을 형성하는 사항으로서, 기본적으로 자료의 특성이 어떠한 것인가를 관찰하고 분석해야 한다는 사실이 강조되어야만 할 것이다.[14] 즉, 연구자들은 자료의 특성에 관한 보다 철저한 분석을 통하여 결측치의 본질에 관한 이론적 지식을 습득하고 실제 분석 자료에서 결측치에 관한 정보를 구하는 것이 무엇보다도 선행되어야 할 과제인 것이다.

<참고문헌>

강민아・김경아. (2006). 행정학 및 정책학 조사연구에서 결측치 발생과 처리 방법에 대한 고찰.「한국행정학보」, 40(2): 31-52.

김태일・서주현. (1998). 행정학분야에서 설문조사를 이용한 연구의 방법론의 문제점 분석.「한국행정학보」, 32(3): 199-203.

박중훈・서성아. (2007).「공직부패의 실태에 관한 설문조사 2005」, 한국행정연구원. 자료서비스기관: 한국사회과학자료원. 자료번호: A1-2005-0040.

Allison, P. D. (2002). *Missing Data*. Sage. Thousand Oaks: CA.

Berk, Richard A. (1983). An Introduction to Sample Selection Bias in Sociological Data. *American Sociological Review,* 48: 386-98.

Dempster, A. P., Laird, N. M., and D. B. Rubin. (1977). Maximum Likelihood Estimation form Incomplete Data vis the EM Algorithm (with Discussion). *Journal of Royal Statistical Association*, B39: 1-38.

Diggle, P., and M. G. Kenward. (1994). Informative dropout in longitudinal data analysis. *Applied Stat,* 43: 49-94.

Gibson, N., and S. Olejnik. (2003). Treatment of Missing Data at the Second Level of Hierarchical Linear Models. *Educational and Psychological Measurement.* 63(2): 204-238.

Ibrahim, J. G., and S. R. Lipsitz. (1996). Parameter Estimation from Incomplete Data in Binomial Regression When the Missing Data Mechanism Is Nonignorable. *Biometics*,. 52: 1071-1078.

14) 이러한 맥락에서 연구자들은 자신들이 자료에서의 결측치의 비율과 그 발생 기제에 관하여 어떠한 가정을 수립하였는가를 명백히 밝혀야 한다는 주장이 제기된다. 이렇게 된다면 설혹 연구자 자신이 그러한 가정의 적절성에 관한 경험적 증거를 제공하지 못한다 할지라도, 가정의 적절성에 관한 최종 판단이 다수의 객관적 연구자 집단에 맡겨져서 문제점의 개선을 위한 공개적인 논의의 장이 개설될 수 있다는 점이 부각될 수 있겠다(Diggle and Kenward, 1994: Ibrahim and Lipsitz, 1996; Pigott, 2001).

King, G., Honaker, J. Joseph, A., and Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation, *American Political Science Association,* 95(1): 49-69.

Laird, N. M., and D. B. Rubin. (1977). Maximum Likelihood Estimation form Incomplete Data vis the EM Algorithm (with Discussion). *Journal of Royal Statistical Association*, B39: 1-38.

Little, R. A. (1992). Regression with missing X's: A Review. *Journal of the American Statistical Association,* 87: 1227-1237.

Little, R. A., and D. B. Rubin. (1987). *Statistical Analysis with Missing Data*. New York, NY: John Wiley and Sons, Inc.5

Meng, X. (1994). Multiple-imputation Inferences with Uncongenial Sources of Input. *Statistical Sciences,* 9: 538-573.

Pigott, T. D. (2001). A Review of Methods for Missing Data. *Educational Research and Evaluation,* 7(4): 353-383.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John and Wiley and Sons, Inc..

_____. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association,* 91: 473-489.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Tanner, M. A. (1993). *Tools for Statistical Inference*. New York: Springer-Verlag.

Tanner, M. A. and W. H. Wong. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion). *Journal of the American Statistical Association,* 82: 528-550.

Tobin, James. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica,* 26: 24-36.

접수일(2011년 06월 05일)
수정일자(2011년 07월 11일)
게재확정일(2011년 07월 30일)