결측값을 가진 경시적 조사 연구 자료에 적용된 혼합효과모형에 관한 고찰

송 주 원*

Ⅰ. 서론

경시적 자료는 한 관측 개체에 관하여 여러 번의 다른 시점에서 반복 측정을 시행한 자료를 의미한다. 사회 과학 분야에서는 각 개체들의 응답이 시간이 변함에 따라 어떻게 변화하는 지 조사해 보는 연구가 많이 시행되므로 이에 관한 적절한 분석은 중요한 것이다. 경시적 자료의 분석에 관한 방법은 여러 가지가 있는데 혼합효과모형(mixed effects model)은 가장 흔히 사용되는 방법 중 하나라고 할 수 있다. 경시적 자료의 특성은 한 개체의 다른 시점에서 관찰된 측정치는 서로 상관되어 있으므로 이 상관 관계를 고려하여야 하며 각 객체의 모형에서의 계수가 다른 경우 이를 포함하여 분석하는 것이 효과적이다. 이런 자료의 분석에 흔히 사용되는 혼합효과모형의 경우 혼합효과(random effects)를 이용하여 위의 특성을 충분히 고려하는 모형을 구축할 수 있으므로 유용하게 사용되어 왔다.

준비가 잘 된 실험에서조차 결측값은 흔히 발견되며 사회과학의 조사 연구 자료의 경우 연구자가 통제할 수 없는 여러 가지 원인으로 인하여 결측값을 가지는 경우가 흔히 발생하므로 결측값을 적절히 처리하는 분석을 시행하는 것은 아주 중요하다고 할 수 있다. 결측값이 존재하는 경우 이를 적절히 처리하지 않는다면 분석 결과는 편향되거나 비효율적인 추정치를 얻을 수 있다는 것은 이미 주지의 사실이다(Little and Rubin 2002). 본 연구에서는 사회 과학 자료의 조사 연구에서 흔히 사용되는 경시적 자료가 결측값을 포함하는 경우 혼합효과모형을 사용하여 이 자료를 분석하려고 할 때 고려해야 할 사항들을 정리해 보고자 한다.

결측값을 가지는 자료의 분석에 있어서 가장 먼저 고려해야 하는 사항은 결측값이 발생하는 과정(missing data mechanism)을 이해하는 것이다. Missing data mechanism은 결측값이 발생

*고려대학교 통계학과

할 확률이 결측값을 가진 관심변수와 나머지 관련 변수와 어떤 관련이 있는 가에 따라 세가지로 구분하는 것이 일반적이다(Little and Rubin 2002). 첫 번째로 가장 강한 가정을 의미하는 Completely missing at random(MCAR)이 있고, 그보다 조금 완화된 missing at random (MAR)이 있으며 위의 두 가정을 만족시키지 않는 경우를 지칭하는 not missing at random (MNAR)이 있다. 위의 가정들은 횡단연구(cross-sectional study)의 경우 명확하게 정의가 된다. 하지만 경시적 자료의 경우 종속변수(dependent variable or response variable)가 여러 시점에서 측정되며 모형이 독립 변수(independent variable or predictor variable)를 포함하므로 관심변수인 종속변수의 한 시점 (또는 여러 시점)에서 결측값이 발생하는 경우 다른 시점의 종속변수 값 및 독립변수 두 가지와 관계를 고려해야 하므로 정의는 좀 더 복잡해지게 되는 것이다. 게다가 결측값이 발생하는 형태도 크게 중간결측(intermediate missing)과 중도탈락(dropout) 두 가지로 흔히 나누어진다. Intermediate missing이란 어느 한 시점(또는 연속된 여러 시점)에서 개체의 관찰치가 결측이지만 그 후 시점에 다시 측정되는 경우를 의미한다. 그에 반하여 dropout은 한 시점에서 그 개체가 연구에서 임의로(또는 다른 여건에 의하여) 배제되는 것을 의미하며 그 시점 이후 모든 시점에서 그 개체의 관찰 값이 측정되지 않는 것이 특징이다. 이런 복잡성으로 인하여 경시적 자료의 missing data mechanism의 정의에 관한 이견이 있어 왔으며 그에 따른 적절한 분석의 어려움으로 이어지는 경향이 있다. 결측값을 포함한 경시적 자료의 경우 missing data mechanism은 여러 연구에서 약간씩 다르게 정의되어 왔으며 많은 연구들이 dropout에 기인한 결측에 초점을 맞추고 있으며 모형 내 독립 변수의 역할은 크게 중요시되지 않고 있다 (예를 들어, Diggle and Kenward 1994). 그 경우를 더 자세히 관련 독립변수와 관련하여 정의한 경우(Little 1995)에 초점을 맞추어 이 연구에서는 경시적 자료에서의 missing data mechanism을 정의하고 적절한 missing data mechanism을 선택하기 위한 유의사항에 관하여 고찰해 보고자 한다. 이를 위하여 횡단 연구에 관한 모의실험(Collins, Schafer, and Kam 2001)에서의 결과에 근거하여 독립변수의 적절한 선택은 missing data mechanism에 영향을 준다는 것을 보임으로써 적절한 혼합효과모형을 적용하는 것을 가능하게 하고자 한다.

Ⅱ. 경시적 자료에서의 missing data mechanism

1. 횡단연구에서의 missing data mechanism

결측값이 발생하는 원인을 Little and Rubin(2002)은 다음과 같이 세가지로 구분하여 정의하고 있다.

(1) 결측값이 발생할 확률의 분포가 자료 값(결측되었거나 측정된 모든 값)에 전혀 의존하지 않는 경우 missing completely at random(MCAR)을 따른다고 정의한다. 이 경우 결측값은 순수하게 임의로(random) 발생한다고 할 수 있다.

(2) 결측값이 발생할 확률의 분포가 측정된 자료 값에는 의존하지만 결측된 자료의 값에는 의존하지 않는 경우 missing at random(MAR)을 따른다고 정의한다. 이 가정은 결측값이 발생할 확률의 분포가 측정된 자료 값에는 의존하는 것을 가능하게 하므로 MCAR보다 약한 가정이라 할 수 있다.

(3) 결측값이 발생할 확률의 분포가 결측된 자료의 값에 의존하는 경우 자료는 not missing at random(NMAR)이라 정의하게 된다. 이 경우는 위의 두 정의가 만족되지 않는 경우를 나타낸다.

Missing data mechanism이 MCAR이나 MAR을 따르고 결측값이 발생할 확률의 분포의 모수가 자료의 분포의 모수와 구별(distinct)될 때 missing data mechanism은 ignorable이라고 정의한다. 여기에서 ignorable이란 자료의 분포의 모수에 관심이 있을 때 결측값이 발생할 확률의 분포를 무시(ignore)할 수 있다는 의미에서 나온다. 모수의 distinctness는 대부분의 경우 만족시키기 쉬우므로 missing data mechanism이 MCAR 또는 MAR을 따른다는 가정은 자료의 분석을 훨씬 용이하게 해 주는 장점을 가지는 것으로서 제안된 결측 자료의 분석 방법들이 가장 많이 전제로 사용하는 조건이다. 이 가정하에서의 여러 가지 형태(연속형, 이산형, 및 혼합형)의 자료의 분석은 Schafer(1997)에 자세히 소개되어 있다.

한편 위의 정의에서는 자료 값의 정의에 독립변수나 종속변수의 구분이 되어 있지 않다는 것을 유의해 볼 만 하다. 즉, 여기에서 의미하는 자료란 독립변수와 종속변수를 함께 포함하는 여러 개의 변수를 가진 형태로 생각하는 것이다. 따라서 결측값은 종속 변수뿐만이 아니라 독립 변수에서도 나타날 수 있으며 이 경우에도 동일한 방법으로 결측값을 처리할 수 있다는 점에서 일반적이라 할 수 있다.

2. 경시적 자료에서의 missing data mechanism

경시적 자료의 경우 missing data mechanism은 dropout에 기인한 경우에 한정하여 정의하는 경우가 종종 발견된다. 임상 실험 자료의 경우 dropout은 결측의 가장 큰 원인이기도 하다. 게다가 이 경우 경시적 자료는 가장 간단한 결측 패턴을 보이게 되며 이 정의는 추후 intermediate missing을 포함하도록 확장하는 것이 가능하게 된다. Diggle and Kenward(1994)는 missing data mechanism을 다음의 세가지로 정의하고 있다.

(1) 만약 dropout의 process와 종속변수의 자료 값이 측정되는 process가 독립이라면 missing

data mechanism은 completely random drop-out(CRD)이라고 정의한다.

(2) 만약 dropout의 process가 그 전 시점에서 측정된 자료 값에 의존한다면 missing data mechanism은 random dropout (RD)라고 정의한다.

(3) 만약 dropout의 process가 그 개체가 만약 중도에 탈락하지 않았다면 측정되었을 (하지만 탈락으로 측정되지 않은) 값에 의존한다면 missing data mechanism은 informative dropout (ID)라고 정의한다.

이 정의에서 CRD는 Little and Rubin(2002)이 정의한 MCAR과 유사하며 RD는 MAR과 ID는 NMAR과 각각 유사함을 알 수 있다. 가장 큰 차이점을 찾아본다면 이 정의에서는 독립변수의 관련성은 missing data mechanism을 정의하는데 포함되지 않는다는 점이다. 한편 Little (1995)는 이 점을 고려하여 다음과 같이 수정된 missing data mechanism을 제시하고 있다.

(1) 만약 dropout의 확률과 종속변수의 자료 값이 측정될 확률은 독립이지만 dropout의 확률은 독립 변수의 값에 의존하는 경우에 missing data mechanism은 covariate-dependent drop-out이라고 정의한다.

(2) 만약 dropout의 확률이 그 전 시점에서 측정된 자료 값 또는 독립 변수에 의존한다면 missing data mechanism은 missing-at-random dropout이라고 정의한다.

(3) 만약 dropout의 확률이 그 개체가 만약 중도에 탈락하지 않았다면 측정되었을(하지만 탈락으로 측정되지 않은) 값에 의존한다면 missing data mechanism은 nonignorable dropout이라고 정의한다.

Little (1995)과 Diggle and Kenward (1994)의 정의는 독립변수를 missing data mechanism을 정의할 때 고려 대상에 포함시키는 가에 관하여 차이가 있다는 것을 알 수 있다.(Davis 2002). 즉, 같은 자료의 경우에도 Little(1995)이 정의한 missing data mechanism하에서 더 강한 가정을 만족할 수 있다는 것이다. 예를 들어 dropout의 확률이 독립 변수에 의존하는 경우 Little (1994)의 정의하에서는 가정 (1)이나 (2)를 충족하게 되지만 Diggle and Kenward(1994)의 정의하에서는 (3)에 포함되는 것이다. 이 문제가 중요한 이유는 가정이 강할수록 분석에 missing data mechanism을 포함하지 않는 단순한 모형이 적합 되어도 되기 때문이다. 예를 들어 missing data mechanism이 MCAR을 따르는 경우 결측값이 포함된 자료를 제외하고 분석하여도 추정치의 편향이 일어나지 않는다고 알려져 있다. 즉, 경시적 자료에서 dropout된 개체들을 제외하고 분석을 실행하여도 된다는 의미이다. 한편, 경시적 자료에 혼합효과 모형을 적용하였을 경우 장점 중 하나는 임의의 시점에서 결측(또는 dropout)이 발생한다고 하더라도 관찰된 자료 값만 가지고 모형을 적합 시켜도 missing data mechanism이 MAR이라면 편향이 발생하지 않는다는 점이다(Littel, et. al., 1996). 이 점은 실제로 혼합효과 모형을 실행하는 연구자들이 가장 많이 의존하는 가정이라 할 수 있다. 하지만 위에서 본 바와 같이 MAR의 가정은 독립

변수를 포함하는 가 아닌 가에 따라 달라지므로 그런 점에서 독립 변수의 역할이 중요하다고 볼 수 있다. 즉, 적절한 독립 변수를 포함시킴으로써 missing data mechanism을 MAR이 만족되도록 할 수 있다면 결측치의 발생 확률에 관한 추가적인 모형 없이 혼합효과모형을 적합하고도 편향되지 않은 결과를 얻을 수 있다는 것을 의미한다.

Ⅲ. 결측값을 포함한 자료의 모형에서 독립변수의 중요성

독립 변수가 missing data mechanism에 미치는 영향은 횡단 자료(cross-sectional data)에 관한 모의 실험을 통하여 보고되었다(Collins, Schafer, and Kam 2001). 이 연구에서는 일반적으로 모형에 포함되는 독립 변수와 종속 변수 이외에 추가 변수(auxiliary variable)의 포함 효과를 횡단 자료의 모의 실험을 통하여 추정하였다. 이를 위하여 다음의 세가지 경우를 고려하였다.

(1) 추가 변수가 종속 변수와 연관되어 있으며 결측값이 발생할 확률과도 연관된 경우,

(2) 추가 변수가 종속 변수와 연관되어 있으나 결측값이 발생할 확률과는 연관되지 않은 경우,

(3) 추가 변수가 종속 변수와 연관되어 있지 않은 경우

위의 (1)과 (2)는 일반적으로 연구자들이 추가 변수를 모형에 포함하는 경우인데 반하여 (3)의 경우는 추가 변수가 모형에 포함되지 않는 경우가 일반적이다.

모의 실험은 정규 분포 하에서 독립 변수, 종속 변수 및 추가 변수에 관한 500 개의 관측 값을 생성하여 실행되었다. 종속 변수는 각각 25%나 50%의 관측 값이 missing data mechanism에 관한 세가지 가정, 즉 MCAR, MAR, and NMAR, 하에서 결측으로 간주되었다. 우선 위의 (1)의 조건하에서 missing data mechanism이 MAR을 따른다고 가정하였을 때 추가 변수를 포함하지 않는 경우 회귀 계수 및 종속 변수의 분산에 편향이 발생함이 나타났다. (2)의 경우 missing data mechanism이 ignorable인 경우는 추가 변수가 포함되는 경우와 포함되지 않는 경우 모두에서 편향이 나타나지 않았지만 missing data mechanism이 nonignorable인 경우에는 추가 변수가 포함될수록 그리고 추가 변수와 종속변수와의 상관 계수가 크면 클수록 편향의 정도는 줄어드는 것으로 나타났다. 마지막으로 (3)의 경우는 추가 변수가 종속 변수 및 자료의 결측 확률 모두와 상관이 없는 경우에 missing data mechanism이 MCAR인 경우에 대하여 시행하였는데 너무 많은 숫자(여기서는 25개 이상)의 불필요한 추가 변수가 포함되지만 않는다면 편향이 발생하지 않음을 보여주고 있다. 여기서 많은 숫자의 불필요한 추가 변수가 모형에 포함되는 경우 추정하여야 하는 모수의 수가 급격하게 늘어나 편향이 나타난 것으로 사려되었다.

이 모의 실험을 통하여 추가 변수가 종속 변수와 관련 정도가 심할수록 추가 변수의 포함이 도움이 된다는 것을 알 수 있었으며 자료의 결측 정도가 심할수록 편향이 더 크게 나타난 것으

로 보이고 있다. 또한 추가 변수가 결측의 확률과 연관이 되어 있는 경우 missing data mechanism이 MAR하에서도 추가 변수를 포함하여 분석을 시행하는 경우가 추가 변수를 포함하지 않고 분석하는 경우보다 더 좋은 결과를 나타내고 있다. 이는 우리가 일반적으로 선호하는 작은 모형(parsimonious model)보다 종속 변수 및 결측의 확률과 연관된 추가적인 변수를 포함하는 큰 모형이 결측치가 존재하는 경우에는 훨씬 더 유익하다는 점을 나타내는 것이다.

경시적 자료의 경우 종속 변수들은 서로 상관되어 있는 것이 일반적이다. 사회과학을 위한 조사 연구에서도 상관계수가 적어도 0.3 또는 0.4 정도 되는 자료를 흔히 접하게 된다. 결측값을 포함하는 변수와 상관 관계가 높은 변수가 모형에 포함되는 경우 missing data mechanism은 ignorable에 가까워지는데 그런 점에서 경시적 자료의 경우 다른 시점에서의 종속 변수 값을 항상 포함하므로 ignorable missing data mechanism을 더 쉽게 만족시킬 수 있을 것으로 기대된다.

Ⅳ. 결론

경시적 자료는 조사 연구에서도 흔히 이용되는 자료의 형태이다. 하지만 위에서 고찰한 바와 같이 결측이 발생하였을 때 missing data mechanism의 정의가 좀 더 복잡해진다는 것을 알 수 있다. 이런 어려움은 경시적 자료에서 흔히 사용하는 혼합효과모형의 경우 종속 변수가 시점에 따라 측정된 측정값들로 이루어져 있을 뿐 아니라 독립 변수도 포함하는 형태를 띄고 있기 때문이다. 게다가 dropout 뿐만 아니라 intermediate missing을 포함하는 경우 결측이 일어나는 패턴도 아주 다양하므로 단순히 dropout에 근거하여 missing data mechanism을 정의하는 경우 독립 변수에 관한 연관성이 언급되지 않은 채 사용되는 것이다.

횡단연구에서 살펴 본 바와 같이 추가 변수의 역할은 missing data mechanism을 결정하는데 중요한 요인으로 작용하는 것이 가능하다. 즉, 추가 변수를 적절히 사용하여ignorable missingness를 만족시킬 수 있다면 추후의 분석은 missing data mechanism에 관한 모형을 포함하지 않고 결측값을 무시한 채 시행될 수 있을 것이다. 이는 추가 변수를 일반적으로 더 많이 포함하는 multiple imputation이 추가 변수를 포함하지 않는 일반 회귀 모형보다 더 나은 결과를 보일 수 있다는 점에서도 알 수 있듯이 missing data mechanism과 관련된 추가 변수의 고려는 중요한 일이라 할 수 있다. 경시적 자료의 경우 종속 변수와 상관되어 있는 종속 변수의 다른 시점에서의 관찰 값이 포함되므로 추가 변수의 포함 여부의 중요성은 횡단연구보다 낮을 수 있을 것으로 예상된다. 하지만, 이 경우에도 추가 변수가 missingness와 연관이 되어 있을 경우 이 변수를 모형에 포함시키는 것은 유익할 것이다.

Missing data mechanism이 MCAR을 따른다고 가정하는 경우 이 가정의 적절함은 각 결측값의 패턴에서 관찰된 값들의 분포의 비교를 통하여 체크할 수 있다(Park and Davis 1993). 경시적 자료에서 결측값이 intermediate missing을 포함하는 경우 결측의 패턴이 다양하게 가능하고 그에 따라 각 패턴에서 표본수가 충분히 크지 않아 위의 테스트를 적용하기 힘들 수도 있다. 결측의 패턴이 다양하여 각 결측 패턴에서 표본수가 크지 않은 경우에 위의 테스트를 확장하는 것은 흥미로울 것이다.

이 연구에서는 결측값이 경시적 자료의 종속 변수에서 발생하는 경우만을 다루었다. 실제로 결측값은 독립 변수에서도 발생할 수 있으며 혼합효과모형을 시행하는 경우 대부분의 프로그램은 독립 변수가 결측된 관찰 값을 제외하는 것이 일반적이다. 이 경우 missing data mechanism이 MCAR(독립 변수에서 결측이 발생할 확률이 다른 독립 변수 및 종속 변수의 값과 독립)이라 하더라도 원 연구 design의 balance가 깨진다거나 자료의 제외로 인한 비효율성이 발생할 수 있다. 그러므로 독립 변수에 결측이 발생할 경우 혼합효과모형은 똑 같은 위험에 놓이게 되는 것이다.

Multiple imputation은 종종 관심 변수 이외에도 결측값이 발생할 확률과 연관된 다른 변수들을 포함하여 시행된다. 또한 multiple imputation은 종속 변수뿐만 아니라 독립 변수도 결측이 발생하는 경우 imputation을 하는 것이 가능하므로 이런 경우 multiple imputation은 좋은 대안으로 가능할 것이다. 즉, 우선 자료에 *multiple imputation*을 시행한 후 혼합효과모형을 적합시키거나 혼합효과모형 하에서 multiple imputation을 시행하는 것이 가능한 것이다.

Missing data mechanism이 독립 변수, 종속 변수, 또는 연관된 추가 변수를 포함하더라고 nonignorable인 경우 모형은 missing data mechanism에 관한 모형을 포함하여야 하며 이에 관한 많은 분석 기법이 제안되어 왔다(몇 가지 예를 들면 Conaway 1992, 1994; Little 1993; Diggle and Kenward 1994). 하지만 이런 기법들은 추가적인 modeling을 요구하므로 실제 조사분석자들이 적용하기에는 부담이 많이 가는 경우가 상당수 존재한다. Missingness와 연관된 추가 변수를 포함하는 경우 기존의 혼합효과모형을 그대로 사용할 수 있으므로 이 점에서 더욱 유용할 것으로 기대된다.

참고문헌

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

Conaway, M. R. (1992), The analysis of repeated categorical measurements subject to nonignorable nonresponse, *Journal of the American Statistical Association*, 87, 817-824.

Conaway, M. R. (1994), Causal nonresponse models for repeated categorical measurements, *Biometrics*, 50, 1102-1116.

Davis, S. D. (2002), Statistical Methods for the Analysis of Repeated Measurements, Springer: New York.

Diggle, P. and Kenward, M. G. (1994), Informative Drop-out in Longitudinal Data Analysis, *Applied Statistics*, 43, 49-63.

Littell, R. C., Milliken, G. A., Stroup, W.W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models, SAS Institute*, Inc.: Cary, NC.

Little, R. J. A. (1993), Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association*, 88, 125-134.

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

Little R. J. A. and Rubin D.B. (2002), *Statistical Analysis with Missing Data* (2nd edn), Wiley: New York.

Park, T., and Davis, C. S. (1993), A Test of the Missing Data Mechanism for Repeated Categorical Data, *Biometrics*, 49, 631-638.

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