형사정책연구 Korean Criminological Review
제29권 제1호(통권 제113호, 2018・봄)

범죄학 연구와 결측 :
실제 자료를 활용한 결측 대치법 비교 분석

라광현*

국 | 문 | 요 | 약

실증 연구를 위한 조사 자료에서 결측값이 존재하는 것은 연구자가 흔히 맞이할 수 있는 상황이다. 사회과학 연구에서 자주 활용되는 자기보고의 경우, 응답자는 실수로 어떠한 문항을 답하지 않을 수도 있으며, 고의적으로 응답하기 불편한 문항을 제외하고 설문을 마무리 할 수 있다. 이러한 결측은 통계분석 결과에서 1종 오류 및 2종 오류가 나타나는 것과 관련이 있다. 이 연구는 범죄학 연구에서 결측 처리의 중요성 및 주요 결측 처리 방법을 논의하였으며, 결측 대치를 통한 해법을 중심으로 한국아동·청소년패널조사 중1패널 6년차 자료를 활용한 분석 결과를 비교하였다.

분석 결과, 연쇄방정식을 활용한 다중대치 및 확률적 회귀대치 자료들은 결측값이 존재하는 원조사 자료에 비하여 완전한 자료에 가까운 결과를 나타냈으며, 원조사 자료의 경우, 회귀분석 결과에서 1종 및 2종 오류의 징후가 모두 나타났다. 이상의 결과를 종합하여 볼 때, 범죄학 연구에서 결측값의 처리는 매우 중요하다고 할 수 있다. 범죄학 연구에서 발생하는 결측은 범죄학 연구의 주요 관심인 범죄 및 일탈과 관련되었을 가능성이 크며, 결측을 적절히 처리하지 않을 경우 범죄학 연구의 중요한 대상(예를 들어, 범죄 및 일탈 성향이 높은 응답자)을 제외시키며 통계분석 결과를 왜곡시킬 우려가 있다. 한편 실무적으로, 조사자료 및 공식통계 등 통계자료를 활용한 실증연구는 증거에 기반을 둔 과학적인 정책 수립을 위한 매우 중요한 자료로 활용될 수 있다는 점에서 범죄 관련 통계 자료를 분석할 때, 결측을 적절하게 처리하는 것은 학술적 가치뿐만 아니라 정책적인 의미를 가진다.

❖ 주제어 : 결측, 결측 대치법, 연쇄방정식을 활용한 다중대치법, 범죄통계, 범죄분석

* 사우스캐롤라이나 주립대학교. 범죄학 및 형사사법학과. 박사과정 수료

Ⅰ. 서론

실증 연구를 위한 조사 자료에서 결측이 존재하는 것은 연구자가 일반적으로 맞이할 수 있는 상황이다. 사회과학 연구에서 자주 활용되는 자기보고의 경우, 응답자는 실수로 어떠한 문항을 답하지 않을 수도 있으며, 고의적으로 응답하기 불편한 문항을 제외하고 설문을 마무리 할 수 있다. 혹은, 패널연구의 경우 특정한 성향을 가진 응답자들이 일반 응답자들에 비해 높은 확률로 탈락(attrition or dropout) 할 수도 있다. 실수로 인한 무응답과 고의적인 무응답은 자료의 외견상 차이를 일으키지는 않는다. 예를 들어, 자기보고의 경우, 실수로 인한 무응답과 고의적인 무응답은 모두 공란으로 남게 될 것이고, 일반적으로 연구자는 이 두 경우를 구분하기 어려우며, 모두 같은 방법 혹은 표식을 사용하여 코딩하게 된다.

하지만, 표본 자료를 활용하여 모집단의 특성을 유추하는 추리통계의 방법론과 관련하여, 두 유형의 무응답은 완전히 다른 의미를 지닌다. 일반적으로, 응답자의 실수로 인한 무응답은 무작위(random)로 발생[1]한다고 볼 수 있으며, 무작위로 발생한 결측은 통계적인 검정력을 약화시킬 수 있다. 한편, 후자의 고의적인 무응답 및 특정 집단의 높은 중도탈락률은 전자와는 달리 체계적인 결측을 발생시키며, 통계적 검정력의 약화와 더불어 통계적인 편향(biased estimation)을 일으킬 우려가 있다. 다시 말하여, 전자의 결측 문제는 통계적 검정력을 약화시켜 1종 오류[2]의 원인이 될 수 있으며, 후자의 결측 문제는 1종 오류와 더불어 편향된 추정으로 인한 2종 오류[3]의 원인이 될 수 있다. 즉, 두 유형의 무응답은 외견상 구분할 수 없지만, 전혀 다른 결과를 발생시키며, 일반적으로 후자의 결측이 더 큰 문제를 일으키게 된다.

후자의 결측 문제는 범죄 관련 연구에 중요한 함의를 가진다. 많은 연구들은 자기보고 설문을 범죄자료 수집을 위한 주요한 방법으로 삼고 있다(Thornberry & Krohn, 2000;

1) 물론, 설문지의 포맷이나 구성이 응답자의 부주의를 유발하여 특정 문항에 대한 무응답율에 영향을 미칠 수 있다.

2) 1종 오류는 잘못된 긍정(false positive)으로서, 귀무가설(null hypothesis)이 참임에도 불구하고, 이를 기각하는 경우이다. 즉, 거짓인 연구가설이 채택될 때 1종 오류가 발생한다.

3) 2종 오류는 잘못된 부정(false negative)으로서, 귀무가설이 거짓임에도 불구하고, 이를 채택하는 경우이다. 다시 말하여, 참인 연구가설이 채택되지 못하는 경우에 2종 오류가 발생한다.

p. 34). 범죄와 관련한 자기보고식 설문조사에서, 응답자들은 자신의 범죄 및 일탈 경험 혹은 범죄피해에 대한 응답을 주저할 수 있으며, 다년간 자료를 수집하는 패널 연구의 경우, 범죄성향이 높은 응답자의 경우, 다른 응답자에 비하여 패널에서 탈락할 위험성이 크다. 전술한 바와 같이, 이러한 결측들은 통계적 검정력을 약화시킬 뿐만 아니라 통계 분석의 결과를 왜곡할 가능성이 있다.

이러한 문제를 해결하기 위하여, 다양한 대치법 및 불편추정량(unbiased estimation)을 산출하는 통계적인 기법(김형민 외, 2016; 송주원, 2010; Van Buuren & Oudshoorn, 1999) 등 결측 문제에 대한 다양한 해법이 제시되고 있으나, 많은 연구들은 단순히 무응답을 제외하고 분석을 실시하고 있는 상황이며(예를 들어, listiwse deletion), 통계적인 오류 가능성에 그대로 노출되고 있다. 이에 이 연구는 범죄학 연구에서 결측 처리의 중요성 및 주요 결측 처리 방법을 논의하고, 결측 대치를 통한 해법을 중심으로 실제 자료를 활용한 결과를 비교하고자 한다.

Ⅱ. 범죄학 연구에서 결측 처리의 중요성

1. 결측 패턴의 유형

결측 패턴은 크게 세 가지 유형으로 분류할 수 있다. 첫째는 완전한 무작위 패턴(Missing Completely at Random; MCAR)이다. MCAR이란 결측이 완전히 독립적 혹은 임의적(random)으로 발생하였다는 것이며, 결측의 발생은 다른 변수에 의해 예측되지 않는다. 예를 들어, 응답자들이 단순 실수로 응답하지 않는 문항들은 MCAR 결측 패턴을 형성할 수 있다. 이러한 경우, 어떠한 결측 처리를 하지 않아도 통계적인 편향이 없이 모수(parameter)를 추정하게 된다. 다만, 이러한 경우는 분석에 활용되는 자료의 수가 감소하기 때문에 통계적 검정력이 약해져 1종 오류의 원인이 될 수 있다.

둘째는 무작위 패턴(Missing at Random; MAR)이다. MAR 결측 패턴을 가지고 있는 자료의 결측은 자료 내에 있는 다른 변수에 대하여 독립적이지 않으며, 다른

변수들을 통해 예측이 가능하다 (Schlomer 외, 2010; p. 2). 예를 들어, 어떤 설문에서 “불안 심리”와 “범죄 행위”를 측정한다고 가정 할 때, 심리적으로 불안한 사람들은 범죄에 대한 자기보고에 응답하지 않을 확률이 높을 수 있다. 이러한 경우 결측 여부는 “불안”이라는 변수에 의해 예측될 수 있다. 즉, MAR이란 통제된 상태에서의 무작위 패턴을 의미한다. 이러한 경우, 결측 대치 및 다른 기법을 통해 결측 문제를 해결하지 않을 경우, 1종 오류와 더불어 편향된 추정으로 인한 2종 오류를 발생시킬 수 있다. 물론, MAR 상태의 자료는 결측 대치 혹은 통계적인 기법을 활용하여 편향되지 않은 모수를 추정할 수 있다.

셋째는 非무작위 패턴(Missing not at Random 혹은 Not Missing at Random; NMAR)이다. NMAR이란 자료 내에 존재하는 결측은 무작위로 발생한 것이 아니며, 자료가 포함하고 있지 않은 변수들을 통해 예측될 수 있다는 것이다. 이는 다시 말하여 자료가 포함하고 있는 변수들로만은 결측 문제를 해결할 수 없다는 것이다. 이러한 경우, 결측은 1종 및 2종 오류를 일으킬 수 있지만, NMAR은 자료가 포함하지 않은 외재적인 요소에 의한 결측 패턴이기 때문에, 이를 실제적으로 혹은 실증적으로 진단하는 것은 매우 어렵다(Schlomer 외, 2010; p. 3).

2. 범죄통계 자료에서의 결측 문제

가. 범죄학 연구에서 결측 자료 처리의 필요성

결측의 패턴이 MCAR이며, 전체에서 5% 미만의 결측을 가지는 자료는 관측된 자료만으로 분석 할 수 있으나(Graham, 2009; Schafer, 1999), 이러한 조건을 충족하는 경우는 설문을 활용한 일반적인 조사자료에서 드물다고 볼 수 있다. 특히, 패널자료와 범죄 및 일탈과 관련한 자기보고식 설문 자료는 전술한 기준을 충족하기가 까다롭다.

통계자료(특히, 자기보고 및 패널 조사)에 존재하는 결측값은 범죄 및 일탈과 여러 가지 관련을 가지고 있다. 첫째, 사람들은 일반적으로 범죄 및 일탈 등 사회적으로 바람직하지 않다고 여겨지는 행동에 대한 응답을 꺼리는 경향이 있다(Huizinga & Elliott, 1986; Tourangeau, Rips, & Rasinski, 2000). 즉, 범죄성향이 매우 높은

집단의 경우 일반적인 통계조사 대상에서 원천적으로 누락될 가능성이 크기 때문에 연구 결과의 일반화에 문제가 발생하기도 하지만[4], 조사에 자발적으로 참여한 응답자들도 범죄 및 일탈에 대한 응답은 성실히 대답하지 않을 수 있다는 것이다. 이러한 경우, 범죄 및 일탈율은 실제보다 과소추정 될 수 있으며, 이를 활용한 추리통계 분석의 결과가 편향 될 수 있다.

둘째, 낮은 자기통제 등 특정한 개인적 특성 및 성향을 가진 사람들은 자신의 범죄 행위에 대한 응답을 기피할 수 있다(Enzmann, 2013). 일례로 Enzmann(2013)의 The Second International Self-Report Delinquency(ISRD-2)를 활용한 연구는 자기통제력이 낮은 청소년들은 약물사용이나 가해경험을 묻는 질문에 높은 무응답률을 나타내는 것을 확인하였다. 이론적으로 낮은 자기통제력은 범죄 및 일탈의 원인이 될 수 있다(Gottfredson & Hirschi, 1990). 따라서, 이러한 경우 전체적인 약물사용율 및 가해율이 실제에 비하여 과소 추정될 것이며, 이 자료를 그대로 회귀분석 등에 활용할 경우, 모수에서 자기통제력과 범죄 및 일탈 관 상관관계가 존재한다고 하더라도, 추리통계의 결과는 계수(coefficient)를 모수보다 작게 추정하거나 유의한 결과를 찾지 못하는 등 결과가 왜곡될 가능성이 매우 높다.

물론 결측 문제는 연구 분야를 막론하고 통계 자료에서 흔히 나타나는 문제이다. 다만, 일반적인 현상을 연구하는 분야(예를 들어, 교육 성취)와 비교하여, 정상에서 벗어난 일탈 현상을 연구하는 범죄학의 경우 관련 자료에서의 결측은 연구 결과에 매우 중대한 영향을 미칠 우려가 있다. 전술한 바와 같이, 일탈의 원인과 결측 발생의 원인이 중복될 수도 있으며(예를 들어, 낮은 자기통제), 패널 연구 등에서는 연구 관심사인 일탈 자체가 탈락(drop out or attrition)이나 결측 발생의 원인이 될 수도 있다. 일례로, 많은 경우에 연구자들은 “불성실한 응답(자)”를 제외한 자료를 분석하는데, 연구에 활용하지 않은 “불성실”한 응답자들에 대한 정보는 범죄와 일탈의 연구에서 중요한 기치를 지니고 있을 수 있다. 이러한 점에서 범죄학 및 형사사법학

4) 예를 들어, 학교를 중심으로 표집을 실시하는 청소년 조사 자료의 경우, 문제 행동으로 인하여 학교를 이탈한 청소년들을 배제하게 되며, 가구 조사의 경우, 노숙인 및 교도소 재소자 등이 연구 대상에서 제외되게 된다. 범죄, 일탈 및 문제행동에 관심을 갖는 범죄학 및 형사사법학에서 이렇게 배제된 집단은 매우 중요한 의미를 가진다. 다만, 이 연구는 수집 된 자료 안에 있는 결측 문제를 연구 범위로 하고 있기 때문에, 전술한 문제는 논외로 한다.

조사 자료의 결측 패턴 분석과 적절한 처리는 매우 중요하다.

나. 범죄통계 자료와 결측의 처리: 미국의 범죄통계 자료 사례

범죄 및 형사사법 연구와 관련된 범죄통계 자료는 크게 공식통계(예를 들어, 대검찰청의 범죄분석)와 자기보고 통계자료(예를 들어, 형사정책연구원의 전국범죄피해조사)가 있다. 이러한 범죄통계 자료에 존재하는 결측값 처리 방식은 범죄학자들에게 중요한 연구의 대상이 되고 있다. 미국의 경우, 대표적인 공식통계자료는 FBI의 Uniform Crime Reporting과 National Incident-Based Reporting System(NIBRS)이 있으며, 이들 자료에는 상당수의 결측이 포함되어 있다 (Thompson, Saltzman, & Bibel, 1999; Lynch & Jarvis, 2008).

UCR의 경우, 많은 법집행 기관들이 FBI에 범죄 통계를 제공하지 않거나 결측이 있는 자료를 제공하는 경우가 있다. 일례로, 2003년의 경우, 65.5%의 법집행 기관만이 FBI에 완전한 자료를 제공하였다 (Lynch & Jarvis, 2008). 이에 FBI는 결측값 및 극단치(outlier)를 범죄통계 추세 및 유사한 관할지역의 자료 등을 활용하여 대치하거나 보정하여 제공한다(Akiyama & Propheter, 2005; Lynch & Jarvis, 2008). 다만, UCR의 결측 문제는 미국 내 각 법집행 기관들의 FBI에 대한 범죄통계 보고가 자발적으로 이루어진다는 점에서 비롯된 것으로 우리의 실정과는 다소 거리가 있다.

UCR이 일정 기간 동안 발생한 범죄 및 체포 등의 요약값만을 제공하는 반면에, 사건 기반 보고시스템인 NIBRS는 개별 사건에 대한 구체적이며 다양한 정보를 제공한다는 점에서 연구자료로서의 가치가 높게 평가받고 있다(Addington, 2007). 다만, NIBRS은 UCR보다 다양하며 구체적인 범죄 관련 정보를 수집하기 때문에, 각 법집행 기관들이 자료를 입력하는 과정에서 많은 결측값을 남기게 된다. 예를 들어, 매사추세츠 주의 NIBRS 보고를 검토한 연구에서 특정 변수(범죄자의 알콜/마약 연루 여부)의 81%가 결측으로 나타나기도 하였다(Thompson, Saltzman, & Bibel, 1999). NIBRS의 경우, FBI는 UCR과는 달리 대치된 자료를 제공하지는 않으나, NIBRS에 존재하는 결측은 연구자들에게 심각하게 받아들여지고 있으며 (Thompson, Saltzman, & Bibel, 1999), 결측 문제를 해결하기 위한 통계적인 방법

들이 제시되고 있다(Haas 외, 2012; LaValle, Haas, & Nolan, 2014).

다음으로, 범죄피해조사 또한 중요한 범죄 연구의 자료로 사용되고 있다. 범죄피해조사와 공식통계 중 어떠한 자료가 연구에 더 적합한가에 대한 논란이 존재하기는 하지만(Berg & Lauritsen, 2016; Lynch & Addington, 2007), 공식통계와 비교하여 범죄피해조사는 어떠한 사회의 실제 범죄현상을 더 정확하게 반영하는 대안적인 자료로 여겨지고 있다(황지태, 2010; Bursik 외, 1993; Mosher Miethe, & Hart, 2010). 다만, 범죄피해조사의 경우, 응답자가 의도적으로 특정 문항에 대한 응답을 하지 않을 우려가 있다. 예컨대, 2014년 및 2015년도 미국 범죄피해 조사(National Crime Victimization Survey; NCVS)의 경우, 평균 32%의 가구가 가계소득을 응답하지 않았다. 이러한 높은 무응답률은 범죄피해 관련 분석에 영향을 미칠 우려가 있기 때문에, NCVS는 가구 소득의 경우 핫 덱 기법 등을 활용하여 대치된 결과를 제공한다(Berzofsky 외, 2014 ; Bureau of Justice Statistics, 2016).

요약하자면, 미국의 경우, 대표적인 범죄 통계 자료들인 UCR, NIBRS, NCVS 모두 자료의 생산 과정에서 결측이 발생한다는 사실이 잘 알려져 있으며 연구자들의 관심을 받고 있다. 따라서, 결측이 발생하는 이유, 결측의 처리방식, 결측과 관련한 자료의 해석 방식 등에 관한 다양한 연구들이 수행되고 있다. 또한 NCVS의 소득 자료 및 UCR의 경우, 자료의 생산기관에서 직접 대치된 자료를 제공하고 있다.

Ⅲ. 주요 결측 대치 방법

결측을 대치하는 방법은 평균대치법, 최근방대치법(nearest neighbor imputation), 콜드덱대치법, 회귀대치법, 핫덱대치법, 기대치 최대화 기법(Expectation maximization; EM) 등으로 다양하다. 이 연구에서는 확률적 회귀 대치법 및 연쇄방정식을 활용한 다중 대치법(Multiple Imputation by Chained Equations; MICE)을 활용하여 실증자료를 분석하였다.

확률적 회귀 대치법의 경우, 일반적으로 EM 및 다중 대치법을 제외한 다른 대치법들보다 우수한 대치 방법으로 제시되고 있다(Enders, 2010). 한편, EM과 다중 대

치법은 둘 다 우수한 결측 문제의 해법으로 제시되고 있으나, 다중 대치법이 EM에 비하여 덜 편향된(less biased) 결과를 산출하는 것으로 알려져 있다(Schlomer & Bauman, 2010). 이에 이하에서는 이 연구에서 활용된 확률적 회귀 대치법 및 다중 대치법의 최신 기법인 MICE 대해 간단히 설명하였다.

1. 회귀 대치법 및 확률적 회귀 대치법

회귀 대치법은 결측이 없는 자료들을 활용하여 회귀계수를 추정하며, 추정된 회귀식을 결측값을 대치하는 데 적용하는 방법이다. 예를 들어, 어떤 자료에 소득에 대한 무응답이 다수 존재할 경우, 먼저 결측을 제외한 자료를 활용하여, 소득을 추정하는 회귀 방정식을 적합 한다. 다음으로, 이 회귀 방정식을 활용하여 결측값을 대치하게 된다. 이 방법은 MCAR 및 MAR 상태의 자료 모두에서 편향되지 않은 평균을 산출하게 된다(Schlomer & Bauman, 2010). 하지만, 이 방법은 대치된 자료를 포함하는 변수의 변량을 감소시키며 변수 간 공변량에 영향을 미치는 문제점을 가지고 있다(Graham 외, 2003).

다시 말하여, 일반 회귀 대치법을 활용할 경우, 동일한 값의 예측 변인을 가지고 있는 결측값은 동일한 값으로 대치되게 된다. 예컨대, 어떠한 연구에서 소득의 결측값들을 X_1, X_2, X_3 이라는 독립변수들을 활용하여 회귀 대치를 실시하였다. 이러한 경우, X_1, X_2, X_3 변수값이 동일한 응답자들의 소득은 모두 같은 값으로 대치되게 되며, 변량 및 공변량에 영향을 미치게 된다. 이러한 이유로 회귀 대치법은 장려되지 않고 있다.

다만, 이러한 문제를 해결하기 위하여, 확률적 회귀 대치법(stochastic regression imputation)을 활용 하여 편향되지 않은 변량을 추정할 수 있다. 확률적 회귀 대치법은 회귀 대치법을 활용하여 추정된 값에 무작위 값을 더하여 최종 대치값을 결정하는 방법이다. 이 때, 무작위 값의 기댓값은 0이며, 변량은 관측치로부터 계산된 변량을 따른다. 전술한 예시를 활용하자면, 독립변수 X_1, X_2, X_3의 변수값이 동일한 응답자들이 10명 존재하며, 회귀 대치법을 통해 예측된 소득이 200만원이라고 가정 할 때, 확률적 회귀 대치법은 응답자 10명의 결측을 200만원에 무작위 값을 더한 값으로

대치한다. 이 때, 10명의 응답자들에게 부여되는 무작위 값 10개의 평균은 0으로 수렴하며, 무작위 값들의 변량은 관측치에서 계산된 변량에 근사(approximate)하게 된다.

2. 연쇄방정식을 활용한 다중 대치법(Multiple Imputation by Chained Equations; MICE)

연쇄방정식을 활용한 다중 대치법(MICE)은 다중 대치법의 일종이다. 일반적인 다중 대치법(Multiple Imputation)은 Rubin(1987)에 의해 개발되었으며, 각각의 결측은 관찰된 변수들을 활용한 회귀분포(distribution of an imputation regression model)로부터 무작위로 추출하는 등의 기법을 통하여 대치된다. 이런 단계를 반복하여 복수의 자료를 생산한다. 통계분석 결과는 예측 후 통합(predict then combine; PC) 기법을 통해 산출된다(Miles, 2015). 다시 말하여, 다중 대치 자료의 분석 알고리즘은 개별 자료를 분석한 후, 이 결과들을 Rubin의 법칙[5](Rubin's rule; Rubin, 1987)을 활용하여 종합하는 방식으로 이루어진다.

연쇄방정식을 활용한 다중 대치법(MICE)은 회귀모형을 활용하여 다수의 대치된 자료(dataset)를 만들어 분석에 활용하는 기법으로써 기존의 다양한 다중 대치법들의 장점을 통합하였으며, 다양한 방식으로 세부 사항을 조절(alter) 할 수 있는 유연한 대치 기법이다(Van Buuren & Oudshoorn, 1999).

MICE 기법으로 대치된 자료를 생산하는 방법은 6단계로 설명할 수 있다(Azur et al., 2011; p. 3). 1단계에서는 평균대치법 등을 활용하여 변수들의 결측값을 대치한다. 2단계에서는 대치된 변수들 중 특정 변수("변수 A")의 대치값을 결측상태로 되돌린다. 3단계에서는 "변수 A"를 종속변수로 하며 나머지 변수들을 독립변수로 하는 회귀분석이 실시된다. 이 때 회귀분석 모형은, 변수의 분포에 따라 일반선형 회귀분석, 로짓분석, 혹은 포아송 회귀분석 등으로 지정 할 수 있다. 4단계에서는 회귀분석 결과를 토대로 "변수 A"의 결측값이 대치된다. 이렇게 대치된 "변수 A"

5) Rubin의 법칙을 통해 산출된 점 추정값(point estimation)은 각 각의 자료에서 산출한 결과의 평균치이며, 표준오차는 자료 내 분산 및 자료 간 분산을 활용하여 계산된다.

는 이후의 단계에서 다른 변수 대치를 위한 독립변수의 역할을 한다. 5단계는 2단계부터 4단계 과정이 “변수 A” 외 나머지 변수들을 위해 반복되는 것이다. 이 과정이 모든 변수들을 위해 1회 실시되는 것을 하나의 사이클(cycle)로 정의한다. 6단계는 복수의 사이클을 실시하여, 대치값이 안정적인 값을 갖도록 갱신하는 것이다. 일반적으로 10회의 사이클이 실시되며, 사이클의 수는 임의로 지정할 수 있다.

1단계에서 6단계의 과정을 거치면 하나의 대치된 자료가 생산된다. MICE는 이러한 과정을 다수 거쳐 복수의 대치된 자료를 생산하며, 통계분석에는 복수의 대치된 자료가 활용된다. 복수의 자료 활용은 대치로 인한 불확실성(uncertainty)을 감소시켜 결과의 신뢰성을 높일 수 있게 된다.

Ⅳ. 연구방법

1. 연구 자료

대치 방법이 자료의 통계 분석 결과에 미치는 영향을 확인하기 위하여, 청소년정책연구원에서 수행한 한국아동·청소년패널조사(KCYPS) 중1패널의 6년차 자료를 활용하여 분석을 진행하였다. 후술할 통계분석에는 총 네 개의 자료가 활용되었다. 첫째는 KCYPS 중1패널 6년차 “원조사자료(raw data)”이며, 이 자료에는 다수의 결측값이 존재한다. 이와 더불어, KCYPS 6년차 자료의 결측을 확률적 회귀대치 및 연쇄방정식을 활용한 다중대치(MICE)를 통하여 2개의 추가적인 자료[6]를 구성하였다. 여러 가지 대치 방법 중 확률적 회귀 대치법과 MICE를 활용한 이유는, KCYPS 6년차 자료의 결측 패턴은 MCAR이라기보다 MAR이라는 가정이 더 적합하기 때

6) 확률적 회귀대치 및 MICE는 STATA 15.0 프로그램을 활용하여 수행되었다. MICE 분석의 경우, 흡연유무는 로짓모형(Logit), 부친 교육수준 및 모친 교육수준, 자기통제는 순서형 로짓 모형(Ordered logit), 음주 횟수 및 평균소득은 연속형 변수를 위한 예측적 평균 매칭 모형(Predictive mean matching for a continuous variable; PMM)으로 설정하였다. 대치된 총 자료의 수는 60개로 설정하였다. 한편, MICE는 본질적으로 복수의 확률적 회귀대치 자료를 산출하는 기법이라는 점에서 (https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/), MICE를 통하여 하나의 자료를 새로 산출하여, 이를 확률적 회귀대치 자료로 활용하였다.

문이다. 마지막으로, 전술한 세 개의 자료를 활용한 통계 분석이 얼마나 정확한지를 확인하기 위하여 결측이 존재하지 않는 현실의 대리자료(proxy)로서 "완전한 자료[7]"를 구성하였다.

2. 분석 방법

각 자료들이 "완전한 자료"를 정확하게 반영하고 있는지를 확인하기 위하여, "완전한 자료", 원조사 자료, 확률적 회귀대치 자료, 다중대치 자료를 활용하여 분석한 결과를 비교하였다. 통계분석에 활용된 변수는 음주횟수, 흡연유무, 부친 학력, 모친 학력, 낮은 자기통제, 가계소득으로써, 이상 네 개의 자료를 활용하여 각 변수들의 평균을 비교하고, 음주횟수를 종속변수로 하는 회귀분석 결과를 비교하였다.

이들 변수들은 본 연구의 목적인 대치 방법에 따른 결과의 비교를 위해 선택되었다. 즉, 회귀분석 모델은 음주 혹은 일탈과 관련한 이론적 배경이나 선행연구를 활용하여 구성된 것이 아니며, 따라서 분석 결과는 범죄학의 이론적 측면에서 가치를 가지지는 않는다.

부친 학력, 모친 학력, 가계소득이 활용된 이유는 이들 변수들은 다른 변수에 비하여 "완전한 자료"를 구성하는데 이점이 있기 때문이다. 예컨대, KCYPS 6년차 자료의 부친 학력이 결측일 경우, 이 변수는 KCYPS 5년차 자료로부터 보완될 수 있다. 물론, 부모 학력 및 가계소득은 시간에 따른 변동이 발생하지만, 자료 안에 존재하는 다른 변수들보다(예컨대, 월평균 교육비, 컴퓨터 사용시간) 변동이 적으며, 주관적인 척도를 사용하는 변수들보다(예를 들어, "휴대전화가 없으면 불편해서 살 수 없다") 안정적[8]으로 측정 될 것으로 판단하였다.

---

7) 원조사자료(청소년패널데이터 6년차)는 결측값을 다수 가지고 있기 때문에, 결측의 실제값을 알 수 없으며, 대치된 자료들이 실제값에 얼마나 근접하였는지를 확인할 수 없다. 이에, 청소년패널데이터 과거년도에 존재하는 변수들을 활용하여, 연구에 활용될 변수들의 결측을 현실에 근접하게 보완하였다. 예컨대, 청소년패널데이터 6년차 자료의 가계소득이 결측일 경우, 이를 청소년패널데이터 5년차 자료를 활용하여 보완하였다. 이렇게 해서 만들어진 자료를 "완전한 자료"로 표기하였으며, 여기서 "완전한 자료"란 현실을 완벽하게 반영한다는 의미가 아니라, 각 대치 방법을 통해 산출한 결과물의 기준점을 제시하기 위한 현실에 가장 가까운 준거자료라는 의미로 사용되었다.

8) 부모 학력의 경우, "고졸", "대졸" 등으로 측정된다. 따라서, 반복 측정시, 학력에 변화가 없다면 과년도의 응답을 반복할 확률이 크다. 이에 비하여, "매우 그렇다", "전혀 그렇지 않다" 등으로 측정

한편, 자기통제의 경우, 이론적인 측면에서 시간의 흐름에 따라 비교적 안정적인 개인적 특성(individual trait)으로 알려져 있다(Gottfredson & Hirschi, 1990). 마지막으로 음주와 흡연의 경우, 한 번 발현시, 다른 일탈 및 비행 행위에 비하여 시간에 따른 변동이 적을 것으로 가정하였다. 즉, 자료에 포함되어 있는 다른 변수들과 비교할 때, 이상의 변수들은 과년도 조사자료를 활용하여 신뢰할만한 "완전한 자료"를 완성할 수 있을 것으로 기대할 수 있다.

V. 결과

1. 평균 비교

먼저, 완전한 자료와 다른 자료 간 변수 평균값을 비교하였다. 음주횟수의 경우, 완전한 자료는 평균 0.618을 나타냈으며, 원조사자료는 0.571, 다중대치는 0.573, 회귀대치는 0.570을 나타냈다. 흡연유무의 평균은 완전한 자료에서 0.114로 나타났으며, 이는 응답자의 11.4%가 흡연 경험이 있음을 나타낸다. 원조사자료의 흡연유무 평균은 0.103, 다중대치 자료는 평균 0.109, 회귀대치 자료는 평균 0.111로 나타났다. 부친 학력 평균[9]의 경우, 완전한 자료 3.023, 원조사 자료 3.031, 다중대치 자료 3.007, 회귀대치 자료 3.016으로 나타났다. 모친 학력 평균의 경우, 완전한 자료 2.770, 원조사 자료 2.782, 다중대치 자료 2.767, 회귀대치 자료 2.771로 나타났다. 낮은 자기통제 평균의 경우, 완전한 자료 2.590, 원조사자료 2.579, 다중대치 자료 2.587, 회귀대치 자료 2.587로 나타났다. 마지막으로 가계소득 평균은 완전한 자료 5048.365, 원조사 자료 5143.740, 다중대치 자료 5054.222, 회귀대치 자료 5058.413으로 나타났다.

이상의 결과들을 비교할 때, 다중대치와 회귀대치가 원조사 자료에 비하여 완전

되는 변수들은 시간이 경과하고 반복 측정시 실제로는 변화가 없더라도, 응답자가 과년도에 선택한 것과 다른 응답항목을 선택할 가능성이 상대적으로 크다.

9) 부친과 모친의 학력 "평균"은 각 자료 간의 비교 목적을 위해 계산되었으나, 이들 변수들은 순위척도(ordinal scale)로 측정되지 않았기 때문에, "평균"의 실제적인 가치는 존재하지 않는다.

한 자료에 근접한 것으로 확인되었다. 회귀대치 자료의 경우, 여섯 가지 변수 평균 중 네 차례 완전한 자료에 가장 근접하였으며, 다중대치 자료의 경우, 여섯 가지 변수 평균 중 세 차례 완전한 자료에 가장 근접한 것으로 나타났다. 예컨대, 가계소득의 경우, 완전한 자료가 약 5,048만원을 나타냈고, 다중대치 자료는 약 5,054만원, 회귀대치 자료는 약 5,058만원으로 나타났다. 즉, 완전한 자료와 비교하여 다중대치 자료와 회귀대치 자료는 약 5만원 ~ 10만원 정도의 오차를 나타낸 것이다. 하지만, 결측값을 대치하지 않은 원조사 자료의 평균은 약 5,143만원으로 완전한 자료와 100만원 가까운 차이를 나타냈으며, 다른 변수들 같은 경우에도 원조사 자료가 완전한 자료에 가장 근접한 경우는 없는 것으로 나타났다.

〈표 1〉 평균 비교

	완전한 자료	원조사자료		다중대치		회귀대치	
	A	B	A-B	B	A-B	B	A-B
음주횟수	0.618	0.571	0.047	**0.573**	0.045	0.570	0.048
흡연유무	0.114	0.103	0.011	0.109	0.005	**0.111**	0.003
부 학력	3.023	3.031	-0.008	3.007	0.016	**3.016**	0.007
모 학력	2.770	2.782	-0.012	2.767	0.003	**2.771**	-0.001
낮은자기통제	2.590	2.579	0.011	**2.587**	0.003	**2.587**	0.003
가계소득	5048.365	5143.740	-95.375	**5054.222**	-5.857	5058.413	-10.048

주: 볼드체는 완전한 자료에 가장 가까운 근사값임.

2. 회귀분석 결과

다음으로 음주횟수를 종속변수로 설정한 회귀분석을 실시하였다. 다른 자료들의 기준이 되는 완전한 자료의 분석 결과, 분석모형은 통계적으로 유의한 것으로 나타났으며(LR chi2(10) = 190.78***), 음주횟수는 흡연유무(b=.903***), 모친 학력(b=-.086*), 낮은 자기통제(b=.158***)로부터 통계적인 영향을 받는 것으로 나타났다. 다시 말하여, 흡연을 하는 청소년들은 그렇지 않은 청소년들에 비해 음주횟수가 증가하며, 자기통제력 혹은 모친 학력이 낮은 경우에도 청소년들의 음주횟수가 증가하는 것으로 나타났다.

결측값을 가지고 있는 원조사자료의 분석결과 역시 전체 모형은 통계적으로 유의한 것으로 나타났으며(LR chi2(10)=108.10***), 흡연유무(b=.814***), 부친 학력(b=.094), 모친 학력(-.129**)이 음주횟수에 영향을 미치는 것으로 나타났다. 이는 흡연을 하는 청소년들은 그렇지 않은 청소년들에 비해 음주횟수가 증가하며, 부친 학력의 증가는 청소년들의 음주횟수를 증가시키지만 모친 학력의 증가는 음주횟수를 감소시킨다는 것을 의미한다.

결측값을 MICE 기법을 활용하여 복수의 데이터세트로 대치한 다중대치 자료의 분석결과는 통계적으로 유의한 것으로 나타났으며(F(5, 8616.0) = 21.31***), 흡연유무(b=.748***), 모친 학력(b=-.129**), 낮은 자기통제(b=.115*)가 음주횟수에 통계적인 영향을 미치는 것으로 나타났다. 이는 완전한 자료 결과와 동일한 것으로 흡연 및 낮은 자기통제와 모친 학력이 음주횟수를 증가시킨다는 것이다.

결측값을 확률적 회귀 대치법을 활용하여 대치한 회귀대치 자료의 분석결과는 통계적으로 유의한 것으로 나타났으며(LR chi2(10)=113.66***), 흡연유무(b=.752***), 부친 학력(b=.099*), 모친 학력(b=-.129**), 낮은 자기통제(b=.106*)가 통계적으로 유의한 것으로 나타났다. 이는 흡연, 부친 학력 증가, 모친 학력 감소, 낮은 자기통제가 음주횟수 증가에 영향을 미친다는 것이다.

각 변수의 통계적인 유의도를 기준으로 볼 때, 다중대치 모형의 결과는 완전한 자료 모형의 분석 결과와 동일한 결과를 나타냈으며, 확률적 회귀대치의 경우, 완전한 자료에서 유의하지 않았던 부친 학력이 유의한 것으로 나타났다. 한편, 결측값 처리를 하지 않은 원조사 자료의 분석 결과는 완전한 자료에서 유의하지 않았던 부친 학력이 유의하였고, 완전한 자료에서 유의했던 자기통제는 유의하지 않은 것으로 나타났다.

즉, 완전한 자료가 모수(parameter)를 정확히 추정한다고 가정할 때[10], 결측을 포

10) 다만, 이 가정은 통계적으로 강한 가정(strong assumption)으로써, 결측값이 전혀 없는 조사자료일지라도, 통계분석과정에서 표본의 수, 표본 선정 방법, 부적절한 통계 모형 등 여러 가지 이유로 인하여 오류가 발생할 가능성이 존재한다. 다시 말하여, 결측값이 없는 조사자료에서도 분석을 통해 산출된 추정값(estimate)이 모수(parameter)와 통계적으로 유의한 차이를 보일 수 있다는 것이다. 즉, 이상의 연구결과에서, 통계적인 우연에 의하여 원조사 자료가 완전한 자료에 비하여 모수(parameter)를 더 정확하게 추정할 가능성은 낮은 확률이나마 존재한다.

함하고 있는 원조사 자료에서는 1종 및 2종 오류가 모두 발견되었으며, 회귀대치 모형에서도 2종 오류가 발견되었다고 할 수 있다.

다중대치 모형의 경우, 완전한 자료와 가장 유사한 결과가 나타났다. 이는 앞선 기술통계(평균) 분석에서 다중대치 자료와 회귀대치 자료가 완전한 자료에 더 근접한 평균값을 가지고 있다는 결과와도 일치한다고 볼 수 있다.

〈표 2〉 회귀분석 결과

	완전한 자료			원조사자료			다중대치			회귀대치		
	b	S.E.	p	b	S.E.	p	b	S.E.	p	b	S.E.	p
흡연유무	0.903***	0.070	0.000	0.814***	0.082	0.000	0.748***	0.078	0.000	0.752***	0.076	0.000
부 학력	0.066	0.038	0.085	0.094*	0.043	0.028	0.088	0.046	0.059	0.099*	0.040	0.014
모 학력	-0.086*	0.042	0.043	-0.129**	0.047	0.006	-0.129**	0.049	0.009	-0.129**	0.044	0.004
낮은자기통제	0.158***	0.042	0.000	0.080	0.047	0.088	0.115*	0.045	0.011	0.106*	0.044	0.016
가계소득	0.000	0.000	0.932	0.000	0.000	0.190	0.000	0.000	0.399	0.000	0.000	0.520
Cons.	-1.029	0.152	0.000	-0.917	0.167	0.000	-0.957	0.167	0.000	-0.949	0.158	0.000
	N = 1,871 LR chi2(10) = 190.78*** Pseudo R2 = .0322			N = 1,655 LR chi2(10) = 108.10*** Pseudo R2 = .0218			N = 1,871 Imputations = 60 Average RVI = .195 Largest FMI = .336 F(5, 8616.0) = 21.31***			N = 1,871 LR chi2(10) = 113.66*** Pseudo R2 = .0203		

주1: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Ⅵ. 논의 및 결론

1. 결과의 요약

이 연구는 범죄학 연구에서 활용되는 조사자료의 결측값 및 결측처리 방법의 통계적인 영향을 KCYPS 중1패널 6년차 자료를 활용하여 분석하였다. 분석에는 KCYPS 과년도 자료를 활용하여 구성한 “완전한 자료”, 결측값이 존재하는 “원조사 자료”, MICE 기법을 활용하여 대치한 “다중대치 자료”, 확률적 회귀 대치 기법을 활용하여 대치한 “회귀대치 자료”가 사용되었다.

결측값이 그대로 존재하는 원조사 자료 및 결측을 처리한 자료들이 얼마나 정확한 근사값을 갖는지 확인하기 위하여, 몇 가지 변수를 활용하여 변수의 평균값 및 회귀분석 결과를 비교하였다.

분석 결과, 다중대치 및 확률적 회귀대치 자료들이 원조사 자료에 비하여 완전한 자료에 가까운 결과를 나타냈다. 평균의 비교에서는 확률적 회귀대치 자료가 완전한 자료와 가장 유사한 결과를 나타냈으며, 회귀분석에서는 다중대치 자료가 완전한 자료와 가장 유사한 결과를 나타냈다. 결측값이 존재하는 원조사 자료의 경우, 회귀분석 결과에서 1종 및 2종 오류의 징후가 모두 나타났다.

주목할 만한 점으로, 다중대치와 확률적 회귀대치의 결과에서, 단순 평균의 비교는 확률적 회귀대치가 완전한 자료에 근접한 값들을 나타냈으나, 회귀분석의 결과에서는 다중대치 기법이 더 우수한 것으로 나타났다. 이는 확률적 회귀대치와 관련한 이론적 논의를 통해 설명이 가능하다. 확률적 회귀대치의 경우, 일반 회귀대치기법보다는 덜 편향된 표준오차(standard error)를 산출하지만, 여전히 과소추정된 표준오차를 산출하는 경향이 있다. 이는 다중회귀대치법과 비교하여 더 높은 1종 오류(잘못된 긍정) 가능성으로 이어진다(Enders, 2010). 이 연구의 결과에서도, 완전한 자료에서 통계적으로 유의하지 않은 부친 학력이 확률적 회귀대치자료에서는 유의한 것으로 나타나 1종 오류의 징후를 나타냈다.

2. 학술적 함의

결측이 완전한 무작위 패턴(Missing Completely at Random; MCAR)인 경우를 제외할 때, 원조사자료(raw data)는 이론적으로 1종 오류 및 2종 오류 모두에 취약하다. 물론, MCAR인 경우에도 결측처리를 한 자료에 비해 응답자 수가 적기 때문에, 1종 오류에 더 크게 노출 될 수 있다. 한편, 확률적 회귀 대치의 경우, 결측값을 대치함으로써 1종 오류의 확률을 낮추며, 적절한 회귀 모형을 사용할 경우 2종 오류 혹은 왜곡된 추정값(biased estimate)이 나타날 확률을 낮출 수 있다. 다만, 회귀 대치를 포함하는 단일값 대치 방법들의 경우, 결측을 하나의 값으로 대치한다는 점에서, 대치값의 불확실성을 파악할 수 없다는 문제를 가진다. 다중 대치는 이러한 단일값 대치의 문제점을 보완하는 기법으로써, Enders (2010)는 어떤 대치법을 사용해야 할 지 확신이 없는 경우, EM 기법이나 다중 대치법을 사용하는 것을 권장하였다.

이 연구는 이상의 이론적 논의들이 실제 데이터에서 어떻게 구현되는지를 분석하였고, 이론적 논의들에 부합하는 결과를 확인할 수 있었다. 즉, 결측값을 가지고 있는 원조사자료를 통계분석에 그대로 활용하는 것은 왜곡된 추정값을 산출할 확률을 높인다. 특히, 범죄학 연구의 주요 변수들(예를 들어, 낮은 자기 통제, 일탈 행위 가담)은 잠재적으로 결측 발생 및 패널 탈락과 관련이 있을 수 있기 때문에, 적절하게 결측을 처리하지 않을 경우, 비행, 범죄 및 범죄피해가 과소 추정될 우려가 있으며, 이들과 다른 변수들 간의 통계적인 영향관계가 왜곡될 우려가 있다.

실제로 평균의 비교에서, 현실을 반영한 완전한 자료에서는 약 11.4%의 응답자가 흡연 경험이 있는 것으로 나타났으나, 원조사 자료에서는 실제보다 적은 10.3%로 나타났다(다중대치의 경우 10.9%, 회귀대치의 경우 11.1%). 가계소득의 경우, 자료 유형에 따른 평균 차이가 매우 큰 것으로 나타났다. 완전한 자료의 경우 가계 평균소득이 5,048만원으로 나타났으나, 결측값이 존재하는 원조사 자료에서는 5,143만원으로 나타나 약 100만원 가까운 차이를 보였다(다중대치의 경우 5,054만원, 회귀대치의 경우, 5,058만원). 이와 같이 원조사 자료가 완전한 자료보다 더 큰 가계소득 및 더 낮은 흡연자 비율을 나타내는 것은, 저소득층 및 흡연자 청소년들이

설문에서 탈락(attrition)하는 경향이 높기 때문으로 볼 수 있다.

이보다 더 큰 문제는, 결측은 회귀분석 등 추리통계분석 결과를 왜곡할 수 있다는 점이다. 예를 들어, 이 연구에서 완전한 자료를 활용한 분석결과는 낮은 자기통제가 음주횟수와 1종 오류 확률 .001 미만의 매우 유의한 통계적인 관계를 가지고 있다고 나타냈으나, 결측값을 가지고 있는 원조사 자료를 활용한 연구결과는 자기통제와 음주횟수 간의 통계적인 관계를 발견하지 못했다. 이러한 결과는 원조사자료의 상대적으로 적은 표본 크기 혹은 자기통제력이 낮은 응답자들이 패널에서 탈락하는 현상 등이 원인이 될 수 있다. 반면, 완전한 자료에서 나타나지 않은 부친 학력과 음주횟수 간의 통계적인 관계가 원조사 자료에서 발견되는 2종 오류의 징후도 나타났다.

이상의 결과를 종합하여 볼 때, 범죄학 연구에서 결측값의 처리는 매우 중요하다고 할 수 있다. 범죄학 연구에서 발생하는 결측은 범죄학 연구의 주요 관심인 범죄 및 일탈과 관련되었을 가능성이 크며, 결측을 적절히 처리하지 않을 경우 범죄학 연구의 중요한 대상(범죄 및 일탈 성향이 높은 응답자)를 제외시키며 통계분석 결과를 왜곡시킬 우려가 있다.

3. 정책·실무적 함의: 정책의 타당성 확보를 위한 결측 처리의 필요성

형사정책은 개인의 자유를 일시적 혹은 장기간에 걸쳐 구속할 수 있으며 시민의 안전과 직결된다는 점에서 정확한 과학적 근거를 가지고 추진되어야 한다. 예컨대, 재소자의 가석방 심사에서 활용되는 재범예측지표는 재범 가능성을 높은 확률로 예측 할 수 있어야 하며, 경찰의 순찰 방법은 범죄억제 효과를 극대화할 수 있는 방법을 활용하여 실시해야 한다.

조사자료 및 공식통계 등 통계자료를 활용한 실증연구는 증거에 기반을 둔 과학적인 정책 수립을 위한 매우 중요한 자료로 활용될 수 있다. 이러한 점에서, 신뢰할 만한 결과를 도출하는 것은 연구자에게 요구되는 학술적 역량일 뿐만 아니라, 시민의 안전 및 개인의 자유와 직결된 사회적인 의무로 볼 수 있다.

이러한 점에서 범죄 관련 통계 자료를 분석할 때, 결측을 적절하게 처리하는 것은

학술적 가치뿐만 아니라 정책적인 의미를 가진다. 특히, 패널조사에서 일탈 성향이 높은 응답자 및 저소득층 응답자의 탈락 가능성이 높다. 따라서, 형사정책 및 청소년보호 정책 수립을 위한 패널 연구에서는 결측 문제를 반드시 적절한 방법을 활용하여 해결해야 한다. 예를 들어, 이 연구에서 현실의 대리(proxy)자료인 완전한 자료의 회귀분석 결과는 낮은 자기통제가 음주횟수를 증가시킨다는 결과를 나타냈으나, 결측치를 가지고 있는 원자료에서는 이러한 영향 관계가 나타나지 않았다. 만약 이 연구결과가 청소년의 음주와 관련한 정책 수립의 자료로 활용된다면, 청소년 음주 문제 해결을 위해 낮은 자기통제와 관련한 정책적인 개입이 필요함에도 불구하고, 편향된 분석 결과에 의하여 낮은 자기통제와 관련한 프로그램은 지지받지 못하게 된다.

한편, 통계 자료의 배포에 있어서, 통계 자료 생산자는 결측 처리 방법에 대한 권장사항을 제시하거나 결측 대치된 자료를 추가적으로 제공할 필요가 있다. Engels와 Diehr(2003, p.968)이 지적한 바와 같이, 결측 대치는 어떠한 자료를 더 완전하게 만드는 방법으로써, 다양한 연구자들에 의해 활용될 규모가 크며 영향력 있는 통계자료(large prospective database)에 적용이 장려된다. 이는 단순히 연구자들의 편의를 위한 것이 아니라, 자료를 활용한 일관성 있으며 신뢰성 높은 연구를 촉진하기 위한 것이다. 앞서 확인한 바와 같이, 결측 처리를 하지 않은 자료에 비하여, 결측 처리를 한 자료들이 더 정확한 결과를 산출하였으며, 결측 처리 방식에 따라서도 연구 결과가 바뀔 수 있다.

이러한 측면에서, 전술한 미국의 범죄통계 자료 외에도 다양한 조사들은 대치된 자료를 제공하고 있다. 이내성(2007)의 연구에 따르면 미국의 Consumer Expenditure Survey 및 일본의 Comprehensive Survey of Living Condition of the People on Health and Welfare 등은 회귀대치 자료를 제공하며, 영국의 Family Resources Survey 및 호주의 Household Expenditure Survey 등은 핫덱 대치법을 활용하여 결측값을 대치한 자료를 제공하고 있다.

대치된 자료를 제공하는 것은 추가적인 자원을 필요로 하기 때문에 모든 통계 자료 배포에 강제할 수는 없지만, 최소한 정책 활용도가 높은 국가승인통계의 경우 이를 활용한 연구결과의 신뢰성 확보를 위하여 결측 문제 해결을 위한 해법을 제공

할 필요가 있다.

4. 연구의 한계

이 연구는 결측값의 처리 방식이 통계분석에 미치는 영향의 차이를 확인하기 위하여 “완전한 자료”를 기준으로 활용하였다. 완전한 자료는 과년도의 자료를 활용하여 구성된 것으로, 이 연구는 완전한 자료가 결측값들을 가장 완벽하게 보완한 자료라고 가정하였다. 이러한 가정을 위하여, 연구에 활용한 변수들 또한 자료 내에 존재하는 다른 변수들에 비하여 시간에 따른 변동이 적으며, 안정적인 척도를 가진 변수들로 선정하였다.

전술한 연구설계에도 불구하고, 완전한 자료는 결측의 실제 값을 완벽하게 구현했다고 볼 수 없다. 이와 더불어, 완전한 자료 및 대치된 자료에서 산출한 결과보다 원조사 자료의 분석 결과가 실제 모수에 더 가까울 가능성도 낮은 확률로 존재한다. 이 연구의 경우, 원조사 자료에 비하여 다중대치 및 회귀대치가 완전한 자료에 가까운 통계학적인 설명과 일치하는 결과를 나타냈지만, 다른 자료를 활용할 경우 원조사 자료가 여타 대치 방법보다 더 정확하게 완전한 자료를 반영할 가능성 역시 낮은 확률로 존재한다. 따라서, 대치된 자료가 원조사 자료에 비해 항상 우월하다고 볼 수는 없다.

하지만 표본이 적절히 선택되었다고 가정할 때, 통계학적으로 적절히 대치된 자료가 원조사 자료에 비해 모수를 더 정확히 예측할 가능성은, 원조사 자료가 더 정확한 모수를 추정할 확률보다 매우 높다. 이 연구에서 활용된 “완전한 자료”의 신뢰성이 완전하게 담보될 수 없음에도 불구하고, 이 연구의 결과는 이러한 통계학적 이론과 일치하며, 이를 통해 뒷받침 될 수 있다.

5. 결어

국내외 일부 연구자들은 결측값 대치를 부적절한 자료의 조작이라고 여기는 경우도 있으며, 적절한 방법을 사용하지 않은 결측값 대치는 자료의 편향을 심화시킬

우려도 있다. 하지만, 많은 자료들은 앞서 원자료 활용이 가능한 기준으로 제시하였던 5% 이상의 결측값을 가지고 있으며, 또한 대부분의 경우 결측 패턴은 완전한 무작위 패턴(MCAR)이 아니다. 따라서, 이러한 자료들을 그대로 활용하여 통계 분석을 실시하는 것은 오히려 편향된 결과를 산출할 확률을 높이는 것이다. 범죄 및 일탈은 일반적으로 스스로 밝히기 꺼리는 행동이며, 범죄성향이 높은 사람들은 패널조사에서 중도 탈락할 확률이 높다는 점에서, 범죄학 및 형사사법 연구에서도 결측 패턴의 분석 및 결측값에 대한 적절한 처리는 매우 중요하다.

이 연구에서 활용된 MICE 기법은 결측 대치 방법 중 가장 우수한 대치 방법의 하나로 여겨지고 있으나, 이러한 기법 외에도 다양한 대치 방법 및 결측 문제의 통계적인 해결방법들이 제시되고 있다. 즉, MICE는 만능 대치법은 아니며, MICE로 대치된 자료를 활용하여 실시하기 어려운 분석 방법도 존재한다. 이 연구에서 MICE 기법을 활용한 이유는, MICE 기법의 우수성을 제시하려는 목적보다는, 결측 대치의 필요성을 실증 자료를 활용하여 증명하기 위한 것으로써, 향후 연구들에서 MICE 기법을 포함하여 자료의 특성에 비추어 가장 적절한 대치 방법을 활용하여 유효한 연구결과를 산출하는데 기여하기를 기대한다.

참고문헌

김형민·함건희·서병태, (2016), 결측 공변량을 갖는 혼합회귀모형에서의 EM 알고리즘, *The Korean Journal of Applied Statistics*, 29(7), 1347-1359.

송주원, (2010). 결측을 포함한 반복측정자료 모형에서 결측자료 메커니즘의 영향, *Journal of the Korean Data Analysis Society*, 12(3) (B), 1463-1472.

이내성, (2007), 회귀분석을 이용한 Imputation 기법활용 연구 -사회통계조사 항목 무응답을 중심으로, 통계청 연구보고서.

황지태. (2010). 범죄피해율과 공식범죄발생률간의 비교분석: 2008년도 주요범죄 암수추정. 형사정책연구원. 형사정책연구, 21(3), 7-51.

Addington, L. A. (2007). Using NIBRS to study methodological sources of divergence between the UCR and NCVS. *Understanding crime statistics: Revisiting the divergence of the NCVS and the UCR*, 225-250.

Akiyama, Y. & Propheter, S. K. (2005). *Methods of data quality control: For Uniform Reporting Programs. Clarksburg, WV*, Criminal Justice Information Services Division, Federal Bureau of Investigation.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research*, 20(1), 40-49.

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25, 464-469.

Berg, M. T., & Lauritsen, J. L. (2016). Telling a similar story twice? NCVS/UCR convergence in serious violent crime rates in rural, suburban, and urban places (1973–2010). *Journal of Quantitative Criminology*, 32(1), 61-87.

Berzofsky, M., Creel, D., Moore, A., Smiley-McDonald, H., & Krebs, C. (2014). Imputing NCVS income data. US Department of Justice, Bureau of Justice Statistics, Washington, DC.

Bureau of Justice Statistics. (2016). Criminal Victimization, 2015.

Enders, C. K. (2010). *Applied Missing Data Analysis*. The Guilford Press.

Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology*, 56, 968-976.

Enzmann, D. (2013). The impact of questionnaire design on prevalence and incidence rates of self-reported delinquency: Results of an experiment modifying the ISRD-2 questionnaire. *Journal of Contemporary Criminal Justice*, 29(1), 147-177.

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime*. Stanford University Press.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549-576.

Graham, J. W., Cumsille, P. E., & Elec-Fisk, E. (2003). Methods for handling missing data. In J.A. Schinka & W. F. Velicer (Eds.). Research Methods in Psychology: Vol. 2. Handbook of Psychology, 87-114. Wiley.

Haas, S. M., LaValle, C. R., Turley, E., & Nolan, J. J. (2012). Improving state capacity for crime reporting: An exploratory analysis of data quality and imputation methods using NIBRS data.

Huizinga, D., & Elliott, D. S. (1986). Reassessing the reliability and validity of self-report delinquent measures. *Journal of Quantitative Criminology*, 2(4), 293-327.

LaValle, C. R., Haas, S. M., & Nolan, J. J. (2014). Testing thevalidity of demonstrated imputation methods on longitudinal NIBRS data. Criminal Justice Statistical Analysis Center.

Lynch, J. P. & Addington, L. A (2007) *Understanding crime statistics: revisiting the divergence of the NCVS and UCR*. Cambridge University Press, Cambridge

Lynch, J. P., & Jarvis, J. P. (2008). Missing data and imputation in the uniform

crime reports and the effects on national estimates. *Journal of Contemporary Criminal Justice*, 24(1), 69-85.

Miles, A. (2015). Obtaining predictions from models fit to multiply imputed data. *Sociological Methods & Research*, 45(1), 175-185.

Mosher, C. J., Miethe, T. D., & Hart, T. C. (2010). *The mismeasure of crime*. Sage.

Pallant, J. (2007). SPSS survival manual, 3rd ed., Open University Press.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, 57(1), 1.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer, J. L. (1999). Multiple Imputation: A primer. *Statistical Methods in Medical Research*, 8(1), 3-15.

Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. *Criminal Justice*, 4(1), 33-83.

Thompson, M. P., Saltzman, L. E., & Bibel, D. (1999). Applying NIBRS data to the study of intimate partner violence: Massachusetts as a case study. *Journal of Quantitative Criminology*, 15(2), 163-180.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Van Buuren, S., & Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden, The Netherlands: TNO Prevention Center.

UCLA Institute for Digital Research and Education 자료. URL: https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/

Data in Criminology and Missingness: Comparison of Missing Imputations utilizing Real-World Data

Ra, Kwang-hyun*

Researchers often find that their datasets have missing values. In self-report surveys that are frequently utilized in social science research, respondents are able not to answer certain questions, or accidentally, they may not answer an item. These missing values are relevant to occurrence of type 1 error and type 2 error. The current study reviewed the significance of proper treatment of missing values and discussed several methods to handle missingness in data. In addition, this study compared results from several imputation methods, utilizing data from the Korean Children & Youth Panel Survey.

It is found that results from multiple imputation and regression-based imputation are similar to those from the proxy data of real world, whereas the row data that include missing values have some symptoms of type 1 and type 2 errors. Given the results, it is imperative to properly handle missing values in criminology research. In survey datasets, missing values may be related to crime and delinquency so that non-treatment of missing values may result in significant subjects in the dataset (e.g., individuals with high criminality), biasing the estimations of statistical analyses. The results also have some practical implications in that empirical findings are often utilized as scientific evidence of policies and practices.

❖ Keyword: Missingness, Imputation, Multiple Imputation by Chained Equations, MICE, Crime Statistics

* University of South Carolina. Department of Criminology and Criminal Justice. Ph.D. Candidate.

투고일 : 2월 28일 / 심사일 : 3월 23일 / 게재확정일: 3월 23일