패널자료의 종단적 결측패턴에 관한 실증분석 연구

손창균[1)]

본 논문에서는 패널조사와 같은 종단면 연구에서 시간의 흐름에 따라 패널의 노후화 등의 원인으로 각 조사주기별로 발생하는 무응답(결측)에 대해 특정한 패널집단을 대상으로 무응답 패턴을 통계모형을 이용하여 분석하였다. 이러한 무응답 패턴분석을 기반으로 결측자료가 존재하는 종단자료의 분석에서 적절한 방법을 선택하여 분석을 수행할수 있으며, 만일 무응답 대체가 필요한 경우 적절한 대체 방법을 결정할 수 있을 것이다. 횡단면 조사와는 달리 이용가능한 보조정보가 각 웨이브별로 다양하게 존재하며, 이와 같은 보조정보를 무응답 대체에 활용할수 있다면, 결측자료가 존재하는 패널 자료에 비해 전통적인 통계분석 방법을 적용하여 표준적인 결과를 산출할 수 있을 것으로 기대된다.

주요용어 : 무응답 오차, 무응답 패턴, 완전임의결측, 임의결측, 비임의결측

1. 서론

자료를 수집하는 시점에 따라 횡단면 조사(cross−sectional survey)와 종단면 조사(longitudinal survey)로 구분할 수 있다. 특정 기간에 한정되어 자료가 수집된 경우는 흔히 횡단면 조사로 통칭되며, 지속적인 시점에 따라 자료가 수집되는 경우는 종단면 조사로 명명된다. 통상적으로 일반 횡단적 조사와 마찬가지로 종단적 조사에서 결측자료는 필연적으로 발생하게 된다. 즉, 횡단면 조사에서와 같이 조차 참여 거부, 부재, 특정항목에 대한 무응답 등으로 인해 결측이 발생할 수 있고, 이와는 별도로 종단조사의 경우 특정 관측시점에서 조사 개체가 결측될 수 있으며, 임상연구와 같은 종단연구에서는 특정 연구시점에서 일부 개체는 자료를 제공하지만, 다른 개체들은 사망이나 실험에서의 탈락 등으로 그렇지 못할 수 있다. 주어진 시점에서 특정 개체들은 일부 연구변수에만 반응하여 불완전한 자료(incomplete data)를 생성할 수도 있다. 또한 가구 또는 시설대상 조사에서는 시간의 흐름에 따라 조사대상의 사망, 응답거부, 주소미상 등의 사유로 발생하는 종단

1) 122-705 서울시 은평구 불광동 진흥로 268 한국보건사회연구원 연구위원
E-mail : chkson@kihasa.re.kr

면 결측치(longitudinal missing data)와 어느 한 시점에서 특정 변수의 값이 결측되는 횡단적 결측치(cross-sectional missing data)이다. 이러한 무응답에 의한 결측자료는 결과적으로 조사 자료의 신뢰성에 큰 영향을 미치게 되며, 통계적인 방법으로 완전자료(complete data)를 구성하여 표준적인 통계분석을 수행하게 된다.

패널조사와 같은 종단적 연구에서는 조사에 참여하는 패널의 노후화 등으로 인해 웨이브무응답(wave nonresponse)이 발생하며, 이러한 웨이브 무응답을 단순히 결측으로 처리하여 패널자료를 분석할 경우 분석대상 자료의 축소에 따른 무응답 편향을 야기하기 때문에 결측치 처리를 위한 별도의 무응답 패턴을 규명할 필요가 있다.

최근까지 종단면 연구에서 결측치를 다루는 다양한 연구논문들이 발표되었고(Demirtas, 2004; Molenberghs et al., 2004; Hogan and Larid, 1997; Gorbein et al., 1992), 이러한 연구들 중에서 Little(1995)은 결측치 연구에 대한 중요한 통계적인 체계를 수립하였고, 최근에 Hogan et al.(2004)는 결측치 연구에 대한 다양한 적용을 소개하였다. 또한 결측치 처리에 관한 연구들로는 Rubin and Little(2002), Diggle et al. (2002), Verbeke and Molenberghs (2000)를 들 수 있다.

본 연구에서는 패널자료 분석 과정에서 발생하는 웨이브 무응답(결측)에 대한 무응답 패턴을 이론적 고찰을 통해 살펴보고, 각각의 결측 패턴에 따른 분석 방법과 결측패턴의 타당성 검토를 위해 실제 종단자료를 활용하여 실증적 연구를 수행하고자 한다. 본 연구에서 활용한 패널 자료는 한국복지패널 1~3차 자료이며, 패널 응답자의 주요변수들에 대한 추정과정에서 결측패턴에 대한 통계적 검증을 실시하였다.

2. 결측 패턴에 관한 이론적 고찰

전통적으로 결측치가 있는 자료에 대한 통계분석 방법에는 우선 완전한 자료만을 선별하여 분석하는 방법이 있으며, 다음으로는 결측자료를 대체하여 완전한 자료로 구성한 후 분석하는 결측치 대체 방법이 있다. 물론 결측지 대체 방법에는 평균대체, 회귀대체, 핫댁대체, 최근방 대체 및 다중대체 등의 다양한 통계적 방법이 존재하며, 실제 자료 분석에서 흔히 사용되는 방법이기도 하다. 이와 더불어 결측치가 존재하는 자료를 이용한 추정방법으로는 최우 추정법, EM, MCEM의 방법으로 추정치를 구할수 있다. 이와 같이 결측치가 존재하는 자료를 분석하기 위한 다양한 방법이 적용될 수 있으나, 각 방법들의 기본 가정은 결측패턴에 대한 가정으로 부터 출발한다.

2.1 결측 패턴의 정의

종단면 자료의 결측패턴을 연구하기에 앞서 먼저 이에 필요한 기호와 메커니즘 등을 정의하고

자 한다. 응답메커니즘으로서 R_{ij}는 만일 개체 i가 시점 j에서 결측이면 1의 값을 가지며, 만일 개체 i가 시점 j에서 관측되면, 0의 값을 가지는 지시변수(indicator variable)이다. 이러한 응답메커니즘은 종속변수 y가 관측되거나 그렇지 않은지를 결정한다. 만일 어떤 관찰연구에서 개체들이 T개의 시점에서 관측되었다면, $T\times 1$ 차원의 완전한 종속변수벡터는 다음과 같다.

$$\boldsymbol{y}_i' = (y_{i1}, y_{i2}, \cdots, y_{i\,T})$$

그러면 임의의 개체에 대한 $T\times 1$ 차원의 결측자료 지시벡터는 다음과 같이 정의할 수 있다.

$$\boldsymbol{R}_i' = (R_{i1}, R_{i2}, \cdots, R_{i\,T}),\ R_{ij} = \begin{cases} 1, & y_{ij} \quad missing \\ 0, & y_{ij} \quad observed \end{cases}$$

응답매체 $\boldsymbol{R}_i$에 근거하여 주어진 개체 i에 대해 완전 종속변수벡터 $\boldsymbol{y}_i$를 관측된 자료 $\boldsymbol{y}_{iobs}$와 결측된 자료 $\boldsymbol{y}_{imis}$ 로 다음과 같이 분해할 수 있다.

$$\boldsymbol{y}_i' = (\boldsymbol{y}_{iobs}, \boldsymbol{y}_{imis})'$$

여기서 $\boldsymbol{y}_i$는 개체 i에 대한 종속변수벡터이며, $\boldsymbol{y}_{iobs}$는 개체 i에 대해 실제로 관측된 종속변수벡터이고, $\boldsymbol{y}_{imis}$은 개체 i에 대해 결측된 종속변수 벡터이다. 한편 $\boldsymbol{X}$를 시간 또는 설명변수 들의 집합으로 공변량 행렬로 정의하자.

1) 완전임의결측(Missing Completely at Random:MCAR)

결측자료에 대한 가장 강한 가정은 어떤 개체가 완전하게 확률적인 원인으로 특정시점에 결측된 것으로서 응답자들의 집합은 원래의 표본에 대해 확률 부차표본으로 고려할 수 있다는 가정이다. 이러한 결측매체를 "완전임의 결측(MCAR)" 이라 한다. 이는 결측자료 지시변수 벡터 $\boldsymbol{R}_i$는 $\boldsymbol{y}_{iobs}$와 $\boldsymbol{y}_{imis}$ 둘 다와 독립임을 의미하며, 다시 말해서 응답확률이 관측종속변수와 비관측종속변수에 좌우되지 않음을 의미한다.

$$\Pr(\boldsymbol{R}_i|\boldsymbol{y}_{iobs}, \boldsymbol{y}_{imis}) = \Pr(\boldsymbol{R}_i),\quad \text{모든 } \boldsymbol{y}\text{에 대해} \tag{1}$$

종단면연구에서는 시간이 흐름에 따라 다양한 원인에 의해 결측자료의 수는 증가하게 되며, 따라서 MCAR에서 이러한 가정을 허용하는 것은 매우 유용하다. MCAR 가정하에서 추론과정은 보통의 일반화 추정방정식(GEE)를 이용하여, 공변량에 대한 조건부 추론을 통해 결측 자료가 있는 불완전 자료의 추론을 하며 이를 "공변량-종속 결측"모형이라 한다.

$$\Pr(\boldsymbol{R}_i|\boldsymbol{y}_{iobs}, \boldsymbol{y}_{imis}, \boldsymbol{X}) = \Pr(\boldsymbol{R}_i|\boldsymbol{X}),\quad \text{모든 } \boldsymbol{y}\text{에 대해} \tag{2}$$

만일 결측패턴이 MCAR이면, 종단면 분석에서 결측된 자료는 모두 제외하고 온전한 자료만을 이용하여 분석가능하기 때문에 분석 모형의 적용이 용이하고, 추정치의 편향은 거의 발생하지 않게 되며, 따라서 표준오차는 적절하게 추정된다.

2) 임의결측(Missing at Random: MAR)

임의결측(MAR)은 완전히 관찰된 공변량 $\boldsymbol{X}$와 관찰된 종속변수 벡터 $\boldsymbol{y}_{iobs}$ 둘 다에 결측치들의 종속성을 허용함으로서 MCAR 가정에 비해 완화된 모형이라고 할 수 있다. MAR은 이러한 조건하에서 결측치들이 관측되지 않은 종속변수벡터 $\boldsymbol{y}_{imis}$과 관련되지 않는다고 가정한다. 다시 말해서, 이를 조건부독립성 가정이라고 생각할 수 있다. 즉, 공변량 $\boldsymbol{X}$와 관찰된 종속변수 벡터 $\boldsymbol{y}_{iobs}$의 조건하에서 결측 $\boldsymbol{R}_i$는 종속변수벡터 $\boldsymbol{y}_{i_{mis}}$에 독립이다.

$$\Pr(\boldsymbol{R}_i|\boldsymbol{y}_{iobs},\boldsymbol{y}_{imis}) = \Pr(\boldsymbol{R}_i|\boldsymbol{y}_{iobs}) \ , \qquad \text{모든 } \boldsymbol{y}_{imis}\text{에 대해} \tag{3}$$

MAR은 결측자료들이 관측자료들인 $\boldsymbol{X}$와 $\boldsymbol{y}_{i_{obs}}$에 연관된다고 가정하는 반면, 비관측자료 $\boldsymbol{y}_{i_{obs}}$과는 추가적으로 관련되지 않음을 가정한다. 결과적으로 종단면 모형에 대한 MAR은 적절한 공변량들이 $\boldsymbol{X}$에 포함되고, $\boldsymbol{y}_i$의 분산-공분산 구조가 적절히 규정되어야 한다. 만일 이들중 하나의 조건이 성립되지 않으면, 주어진 분석은 기본적인 MAR 매체와 일치하지 않는다. MAR 가정하에서는 우도에 근거한 추론이 가능한데, 관측치와 공변량에에 대한 조건부 추론으로서 다음과 같은 우도모형을 고려한다.

$$\Pr(\boldsymbol{R}_i|\boldsymbol{y}_{iobs},\boldsymbol{y}_{imis},\boldsymbol{X}) = \Pr(\boldsymbol{R}_i|\boldsymbol{X},\boldsymbol{y}_{iobs}) \ , \qquad \text{모든 } \boldsymbol{y}_{imis}\text{에 대해} \tag{4}$$

MCAR과 MAR 모형 둘 다는 결과적으로 공변량 $\boldsymbol{X}$에 의존하는 결측을 허용하는 모형이며, 종단면 연구에서 결측치에 대한 예측으로 적절한 공변량 $\boldsymbol{X}$를 고려하는 것이 매우 중요하다. 예를 들어 가구소득 변수의 결측이 다른 가구의 소득에 영향을 받기 보다는 가구원의 결혼상태나 지역적 특성에 영향을 받는 경우 가구소득의 결측은 MAR일 수 있다. MAR에 대해 관심변수 Y와 결측치인 Y간의 검정은 불가능하지만, 보조변수 X와 결측인 Y간의 검정은 가능하다는 측면에서 공변량 X의 고려가 중요하다.

3) 비임의결측(Missing Not at Random: MNAR)

비임의결측(MNAR)은 결측이 관측자료($\boldsymbol{X}$와 $\boldsymbol{y}_{i_{obs}}$)를 고려한 후 비관측된 종속변수벡터 $\boldsymbol{y}_{imis}$과 연관된 경우를 의미한다. 여기서는 비관측된 값(즉, $\boldsymbol{y}_{imis}$의 값)과 결측 $\boldsymbol{R}_i$간의 관계를 의미한다. MAR과 MNAR간의 차이점이 비관측된 자료 $\boldsymbol{y}_{imis}$을 포함하는지 여부이기 때문에 MNAR 대비 MAR을 확신하거나 기각할 수 있는 방법은 없다. 특정한 MNAR 모형에 비교하여 특정한 MAR

모형을 확신하거나 기각할 수 있지만, 이것이 보다 일반적인 모형을 고려하는 것은 아니다. 통상적으로 MNAR의 이용은 자료가 심각하게 MAR에 위배된다고 의심되는 경우로 한정하며, 이 경우 MNAR 결측의 가정하에서 다양한 민감성 분석을 실시하는 것이 유용하다

2.2 결측패턴에 대한 검증

1) MCAR 에 대한 검증

가장 간단한 가정으로 2개의 시점(two-time points)을 가진 자료를 가정하여 첫째 시점에서는 모든 자료가 관측되고, 두 번째 시점에서는 일부 자료가 결측인 경우를 가정하자. 이를 위해 D_i인 지시변수를 고려하여 만일 개체들이 두 시점에서 모두 관측되면, $D_i=0$이라 하고, 만일 개체가 첫 번째 시점에만 관측되면 $D_i=1$ 이라 하자. 그러면 지시변수 D_i에 의해 판별된 두 그룹의 관측값을 비교하여 만일 MCAR이 성립하면 관측값은 차이가 나지 않아야 한다. 즉, MCAR은 두 그룹 $D_i=0$와 $D_i=1$간의 관측치 y가 평균적으로 차이가 나지 않아야 하기 때문에 이에 대해 간단한 t-검정을 고려할 수 있다. 이를 보다 일반화 하면, 다음과 같은 회귀모형을 고려할 수 있다.

$$y_{i1}=\beta_0+\beta_1 D_i+\beta_2\boldsymbol{x}_i+\epsilon_i \tag{5}$$

여기서 β_2는 $\boldsymbol{x}_i$의 공변량의 회귀계수 벡터이다. 만일 이탈지시변수와 공변량간의 상호작용을 고려한다면, 다음과 같은 모형을 고려할 수 있다.

$$y_{i1}=\beta_0+\beta_1 D_i+\beta_2\boldsymbol{x}_i+\beta_3(D_i\times\boldsymbol{x}_i)+\epsilon_i \tag{6}$$

여기서 β_2는 $\boldsymbol{x}_i$의 공변량의 회귀계수 벡터이다. 이모형으로부터 만일 결측이 MCAR이면 $\beta_1=\beta_3=0$ 가 성립한다.

한편 Ridout(1991)은 이탈변수에 대해 다음과 같은 로지스틱 모형을 고려하였다.

$$\log\left[\frac{P(D_i=1)}{1-P(D_i=1)}\right]=\boldsymbol{\alpha}_0+\boldsymbol{\alpha}_1 y_{i1}+\boldsymbol{\alpha}_2\boldsymbol{x}_i+\boldsymbol{\alpha}_3(y_{i1}\times\boldsymbol{x}_i) \tag{7}$$

여기서 $\boldsymbol{\alpha}_2$와 $\boldsymbol{\alpha}_3$는 공변량 $\boldsymbol{x}_i$와 y_{i1}간의 상호작용에 대한 회귀계수벡터를 나타낸다. 따라서 만일 결측패턴이 MCAR이면 $\boldsymbol{\alpha}_1=\boldsymbol{\alpha}_3=0$가 성립한다. 다시 말해서 가설 $\boldsymbol{\alpha}_1=\boldsymbol{\alpha}_3=0$을 기각하면, 결과적으로 MACR 가정을 기각하게 된다. 한편 결측패턴이 MAR 이면 일반화추정방정식 보다는 평균과 분산-공분산 구조가 적절하다면 혼합회귀모형을 이용한 분석이 타당하다. 또한 MCAR과 MAR간의 모형 선택은 관측치 $\boldsymbol{y}_{i_{obs}}$에 의존하는지 여부를 판정함으로서 가능하며, MCAR은 $\boldsymbol{y}_{i_{obs}}$에 의존하지 않으며, MAR은 $\boldsymbol{y}_{i_{obs}}$에 의존하기 때문에 결과적으로 모형에 $\boldsymbol{y}_{i_{obs}}$를 투입하여 분석함

으로서 MCAR에 대한 검증이 가능하다.

2) NMAR 에 대한 검증

자료의 결측 패턴이 무시할 수 없는 결측, 즉 비임의결측(NMAR)일 경우 표준적인 통계모형을 이용한 분석은 심각한 편향을 발생시킨다. 그러나 MAR과 NMAR에 대한 표준적인 통계적 검정은 존재하지 않으며, 실제 관측자료에서 MAR인지 NMAR인지는 전혀 알 수 없다.

NMAR에 대한 일반적인 모형은 선택모형(selection model)과 패턴혼합모형(pattern mixture model)로 구분할 수 있다. 전자는 종단면자료와 결측 과정 둘 다를 모형화 한 것이며, 후자는 종단면 모형에 있는 결측자료의 패턴정보를 이용한다. NMAR을 검증하기 위해서는 주어진 자료에 대하 다양한 형태의 모형을 고려해야 한다. Ten Have et al.(1998)에 따르면 혼합효과 선택모형은 종단면 관측벡터 $\boldsymbol{y}_i$를 다음과 같이 정의하였다.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{v}_i + \epsilon_i \tag{8}$$

여기서 $\boldsymbol{v}_i$는 확률 개체효과 벡터이다.

한편 이탈회수를 나타내는 변수 D_i에 대해, 시점에 대한 보충적 로그-로그 회귀모형은 다음과 같이 정의할 수 있다(Wu and Caroll, 1998).

$$\log(-\log(1-P(D_i=j|D_i \geq j)) = \boldsymbol{W}_i\alpha + \boldsymbol{v}_i\alpha_i^* \tag{9}$$

여기서 $\boldsymbol{W}_i$는 이탈시점의 예측값을 포함하며, 이는 $\boldsymbol{X}_i$에 일부 또는 전부가 포함될 수 있다. 추가적으로 결측은 확률개체효과인 $\boldsymbol{v}_i$에 의존하며, 이는 종속변수 $\boldsymbol{y}_i$의 관측값과 비관측 값 둘 다를 특성화 한다. 회귀계수 α_i^*는 0이 아니며, 결측이 $\boldsymbol{y}_{i_{mis}}$ 과 $\boldsymbol{y}_{i_{obs}}$ 에 의존한기 때문에 무시할 수 없는 모형이다. 결과적으로 이 모형은 결측시점이 개인의 속성에 영향을 받는 것을 의미하며, 만일 α_i^*가 모두 0이라는 가설을 기각하면, 무시할 수 없는 무응답 모형(nonigrorable nonresponse model)이다.

하나의 개체에 대한 주변 우도는 다음과 같이 정의할 수 있다.

$$f(\boldsymbol{y}_i, D_i) = \int_v f_y(\boldsymbol{y}_i|\boldsymbol{v}) f_D(D_i|\boldsymbol{v}) f(\boldsymbol{v}) d\boldsymbol{v} \tag{10}$$

여기서 $f(\boldsymbol{v})$는 확률효과의 밀도함수이며, 이들은 평균이 0 이고, 분산-공분산행렬이 Σ_v인 다변량 정규분포를 따른다.

그러면 N개의 표본개체에 대한 주변우도는 주변우도의 합으로 다음과 같다.

$$\log L = \sum_{i=1}^{N} f(\boldsymbol{y}_i, D_i) \tag{11}$$

만일 확률효과가 정규분포를 따르지 않는 경우에는 $\boldsymbol{v}_i = S\theta_i$로 표준화 하여 적용할수 있고, 분산-공분산은 $\Sigma_v = SS'$ 이며 여기서 S는 하삼각 행렬이다. 그러면 식(10)은 다음과 같이 재 표현된다.

$$f(\boldsymbol{y}_i, D_i) = \int_{\theta} f_y(\boldsymbol{y}_i|\boldsymbol{\theta}) f_D(D_i|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{12}$$

3. 결측패턴에 대한 통계적 검증

3.1 웨이브간 결측치에 대한 기초분석

본 연구는 한국복지패널자료 중에서 2006~2008년의 3개년도 자료를 이용하여 종단적인 결측패턴을 통계적 방법으로 검증하고자 한다. 한국복지패널은 2006년에 1차 웨이브 조사를 통해 7,072가구(14,463명)를 패널로 구축하였고, 2008년 3차 웨이브에서는 총 6,314가구에 대해 조사가 완료되었고, 이중 원표본 유지율을 산정해 보면 86.7%로 국내 유일의 높은 패널유지율을 나타내었다. 그러나 1차 웨이브 이후 지속적으로 패널의 탈락으로 3차 웨이브에서는 약 758가구(1,533명)이 소실된 것으로 나타나 1차 웨이브부터 3차웨이브까지의 자료를 결합하여 분석할 경우 웨이브 결측이 발생하게 된다.

<표 1>에서는 결측패턴에 따른 빈도 및 백분율을 분석한 결과를 나타내고 있다. 1, 2, 3차 웨이브 중 1차 웨이브 이후 각 웨이브에서 응답한 경우에는 "O", 그렇지 않은 경우에는 "M"으로 나타내고, 만일 결측패턴이 "OMM" 이면, 1차 웨이브 이후 2, 3차 웨이브에서 결측인 경우를 의미한다. 한편 1~3차 웨이브까지 모두 조사(응답)된 패널은 "OOO" 로 나태내며 이에 속한 패널은 총 15,614명이며, 1, 2차 웨이브에서는 응답하고, 3차 웨이브에서 결측인 "OOM"은 1,410명, 1차와 3차 웨이브에서는 응답하고, 2차 웨이브만 결측인 경우를 나타내는 "OMO"은 58명 등으로 나타났다. 한편 1차 웨이브에서 결측이고, 2, 3차 웨이브에서 응답한 "MMO" 또는 "MOM", "MOO"의 경우는 1차 웨이브에 조사되지 않았던 가구원이 2, 3차 웨이브에 새롭게 조사된 경우를 나타내며, 이 경우 각각 532명, 44명, 410명 등으로 총 986명으로 분석되었다.

〈표 1〉 결측패턴

결측패턴	빈도수(명)	백분율(%)
MMO	532	2.68
MOM	44	0.22
MOO	410	2.07
OMM	1,774	8.94
OMO	58	0.29
OOM	1,410	7.11
OOO	15,614	78.69

〈표 2〉 지역1과 성별2에 따른 결측패턴 (단위 : 명)

지 역	MMO		MOM		MOO		OMM		OMO		OOM		OOO	
	M	F	M	F	M	F	M	F	M	F	M	F	M	F
1	58	47	5	3	38	30	221	237	7	10	165	161	1,301	1,524
2	81	60	8	7	70	53	282	260	9	7	250	190	1,934	2,163
3	117	76	11	4	73	71	286	293	6	11	251	244	2,445	2,719
4	37	35	2	4	39	24	87	76	2	4	93	74	1,322	1,706
5	15	6	-	-	6	6	14	18	1	1	12	15	230	270
계	308	224	26	18	226	184	890	884	25	33	726	684	7,232	8,382
p-값	0.443		0.369		0.624		0.599		0.764		0.753		0.016	
CramerV	0.084		0.269		0.079		0.039		0.176		0.037		0.028	

주1) 1=서울, 2=대도시, 3=중소도시, 4=농어촌, 5=도농지역,
주2) M=남성, F=여성

〈표 2〉에서는 패널 응답자의 거주지역별, 성별에 따른 결측패턴을 나타내고 있으며, 결측패턴에서 1회 이상의 결측(신규진입자 포함)을 나타내는 패널은 4,228명으로 나타났다. 서울지역의 경우 1회 이상 결측인 패널은 982명이며, 서울을 제외한 대도시지역(광역시)는 1,277명, 중소도시(도지역의 시군구)는 1,443명, 농어촌지역은 477명, 도농지역은 94명으로 도시지역 패널의 결측 빈도가 상대적으로 높게 나타났다. 한편 1회 이상 결측된 패널에 대해 지역에 따른 성별의 차이를 살펴보면, 대도시지역을 제외하고(남:31.8%, 여:28.5%) 지역별로 성별에 따른 차이는 거의 나타나지 않았다. 한편 1~3차까지 모두 응답한 패널의 지역에 따른 패널의 성별차이를 분석한 결과 3년간 조사에 참여한 패널(OOO)의 거주지역에 따라 응답자의 성별에 차이가 나타났다($\chi^2=12.2055$, $p<0.05$).

응답패턴별로 지역과 성별간의 연관성을 파악해보면, "MOM"으로 Cramer-V값이 0.269로나

타나 2차 웨이브에서만 응답한 경우 성별과 지역간 연관성이 큰 것으로 나타났다. 다음으로 “OMO"가 Cramer－V값이 0.176으로 2차 웨이브에서 무응답인 경우 지역과 성별 간에 연관성이 있는 것으로 분석되고, 그 외에서는 연관성이 매우 약하게 나타나 결측패턴이 성별과 지역에 따라 차이가 없는 것으로 나타났다.

〈표 3〉에서는 지역과 가구 유형(소득수준)에 따른 결측패턴의 연관성을 분석한 결과이다. 결측패턴이 “OMO"인 경우 Cramer－V값이 0.598로서 지역과 가구 유형에 따른 결측패턴이 연관성이 높게 나타났는데, 1차 응답, 2차 결측 후 3차에 재응답한 경우로서 저소득 가구에 비해 일반가구의 재응답률이 높았고, “MOM"인 경우 Cramer－V값이 0.388로 연관성이 높게 나타났다.

〈표 3〉 지역*과 가구유형에 따른 결측패턴** **(단위 : 명)**

Region	MMO		MOM		MOO		OMM		OMO		OOM		OOO	
	H	L	H	L	H	L	H	L	H	L	H	L	H	L
1	90	15	7	1	61	7	343	115	7	10	268	58	2,157	668
2	113	28	12	3	104	19	382	160	15	1	282	113	2,878	1,219
3	169	24	12	3	125	19	458	121	11	6	388	107	3,698	1,466
4	47	25	2	4	48	15	107	56	0	6	98	69	1,527	1,501
5	15	6	-	-	9	3	23	9	2	0	17	10	337	163
계	434	98	33	11	347	63	1,313	461	35	23	1,053	357	1,0597	5,017
p-값	0.0005		0.085		0.186		0.0015		0.0004		<0.0001		<0.0001	
CramerV	0.195		0.388		0.123		0.099		0.598		0.169		0.189	

주1) 1=서울, 2=대도시, 3=중소도시, 4=농어촌, 5=도농지역,
주2) H=일반가구, L=저소득 가구

3.2 결측패턴에 대한 가설검정

앞 절에서 살펴본 바와 같이 패널응답자들의 결측패턴에 대한 가설검증을 위해, 인구사회학적 변수를 토대로 다음과 같이 변수들을 재 정의하였다. 즉, 결측확률이 관심변수와 연관이 있는지 또는 그외 설명변수와 관심변수간에 연관성이 있는지에 대한 검증을 통해 결측자료의 패턴을 식별할 수 있기 때문이다. 이를 위해 가설검증 모형에 사용하기위한 변수를 다음의 〈표 3〉과 같이 정의하였다. "drop"은 결측여부에 대한 지시변수로서 1~3차 까지 1회 이상의 결측이 있는 경우에 “1”, 그렇지 않은 경우에는 “0”을 값으로 재코딩하였다.

1~3차에 걸쳐 웨이브 무응답의 결측패턴을 파악하기 위해 먼저 모형 식(7)을 적용하여 각 독립

변수들에 대해 소득변수에 대한 상호작용($y_{ij} \times x_{ij}$)항에 대한 유의성 검정을 수행한 결과 다음의 〈표 4〉와 같다. 모형1은 소득(y_i)와 나머지 설명변수($\boldsymbol{X}_i$:가구원수, 연령, 성별, 가구주)들간의 상호작용을 고려한 모형이다. 모형2는 소득(y_i)와 모형1에서 고려한 설명변수($\boldsymbol{X}_i$:가구원수, 연령, 성별, 가구주)외에 교육정도, 경활상태를 추가하여 상호작용을 고려한 모형이다. 모형3은 소득(y_i)와 모형1과 2에서 고려한 설명변수 외에 장애유형, 거주지역을 추가하여 상호작용을 고려한 모형이다.

각각의 모형에서 소득변수의 추정계수들이 모두 통계적으로 유의하지 않게 나타나, 소득과 결측패턴간에는 유의미한 관계가 나타나지 않아 귀무가설 H_0: $\alpha_1 = 0$을 채택함으로 나머지 변수들간의 상호작용에 대한 검토는 불필요하였다. 결과적으로 소득(y_i)과 웨이브 결측간에는 서로 독립적으로 작용하고 있기 때문에 MCAR이라는 주장을 받아들일 수 있다.

〈표 3〉 변수정의

Variable	Description
drop	missing indicator(missing =1, observed=0 : outcome variable)
log_cin	income(with logarithm)
f_num	number of family
age	age
gender1	sex(male=1, female=0)
h_head1	head of household1(male=1, other=0)
h_head2	head of household2(female=1, other=0)
edu_new1	education1(Middle=1, other=0)
edu_new2	education2(High=1, other=0)
edu_new3	education3(University=1, other=0)
edu_new4	education4(Graduate=1, other=0)
eco01	economical status1(Salary=1, other=0)
eco02	economical status2(Own business et al. =1, other=0)
eco03	economical status3(Non-economic activity=1, other=0)
dis_lev1	disability level1(higher=1, other=0)
dis_lev2	disability level1(Lower=1, other=0)
reg1	region1(Seoul=1, other=0)
reg2	region2(Matropolitan=1, other=0)
reg3	region3Urban=1, other=0)
reg4	region4(Rural=1, other=0)
reg5	region5(half urban=1, other=0)

〈표 4〉 GEE Analysis for Logistic Model

Model1		Model2		Model3	
Parameter	Estimate	Parameter	Estimate	Parameter	Estimate
Intercept	-1.74***	Intercept	-18.22***	Intercept	-18.97***
log_cin	0.00	log_cin	0.02	log_cin	0.08
f_num	0.27*	f_num	0.33**	f_num	0.30*
age	-0.30***	age	-0.12	age	-0.09
gender1	-0.39	gender1	-0.26	gender1	-0.28
h_head1	0.60	h_head1	0.47	h_head1	0.46
h_head2	-0.52	h_head2	-0.50	h_head2	-0.59
log_cin*f_num	-0.04*	edu_new1	-1.96***	edu_new1	-1.79***
log_cin*age	0.03**	edu_new2	-0.41	edu_new2	-0.33
log_cin*gender1	0.10**	edu_new3	0.00	edu_new3	0.01
log_cin*h_head1	-0.14**	edu_new4	0.00	edu_new4	0.00
log_cin*h_head2	0.07	eco01	17.19***	eco01	17.10***
		eco02	14.26***	eco02	14.54***
		eco03	16.49	eco03	16.51
		log_cin*f_num	-0.04**	dis_lev1	-0.12
		log_cin*age	0.01	dis_lev2	-1.05
		log_cin*gender1	0.08*	reg1	0.42
		log_cin*h_head1	-0.12*	reg2	1.42
		log_cin*h_head2	0.07	reg3	0.99
		log_cin*edu_new1	0.22***	reg4	-1.05
		log_cin*edu_new2	-0.01	reg5	0.00
		log_cin*edu_new3	-0.03	log_cin*f_num	-0.04*
		log_cin*edu_new4	0.00	log_cin*age	0.01
		log_cin*eco01	-0.09*	log_cin*gender1	0.08*
		log_cin*eco02	0.25***	log_cin*h_head1	-0.12*
		log_cin*eco03	0.00	log_cin*h_head2	0.08
				log_cin*edu_new1	0.20***
				log_cin*edu_new2	-0.02
				log_cin*edu_new3	-0.03
				log_cin*edu_new4	0.00
				log_cin*eco01	-0.08*
				log_cin*eco02	0.22***
				log_cin*eco03	0.00
				log_cin*dis_lev1	-0.01
				log_cin*dis_lev2	0.10
				log_cin*reg1	-0.01
				log_cin*reg2	-0.15
				log_cin*reg3	-0.10
				log_cin*reg4	0.13
				log_cin*reg5	0.00
Scale	1.00	Scale	1.00	Scale	1.00

4. 결론

본 연구에서는 한국복지패널의 1~3차 웨이브 자료를 이용하여 패널응답자의 결측패턴에 대해 웨이브별 결측이 설명변수에 따라 다르게 나타나는지에 대해 분석하였다. 실제 자료를 이용하여 결측 패턴의 통계적 검증 방법을 적용하여 개인의 특성에 따라 결측패턴을 정의하였다. 소득, 가구원수, 교육 등의 변수는 웨이브 무응답과는 독립(MCAR)이며, 경활상태는 웨이브 무응답에 영향을 주는 것(MAR)으로 나타났다.

결과적으로 가구원의 결측패턴이 소득, 가구원수 및 교육정도에는 무관하게 발생함으로서 완전임의결측(MCAR) 매커니즘이 적절하며, 따라서 웨이브 무응답을 제거하고 분석하여도 무방한 것으로 나타났으며, 경활상태는 MAR 로서 결측치에 대한 대체후 분석이 필요한 것으로 나타났다.

참고문헌

Allison, P. D. (1999), Logistic Regression Using The SAS Sysytem, SAS press.

Agresti, A. (2002), Categorical Data Analysis, 2nd ed. John Wiley and Sons Inc, New York.

Allison, P. D. (2002), Fixed Effects Regression Methods for Longitudinal Data Using SAS, SAS press.

Demirtas, H. (2004), Modeling Incomplete Longitudinal data, Journal of Modern Applied Statistical methods, Vol.3, pp. 305-321.

Demirtas, H. and Schafer, J. L.(2003), On the performance of random-coefficient pattern-mixture models for nonignorable dropout, Statistical in Medicine, Vol. 22, pp.2553-2575.

Diggle. P. J. Kenward, M. G.(1994), Informative drop-out in longitudinal data analysis (with discussion), Applied Statistics, Vol.4, pp.49-93.

Fees, E. W. (2004), Longitudinal and Panel Analysis and Applications in the Social Science, Cambridge University Press, New York.

Hedeker, D and Gibbons, R. D. (2006), Longitudinal Data Analysis, John Wiley and Sons Inc, New York.

Hsiao, C. (2003), Analysis of Panel Data, Cambridge University Press, New York.

Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P.(1989), Panel Survey, John and Wiley Sons Inc, New York.

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

Singer, J. D., and Willett, J. B.(2003), Applied Longitudinal Data Analysis : Modeling Chance and Event Occurrence, Oxford University Press, New York.

한국복지패널 기초분석 보고서 (2006, 2007, 2008), 각 년도, 한국보건사회연구원

손창균, 류제복(2011), 결측자료 패턴에 대한 분석－한국복지패널을 중심으로－, Proceedings of the Korean Data Analysis Society, pp.359－366.