연구논문

불완전한 반복측정 자료의 보정방법

Methods for Handling
Incomplete Repeated Measures Data

우해봉* · 윤인진**
Haebong Woo · In-Jin Yoon

사회조사 자료를 활용한 통계분석에 있어서 불완전 자료의 문제는 거의 모든 연구자들이 경험하는 하나의 보편적인 문제이다. 불완전 자료의 문제는 특히 패널조사와 같은 종단적 자료를 활용한 연구에 있어서 중요한 이슈가 된다. 본 연구의 목적은 최근까지 이루어진 불완전 자료에 대한 보정방법을 소개하는 것이다. 특히, 본 연구는 패널자료에서 발생한 불완전 자료의 처리에 대한 관심이 부족한 점을 고려하여 최근까지 이루어진 보정방법들을 반복측정 패널자료 분석에 적용하는데 초점을 맞춘다.

첫째, 본 연구는 불완전 자료에 대한 적절하지 못한 사후처리는 분석결과에 있어서 유의미한 차이로 이어질 수 있음을 시사한다. 특히, 분석결과는 반복측정 자료를 사용하는 연구의 경우 불완전 자료의 발생은 궤적의 초기값보다는 시간의 경과에 따른 궤적의 변화를 적절히 추정하는데 문제를 가질 수 있음을 시사하고 있다.

둘째, 분석결과는 완전제거법이나 평균대체법이 EM, FIML, MICE 방법들에 비해 불완전 자료의 처리효과가 상대적으로 떨어짐을 보여준다. 특히, 완전제거법이나 평균대체법과 같은 방법에 비해 최대우도법이나 다중대체법이 갖는 상대적 우위는 MCAR 가정에 비해 보다 현실적인 가정이라고 할 수 있는 MAR 조건하에서 크게 나타난다. 본 연구의 분석결과는 또한 비록 결측치의 발생기제가 MNAR 상황이라고 하더라도 연구자가 결측치의 발생과 관련된 변수들을 보정과정에서 적절하게 활용하면 편의의 상당부분을 감소시킬 수 있음을 시사한다.

주제어: 불완전 자료, 패널조사, 최대우도법, 다중대체법

Problems of incomplete data are pervasive in statistical analysis. In

* 국민연금연구원 부연구위원
** 교신저자(corresponding author): 고려대학교 사회학과 교수 윤인진.
E-mail: yoonin@korea.ac.kr

particular, incomplete data have been an important challenge in repeated measures studies. The objective of this study is to give a brief introduction to missing data mechanisms and conventional/recent missing data methods and to assess the performance of various missing data methods under ignorable and non-ignorable missingness mechanisms. Given the inadequate attention to longitudinal studies with missing data, this study applied recent advances in missing data methods to repeated measures models and investigated the performance of various missing data methods, such as FIML (Full Information Maximum Likelihood Estimation) and MICE(Multivariate Imputation by Chained Equations), under MCAR, MAR, and MNAR mechanisms.

Overall, the results showed that listwise deletion and mean imputation performed poorly compared to other recommended missing data procedures. The better performance of EM, FIML, and MICE was more noticeable under MAR compared to MCAR. With the non-ignorable missing data, this study showed that missing data methods did not perform well. In particular, this problem was noticeable in slope-related estimates. Therefore, this study suggests that if missing data are suspected to be non-ignorable, developmental research may underestimate true rates of change over the life course. This study also suggests that bias from non-ignorable missing data can be substantially reduced by considering rich information from variables related to missingness.

Key words: incomplete data, panel survey, full information maximum likelihood estimation, multiple imputation

Ⅰ. 서론

사회조사 자료를 활용한 통계분석에 있어서 불완전 자료의 문제는 거의 모든 연구자들이 경험하는 하나의 보편적인 문제이다. 또한 통계자료를 사용하는 연구자의 입장에서 이러한 불완전 자료의 문제는 대부분 연구자의 통제범위를 벗어난 요인들과 관련된다. 불완전 자료의 문제는 종단적(longitudinal) 자료를 활용한 연구에 있어서 특히 중요한 이슈가 된다. 횡단적(cross-sectional) 조사의 경우 연구자는 불완전 자료의 규모가 작을 경우 분석과정에서 결측치를 가진 사례를 제거함으로써 문제를 해결할 수 있지만, 장기간에 걸친 패널조사와 같은 종단적 자료의 경우 분석과정에서 불완전 정보를 제공하는 사례를 매

회에 걸쳐 모두 제거하는 것은 결과적으로 최초 추출된 표본과는 매우 속성이 다른 표본으로 이어질 위험을 내포하고 있다.

불완전 자료의 문제가 대부분의 사회조사에 있어서 빈번하게 발생하는 현상이지만, 자료를 분석하는 연구자들은 대체로 이러한 불완전 자료가 무작위적으로 발생한다는 가정 하에 분석을 수행하는 경향이 있다. 또한 불완전 자료의 발생기제에 대한 충분한 고려 없이 불완전한 정보를 제공하는 사례를 분석에서 제외하거나 평균/중앙값 등으로 대체함으로써 불완전 자료의 문제를 비체계적으로 처리하는 경향이 있었다. 연구자가 제시한 분석결과가 불완전 자료의 문제로 인해 그 일반화에 있어서 어떠한 영향을 받을 수 있는가에 대한 논의는 더욱 더 찾아보기 힘들다.

불완전 자료의 체계적 처리에 관한 사회과학자들의 상대적 낮은 관심에도 불구하고 최근 수십 년 동안 불완전 자료의 처리에 관한 통계적 이론과 그 처리방법은 상당한 수준의 진척을 이루었다. 그러나 불완전 자료의 처리와 관련된 이론적 그리고 기술적 진척에도 불구하고 현재까지 불완전 자료와 관련된 논의는 대체로 횡단적 자료에 국한되는 모습을 보였음이 지적될 필요가 있다. 결과적으로 종단적 자료의 경우 불완전 자료의 발생이 분석결과에 어떠한 함의를 갖는가에 대한 논의는 상대적으로 찾아보기 힘들다(예를 들면, Hedeker & Gibbons 1997; Wothke 2000). 국내의 경우도 불완전 자료의 처리에 관한 관심이 증가하고 있지만(강민아 · 김경아 2006; 김규성 2000; 김덕준 2007), 종단자료의 측면에서 불완전 자료의 발생과 그 처리방법에 관한 논의는 상대적으로 미미한 수준이다. 기존의 논의를 확장하여, 첫째, 본 연구는 반복측정된 종단자료에서 발생한 불완전 자료가 분석결과에 어떠한 효과를 갖는가를 탐색한다. 둘째, 본 연구는 종단자료의 분석과정에서 불완전 정보를 제공하는 사례를 제거하는 방식이나 평균으로 대체하는 방식에 대비해 최대우도법이나 다중대체법과 같은 보다 체계적인 접근법들이 보이는 수행결과를 비교한다. 나아가 본 연구는 국내에서 증가하고 있는 패널자료에 대한 관심을 고려하여 종단자료에서 발생한 불완전 자료의 처리방법과 관련된 함의를 도출하고자 한다.

Ⅱ. 불완전 자료의 발생기제와 보정방법

1. 불완전 자료의 발생기제

사회과학에서 종단적 연구가 차지하는 비중은 매우 크다. 패널조사와 같은 종단적 조사의 경우 개인들은 중요한 변수들에 있어서 시간의 경과에 따른 변화를 반복적으로 응답하게 된다. 횡단조사에 대비한 이러한 패널조사의 특성을 충분히 활용하기 위해서는 장기간에 걸쳐 응답자들의 변화를 정확히 추적하는 작업이 요구되지만, 시간의 경과에 따른 불완전 자료의 발생은 이러한 목적 달성에 중대한 제약을 부과하는 경향이 있다. 패널조사의 경우 불완전 자료는 상대적으로 조사과정이 잘 통제된 경우에도 빈번하게 발생하는 경향을 보이는데, 패널자료를 활용한 분석의 경우 분석과정에서 불완전 자료에 대한 특별한 고려가 이루어지지 않을 경우 분석 대상 사례 수를 상당한 수준으로 감소시킬 위험이 있다. 결과적으로 패널자료의 경우 불완전 자료의 처리와 관련된 충분한 관심이 주어지지 않을 경우 분석결과의 타당성이 상당한 수준으로 훼손될 개연성이 존재한다고 할 수 있다.

패널조사를 포함한 사회조사에서 발생하는 불완전 자료를 적절하게 처리하기 위해서는 불완전 자료의 발생 원인에 대한 이해가 반드시 필요하다. 이렇게 불완전 자료의 발생기제에 대한 이해가 요구되는 것은 불완전 자료를 처리하는 통계적 방법들이 불완전 자료의 발생기제에 의존하는 것과 관련된다. 이러한 측면에서 Little & Rubin(2002)은 결측치 발생 유형(missing-data pattern)과 결측치 발생기제(missing-data mechanism)를 구분하는데, 결측치 발생 유형이 자료행렬에서 어떤 변수의 값이 관측 혹은 결측되었는가를 나타내는 반면 결측치 발생기제는 결측치의 발생과 자료에 있는 변수들 사이의 관계를 지시한다.

결측치 발생기제를 보다 구체적으로 설명하기 위해 Y를 $n \times k$ 차원의 자료행렬이라고 하자. 여기에서 n은 관측사례의 수를 그리고 k는 변수의 수를 각각 나타낸다. 결측치가 존재하는 경우 Y의 관측된 부분을 Y_{obs} 그리고 결측된 부분을 Y_{mis}로 표기하면, 전체 자료 Y는 관측된 값들과 결측된 값들로

구성된다($Y=(Y_{obs}, Y_{mis})$). 또한 불완전 자료의 발생 유형을 기술하기 위해 Y를 구성하는 변수 값이 관측되면 1 그리고 결측이면 0의 값을 갖는 $n\times k$ 차원의 더미변수들의 행렬을 R, 분석모형에서 추정될 모수의 벡터를 θ, 그리고 결측치 발생과 관련된 모수의 벡터를 ξ라고 하자.

불완전 자료의 발생과 관련하여 $f(R|Y_{obs}, Y_{mis}, \xi)=f(R|Y_{obs}, \xi)$의 조건을 충족하는 불완전 자료 발생기제가 MAR(Missing At Random)이다. 다시 말하면, MAR 조건은 결측치의 발생이 관측되지 않은 값에 의존하지 않음을 의미한다. MCAR(Missing Completely At Random)은 MAR의 특수한 경우인데, 결측치의 발생이 관측된 값(Y_{obs})이나 결측된 값(Y_{mis}) 모두에 의존하지 않는 불완전 자료 발생기제, 다시 말하면, $f(R|Y_{obs}, Y_{mis}, \xi)=f(R|\xi)$의 조건을 충족하는 불완전 자료 발생기제이다. MCAR 상황에서의 결측치 발생은 다른 모든 변수들과 독립적이지만, MAR의 경우 결측치의 발생은 「관측된」 다른 변수들에 의존할 수 있다. 그러므로 MAR 조건은 항상 분석모형에서 고려되는 변수에 조건적인데, MAR 조건이 충족되는 경우 Y에 있어서의 결측치 발생 확률은 결측치 발생과 관련된 다른 변수들을 분석모형에서 통제할 경우 더 이상 Y에 의존하지 않는다.

결측치 보정에 있어서 MAR의 중요성은 분석모형의 모수(θ)에 관한 우도추정(예를 들면, θ에 대한 최대우도값) 과정에서 결측치 발생을 명시적으로 모형화할 필요가 있는가와 관련이 있다(Schafer 1997: 10-13).[1]

$$f(R, Y_{obs}|\theta, \xi)=\int f(R, Y|\theta, \xi)\,dY_{mis}$$

$$=\int f(R|Y, \xi)f(Y|\theta)\,dY_{mis}=f(R|Y_{obs}, \xi)\int f(Y|\theta)dY_{mis}$$

$$=f(R|Y_{obs}, \xi)f(Y_{obs}|\theta) \qquad (1)$$

---

1) 분석모형에서의 모수(θ)와 결측치 발생 모수(ξ) 사이의 독립성 가정이 추가적으로 필요하지만, 현실적으로 이 조건은 대부분 충족되는 것으로 논의된다. 분석모형의 모수 θ는 관측된 자료(Y_{obs})의 모수가 아닌 전체 자료($Y=(Y_{obs}, Y_{mis})$)에서의 모수를 의미한다.

방정식 (1)은 MAR 조건하에서 관측된 자료(Y_{obs})와 결측치 발생(R)의 결합분포(joint distribution)가 모수 ξ와 θ의 부분으로 분해됨을 보여준다. 그러므로 MAR (그리고 θ와 ξ의 독립성) 조건이 충족되는 경우 분석모형의 모수(θ)에 관한 추정은 결측치 발생과 관련된 모형을 별도로 고려함 없이 이루어질 수 있는 것이다.[2] 이러한 조건들이 충족되지 않은 결측치 발생기제가 NI(Non-Ignorable) 혹은 MNAR(Missing Not At Random)이다. MAR 가정과 달리 결측치 발생기제가 무시할 수 없는(Non-Ignorable) 경우 전체 자료(Y)와 결측치 발생(R)의 결합분포를 모형화하는 것이 필요하다.

2. 불완전 자료의 보정방법

1) 완전제거법

완전제거법(Listwise Deletion)은 결측치의 처리와 관련하여 가장 빈번하게 사용되어 온 방법 중의 하나인데, 이 방법은 분석모형에 사용된 변수 중 결측치를 하나라도 가진 모든 사례를 분석에서 제거하는 방법이다. 완전제거법은 결측치의 발생이 MCAR의 조건을 충족한다는 가정에 기초하고 있다. 비록 개별 변수의 결측치 비율이 낮더라도 모형에 사용된 변수의 수가 많을 경우, 혹은 장기간에 걸친 패널자료의 경우, 완전제거법의 사용은 분석에서 사용될 수 있는 유효사례의 수를 유의미하게 감소시킬 수 있다. 일반적으로 완전제거법을 사용한 분석은 보수적인 분석결과를 제시한다고 볼 수 있는데, 이는 유효표본 사례 수의 감소가 표준오차의 증가로 이어지기 때문이다. 그러나 완전제거법은 동시에 표준오차의 증가로 인해 제2종 오류(Typle II)의 개연성을 증가시킨다(Acock 2005).[3]

---

2) 방정식 (1)에서 불완전 자료 발생기제를 제외한 $f(Y_{obs}|\theta)$(혹은 $L(\theta|Y_{obs})$)가 관측자료 우도(observed-data likelihood) 혹은 결측치 발생기제를 무시한 우도(likelihood ignoring missing-data mechanism)로 불린다(Schafer 1997).

3) 결측치를 지닌 모든 사례를 분석에서 제외하는 대신 활용 가능한 모든 정보를 사용하는 부분제거법(Pairwise Deletion)이 사용되기도 한다. 부분제거법은 모든 변수들을 동시에 고려하는 대신 두 변수 간의 관계에서 완전한 정보를 제공하는 사례들로부터 공분산을 계산한다. 부분제거법의 경우 공분산이 각각 다른 표본 수에 기초해

2) 평균/회귀대체법

평균대체법(Mean Imputation)은 결측치를 갖는 변수의 값을 해당 변수의 관측된 값들의 평균으로 대체하는 방법이다. 비록 분석에서 활용될 유효사례의 수를 증가시키는 장점이 있지만, 결측치를 갖는 변수와 다른 변수들 사이에서 관측되는 연관성을 고려하지 않는 단점이 있다. 일반적으로 평균대체법은 결측값을 갖는 변수의 분산을 과소평가하는 경향이 있는데, 이는 상관이나 회귀계수의 과소평가로 이어질 개연성이 크다(Kline 1998; Schafer & Graham 2002). 회귀대체(Regression Imputation)는 결측치를 지닌 변수와 관련된 다른 변수들을 사용하여 결측치를 지닌 변수의 값을 예측하는 방식을 취한다. 결측치를 지닌 변수와 다른 변수들 사이에서 관측되는 연관성을 고려한다는 측면에서 평균대체법에 비해서는 진일보한 방법으로 평가된다. 그러나 회귀대체법은 변수들 사이의 연관성을 과대평가하기에 혹은 보정과정에서 내재적으로 존재하는 불확실성을 고려하지 않기에 상관이나 공분산을 사용하는 분석의 경우에 권장되지 않는다(Graham & Hofer 2000).

3) EM 알고리듬

EM(Expectation-Maximization) 알고리듬은 불완전 자료 문제를 해결하는데 널리 활용되는 방법인데, EM 알고리듬의 적용에 있어서 결측된 정보는 결측자료, 알려지지 않은 모수, 혹은 양자 모두일 수 있다. 기본적으로 EM 알고리듬은 결측자료(Y_{mis})와 모수(θ) 사이의 상호의존성을 활용하는데, 이는 θ는 Y_{mis}를 그리고 Y_{mis}는 θ를 추정하는데 유용한 정보를 제공한다는 점과 관련된다. EM 알고리듬은 두 단계로 구성되어 있는데, 첫째, E 단계에서는 θ의 초기 추정치를 통해 Y_{mis}를 보정하며, 둘째, M 단계에서는 관측된 값(Y_{obs})과 E 단계에서 보정된 값을 모두 사용하여 새로운 θ를 추정한다. EM 알고리듬은 모수 추정치에서의 변화가 충분히 작을 때까지 이러한 E 단계와 M 단계를 반복한다(Dempster, Laird & Rubin 1977; Little & Rubin

---

계산되기에 비양정치(nonpositive definite) 행렬의 발생이나 다중공선성(multicollinearity)의 문제를 야기하는 경향이 있다(Kline 1998).

2002).

일반적으로 EM 알고리듬을 활용한 불완전 자료의 보정은 결측치를 보정 값으로 대체하여 완전한 자료를 생성하기보다는 EM 알고리듬을 통해 생성된 평균, 표준편차, 상관 등을 추가적 모형분석에 활용함으로써 이루어진다. 비록 EM 알고리듬에 기초한 모수 추정치가 불편 추정치이고 효율적이지만 EM 알고리듬을 통해 생성된 공분산 등을 사후적으로 분석모형에 사용하는 방법은 편향된 표준오차를 생성할 개연성이 있다. 일반적으로 EM 알고리듬은 분석모형에서의 표준오차를 과소평가하며 추정의 정확성을 과대평가하는 경향이 있음이 지적된다(Allison 2002).

4) 최대우도법

EM 알고리듬을 통한 불완전 자료의 처리는 불완전 자료의 보정모형과 추후에 이루어지는 분석모형이 분리됨으로 인해 표준오차가 과소평가되는 문제가 있음이 지적된다.[4] EM 방법이 갖는 이러한 문제는 불완전 자료에 관한 모형과 연구자의 관심대상인 분석 모형을 동시에 추정하는 모형에 기반을 둔 방법(model-based method)을 사용함으로써 보다 효율적으로 처리될 수 있다. FIML(Full Information Maximum Likelihood)로 불리는 이 방법은 EM 알고리듬과 달리 결측치를 보정 값으로 대체하는 대신 MAR 가정하에 사례별 관측자료의 우도(observed-data likelihood)를 극대화함으로써 직접적으로 분석모형의 모수 추정치를 구한다(Allison 2002; Schafer & Graham 2002).

현재 FIML 방법은 구조방정식 모형의 맥락하에서 사용되어지고 있는데, FIML은 일반적으로 결측치를 지닌 모든 변수들의 다변량 정규분포를 가정한다. FIML 접근법은 결측치 발생 유형의 수에 제한을 받지 않으며 결측치를 모형화하기 위한 복잡한 절차를 필요로 하지 않는 장점이 있다. 특히, FIML 방법을 통한 불완전 자료의 처리는 연구자에 의해서 설정된 분석모형이 타당

4) 연구자에 따라서는 EM과 최대우도법(FIML)을 동일한 의미로 사용하기도 한다. 비록 EM 알고리듬이 FIML 절차에서 활용되기도 하지만, 본 연구의 경우 EM 알고리듬은 불완전 자료 보정 모형과 연구대상 분석 모형을 분리한다는 측면에서 FIML과는 구분되는 것으로 파악한다.

할 경우에 권장된다(Graham & Hofer 2000).

5) 다중대체법

FIML의 경우처럼 다중대체법(multiple imputation)도 MCAR 대신 덜 엄격한 MAR 가정에 기초하고 있다. 다중대체법은 단순대체법의 장점을 보유하는 동시에 보정과정에서 나타나는 확률적 변이를 명시적으로 고려하는 특성을 보인다. 다중대체법의 경우 보정과정에서의 불확실성은 복수(multiple)의 자료 세트를 생성함으로써 반영되는데, 생성된 복수의 자료 세트 사이의 변이가 보정과정의 불확실성을 나타낸다(Graham & Hofer 2000; Schafer 1997). 이러한 결측치가 없는 복수의 완전한 자료 세트들이 생성된 후 표준적인 통계분석이 생성된 보정자료 세트들의 각각에 대해서 이루어진다.

다중대체법을 통해 m 세트의 보정된 자료를 생성할 경우 최종 모수 추정치는 m개의 모수 추정치들의 평균이다. 표준오차는 두 가지의 구성요소를 포함하고 있는데, 각각의 보정세트내의 분산과 m 세트의 보정자료 사이에 존재하는 분산이 그것이다(Rubin 1987). 비록 FIML 방법의 경우에도 분석모형에서 보조변수(auxiliary variable)를 고려할 수 있지만, 다중대체법의 경우 연구자가 보정모형과 분석모형을 다르게 설정할 수 있다는 점에서 더욱 큰 유연성을 가지고 있다고 할 수 있다(Schafer 2003).

6) 선택편의 모형/패턴혼합 모형

결측치의 발생기제가 MNAR일 경우 현재 두 가지 방법이 결측치를 모형화하기 위해서 사용되어질 수 있는데, 선택편의 모형(selection bias model)과 패턴혼합 모형(pattern-mixture model)이 그것이다. 선택편의 모형과 패턴혼합 모형 모두 자료(Y)와 결측치 발생 분포(R)를 동시에 모형화($f(Y, R)$)한다. 선택편의 모형은 우선 결측치를 포함한 전체 자료의 주변분포($f(Y)$)를 설정한 다음 이러한 자료를 조건부로 한 결측치 발생의 분포($f(R|Y)$)를 명시화한다. 비록 이 방법이 연구자들, 특히 계량경제학자들, 사이에서 많이 활용되지만 전체 자료(Y)의 분포에 대한 가정에 매우 민감하다는 점이 문제로 지적되고 있다(보다 깊이 있는 논의는 Kenward(1998)와 Little & Rubin(2002)

을 참조).

선택편의 모형에 비해 패턴혼합 모형은 결측치 발생 유형의 주변분포($f(R)$)와 결측치 발생 유형을 조건부로 한 전체 자료의 분포($f(Y|R)$)를 명시화한다. 패턴혼합 모형의 경우 관측된 자료로부터 직접 추정되어질 수 없는 조건부 분포($f(Y|R)$)와 관련된 모수에 대한 가정이 필요하다. 그러므로 무시할 수 없는 결측치와 관련하여 패턴혼합 모형을 사용하는 경우 가정 위배에 대한 민감도 분석이 이루어질 필요가 있다(Allison 2002). 또한 연구자들은 일반적으로 각각의 결측치 발생 유형별 분포($f(Y|R)$)보다는 전체 자료의 주변분포($f(Y)$)와 관련된 모수 추정이 관심의 초점이다. 결과적으로 일반적인 패턴혼합 모형의 경우 연구자가 원하는 결과를 얻기 위해서는 결측치 발생 유형별 결과를 평균하는 추가적인 작업이 필요하다(Hedeker & Gibbons 1997; Little 1995 참조).

Ⅲ. 분석방법

1. 자료 및 분석모형

1) 자료

본 연구에서 사용되는 자료는 미국의 HRS(Health and Retirement Study)이다. HRS는 미국 중고령층의 건강, 경제상태, 가족관계, 근로활동 및 은퇴와 관련된 대표적인 패널조사이다. 현재 HRS로 지칭되는 패널조사는 상호 관련되었지만 개별적으로 진행되어 온 두 조사에서 비롯되었는데, 우선, 1992년에 시작된 초기의(원래의) HRS는 근로활동기에서 은퇴기로 전이하는 개인들(51-61세)을 추적하기 위해서 설계된 조사였다. 이에 비해 AHEAD(Asset and Health Dynamics among the Oldest Old)는 70세 이상 개인들이 은퇴 이후에 경험하는 건강, 가족, 경제상태 등과 관련된 사항들을 조사하기 위해서 설계된 조사였다. 원래의 HRS는 1992년, 1994년, 1996년에 그리고 AHEAD는 1993년과 1995년에 각각 조사가 이루어졌다. 상호 밀접하게 관련된 이 두 조사는 1998년에 HRS라는 명칭으로 통합되어 현재까지 격년으로 조사가 이루

어지고 있다. 또한 이 두 조사가 통합된 1998년의 경우 새로운 표본이 추가됨으로써 HRS는 미국에서 50세 이상의 개인들을 대표하는 패널조사로서 자리를 잡게 되었다(HRS에 관한 보다 자세한 사항은 http://hrsonline.isr.umich.edu/ 참조).

본 연구의 분석대상은 HRS가 최초 시작된 1992년(1회차)부터 2004년(7회차)까지 12년에 걸쳐 분석모형에 포함된 모든 변수들에 있어서 「완전한」 정보를 제공하고 있는 1940년에 출생한 코호트 성원 540명이다. 본 연구는 이러한 결측치가 존재하지 않는 완전한 자료로부터 각각의 불완전 자료 발생기제 시나리오(MCAR, MAR, MNAR)에 기초한 가상의 자료를 생성한다. 본 연구에서 결측치가 없는 완전한 자료는 여러 보정방법들의 수행결과를 비교하는데 있어서 준거점으로서의 역할을 한다.[5]

본 연구에서는 예측변수들로 성, 결혼상태, 인종/민족, 교육수준, 가구소득을 사용한다. 결혼상태와 가구소득의 경우 기간의 경과에 따라 변화하는(time-varying) 변수들이지만, 본 연구에서는 분석의 간명성 차원에서 HRS가 최초 시작된 1992년의 결혼상태와 가구소득을 사용한다. 성(여성=1), 결혼상태(유배우=1), 인종/민족(백인=1) 변수는 모두 더미변수들이며, 교육수준과 가구소득은 연속변수로 처리된다.

본 연구의 결과변수는 우울증(depression)인데, 우울증의 증상을 측정하기 위해 CES-D(Center for Epidemiological Studies Depression Scale)가 사용되었다. HRS가 시작된 1992년 이후 우울증과 관련 다음의 8개 항목이 반복적으로 조사되었다(예=1; 아니오=0). (1) you felt depressed, (2) you felt that everything you did was an effort, (3) your sleep was restless, (4) you were happy, (5) you felt lonely, (6) you enjoyed life, (7) you felt sad, (8) you could not get going. 본 연구에서는 위의 8개 항목을 합산하여 7회에 걸친 우울증상 반복측정치를 도출하였다(범위: 0 - 8).[6]

---

5) 완전한 자료에 대비한 여러 보정방법들의 수행결과를 비교하고자 하는 이러한 접근은 본 연구에서 사용되는 분석모형의 복잡성과 이로 인한 모의실험의 어려움과도 관련된다. 결과적으로 본 연구에서 도출된 결과는 결측치를 포함하고 있는 원래의 HRS 1940년생 출생 코호트 전체를 대상으로 한 분석결과와는 구분되어야 한다.

2) 분석모형

결측치 처리와 관련하여 본 연구는 반복측정 변수를 활용한 분석에 초점을 맞춘다. 구체적으로 본 연구는 기간의 경과에 따른 우울증의 궤적을 살펴보기 위해 잠재성장곡선 모형(Latent Growth Curve Model)을 사용한다. 잠재성장곡선 모형은 다층구조의 차원에서 개인내 모형(within-respondents model)과 개인간 모형(between-respondents model)으로 다음과 같이 표시될 수 있다.

$$Y_{ij} = \pi_{0i} + \pi_{1i}(AGE_{ij} - 52) + \epsilon_{ij}, \tag{2}$$

$$\pi_{0i} = \gamma_{00} + \sum_{1}^{k}\gamma_{0k}X_{ki} + \zeta_{0i}$$

$$\pi_{1i} = \gamma_{10} + \sum_{1}^{k}\gamma_{1k}X_{ki} + \zeta_{1i}. \tag{3}$$

개인내 모형 (2)에서 표시된 것처럼 본 연구에서 사용되는 시간의 경과를 표시하는 기본적인 축은 연령(AGE)이다.[7] 이는 우울증과 같은 건강의 경우 (역법상의 시간보다는) 연령이 근본적인 역할을 수행한다는 점과 관련된다. Y_{ij}는 개인 i의 연령 j때의 우울증상의 값을, γ_{00}는 52세 때의 준거집단($X_{ki} = 0$)의 우울증상의 값을 표시하는 절편(intercept)을, γ_{10}는 준거집단($X_{ki} = 0$)에 대한 연령의 선형효과를 표시하는 기울기(slope)를, γ_{0k}는 예측변수들(X_{ki})에 기인한 절편에서의 차이를, γ_{1k}는 예측변수들(X_{ki})에 기인한 기울기에서의 차이를 각각 나타낸다.[8] 모형 (2)에서 표시되었듯이 본 연구에서는 12

---

6) 네 번째와 여섯 번째 항목은 역순으로 리코딩되었다. HRS의 1992년 우울증 측정은 4개의 범주로 구성되어 있기에 이분변수로 재구성되었다. 보정방법을 소개하는 본 연구에서는 측정방법 변화 및 신뢰도 등과 관련된 논의는 생략하기로 한다.

7) 1940년 출생 코호트는 만 연령으로 환산할 경우 1992년에 52세이며 2004년에 64에 도달한다. 1992년의 연령을 표시하기 위해 모형 (2)에서 52세로의 중심이동(centering)이 이루어졌다.

8) 잠재성장곡선 모형에 대한 보다 자세한 논의는 Singer & Willett(2003)을 참조하기 바란다.

년(52세~64세)에 걸친 우울증상에 있어서의 변화를 선형으로 표상하고 있다.[9]

2. 불완전 자료의 유형 및 보정방법

1) 불완전 자료의 유형

본 연구에서 설정된 분석모형처럼 종단자료를 활용한 분석의 경우 분석모형의 복잡성으로 인해 체계적인 모의실험을 수행하는 것은 매우 어려운 작업이다. 비록 불완전 자료의 처리와 관련된 분석결과의 일반화를 위해서는 체계적인 모의실험이 바람직하지만, 앞에서 언급했듯이 본 연구에서는 불완전 자료 발생의 영향 및 처리방법들의 효과를 비교해 보기 위해 결측치가 없는 완전한 자료로부터 MCAR, MAR, MNAR의 상황에 해당하는 가상의 자료 세트를 생성한다. 본 연구에서 고려되는 불완전 자료는 반복측정된 결과변수(우울증)에서의 결측치이다.[10] 또한 횡단자료에 비해 종단자료의 경우 현실적으로 관측되는 불완전 자료의 발생 유형도 매우 다양하다. 본 연구에서는 패널조사에서 발생할 개연성이 상대적으로 높은 불완전 자료 발생 상황에 초점을 맞추어 MCAR, MAR, MNAR 조건을 각각 충족하는 자료 세트를 생성한다.

<표 1>은 각각의 불완전 자료 발생기제하에서 생성된 자료의 결측치 발생 유형을 나타내고 있다. 첫째, MCAR 상황과 관련 우울증상에서의 결측치는 매회차별로 무작위적으로 발생되었다. 패널조사의 경우 회차가 증가함에 따라 패널탈락이 증가하는 것이 일반적인 패턴인데, 이러한 경향을 반영하여 결측치 발생 비율은 1992년(52세) 0%, 1994년 5%, 1996년 10%, 1998년 15%, 2000년 20%, 2002년 25%, 2004년(64세) 30%로 증가시켰다. 이러한 조건하에서 7회에 걸쳐 우울증상에 관해 완전한 정보를 제공하는 개인은 총 150명이며, 완전제거법을 사용할 경우 이는 대상표본 수에 있어서 72% 감소에 해당한다.

---

9) 연령과 우울증과의 관계에 대한 본 자료의 탐색적 분석결과는 선형관계가 적절함을 보여준다(연령과 우울증 사이의 관계에 대한 보다 깊이 있는 논의는 Mirowsky & Ross(1992)참조).

10) 다중대체법의 경우 예측변수에서의 결측치 문제는 결과변수와 동일하게 다루어질 수 있다. 최대우도법(FIML)의 경우 예측변수에서 발생한 결측치 문제는 결측치를 지닌 변수의 분포에 대한 가정을 도입함으로써 해결될 수 있다.

〈표 1〉은 또한 MCAR 조건하에서 총 29개의 결측치 발생 유형이 관측됨을 나타내고 있다.

둘째, MAR 조건하의 결측치 발생은 패널 회차의 증가에 따라 결측 비율이 증가하는 동시에 한 번 탈락한 패널성원은 복귀하지 않는 단조형(monotone) 결측치 발생패턴을 상정한다. 구체적으로 1994년(2회차)부터 「이전」 조사에서 3개 이상의 우울증상을 보고한 개인들의 40%가 이후의 조사에서 결측치를 갖는 것으로 상정된다. 예컨대, 1992년(52세)에 3개 이상의 우울증상을 보고한 개인들의 40%가 1994년-2004년의 우울증상에서 결측치를 가진다. 마찬가지로 1994년(54세)에 결측치를 갖지 않는 나머지 패널성원 중에서 3개 이상의 우울증상을 보고한 개인들의 40%가 1996년-2004년의 우울증 조사에서 결측치를 갖는다. 이러한 결측치 발생 시나리오는 패널조사에서 흔히 발생하는 유형이라고 볼 수 있는데, 어떤 시점에서 낮은 점수(예를 들면, 나쁜 건강상태)를 보고한 후 (사망 등의 이유로) 이후의 조사에 참여하지 않는 경우가 여기에 해당한다. 〈표 1〉은 이러한 결측치 발생 시나리오에서 표본의 감소가 대략 30% 수준에 이르고 있음을 보여준다.

마지막으로 MNAR 상황에서는 1994년(2회차)부터 각각의 조사시점에서 3개 이상의 우울증상을 보인 개인들의 40%에 대해서 결측치를 발생시키며, MAR과 동일하게 한 번 탈락한 패널성원은 이후 조사에서 복귀하지 않는 것으로 상정된다. 위에서 언급된 MAR의 경우 현 시점에서의 결측치 발생은 이전 조사에서 「관측된」 우울증상과 연관된다. 이에 비해 MNAR 상황의 경우 현 시점에서의 결측치 발생은 현 시점의 우울증상과 연관된다. 예컨대, 1994년(54세)에 3개 이상의 우울증상을 보고한 개인들의 40%에 대해서 1994년(54세)의 우울증상 값을 삭제함으로써 결측치를 발생시킨다.

2) 불완전 자료의 처리방법

불완전 자료의 처리와 관련하여 완전제거법, 평균대체법, EM 알고리듬, 최대우도법(FIML), 다중대체법(Multiple Imputation)이 사용된다. EM 알고리듬의 사용과 관련하여 본 연구는 결측치에 대한 직접적인 보정 대신 EM 알

〈표 1〉 결측치 발생기제별 결측치 발생 유형(HRS 2회차-7회차)

MCAR			MAR			MNAR		
유형	변수	사례	유형	변수	사례	유형	변수	사례
111111	0	150	111111	0	380	111111	0	375
111110	1	82	100000	5	36	000000	6	40
111101	1	65	110000	4	32	100000	5	35
111011	1	52	111000	3	28	111000	3	29
110111	1	32	111100	2	25	110000	4	27
101111	1	21	111110	1	20	111100	2	20
111100	2	20	000000	6	19	111110	1	14
110110	2	13						
011111	1	8						
111010	2	8						
111001	2	8						
111000	3	8						
110101	2	7						
101101	2	7						
011110	2	7						
110000	4	7						
110011	2	6						
101110	2	6						
100111	2	6						
101010	3	6						
011101	2	4						
011011	2	4						
100000	5	4						
101011	2	2						
110001	3	2						
010101	3	2						
100011	3	1						
011100	3	1						
000111	3	1						

주: N=540; 결측치 발생 유형(1=관측; 0=결측); 변수는 결측치가 발생한 변수의 수를 의미; HRS 1회차(1992년)의 경우 결측치 없이 모두 관측(1)임.

고리듬을 통해 생성된 평균 및 공분산 행렬을 이용하여 사후적으로 분석에 활용하는 방식을 사용한다. 최대우도법(FIML)을 통한 불완전 자료의 처리는 현재 구조방정식 모형을 추정하기 위해 특수화된 프로그램들에 의해서 제공되고 있는데, 본 연구에서는 이들 프로그램들을 통하여 잠재성장곡선 모형을 추정하는 방식을 취한다.[11]

다중대체법의 적용도 여러 방법들에 의해서 이루어질 수 있지만 본 연구에서 사용하는 방법은 MICE(Multivariate Imputation by Chained Equations)이다(Royston 2004, 2005; Van Buuren, Boshuizen & Knook 1999; Van Buuren & Oudshoorn 2000)이다. MICE는 결측치를 가진 변수 전체에 대한 하나의 모형(예를 들면, 다변량 정규분포 모형)을 구성하는 대신 결측치를 보유한 개별 변수의 속성에 따라 여러 유형의 회귀모형을 추정한다(예를 들면, 일반회귀 모형, 로짓 모형, 다항로짓 모형, 서열로짓 모형 등). 비록 본 연구가 연속형 변수에서 발생한 결측치에 초점을 맞추지만 MICE는 다양한 유형의 변수에서 발생한 결측치가 존재하는 경우에 더욱 적합성을 갖는다고 평가할 수 있다. 다중대체법과 관련해서 본 연구에서는 총 10 세트의 보정자료를 생성한다.[12]

무시할 수 없는(Non-Ignorable) 결측치의 처리에서는 패턴혼합 모형을 예시적인 차원에서 추가적으로 사용한다. <표 1>의 경우 MNAR 상황에서 모두 7개의 결측치 발생 유형을 보여주고 있지만, 패턴혼합 모형의 경우 관측된 모든 유형을 고려할 필요는 없다. 본 연구에서는 분석의 간명화와 사례 수를 고려하여 우울증상의 측정과 관련 7회에 걸쳐 완전한 정보를 제공하는 유형(n=375)과 최소한 1개의 결측치를 제공하는 유형(n=165)을 구분한다. 본 연구에서는 또한 완전한 정보를 제공하는 집단과 불완전 정보를 제공하는 집단

11) 본 연구에서 EM 알고리듬을 통한 공분산 행렬의 생성은 SAS를 통해 이루어졌다. 만일 EM 알고리듬을 통해 결측치를 직접 보정하고자 할 경우 EMCOV 프로그램이 활용될 수 있다. 최대우도법(FIML)의 적용과 관련 현재 AMOS, EQS, LISREL, Mplus, Mx 등 구조방정식 모형을 추정하기 위해 특수화된 프로그램들이 FIML 절차를 제공하고 있다. 비록 잠재성장곡선 모형의 추정이 다양한 방식으로 이루어질 수 있지만, 최대우도법을 활용하고자 할 경우 구조방정식 모형을 추정하기 위해 특수화된 프로그램들을 사용할 필요가 있다.

12) 현재 R, S-Plus, Stata 등의 프로그램들이 MICE 보정 절차를 제공하고 있는데, 본 연구에서는 Stata 모듈을 사용하여 보정 작업을 수행하였다. 또한 결측치가 보정된 다중의 자료 세트들을 사용하여 잠재성장곡선 모형의 추정치를 얻기 위해 Mplus를 사용하였다. MICE 외에 널리 활용되고 있는 다중대체법으로는 DA(Data Augmentation) 알고리듬이 있는데, SAS의 PROC MI나 Schafer의 NORM이 이에 기초하고 있다.

은 우울증의 궤적에 있어서 절편(intercept)과 기울기(slope)에서만 차이를 보인다는 것을 가정한다. 구체적으로 본 연구에서는 관측기간인 12년 동안 완전한 정보를 제공하는 집단에 비교할 때 불완전 정보를 제공하는 집단은 우울증의 초기값(절편)이 높은 동시에 우울증의 변화(기울기)에 있어서도 빠른 증가를 보인다는 것을 가정한다.[13]

Ⅳ. 분석결과

1. MCAR 상황

〈표 2〉는 MCAR 상황에서 도출된 비조건부(예측변수가 포함되지 않은) 잠재성장곡선 모형의 추정결과를 보여주고 있다. 불완전 자료 처리방법들의 효과를 살펴보기 위해 〈표 2〉는 또한 준거점으로서 결측치가 없는 완전한 자료를 사용한 모형의 절편과 기울기의 추정치 및 표준오차를 보여주고 있다. 결측치가 없는 완전한 자료를 사용한 추정결과는 52세 때의 우울증상(절편)이 .985임을 보여주고 있다. 또한 기울기 추정치는 관측기간 12년에 걸쳐 우울증상이 연 .047만큼 증가하고 있음을 보여주고 있다. 〈표 2〉는 또한 절편(1.293)과 기울기(.007)의 분산 추정치 모두 통계적으로 유의미함으로써($\alpha = .05$), 최초 관측시점(52세)에서의 우울증상(절편) 및 연령의 증가에 따른 우울증상의 변화(기울기)에 있어서 개인별 유의미한 차이가 존재함을 보여주고 있다.

〈표 2〉는 또한 결측치의 처리와 관련된 여러 방법들의 적용결과를 보여주고 있다. 결측값이 없는 완전한 자료에서 얻어진 추정치와 비교할 때 불완전 자료 처리방법들은 전반적으로 그 적용결과에 있어서 큰 차이를 보이지 않는

13) 모형에서 사용된 절편과 기울기의 값은 불완전 정보를 제공하는 집단이 완전한 정보를 제공하는 집단에 비해 최초 관측시점의 우울증상이 1.371만큼 높으며, 우울증의 변화와 관련해서도 연 .097만큼 빨리 증가하는 것으로 설정된다. 이러한 절편과 기울기의 값은 잠재성장곡선 모형에서 두 집단 간의 차이를 비교함으로써 도출되었다. 패턴혼합 모형의 경우 설정된 가정에 대한 민감도 분석이 이루어질 필요가 있지만 패턴혼합 모형을 예시적으로 사용하고 있는 본 연구에서는 이를 생략하기로 한다.

〈표 2〉 비조건부 잠재성장곡선 모형 추정치와 표준오차: MCAR

	절편		기울기	
	평균	분산	평균	분산
결측치 없음	.985	1.293	.047	.007
	(.060)	(.126)	(.006)	(.001)
완전제거법	.840	.765	.058	.008
	(.091)	(.162)	(.011)	(.002)
평균대체법	.983	1.213	.047	.004
	(.059)	(.121)	(.005)	(.001)
EM 알고리듬	1.002	1.199	.043	.006
	(.059)	(.119)	(.006)	(.001)
FIML	.969	1.214	.049	.006
	(.069)	(.216)	(.008)	(.002)
MICE	1.004	1.232	.043	.006
	(.060)	(.124)	(.007)	(.002)

다. 그러나 다른 방법들과 비교할 때 완전제거법의 경우 결측치가 없는 완전한 자료를 사용한 경우에 비해 우울증 궤적의 절편과 기울기에서 상대적으로 큰 차이를 보인다. 완전제거법은 또한 이용 가능한 사례 수의 감소로 인해 다른 보정방법들에 비해 표준오차가 상대적으로 높은 패턴을 보인다.[14]

〈표 2〉에서 제시된 결과는 연령의 증가에 따른 우울증 궤적의 초기값과 변화에 있어서 개인간 유의미한 차이가 있음을 보였다. 〈표 3〉은 이러한 관측된 이질성을 설명하기 위해 사회인구학적 변수들을 모형에 포함한 조건부 잠재성장곡선 모형의 추정치와 표준오차를 보여주고 있다. 결측값이 없는 완전한 자료는, 첫째, 우울증 궤적의 초기값(절편)과 관련하여 성을 제외한 다른 모든 예측변수들이 유의미하게 부적(−)으로 연관되어 있음을 보여준다(α = .01). 결측치가 없는 완전한 자료는 또한 성과 교육수준이 연령의 증가에 따른 우울

14) 본 연구가 체계적인 모의실험의 결과가 아니기에 완전제거법과 관련된 위의 분석결과가 특수한 표본에 기인한 결과 혹은 완전히 우연에 기인한 결과일 수 있다는 점에서 분석결과를 일반화하는데 한계가 존재한다. 또한 본 연구의 경우 MCAR 조건에서 설정된 유효사례의 감소(72%)가 매우 크다는 점도 고려될 필요가 있다. 모의실험에 기초한 기존연구(Wothke 2000)는 최대우도법과 같은 방법에 비해 완전제거법의 경우 추정치들의 분산이 상대적으로 매우 크다는 점을 보여주고 있다.

증의 변화(기울기)를 설명하는 유의미한 변수들임을 보여주는데, 여성(.029)이 12년간의 관측기간 동안 우울증에서 상대적으로 빠른 증가를 보였다(α = .05). 이와 반대로 교육의 효과와 관련하여 모수 추정치(−.005)는 교육수준이 높을수록 우울증이 상대적으로 완만하게 증가하고 있음을 보여주고 있다(α = .05).

불완전 자료의 처리방법들의 적용결과와 관련하여 〈표 3〉은 비조건부 잠재성장곡선 모형의 분석결과처럼 모수 추정치에 있어서 완전제거법에 기초한 결과와 다른 방법들을 적용한 결과 사이에 상대적으로 큰 차이가 있음을 보여 준다. 평균대체법의 경우 우울증 초기값(절편)의 경우 대체로 EM, FIML, MICE 등과 유사한 결과를 보이지만 우울증상의 변화(기울기)와 관련해서는 성이나 교육수준의 효과 모두 통계적으로 유의미하지 않은 것으로 나타나고 있다. 전반적으로 〈표 3〉은 EM, FIML, MICE 방법이 완전제거법이나 평균대체법에 비해 결측치가 없는 완전한 자료에서 추정된 값과 유사한 결과를 보이며, 특히 EM과 MICE의 적용결과가 상당히 유사한 것으로 나타나고 있다.

2. MAR 상황

〈표 4〉는 MAR의 상황에서 불완전 자료 처리방법들의 적용결과를 나타내고 있다. 우울증 궤적의 절편의 경우 완전제거법을 제외한 나머지 방법들은 대체로 유사한 결과를 보이고 있다. MCAR 상황에 비해 유효사례의 수가 많이 증가했음에도 불구하고 완전제거법은 결측치가 없는 완전한 자료로부터 도출된 추정치들과는 상당히 다른 결과를 보여주고 있다. 결측치가 없는 완전한 자료에서 얻은 결과와 완전제거법을 적용한 결과 사이의 차이는 MCAR 상황에 비해 MAR 상황에서 더욱 현저하게 나타난다. 〈표 4〉는 또한 MCAR 상황에 비해 MAR 상황에서 우울증 궤적의 기울기 추정에 있어서 불완전 자료 처리방법들의 적용효과가 상대적으로 떨어지고 있음을 보여주고 있다.

조건부 잠재성장곡선 모형 추정결과를 보여주는 〈표 5〉의 경우 완전제거법과 평균대체법은 다른 처리방법들과 상당히 다른 추정결과를 보여주고 있다. 특히, 이러한 차이는 기울기에 대한 예측변수들의 효과에서 더욱 크게 나타난

〈표 3〉 조건부 잠재성장곡선 모형 추정치와 표준오차: MCAR

		상수항	성	결혼 상태	인종/ 민족	교육 수준	가구 소득
결측치 없음	절편	3.151	.096	-.355	-.482	-.113	-.003
		(.293)	(.116)	(.134)	(.130)	(.021)	(.001)
	기울기	.070	.029	.015	.007	-.005	.000
		(.032)	(.013)	(.015)	(.014)	(.002)	(.000)
완전제거법	절편	2.671	.167	-.378	-.280	-.104	-.002
		(.420)	(.168)	(.209)	(.198)	(.031)	(.002)
	기울기	.111	.036	-.031	-.009	-.004	.000
		(.054)	(.022)	(.027)	(.026)	(.004)	(.000)
평균대체법	절편	3.146	.109	-.373	-.474	-.112	-.003
		(.286)	(.113)	(.131)	(.127)	(.020)	(.001)
	기울기	.013	.019	.020	.013	-.001	.000
		(.029)	(.111)	(.013)	(.013)	(.002)	(.000)
EM 알고리듬	절편	3.202	.124	-.374	-.511	-.114	-.003
		(.288)	(.114)	(.132)	(.128)	(.020)	(.001)
	기울기	.060	.026	.014	.006	-.004	.000
		(.031)	(.012)	(.014)	(.014)	(.002)	(.000)
FIML	절편	3.141	.099	-.367	-.490	-.112	-.003
		(.400)	(.111)	(.163)	(.147)	(.024)	(.001)
	기울기	.067	.031	.015	.002	-.004	.000
		(.039)	(.013)	(.016)	(.016)	(.003)	(.000)
MICE	절편	3.210	.122	-.370	-.512	-.114	-.003
		(.292)	(.115)	(.133)	(.129)	(.021)	(.001)
	기울기	.056	.027	.012	.005	-.004	.000
		(.033)	(.013)	(.016)	(.015)	(.002)	(.000)

〈표 4〉 비조건부 잠재성장곡선 모형 추정치와 표준오차: MAR

	절편		기울기	
	평균	분산	평균	분산
결측치 없음	.985	1.293	.047	.007
	(.060)	(.126)	(.006)	(.001)
완전제거법	.513	.284	.028	.003
	(.041)	(.053)	(.005)	(.001)
평균대체법	1.014	1.125	-.009	.008
	(.058)	(.118)	(.006)	(.001)
EM 알고리듬	1.027	1.394	.036	.004
	(.061)	(.126)	(.005)	(.001)
FIML	.954	1.183	.053	.005
	(.069)	(.228)	(.011)	(.002)
MICE	1.022	1.431	.038	.004
	(.064)	(.148)	(.007)	(.002)

다. 예를 들면, 평균대체법의 경우 성의 효과는 유의미하지 못한 반면, 교육의 기울기에 대한 효과는 오히려 반대방향으로 나타나고 있다. 〈표 3〉의 MCAR 상황에 비교할 때 〈표 5〉의 MAR 상황에서 관측되는 또 다른 차이는 EM, FIML, MICE의 수행결과(특히, 기울기에 대한 효과)가 상대적으로 떨어진다는 점이다. 예컨대, 〈표 5〉의 경우 어떤 불완전 자료 처리방법에서도 교육의 기울기에 대한 효과는 유의미하게 나타나지 않고 있다($\alpha = .05$).

〈표 5〉 조건부 잠재성장곡선 모형 추정치와 표준오차: MAR

		상수항	성	결혼 상태	인종/ 민족	교육 수준	가구 소득
결측치 없음	절편	3.151	.096	-.355	-.482	-.113	-.003
		(.293)	(.116)	(.134)	(.130)	(.021)	(.001)
	기울기	.070	.029	.015	.007	-.005	.000
		(.032)	(.013)	(.015)	(.014)	(.002)	(.000)
완전제거법	절편	1.363	.031	-.061	-.246	-.041	-.002
		(.233)	(.080)	(.098)	(.096)	(.016)	(.001)
	기울기	.033	.016	-.011	.015	-.002	.000
		(.031)	(.011)	(.013)	(.013)	(.002)	(.000)
평균대체법	절편	3.027	.128	-.322	-.414	-.107	-.003
		(.283)	(.112)	(.129)	(.126)	(.020)	(.001)
	기울기	-.156	.005	.024	.033	.007	.000
		(.030)	(.012)	(.014)	(.014)	(.002)	(.000)
EM 알고리듬	절편	3.293	.095	-.356	-.454	-.121	-.003
		(.300)	(.119)	(.137)	(.133)	(.021)	(.001)
	기울기	.007	.024	-.001	.019	.000	.000
		(.027)	(.011)	(.013)	(.012)	(.002)	(.000)
FIML	절편	3.003	.075	-.318	-.416	-.108	-.003
		(.404)	(.107)	(.158)	(.140)	(.024)	(.001)
	기울기	.074	.028	-.009	.012	-.003	.000
		(.043)	(.012)	(.015)	(.016)	(.003)	(.000)
MICE	절편	3.283	.078	-.355	-.449	-.120	-.003
		(.313)	(.121)	(.138)	(.137)	(.022)	(.001)
	기울기	.006	.028	.002	.016	.000	.000
		(.040)	(.012)	(.015)	(.015)	(.002)	(.000)

3. MNAR 상황

〈표 6〉은 MNAR 상황에서 불완전 자료 처리방법들의 적용결과를 제시하고 있다. 최대우도법이나 다중대체법의 경우 MAR 가정에 기초하고 있기에 FIML이나 MICE의 적용결과가 MAR 상황에 비해 MNAR 상황에서 상대적으로 떨어지는 것은 예측될 수 있다. 특히, MCAR이나 MAR 상황과 비교할 때 불완전 자료 처리방법들은 우울증 궤적의 기울기와 관련된 추정에서 적용효과가 상대적으로 크게 떨어지며, 이에 비해 패턴혼합 모형의 경우 결측치가 없는 완전한 자료에서 얻은 추정치에 비해 절편의 모수 추정치들은 과소평가되는 반면 기울기의 모수 추정치들은 과대평가되는 모습을 보인다.

MNAR 상황에서의 조건부 잠재성장곡선 모형 추정결과를 나타내는 〈표 7〉의 경우 전반적으로 다른 불완전 자료 처리방법들에 비해 패턴혼합 모형이 결측치가 없는 완전한 자료와 상대적으로 유사한 결과를 보인다.

비록 MNAR 상황에서 EM, FIML, MICE의 적용효과가 MAR 상황에 비해서 떨어지지만, 이들 방법에 기초한 모수 추정치들은 완전제거법이나 평균대체법에 비해 상대적으로 결측치가 없는 완전한 자료에서 추정된 결과와

〈표 6〉 비조건부 잠재성장곡선 모형 추정치와 표준오차: MNAR

	절편		기울기	
	평균	분산	평균	분산
결측치 없음	.985	1.293	.047	.007
	(.060)	(.126)	(.006)	(.001)
완전제거법	.582	.319	.016	.001
	(.045)	(.062)	(.005)	(.001)
평균대체법	.934	.776	-.014	.004
	(.051)	(.091)	(.005)	(.001)
EM 알고리듬	.923	.899	.014	.001
	(.053)	(.095)	(.004)	(.001)
FIML	.895	.898	.016	.001
	(.058)	(.184)	(.008)	(.001)
MICE	.915	.900	.016	.001
	(.055)	(.133)	(.005)	(.001)
패턴혼합	.900	1.065	.079	.011
	(.057)	(.122)	(.007)	(.002)

〈표 7〉 조건부 잠재성장곡선 모형 추정치와 표준오차: MNAR

		상수항	성	결혼 상태	인종/ 민족	교육 수준	가구 소득
결측치 없음	절편	3.151	.096	-.355	-.482	-.113	-.003
		(.293)	(.116)	(.134)	(.130)	(.021)	(.001)
	기울기	.070	.029	.015	.007	-.005	.000
		(.032)	(.013)	(.015)	(.014)	(.002)	(.000)
완전제거법	절편	1.736	-.039	-.142	-.372	-.047	-.002
		(.237)	(.087)	(.106)	(.105)	(.016)	(.001)
	기울기	.038	.018	.003	.010	-.004	.000
		(.027)	(.010)	(.012)	(.012)	(.002)	(.000)
평균대체법	절편	2.729	.115	-.123	-.404	-.102	-.003
		(.248)	(.098)	(.113)	(.110)	(.017)	(.001)
	기울기	-.108	-.001	.013	.025	.004	.000
		(.026)	(.010)	(.012)	(.012)	(.002)	(.000)
EM 알고리듬	절편	2.845	.115	-.205	-.445	-.105	-.003
		(.257)	(.102)	(.118)	(.114)	(.018)	(.001)
	기울기	.009	.014	.016	.007	-.002	.000
		(.024)	(.009)	(.011)	(.010)	(.002)	(.000)
FIML	절편	2.774	.095	-.206	-.422	-.102	-.003
		(.363)	(.098)	(.136)	(.130)	(.022)	(.001)
	기울기	.018	.017	.017	.005	-.003	.000
		(.037)	(.011)	(.013)	(.014)	(.002)	(.000)
MICE	절편	2.835	.117	-.196	-.444	-.106	-.003
		(.267)	(.102)	(.122)	(.120)	(.018)	(.001)
	기울기	.019	.015	.014	.005	-.002	.000
		(.028)	(.010)	(.014)	(.015)	(.002)	(.000)
패턴혼합	절편	2.882	.110	-.273	-.443	-.107	-.003
		(.276)	(.108)	(.127)	(.122)	(.019)	(.001)
	기울기	.156	.024	.005	-.017	-.006	.000
		(.037)	(.013)	(.016)	(.016)	(.002)	(.000)

유사한 모습을 보인다. MAR 조건이 충족되지 못한 상황에서도 관측되는 이런 결과는 완전제거법이나 평균대체법과 같은 방법의 경우 비록 불완전하지만 이용가능한 모든 정보를 충분히 활용하지 못하는 점과 관련있다고 볼 수 있다.

MNAR의 경우 비록 우울증상에서의 결측치 발생이 관측되지 않은 값들에 의존하지만, 이들 우울증에서의 결측치들은 동시에 개인의 (관측된) 과거 우울증 수준이나 교육수준과 같은 사회인구학적 요인들과 밀접하게 연관되어 있음

이 지적될 필요가 있다. 결과적으로 MNAR 상황에서도 완전제거법이나 평균대체법에 비해 비록 불완전하지만 관측된 모든 자료를 활용하는 EM, FIML, MICE와 같은 방법들이 결측치가 없는 완전한 자료에서 얻은 결과와 상대적으로 유사한 결과를 산출할 것을 기대할 수 있는 것이다(이에 대한 모의실험의 예는 Collins et al.(2001)과 Schafer(1997)를 참조).

Ⅴ. 결론

본 연구의 주된 목적은 반복측정된 패널자료에서 불완전 자료의 발생이 갖는 효과에 대한 탐색과 여러 불완전 자료 처리방법의 수행효과를 상호 비교하는 것이다. 앞에서도 지적하였듯이 본 연구가 체계적인 모의실험의 결과가 아니기에 분석결과의 일반화에 있어서 이에 대한 주의가 필요하며, 패널조사에서 발생하는 다양한 결측치 발생 유형을 고려하는 보다 체계적인 후속 연구가 필요하다.

이러한 사항을 염두에 두고 전체적인 결과를 요약하면, 첫째, 본 연구는 불완전 자료에 대한 적절하지 못한 사후처리는 분석결과에 있어서 유의미한 차이로 이어질 수 있음을 시사한다. 분석결과는 MCAR 상황에서 MAR 그리고 MNAR 상황으로 조건이 변경됨에 따라 완전제거법과 같은 비체계적인 불완전 자료 처리방법을 사용한 분석결과가 결측치가 없는 완전한 자료로부터 도출된 결과와는 상이할 개연성이 상대적으로 높아짐을 시사하고 있다.

특히, 분석결과는 반복측정 자료를 사용하는 연구의 경우 불완전 자료의 발생은 궤적의 초기값보다는 시간의 경과에 따른 궤적의 변화(본 연구의 경우 기울기)를 적절히 추정하는데 심각한 문제를 가질 수 있음을 시사하고 있다. 이러한 점은 일반적으로 패널조사 초기의 경우 결측치의 발생 수준이 낮지만 패널조사의 회차가 증가함에 따라 결측치의 비중이 점차 증가한다는 점과 관련된다고 볼 수 있다.

둘째, 분석결과는 완전제거법이나 평균대체법과 같이 (비록 불완전하지만) 활용 가능한 정보를 적절히 사용하지 못하는 비체계적인 방법들이 최대우도법

이나 다중대체법에 비해 불완전 자료에 대한 처리효과가 상대적으로 떨어짐을 보여준다. 특히, 완전제거법이나 평균대체법과 같은 방법에 비해 최대우도법이나 다중대체법이 갖는 상대적 우위는 MCAR 가정에 비해 보다 현실적인 가정이라고 할 수 있는 MAR 조건하에서 상대적으로 크게 나타난다.

본 연구의 경우 MNAR 상황의 경우 패턴혼합 모형이 상대적으로 좋은 결과를 보였지만 이미 언급되었듯이 패턴혼합 모형은 관측된 자료로부터 도출될 수 없는 가정에 기초하고 있으며, 본 연구처럼 반복측정 자료의 경우나 분석모형에 많은 모수가 포함될 경우 민감도 분석은 더욱 어려워지는 경향이 있음도 지적될 필요가 있다.

선택편의 모형이나 패턴혼합 모형이 갖는 가정 및 분석수행상의 문제를 고려하더라도 무시할 수 없는(Non-Ignorable) 결측치 발생이 의심되는 상황에서 최대우도법이나 다중대체법을 사용하는 것이 과연 적절한가에 대한 의문이 제기될 수 있다. 본 연구의 분석결과는 MNAR 상황에서도 최대우도법이나 다중대체법은 완전제거법이나 평균대체법에 비해 적용효과가 상대적으로 높음을 보여준다. 비록 종단설계는 아니지만, 최근의 모의실험 결과(예를 들면, Collins et al. 2003)도 MNAR 상황에서 다중대체법과 같은 방법이 효과적으로 사용될 수 있으며, 불완전 자료의 발생기제에 대한 잘못된 가정의 영향은 상대적으로 미미한 수준임을 시사하고 있다.

또한 본 연구의 경우 최대우도법이나 다중대체법, 특히 다중대체법의 장점이 저평가된 측면이 있음을 지적할 필요가 있다. 이러한 점은, 비록 본 연구에서는 활용되지 못했지만, 일반적으로 다중대체법의 경우 분석모형에서는 사용되지 않지만 결측치의 발생과 연관된 보조변수(auxiliary variable)를 연구자들이 보정과정에서 활용할 수 있다는 점과 관계된다. 비록 결측치의 발생기제가 MNAR 상황이라고 하더라도 연구자가 결측치의 발생과 관련된 변수들을 보정과정에서 적절하게 활용하면 편의의 상당부분을 감소시킬 수 있을 것을 기대할 수 있다. 무시할 수 없는 결측치 발생상황이 의심되더라도 부적절하게 명시화된 선택편의 모형이나 패턴혼합 모형을 사용하는 대신 MAR 가정하에 결측치의 발생과 관련해서 유의미한 정보를 제공하는 보조변수들을 보정과정에서 활용하는 것이 보다 바람직한 대안이 될 여지가 여기에 있는 것이다.

참고문헌

강민아 · 김경아. 2006. "행정학 및 정책학 조사연구에서 결측치 발생과 처리방법에 대한 고찰." 《한국행정학보》 40(2): 31-52.

김규성. 2000. "무응답대체방법과 대체효과." 《조사연구》 1(2): 1-14.

김덕준. 2007. "행정조사분석에서의 결측치 처리방법: NORM을 활용한 다중대체법을 중심으로." 《현대사회와 행정》 17(1): 107-128.

Acock, A. C. 2005. "Working with Missing Values." *Journal of Marriage and the Family* 67(4): 1012-1028.

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

Collins, L. M., J. L. Schafer, & C. Kam. 2001. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6(4): 330-351.

Dempster, A. P., N. M. Laird, & D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society(Series B)* 39(1): 1-38.

Graham, J. W. & S. M. Hofer. 2000. "Multiple Imputation in Multivariate Research." *Modeling Longitudinal and Multiple Group Data: Practice Issues, Applied Approaches and Specific Examples*, (Eds.) T. D. Little, K. U. Schnabel, and J. Baumert, Mahwah, NJ: Lawrence Erlbaum Publishers, pp. 201-218.

Hedeker, D. & R. D. Gibbons. 1997. "Application of Random-Effects Pattern-Mixture Models for Missing Data in Longitudinal Studies." *Psychological Methods* 2(1): 64-78.

Kenward, M. G. 1998. "Selection Models for Repeated Measurements with Nonrandom Dropout: An Illustration of Sensitivity." *Statistics in Medicine* 17(23): 2723-2732.

Kline, R. B. 1998. *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press.

Little, R. J. A. 1995. "Modeling the Drop-Out Mechanism in Repeated-Measures Studies." *Journal of the American Statistical Association* 90(4): 1112-1121.

Little, R. J. A. & D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. New

Jersey: John Wiley & Sons.

Mirowsky, J. & C. E. Ross. 1992. "Age and Depression." *Journal of Health and Social Behavior* 33(3): 187-205.

Royston, P. 2004. "Multiple Imputation of Missing Values." *Stata Journal* 4(3): 227-241.

Royston, P. 2005. "Multiple Imputation of Missing Values: Update." *Stata Journal* 5(2): 1-14.

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

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L. 2003. "Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ." *Statistica Neerlandica* 57(1): 19-35.

Schafer, J. L. & J. W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7(2): 147-177.

Singer, J. D. & J. B. Willett. 2003. *Applied Longitudinal Data Analysis*. New York: Oxford University Press.

Van Buuren, S., H. C. Boshuizen, & D. L. Knook 1999. "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine* 18(6): 681-694.

Van Buuren, S. & C. G. M. Oudshoorn. 2000. "Multiple Imputation by Chained Equations." Leiden: TNO Prevention and Health. Report PG/VGZ/00.038.

Wothke, W. 2000. "Longitudinal and Multigroup Modeling with Missing Data." *Modeling Longitudinal and Multiple Group Data: Practice Issues, Applied Approaches and Specific Examples*,(Eds.) T. D. Little, K. U. Schnabel, and J. Baumert, Mahwah, NJ: Lawrence Erlbaum Publishers, pp. 219-240.

[접수 2008/4/5, 수정 2008/6/30, 게재확정 2008/7/4]