한국인구학 제21권 제1호(1998): 162~183 한국인구학회

불완전한 자료에 대한 보완기법

EM 알고리듬과 2단계(Two Stage) 모델

박경숙*

여기서는 많은 수의 비관측사례로부터 발생할 수 있는 표본의 편의(bias) 문제를 탐구한다. 이 연구는 본래 일본 후생성이 1989년 실시한 〈가족주기와 가구형태에 대한 인구학적 조사〉 자료를 이용하여 노인부모와 자녀간 근접성을 분석하는 목적에서 이루어졌다. 그런데 〈가족주기와 가구형태에 대한 인구학적 조사〉는 노인부모를 대상으로 한 조사가 아니라 전체 가구 일반에 대한 조사이기 때문에 노인부모에 대한 많은 정보를 손상하고 있었다. 또한 본 조사는 가구주를 통하여 가족원에 대한 정보를 획득하는 방식으로 설계되었기 때문에 가족원에 대한 정보가 완전하지 못하였다. 나아가 비관측사례의 유형을 보면 여러 항목들이 동시적으로 관측되지 않고 있었다. 이와 같이 복합적 메커니즘에서 발생한 비관측 사례는 분석의 편의를 초래할 위험이 크다. 우선, 많은 수의 비관측사례로 표준오차를 잘못 추정할 소지가 크다. 더욱이 사례들이 선택적으로 관측되지 않았다면 관측된 자료에 따른 추정은 심각한 편의를 포함할 수 있다.

이와 같이 손상된 자료로부터 발생할 수 있는 추정 편의를 개선하기 위하여 여기서는 두 가지 기법을 활용하였다. 첫째, 관측치와 공변인간의 관계에 기초하여 비관측사례를 추정하는 방법으로 EM 알고리듬을 활용하였다. 둘째, 관찰의 선택성에서 비롯된 추정 편의를 개선하기 위하여 이단계(two stage) 모형을 활용하였다.

1. 서론

사회과학에서 조사연구는 사회적 현상의 제 측면에 대한 이해를 촉진하기 위한 자료들을 수집하는데 널리 이용되는 방법이다. 오늘날 실시되고 있는 거의 대부분

* 서울대학교 사회과학연구원 사회발전연구소

의 조사는 전수조사가 아니라 표본조사이다. 통계학과 확률이론에 바탕을 둔 표본추출이론은 표본값이 모집단의 추정치로서 대표성을 유지하는 것을 중요한 원칙으로 하고 있다. 그런데 표본조사자료의 왜곡이나 손상에 따라 표본의 확률적 속성이 유지되기 어려운 경우가 흔히 발생한다. 예를 들어 소득 조사의 경우 많은 조사대상들이 소득에 대한 질문에 응답하지 않거나 허위적으로 응답하는 경우가 흔하다. 만일 관측되지 못한 소득의 사례가 상당히 많고, 특히 고소득층의 응답률이 저조하고, 허위적 보고가 심각하다면, 관측된 소득에 기초하여 인과적 추정을 시도하는 것은 큰 오류를 범하기 쉽다. 불완전한 자료의 또 다른 예로 동일 대상을 다른 시점에 반복하여 조사하는 종단적 연구에서도 자료손상의 위험은 크다. 조사기간동안 상당한 비율의 표집 대상들이 거주지를 이동할 수 있으며, 노인조사의 경우에는 사망이나 입원 등의 이유로 후차적 조사에서 재조사되지 않는 경우가 많다. 결과적으로 조사 완료 후까지 남아있는 대상의 사례 수는 제한적일 뿐 아니라 이들 남아 있는 대상과 조사중간에서 관찰 중지된 사례의 선택성 문제가 심화된다. 다른 한 예로 부모-자녀간 세대관계에 대한 연구는 흔히 가구조사의 가족표(family roster)를 이용하는데 가족표를 통하여 구성된 자녀의 수는 실제 자녀의 수에 미치지 않는 경우가 흔하다.

이와 같이 표본설계나, 문항의 특성에 따라 발생하게된 자료 손상의 문제와 그에 대한 보완기법에 대한 연구는 이제까지 충분하게 이루어지지 못하였다. 관행적으로 이루어진 자료처리 방법은 비응답사례를 찾아내어 그 응답을 메꾸는 것이거나 조사분석시 이들 손상된 자료를 배제하는 것이었다. 만일 자료 손상의 정도가 미미한 경우라면 이와 같은 방법이 큰 문제가 되지는 않는다. 그러나 비응답사례의 수가 많거나 관측의 선택성이 의심될 경우라면 관측된 사례만을 통한 추정은 오류를 가질 수 있다.

여기서는 일본 노인부모가 그들의 자녀와 어느 정도 함께 살고 있거나 가까이 살고 있는지를 살펴보고 이와 같은 부모-자녀간의 근접성이 자녀의 특성에 따라 어떻게 다른지를 연구하는 과정에서 직면하게 된 자료손상의 문제점과 이에 대한 보완 기법을 제시하는 목적을 가진다. 이 연구에서 활용한 자료는 일본 후생성이 1989년 수행한 〈가족주기와 가구형태에 대한 인구학적 조사〉로 전국단위의 일반 가구조사이다.

2. 비응답사례의 메커니즘

먼저 〈가족주기와 가구형태에 대한 인구학적 조사〉를 이용하여 노인부모와 자녀와의 지리적 근접성을 연구할 때 왜 자료 손상 문제가 발생하였는지를 간략하게 설명하고자 한다.

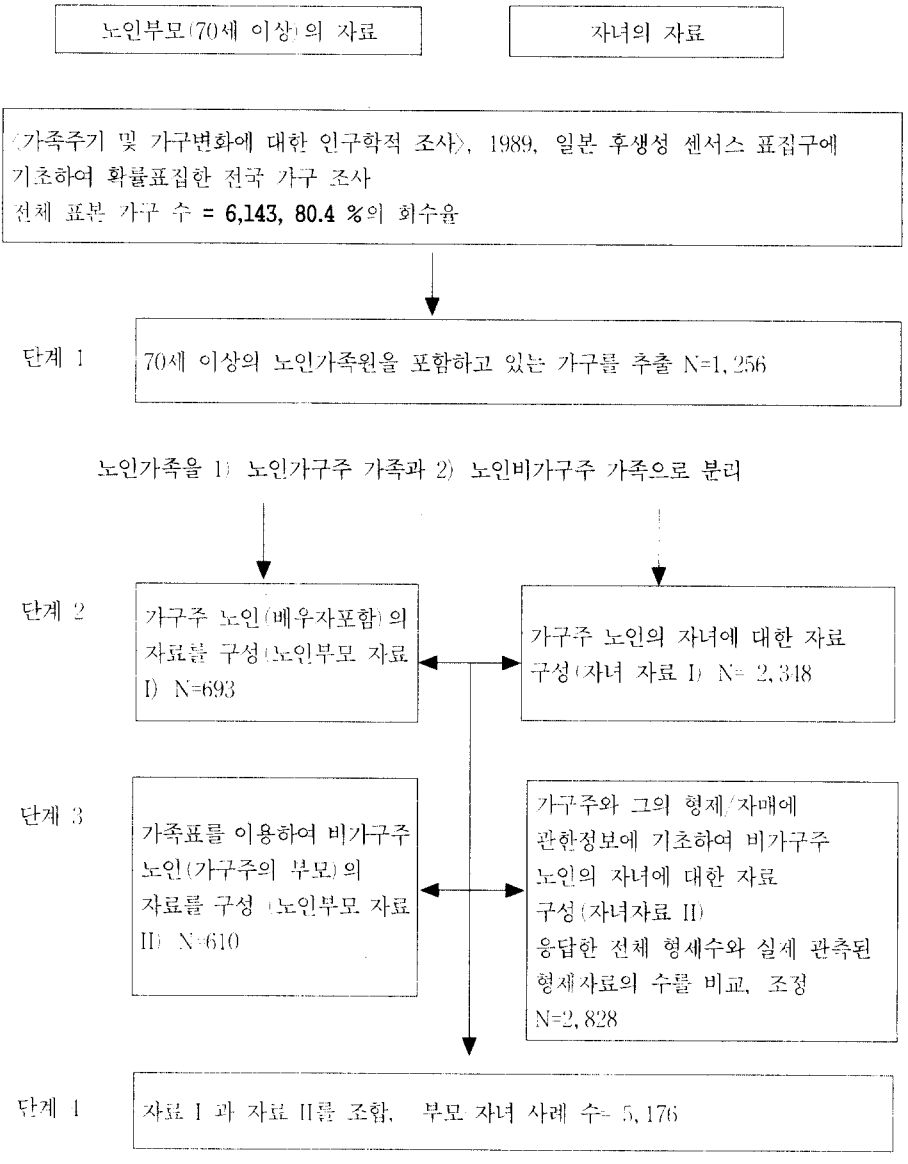
〈표 1〉은 노인부모-자녀의 조응된 사례를 구성하게 된 절차를 도표화한 것이다. 〈가족주기와 가구형태에 대한 인구학적 조사〉는 일본 후생성이 매 2-3년마다 정기적으로 일반 가구의 특성과 가족관계를 파악하기 위하여 가구주 대상으로 실시하는 전국단위의 가구표본조사이다. 이 연구가 활용한 조사는 1989년 실시된 것으로서 80.4%의 회수율을 통하여 총 6,143가구의 조사가 완료되었다. 본 분석에서는 70세 이상의 노인부모와 그들의 자녀에 대한 자료를 새로이 구축하기 위해서 전체표본가구 중 70세 이상의 노인을 포함하고 있는 가구만을 추출하였는데 70세 이상의 노인을 포함하고 있는 가구는 전체 1,256가구에 이르렀다.

다음 단계로 추출된 가구를 노인부모의 가족내 위치를 중심으로 노인부모가 가구주인 경우와(노인가구주) 가구주의 부모의 경우로(노인비가구주) 구분하였다. 이와 같이 자료를 구분한 것은 본 조사에서는 노인부모의 자녀에 대한 정보를 노인의 가족 내의 위치에 따라 상이한 가족표에서 추출할 수 있도록 설계되어 있기 때문이다. 전체 노인가구 중 노인부모가 가구주인 경우는 692 가구였으며 노인부모가 가구주의 부모인 경우는 610 가구였다. 이 두 가구형태에 상응하여 노인의 성, 연령, 건강상태 등에 대한 자료를 구축하였다.

다음으로 노인부모의 자녀에 대한 정보는 노인가구주 자료의 경우 자녀에 대한 가족표를 이용하였고 노인비가구주 자료의 경우 가구주 자녀 자신과 가구주의 형제에 관한 정보를 이용하여 구성하였다. 가구주 노인의 자녀로서 추출된 사례는 전체 2,348명이며 비가구주 노인의 자녀로서 추출된 사례는 전체 2,282명에 이른다.

다음 단계로 이와 같이 추출된 자녀의 수가 실제 노인부모의 자녀 총수와 일치하는지를 검토하였다. 자녀의 총수로 간주된 값은 본 조사가 다른 문항에서 묻고 있는 가구주의 자녀수와 형제수에 대한 응답에 기초하여 산정되었다. 비가구주 노인의 자료에서, 노인의 실제 자녀수는 가구주의 형제수에 가구주 자신을 합산한 값으로 측정하였다. 이와 같이 산정된 자녀 총수와 비교하여 가족표를 통하여 관측된

〈표1〉 노인부모-자녀간 자료 구축과정



〈표2〉 노인부모-자녀간 지리적 근접성의 비관측률

	전체 노인-부모 표본	가구주	
		노인	자녀
		노인-부모 표본	노인-부모 표본
부모-자녀 근접성의 사례수	5,176	2,348	2,828
부모-자녀 근접성의 비관측률(%)	22.3	18.0	25.8

자녀의 수가 작을 때에는 그 차이만큼의 사례를 비응답으로 코딩하였다. 마지막으로 노인부모의 사례와 자녀의 사례를 연결하여 구축한 부모-자녀의 전체 사례수는 5,176이다.

이와 같이 구축된 부모-자녀의 자료를 이용하여 본 연구는 노인부모와 자녀간 지리적 근접성이 가족내 그리고 가족별로 어떻게 분화되어 있는가를 살펴보고자 하였다. 도시화와 가족관계의 변화에 따라 세대간 동거율이 저하되고 있는 한편 일정의 지리적 거리내에서 세대간 빈번한 교류형태를 유지하는 현상이 증가하고 있는 것에 비추어 본 연구에서는 노인부모-자녀간 지리적 근접성이 자녀의 특성에 따라 어떻게 다른지를 살펴보고자 하였다.

그런데 〈표2〉에서 볼 수 있듯이 노인부모-자녀간 지리적 근접성에 대한 항목에서 전체 사례 중 22.3%가 관측되지 못하였다. 또한 노인부모가 가구주일 경우와 비가구주일 경우에 관측률의 차이가 크게 나타났는데, 노인부모가 가구주일 경우 부모-자녀간 근접성의 사례에서 18.0%가 관측되지 못한 반면, 노인부모가 비가구주일 경우 25.8%의 사례가 관측되지 못하였다. 여기서는 제시하지 않았지만 부모-자녀간 근접성 이외에도 자녀의 성, 결혼상의 지위, 연령 등의 항목에서도 많은 비응답사례가 나타났으며 노인부모가 가구주일 경우에 비하여 비가구주일 경우 비관측률이 훨씬 크게 나타났다. 비가구주 노인의 자료에서 관측률이 상대적으로 저조한 것은 노인의 자녀에 대한 정보는 가구주 자녀가 응답한 형제에 대한 정보에 기초하는데, 가구주 자녀는 형제에 관한 정보를 모두 응답하지 않는 경우가 많기 때문으로 이해된다.

노인비가구주 자료에서 부모-자녀간 근접성에 대한 관측률은 〈표 3〉에서 보는 것

〈표3〉 자녀의 특성에 따른 부모-자녀간 근접성의 비관측률의 차이

	부모-자녀간 근접성의 비관측률(%)	
	노인가구주 N=2,348	노인비가구주 N=2,828
한계분포	18.0	25.8
자녀의 특성		
성		
남	15.3	5.5
여	16.0	11.1
출생년도		
1940년 이전	17.9	9.4
1940~1950년	16.2	5.5
1950년 이후	13.8	12.6
결혼상의 지위		
혼인	7.2	7.6
비혼인	5.6	9.5
출생순위		
장자	15.7	5.6
기타	15.9	9.0
형제수		
0	7.4	0.0
1	12.1	43.3
2	19.3	31.0
3	18.4	27.2
4	16.9	28.5
5	17.5	18.8
6	24.0	23.8
7+	25.7	26.8

참고:

처럼, 그 자녀의 성, 결혼상의 지위, 출생순위, 그리고 형제수에 따라 유의한 차이를 보이고 있다. 노인부모의 자녀 중 아들보다 딸에서 비관측률이 크다. 자녀의 결혼상태에 관련하여 결혼한 자녀에 비하여 결혼하지 않는 자녀에서 비관측률이 높게 나타난다. 출생순위의 경우 맏의 자녀에서 비관측률은 상대적으로 낮게 나타난다.

자녀수별 관측률의 차이도 뚜렷하다. 자료가 구축된 특성상 한 명의 자녀를 둔 비가구주노인은 완전하게 관찰되고 있다. 한편 자녀가 두 명일 경우 43.3%의 사례에서 부모-자녀간 근접성이 관측되지 않고 있으며, 세 명의 자녀를 가지고 있는 경우 31.0%의 사례에서 부모-자녀간 근접성이 관측되지 않고 있다. 따라서 두 명 이상의 자녀를 둔 가구에서 대략 한 명의 자녀에 대한 정보가 관찰되지 않고 있다고 풀이할 수 있다.

이렇게 노인부모-자녀간 근접성에 대한 응답률이 공변인의 관측된 값에 따라 다르게 나타나고 있지만, 유의할 점은 근접성에 대한 관측률은 공변인 자체의 관측률과 보다 큰 연관을 가지고 있는 것이다. 〈표 3〉에서 볼 수 있듯이 근접성에 대한 비관측률의 한계분포와 공변인의 값에 따른 관측율의 행분포가 상당한 차이를 가지는데 이것은 비응답사례가 여러 변인에서 동시적으로 발생하고 있는 것을 반영하는 것이다.

비관측사례의 유형에 관련하여 또 하나 지적할 것은 비관측사례가 응답자의 속성에 영향을 받을 수 있는 점이다. 사실 〈표 3〉에서 제시한 것처럼, 부모-자녀간 근접성, 자녀의 성, 연령, 결혼상태 등 여러 항목이 동시적으로 관측되지 않는 경향은 특정 가구내의 정보가 선택적으로 관찰되지 않는 것을 의미한다. 부모-자녀의 사례는 가족별로 자녀의 수에 따라 층화된 것임을 고려할 때, 가족의 특성이 사례의 관찰가능성에 영향을 줄 수 있는 것이다. 특히 주정보원, 다시 말하여 가구주나 그 배우자에 관련된 특성에 주의할 필요가 있다. 예를 들어, 응답자의 건강상태는 조사응답률에 유의한 영향을 미칠 수 있다. 교육수준이 낮은 응답자는 질문의 의미를 이해하기 어려울 수 있으며 이에 따라 조사에 낮은 참여도를 지닐 수 있다.

이와 같은 고려 속에서 〈표4〉에서 보는 것과 같이 노인부모와 자녀와의 근접성이 가구주와 배우자의 특성에 따라 어떻게 다른지를 살펴보았다. 가구주의 연령에 따르면, 60세 이상의 가구주의 응답률이 그보다 나이가 적은 가구주에 비해 높게 나타난다. 이는 가구주가 고령일수록 그들의 자녀 및 형제에 대하여 응답하는 경향이 높은 것을 가리킨다. 가구주의 건강요인에 따라서는 응답률이 유의하게 다르지 않다. 가구주의 교육수준이 높을수록 응답률이 높게 나타난다. 배우자의 연령, 교육수준에 따른 응답률의 차이는 가구주의 경우와 비슷한 형태로 나타나고 있다. 마지막 칼럼에 제시된 비응답항목의 수는 노인부모-자녀간 근접성에 영향을 미친다고 고려된 공변인 중 많은 비응답사례를 포함하고 있는 노인의 건강과 자녀의 성,

〈표4〉 가구주의 특성에 따른 응답률의 차이

	가구주의 결혼상태			
	혼인		비혼인	
	사례수	부모-자녀간 근접성의 응답률(%)	사례수	부모-자녀간 근접성의 응답률(%)
가구주의 특성				
연령				
60세 미만	2134	74.3	318	74.8
60세 이상	2103	79.8	621	84.1
건강상태				
빈약	589	79.8	191	79.6
양호	3648	76.6	748	81.3
교육수준				
8년 미만	2011	76.1	556	80.9
8-11년	1473	77.8	262	79.2
12년 이상	624	79.5	73	82.2
배우자의 특성				
연령				
60세 미만	2272	75.8	-	-
60세 이상	1965	78.5	-	-
건강상태				
빈약	640	77.7	-	-
양호	3597	76.9	-	-
교육수준				
8년 미만	2046	76.6	-	-
8-11년	1636	78.8	-	-
12년 이상	333	77.2	-	-
비응답항목수				
0	2828	94.1	602	94.7
1	762	73.5	210	80.9
2	177	23.2	45	40.0
3	424	0.2	76	2.6
4	46	0.0	6	0.0
전체	4237	77.0	939	80.9

결혼상의 지위, 그리고 연령 등의 요인이 어느 정도로 동시적으로 관찰되지 않는가를 측정한 값이다. 비응답항목의 수가 "0"인 것은 이 들 공변인들의 값이 모두 관찰되는 경우를 가리키며 "4"는 이 들 공변인들의 값이 모두 관찰되지 않은 경우를 가리킨다. 표에서 분명하게 알 수 있듯이 공변인 중 비응답 항목의 수가 커짐에 따라 부모-자녀간 근접성도 관측되지 않는 경향이 강하다. 이는 앞에서도 지적한 것처럼, 여러 항목이 동시적으로 관측되지 않는 경향을 반영한다.

종합하면 〈가족주기와 가구형태에 대한 인구학적 조사〉에서 발생한 자료손상 구조는 복합적이다. 첫째, 가구주 중심의 조사 설계방식에 따라 비가구주 노인과 그들 자녀의 정보를 추출하는데 제약이 따랐다. 둘째, 지리적 근접성에 대한 관측률은 자녀의 성, 결혼상의 지위, 출생순위 등의 공변인의 관측된 값에 따라 유의한 차이를 보이고 있다. 셋째, 지리적 근접성에 대한 관측률은 응답자의 조사에의 참여도와 같은 외재적 특성에 따라 영향을 받고 있다.

불완전한 자료의 문제를 개선하기 위해 적절한 기법을 선택하는데 무엇보다 중요한 기준은 자료가 손상된 메커니즘이다. Little과 Rubin(1987)은 손상된 자료의 구조를 MCAR(Missing completely at Random), MAR(Missing at Random), 그리고 MNAR(Missing not at Random)로 유형화하였다. 노인의 자녀와의 지리적 근접성(y)과 자녀의 성(x)과의 관련을 예로 들어보자. MCAR 조건에서는 지리적 근접성을 관찰할 수 있는 확률이 지리적 근접성의 실재적 값과 자녀의 성별특성에 독립적이다. 이와 같은 경우, 비관측된 사례는 완전한 의미에서 무작위적이며 지리적 근접성의 관측된 사례와 비관측된 사례는 모두 지리적 근접성이라는 모집단 분포의 무작위적 하위표본을 구성하게 된다. 따라서 MCAR의 조건에서는 관측된 근접성의 사례만을 이용하거나 단순 형태의 imputation 기법을 통하여 공변인들의 계수를 일관되게 추정하는데 큰 무리가 따르지 않는다. 하지만 이 경우에도 표준오차는 과대평가될 경향이 큰 데 비응답사례 때문에 표본크기가 감소하기 때문이다.

MAR 조건에서는 지리적 근접성을 관찰할 수 있는 확률은 그 자녀의 성에 따라 다르지만, 근접성의 실재적 값에는 독립적인 성격을 가진다. 이 경우, 근접성의 관측된 값은 반드시 무작위적 표본의 속성을 지니지는 않는다. 대신, 근접성의 관측된 값은 공변인(여기서는 성)의 값에 따라 층화된 각각의 하위 집단내에서 무작위적 표본의 값을 가진다. 〈표 3〉에서도 살펴보았듯이, 부모-자녀간 근접성에 대한 관측률은 그 자녀가 딸보다 아들인 경우에 높게 나타나고 있는데, 이는 근접성의 응

답률이 어느정도 공변인의 관측된 값에 따라 층화되고 있음을 반영한다. Little과 Rubin(1987)은 MCAR과 MAR 상태의 불완전한 자료에서는 우도 혹은 무작위 추정 기법을 이용하거나 표준편차를 재조정함으로써 모수를 일관되게 추정할 수 있다고 주장한다.

한편 MNAR 조건에서는 부모-자녀간 근접성이 관측될 확률이 그 잠재적인 값에 영향을 받는데 예를 들자면 멀리 떨어져 있는 자녀의 거리정도를 관찰할 수 있는 확률은 가까이 사는 자녀의 근접성을 관찰할 수 있는 확률보다 약한 경우를 상정해 볼 수 있다. 이 경우 비응답사례의 메커니즘을 간과하고 관측된 자료만을 이용하여 공변인의 계수를 추정하는 것은 오류를 가질 위험이 크다.

정리해보면 노인부모-자녀간 근접성의 자료가 손상된 구조는 MAR 조건과 MNAR 조건의 특징을 공유하고 있다. MAR 조건과 관련하여, 부모-자녀간 근접성을 관측할 수 있는 경향은 자녀의 성, 결혼상태, 출생순위 등에 따라 유의한 차이를 가지고 있다. MNAR 조건과 관련하여, 근접성에 대한 관측률이 가구주의 특성과 다른 공변인들의 비응답 유형에 밀접하게 연관되어 있는 것을 지적할 수 있다. 다음 두 절에서는 각각의 조건에서의 자료 손상 상태를 개선하기 위하여 EM알고리듬과 2단계(Two Stage) 모델을 활용한다.

3. MAR 조건에서의 자료손상과 EM 알고리듬 자료보완기법

우선 부모-자녀간 근접성에 대한 관측률이 자녀의 성별요인과 같은 공변인의 값에 따라 다르지만 근접성의 실재적 값에는 독립적인 상태(MAR)에서의 자료 보완기법을 모색해 본다. 여기서는 특정 모델을 가정한 뒤 관찰된 자료의 우도 함수를 최대화하는 "우도에 기초한 추정방법(likelihood-based inference method)"을 활용하였다. EM 알고리듬은 불완전한 자료에 대한 우도추정에 유용한 기법이다(Little and Rubin, 1987:129). EM 알고리듬의 과정을 간략하게 설명하면, 먼저 E(expectation) 단계로, 관측된 지리적 근접성의 사례들을 다른 공변인의 값에 따라 회귀시킨 다음 근접성의 확률누적분포를 추정한다. 그런 뒤에 비관측된 근접성의 잠재적 값을 공변인의 추정계수를 이용하여 무작위로 추정한다. 다음 M(maximization) 단계로, 근접성의 관측된 값과 추정된 값을 모두 공변인들로 회

〈표5〉 불완전한 자료 결과와 EM 알고리듬을 통한 최우도 추정치의 비교

A. 노인 가구주

노인부모-자녀간 근접성	불완전한 자료 결과			EM 알고리듬 결과		
	근거리 vs 동거	원거리 vs. 동거		근거리 vs. 동거	원거리 vs. 동거	
자녀의 특성	beta	beta	t	beta[†]	beta[†]	
			2			
출생순위						
장자(준거: 기타)	-0.43 **	-0.63 ***		-0.38 *	-0.61 ***	
성						
아들(준거: 딸)	-0.78 ***	-0.81 ***		0.76 ***	-0.80 ***	
결혼상의 지위						
비혼인(준거:혼인)	-2.51 ***	-2.44 ***		-2.49 ***	-2.45 ***	
형제수						
두명이하(준거: 두 명 이상)	-0.71 ***	0.87 ***		-0.65 ***	-0.83 ***	

B. 노인 비가구주

노인부모-자녀간 근접성	불완전한 자료 결과			EM 알고리듬 결과		
	근거리 vs. 동거	원거리 vs. 동거		근거리 vs. 동거	원거리 vs. 동거	
자녀의 특성	beta	beta	t	beta[†]	beta[†]	
			2			
출생순위						
장자(준거: 기타)	-1.24 ***	-1.51 ***		-1.27 ***	-1.05 ***	
성						
아들(준거: 딸)	-2.76 ***	-2.62 ***		-2.79 ***	-1.84 ***	
결혼상의 지위						
비혼인(준거:혼인)	-1.07 ***	-0.97 ***		1.13 ***	-1.61 ***	
형제수						
두명 이하(준거: 두명이상)	0.86 ***	1.26 ***		-0.77 ***	0.55 ***	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

참고: t는 EM 알고리듬의 회수를 가리킨다.

귀하여 새로운 추정치를 계산한다. 이러한 알고리듬을 모수의 추정치가 수렴될 때까지 반복한다.

앞에서도 살펴보았듯이, 부모-자녀간 지리적 근접성에 대한 관측률은 가구주가 노인부모일 경우와 그렇지 않은 경우에 상당히 다르기 때문에, 이 알고리듬을 두 집단에서 개별적으로 실시하였다. 〈표5〉는 불완전한 자료를 그대로 이용한 경우와 EM알고리듬을 이용한 경우에 분석된 공변인의 효과를 요약한 것이다. 종속변수는 노인과 자녀간의 지리적 근접성으로 동거, 가까이 사는 경우(동일 시내), 멀리 떨어져 사는 경우로 삼분류하였으며, 동거를 준거집단으로 설정하였다. EM 분석결과에서 제시된 t는 반복된 EM 알고리듬의 회수를 가리키며, bt는 마지막 알고리듬에서 산출된 공변인들의 회귀계수를 가리킨다.

불완전한 자료와 EM 알고리듬을 통하여 분석된 회귀계수들의 효과를 비교해보면, 노인가구주의 경우 두 분석의 결과가 유의한 차이를 가지지 않는 것으로 나타난다. 그러나 노인비가구주의 경우, 부모-자녀의 원거리 경향에 대한 추정과 관련하여 두 모형에서 추정된 계수의 크기가 유의한 차이를 가지고 있는 것을 알 수 있다. 물론 계수의 통계적 유의성과 관련하여 두 모델은 차이가 크지 않아 각 변인들의 통계적 유의성을 판단하는데 있어서 오류의 정도는 심각하지 않다고 이해될 수 있다. 그러나, 각 변인의 상대적 효과(계수의 크기)가 두 모델에서 차이가 큼은 불완전한 자료를 활용하여 변인의 상대적 효과를 분석할 때 편의된 해석으로 이를 위험이 있음을 가리킨다.

EM 알고리듬을 통하여 추정된 회귀계수들의 효과를 간략하게 설명하면 자녀와의 동거율은 자녀가 맏일 경우, 딸보다 아들인 경우, 자녀가 결혼한 경우보다 결혼하지 않은 경우에 유의하게 강하다. 이러한 경향은 일본의 전통적인 장자 중심의 가족부양의 관행을 잘 반영하는 것이며, 결혼이전 부모의 집으로부터 분가의 경향이 강한 미국사회와는 달리, 자녀의 분가가 결혼시기까지 지연되는 동양권 사회의 독특한 청년기 생애사를 반영한다.

자녀수의 효과를 살펴보면 자녀수가 적을수록 해당자녀와 동거하는 경향은 가까이 살거나 멀리 사는 경향에 비하여 더 큰 것으로 나타나고 있다. 이러한 자녀수의 효과는 분석단위와 관련하여 쉽게 이해가 된다. 본 연구는 부모와 그의 모든 자녀가 일대일 조응된 부모-자녀사이의 근접성을 관찰사례로 파악하고 있기 때문에 분석단위는 노인부모가 아니라 해당 자녀가 된다. 부모의 입장에서는 자녀 수가 많을

〈표6〉 EM 알고리듬을 통한 최우도추정 전후의 부모-자녀간 근접성의 분포

부모-자녀간 근접성	전체		노인가구주		노인비가구주	
	추정전	추정후	추정전	추정후	추정전	추정후
동거	29.5	24.1	20.8	18.1	36.1	28.3
근거리	26.1	21.5	28.9	25.1	24.0	19.0
원거리	44.4	54.4	50.2	56.8	39.6	52.6
전체	100.0	100.0	100.0	100.0	100.0	100.0
N	4,023	5,176	1,925	2,348	2,098	2,828

수록 적어도 한 자녀와 함께 살거나 가까이 살 경향이 증가하지만, 자녀의 입장에서는 형제가 많을수록 해당 자녀가 부모와 함께 살거나 가까이 살 경향은 확률적으로 감소할 것이다.

〈표6〉은 알고리듬을 통하여 추정된 근접성의 분포를 추정전의 분포와 비교한 것이다. 전체 표본의 경우, 추정 전과 비교하여 추정 후 부모와 동거하는 자녀의 비율은 29.5%에서 24.1%로 감소하였으며 부모와 가까이 사는 비율도 26.1%에서 21.5%로 감소한 반면, 부모와 멀리 떨어져 사는 경우는 44.4%에서 54.4%로 크게 증가하였다. 이와 같은 원거리 자녀의 분포가 크게 증가한 것은 지리적 근접성에 대한 비관측률이 가까이 사는 자녀에 비하여 멀리 떨어져 사는 자녀에서 훨씬 큼을 가리킨다. 전체표본을 노인부모가 가구주일 경우와 비가구주일 경우로 분리하여 살펴본 경우도 추정 전후 비슷한 형태의 분포도 변화가 확인된다. 특히, 노인부모가 비가구주일 경우, 추정전후 멀리 떨어져 사는 자녀의 비율이 크게 변화되는 것을 볼 수 있다.

4. MNAR 조건에서의 자료손상의 선택성(Selection)과 2단계(Two Stage) 모델

앞의 절에서 활용된 EM 알고리듬은 비응답사례가 '약한 의미에서 무작위'적일 경우 손상된 자료의 문제를 개선하는데 유익한 방법이다. 이 방법은 추정치들의 변

이를 고려함으로써 단순 imputation 방법보다 효율적이고 일관된 추정치를 계산하는데 장점을 가진다. 그러나 관측률이 관측치의 실재 값에 영향을 받을 경우 EM 알고리듬은 관측의 '선택성 오류'(selection bias)를 개선하는데 제약을 가진다. 만일 비관측률이 그 관측치의 잠재적 값에 영향을 받을 경우, 관측된 값에 기초하여 우도를 무작위적으로 추정하는 방법은 오류를 개선하기보다 악화시킬 수 있는 위험을 지니고 있기 때문이다. 앞의 절에서 살펴본 것과 같이, 본 자료에서는 여러 문항에서 비응답사례가 동시적으로 나타나는 경향이 강하고, 또 EM 추정 전후의 근접성의 분포가 상당히 다른데 이는 비응답률이 그 변인의 값에 따라 유의하게 다르지 않은가에 대한 의혹을 남게한다.

관찰의 선택성은 〈식 1〉의 결과모형과 〈식 2〉의 선택성 모형이 다음과 같은 두 조건에서 서로 강하게 연관되는 것에 따른다. 첫째, 두 모형의 에러(u)가 독립적이지 않다. 둘째, 두 모델에 포함된 공변인들이 서로 연관되어 있다. 만일 두가지 조건에서의 연관성이 통계적으로 유의하지 않으면, 자료가 크게 손상된 경우에도 일반적인 선형 혹은 비선형 회귀식은 비편의된(unbiased) 추정치를 계산할 수 있다 (Lee, 1982).

선행 연구들은 계수와 표준오차의 추정에서의 오류를 개선하기 위하여 다양한 모형들을 발전시켜왔다(Tobin, 1958; Heckman, 1976, 1979; McDonald and Moffitt, 1980; Berk, 1983; Ronceek, 1992). 이중 Heckman's 이단계 모형(1976, 1979)은 그 방법이 용이하고 여러 형태의 선택성 관찰 상황에서 일관되고 효율적인 추정치를 계산할 수 있는 장점이 있기 때문에 널리 활용되고 있다.

선택(Selection) 모형과 결과(Outcome) 모형의 연관

$$y1i^{*} = X1ib1i + u1i \quad \text{if } y2i^{*} > 0 \quad \cdots\cdots \text{(식 1)}$$

$$y2i^{*} = X2ib2i + u2i \qquad \cdots\cdots \qquad \text{(식 2)}$$

Heckman의 이단계 모형은 선택성 오류(selection bias)를 정규 회귀모형에서의 규정오류(specification error)로 다루고 있다. 다시 말하여 모델이 선택성(selection)

지표를 적절하게 규정되지 못하면, 오차(u)의 평균은 0이 아니고 회귀계수도 정확하게 추정되지 못하는 것이다. 이러한 선택성 오류(selection bias)의 개선책은 선택성 관찰(selective observation)에 대한 최적의 지표(λ)를 구성하여 그것을 〈식 3〉과 같은 결과모형식 안에서 통제하는 것이다.

관측의 선택성(selection) 지표(λi)를 포함한 결과(Outcome) 모형

$$y1i = X1ib1i + \rho\lambda i + e1i \quad \text{where } E(e1i) = 0 \quad \cdots\cdots \text{(식 3)}$$

$$\lambda i = \frac{\varphi(X2ib2i)}{\Phi(X2ib2i)}$$

여기서 람다(λi)는 역 Mill's ratio를 가리키는데 이는 관측률의 정규분포(Φ)에 대한 밀도함수(φ)의 비를 의미한다.

그런데 Heckman의 모형을 활용하는데 몇 가지 주의할 사항이 있다. 우선 이 모형은 선택성 모형과 결과모형이 동일한 공변인을 포함하거나 선택성 모형의 공변인이 결과모형에 포함된 공변인의 하위집단으로 구성될 경우 심각한 규정성 문제(identification problem)를 가지게 된다. 또 다른 문제는 선택성 모형에 포함된 공변인들이 모두 가변수들(dummy variables)로 이루어질 때, 추정된 람다는 한정된 분포를 가지게 되며 모형간의 상관도가 상수의 효과와 혼재되게 된다. 따라서, 모델의 적정성을 제고하기 위하여, 선택성 모형의 공변인들은 관측률에 대하여 강한 설명력을 지녀야 하며, 이 들 공변인은 결과모형의 공변인과 유의한 관계를 지녀서는 안 된다(Little and Rubin, 1987:230).

2단계 모형을 활용하는데 또 하나 유의할 사항은 종속변수의 속성이다. 선택성 오류(selection bias)에 대한 대부분의 선행연구들은 관찰이 중지(censored)되거나 단절된(truncated) 선형종속변수에 초점을 두어왔다. 한편 명목변수나 제한된 서열변수에 대한 연구는 상대적으로 취약한 편이지만 Lee 등의 일부 경제학자들은 다범주 종속변수의 분석에서 잘못 규정된 변인으로부터 파생할 수 있는 오류를 분석하였다(Lee, 1982). 본 연구에서는 부모-자녀간 지리적 근접성의 선택성 관찰(selective observation)의 문제를 개선하기 위하여 Lee의 연구를 원용한다.

선택성 관찰(selective observation)의 문제를 간과함으로써 잘못 규정된 근접성의 한 범주의 확률은 다음과 같이 수식화할 수 있다.

$$p(y=i \mid x)=\frac{Exp(\alpha i0+x\alpha i1)}{1+\sum exp(\alpha i0+x\alpha i1)},\ i=1,2 \cdots\cdots(식\ 4)$$

여기서 i는 부모-자녀간 근접성의 값으로 1은 근거리, 2는 원거리를 가리키면 0은 기준범주(basic category)로서 동거를 가리킨다. 이와 같은 확률모형에 상응하는 로지스틱 모형은

$$ln=\frac{p(y=i \mid x)}{p(y=0 \mid x)}=\alpha i0+x\alpha i1,\ i=1,\ 2\cdots\cdots(식\ 5)$$

이다. 여기서 근접성에 대한 관측률이 근접성의 범주와 공변인의 값에 영향을 받는 가장 심각한 선택성 오류(selection bias) 상황을 가정해보자. 이에 상응하는 근접성에 대한 관측률은

$$p(r=1 \mid x, y=i)=\frac{exp(\delta 0+x\delta 1+\beta i)}{1+exp(\delta 0+x\delta 1+\beta i)},\ i=1,2\cdots\cdots(식\ 6)$$

이다. 여기서 δ는 공변인 x의 한 단위 증가에 따른 관측률의 변화의 로그 odds값을 가리키며 β는 준거가 되는 근접성과 비교되는 특정 근접성간의 관측률의 로그 odds값이다. 여기서의 관심은 근접성의 관측률이 근접성의 범주와 공변인에 영향을 받는 것을 간과할 때 어떻게 추정의 오류가 발생하는가 하는 점이다.

Bayesian의 이론에 따르면(Lee, 1982), 선택성 관측을 전제로 할 때 특정 근접성의 로지스틱 확률은 관측률의 Bayesian 함수와 공변인의 정규 로짓 함수($\alpha i0+x\alpha i1$)로 구성된다.

$$\begin{aligned} ln\frac{p(y=i \mid x, r=1)}{p(y=0 \mid x, r=1)} &= ln\frac{p(r=1 \mid x, y=i)}{p(r=1 \mid x, y=0)}+ln\frac{p(y=i \mid x)}{p(y=0 \mid x)} \cdots\cdots(식\ 7) \\ &= ln\frac{1+exp(\delta 0+x\delta 1)}{1+exp(\delta 0+x\delta 1+\beta i)}+\alpha i0+x\alpha i1 \end{aligned}$$

〈표7〉 관측률[a]의 선택성(selection)에 대한 요인분석과 선택성 지표(λ)

가구주의 결혼상의 지위	혼인(N=4,237)			비혼인(N=939)		
공변인	b	z		b	z	
가구주의 특성						
연령						
60세 이상(준거: 60세 미만)	0.40	2.02	*	0.49	1.95	*
건강상태						
빈약(준거: 양호)	-0.09	-0.58		-1.12	-4.08	***
교육수준						
고등학교 이상(준거: 〈고등학교)	0.12	0.89		-0.15	-0.62	
배우자의 특성						
연령						
60세 이상(준거: 60세 미만)	-0.60	-3.01	**	-	-	
건강상태						
빈약(준거: 양호)	-0.08	-0.56		-	-	
교육수준						
고등학교 이상(준거: 〈고등학교)	0.08	0.84		-	-	
비응답항목수						
한개	-1.72	-14.82	***	-1.60	-6.05	***
두 개 이상(준거: 없음)	-5.51	-30.30	***	-4.77	-15.86	***
상수	2.80	21.22	***	2.98	10.98	***
Pseudo R2		0.47			0.40	
Inverse Mill's Ratio(λ)						
mean		0.24			0.19	
std		0.31			0.27	

참고: a) 종속변수의 이항적 사건은 부모 자녀간 근접성이 관찰된 경우를 1, 관찰되지 않은 경우를 0로 하였다. λ는 부모 자녀간 응답률에 대한 로짓분석을 통하여 추정된 근접성의 관측률의 누적분포(Φ)와 밀도함수(φ)의 비(φ/Φ)를 계산한 값이다. Φ와 φ는 다음과 같은 공식에 따라 추정되었다.

$\Phi = e^{xb}/(1+e^{xb})$,

$\varphi = \Phi' = e^{xb}/(1+e^{xb})^2$

〈표8〉 선택성관찰 지표(inverse Mill's Ratio)를 통제한 결과모형(Outcome Model)

노인부모-자녀간 근접성	근거리 vs. 동거			원거리 vs. 동거		
자녀의 특성	beta	S.E.		beta	S.E.	
출생순위						
장자(준거: 기타)	-0.83	0.11	***	-1.05	0.10	***
성						
아들(준거: 딸)	1.90	0.11	***	-1.84	0.10	***
결혼상의 지위						
비혼인(준거:혼인)	-1.70	0.14	***	-1.61	0.12	***
형제수						
두명 이하(준거: 두명이상)	-0.42	0.12	***	-0.55	0.10	***
선택성 지표(λ)	-0.05	0.01	***	-0.46	0.01	***
상수	2.38	0.18	***	2.92	0.17	***
Pseudo R2						0.11

*** P 〈0.001

위 식에서 선택성 오류(selection bias)의 방향은 근접성의 범주들에 대한 상대적인 관측률의 비(relative risk ratio)에 달려있음을 알 수 있다. 첫째, 상대적 관측률의 비가 1이 아닐 경우, 다시 말하여 관측률이 근접성의 범주간에 동일하지 않을 경우, 상수의 계수(αio) 는 오류를 가지게 된다. 둘째, 공변인(x)이 범주 값의 확률과 관측률에 동시에 영향을 주면 공변인의 계수(αi1)의 신뢰성이 문제가 된다.

이와 같은 다범주 종속변수의 선택성 오류의 발생 메커니즘을 숙지하면서 고안한 2단계 모형은 〈표7〉과 〈표8〉과 같다. 먼저 〈표7〉은 선택성 모형을 요약한 것인데, 근접성에 대한 관측률을 추정하기 위하여 가구주와 배우자의 연령, 건강상태, 교육수준, 그리고 비응답문항의 수를 공변인으로 측정하였다. 위의 요인들은 선택성 관찰(selective observation)의 좋은 지표로서의 두가지 조건을 만족하고 있다. 첫째, 위 공변인은 근접성에 대한 관측률을 강하게 설명하고 있다. 공변인들의 설명력(R2)은 40%를 넘고 있다. 둘째, 위 공변인은 〈표8〉에서 제시한 주 결과모형의 공변인과 유의한 관계를 가지지 않고 있다.

선택성 모형에서 관측률에 영향을 미치는 요인들의 효과를 간략하게 설명하면 가구주 및 배우자가 60세 이상의 경우 그보다 나이가 적은 경우에 비해 관측률이

유의하게 높게 나타난다. 비혼인 상태의 가구주의 건강상태가 양호하지 않은 경우 관측률이 통계적으로 유의하게 낮게 나타난다. 마지막으로 비응답사례의 수의 효과는 비응답이 여러 문항에서 동시적으로 발생하는 경우의 효과를 가리킨다.

이와 같은 공변인들의 회귀계수를 통하여 정규 누적분포(Φ)와 밀도함수(φ)를 구한 다음 두 함수의 비로서 역의 Mill's Ratio(λ)를 구하였다.

다음 단계로 결과모형에 이 람다값을 다른 공변인들과 함께 통제하면서 노인부모-자녀간의 근접성의 로짓값을 분석하였다. 〈표8〉은 분석결과를 요약한 것이다. 람다값의 효과를 먼저 살펴보면 동거의 경우에 비하여 근거리나 원거리에 대한 람다값이 통계적으로 유의하게 낮은데, 이는 근거리나 원거리에 대한 관측률이 동거에 대한 관측률에 비하여 유의하게 작음을 의미한다. 이와 같은 범주간 관측률의 차이를 통제한 후 공변인의 효과를 보면 앞의 EM 분석에서와 유사한 형태로 나타나고 있다. 부모-자녀간 동거율은 자녀가 맏일 경우, 딸보다는 아들인 경우, 결혼한 자녀보다 결혼하지 않은 자녀인 경우에 강하며 형제수가 적을수록 부모와의 동거율이 강하게 나타나고 있다.

5. 결론

여기서는 노인부모와 자녀간 지리적 근접성을 연구하는데 있어서 직면하였던 자료손상의 문제를 분석하고 이에 대한 자료보완 기법을 개발하였다. 불완전한 자료의 문제점을 개선하기 위해 먼저 고려해야 할 사항은 자료가 손상된 메커니즘이다. 이 연구가 당면하였던 자료 손상메커니즘은 복합적이다. 우선 본 연구가 활용한 자료는 노인부모를 대상으로한 조사가 아니라 가구 일반의 특성에 대한 조사에 기초한 것이기 때문에, 노인부모에 대한 충분한 자료를 확보하는데 어려움이 컸다. 또한 가구주 중심의 조사설계방식에 따라 비가구주 노인과 그들 자녀에 관한 정보의 상당수가 관측되지 못하였다. 나아가 비관측된 사례가 여러 항목에서 동시적으로 나타나고 있는데 이는 특정 가구에 대한 정보가 선택적으로 관측되지 않고 있는 것을 의미한다.

이와 같이 복합적인 형태를 띤 자료 손실 구조를 고려하여 EM 알고리듬과 이단계 모형을 활용하여 추정의 적합도를 개선하도록 하였다. EM 알고리듬은 종속변

수에 대한 비관측률이 그 변인의 실제적 값에 독립적이지만 공변인의 값에 따라 다를 경우 발생할 수 있는 추정편의를 개선하는데 유용한 기법이다. 이 방법은 단순 imputation 방법과는 달리 추정치들의 변이를 고려함으로써 효율적이고 일관된 추정치를 계산하는데 이점을 가진다. EM 분석결과를 불완전한 자료의 분석결과와 비교하면, 비가구주노인의 자녀와의 원거리 경향에 대한 공변인의 효과가 차이가 큰 것을 확인할 수 있었다. 또한 EM 알고리듬에 따라 추정된 근접성의 분포를 추정전 관측된 근접성의 분포와 비교하면, 원거리의 비율이 두 분석결과에서 차이가 크게 나타났다. 이는 멀리 떨어져 사는 자녀는 함께 살거나 가까이 사는 자녀에 비해 관측되지 않은 경향이 강하며, 이에 따라 관측된 사례에 기초한 원거리 추정치가 편의를 가질 위험이 큼을 시사한다.

이단계모형(two stage models)은 근접성의 선택적 관찰에서 파생되는 추정치의 오류를 개선하는데 유용한 모형이다. 1단계로 관찰의 선택성에 유의한 영향을 미치는 요인으로서 조사 대상자인 가구주와 배우자의 연령, 교육수준, 건강상태 등의 효과를 분석하였다. 이들 요인은 근접성에 대한 관측률에 상당히 큰 설명력을 가지고 있음이 분석결과 확인되었다. 2단계로 선택성의 모형(selection model)에서 추정된 공변인의 계수값에 기초하여 선택성 관측률 지표인 Mill's ratio를 구한 다음 이 지수의 효과를 근접성의 결과모형(outcome model)에서 통제하였다. 통제된 Mill's ratio는 동거, 근거리, 원거리에서 유의한 차이를 가지고 있음이 확인되었다.

마지막으로 본 연구의 제한점과 관련하여 지적하고 싶은 것은 손상된 자료의 표본 편의를 개선하는데 EM 알고리듬과 이단계 모형의 상대적 유효성을 판단하지 못한 점이다. 두 기법은 자료 손실 구조를 상당히 다르게 전제하고 있는데, 자료가 손상된 메커니즘을 보다 충분히 검토하지 못하였고 자료손상 구조가 해당 기법의 기본전제들과 어느 정도 조응하는지를 좀더 체계적으로 비교하지 못하였다.

참고문헌

Berk, Richard(1983), "An Introduction to Sample Selection Bias in Sociological Data", *American Sociological Review* 48: 386-398.

Heckman, James(1979), "Sample Selection Bias as a Specification Errors", *Econometrica* 47: 153-161.

_____(1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models", *Annals of Economic and Social Measurement* 5: 475-492.

Maddala, G. S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

McDonald, John and Robert Moffitt(1980), "The Uses of Tobit Analysis", *The Review of Economics and Statistics* 62:318-321.

Lee, L. F. (1982), "Specification Error in Multinomial Logit-Models-Analysis of the Omitted Variable Bias", *Journal of Econometrics* 20: 197-209.

Little, Roderick and Donald Rubin(1987), *Statistical Analysis with Missing Data*, New York: John Willely & Sons.

Ronceck, D. W. (1992), "Learning More from Tobit Coefficients: Extending a Comparative Analaysis of Political Protest", *American Sociological Review* 57:503-507.

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

abstract

EM Algorithm and Two Stage Model for Incomplete Data

Keong-Suk Park

This study examines the sampling bias that may have resulted from the large number of missing observations. Despite well-designed and reliable sampling procedures, the observed sample values in DSFH (Demographic Survey on Changes in Family and Household Structure, Japan) included many missing observations. The head administerd survey method of DSFH resulted in a large number of missing observations regarding characteristics of elderly non-head parents and their children. In addition, the response probability of a particular item in DSFH significantly differs by characteristics of elderly parents and their children. Furthermore, missing observations of many items occurred simultaneously. This complex pattern of missing observations critically limits the ability to produce an unbiased analysis. First, the large number of missing observations is likely to cause a misleading estimate of the standard error. Even worse, the possible dependency of missing observations on their latent values is likely to produce biased estimates of covariates.

Two models are employed to solve the possible inference biases. First, EM algorithm is used to infer the missing values based on the knowledge of the association between the observed values and other covariates. Second, a selection model was employed given the suspicion that the probability of missing observations of proximity depends on its unobserved outcome.