The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.6.933
25(6), 933–945

An Analysis of Response Pattern and Panel Attrition in KLIPS(Korean Labor and Income Panel Study)

Ki-Seong Nam[1] · Young-Min Chun[2]

[1]Center for Employment Survey and Analysis, Korea Employment Information Service
[2]Center for Employment Survey and Analysis, Korea Employment Information Service

(Received July 16, 2012; Revised September 20, 2012; Accepted October 16, 2012)

Abstract

In this paper we used the KLIPS(Korean Labor and Income Panel tudy) data that surveyed from 2006(wave 9) to 2009(wave 12). Other previous studies are concerned with the panel attrition in the early wave, but this study classifies the response pattern and investigates some factors that influence panel attrition when the panel tends to stabilize. It was revealed that panel attrition was influenced by relocation and housing type through the logit model. Besides it was appeared that panel attrition was affected by the monthly living expenses and the overall household income through the decision tree.

Keywords: CAPI, decision tree, logit model, panel attrition.

1. 서론

고정된 조사대상의 전체를 패널(panel)이라 하며, 초기에는 시장조사에서 소비자의 소비행동과 소비태도의 변화 과정을 분석하기 위해서 이용되었는데, 최근에는 여론의 형성과정과 변동과정의 연구에 이용되기도 하고, 직업이동의 궤적을 밝혀내기 위해서 이용되는 등 응용범위가 넓다. 조사대상자를 매년 바꿔가며 실시하는 일반통계 조사와는 달리 동일한 가구를 매년 조사하는 통계조사방법으로 일반통계조사는 일정지역의 거주자를 대상으로 삼아 조사를 하고 있으나 패널조사는 한 지역에 살던 사람이 다른 지역으로 이사를 가더라도 다음 조사 때 추적해서 조사한다. 이 때문에 조사대상인 개인이나 가구가 어떻게 변했는가를 쉽게 알 수 있고 그런 점에서 더 과학적이다. 패널조사 방식에 따른 통계를 이용하면 총량변화 속에 내재된 미시적 부분까지도 분석할 수 있는 장점이 있다. 이런 패널 조사는 1968년 미국 미시간대학의 조사연구소가 처음 시작했다. 유럽에서도 독일이 1984년에 데이터를 작성한 이래 다른 나라에서도 조사가 시작됐다.

패널조사는 일정기간을 두고 조사를 하기는 하지만 동일 대상자에게 질문을 하게 되므로, 반복할 때마다 표본의 수가 감소하는 것이 문제가 된다. 즉, 이것이 패널 탈락을 의미한다. 또한 조사기획과 조사과정이 잘 이루어졌을 지라도 무응답은 발생하며, 무응답은 조사 차수가 지남에 따라 조사대상의 탈락으로 발생하는 종단면 결측치와 어느 한 시점에서 특정 변수의 값이 결측되는 횡단면 결측치가 있다. 이러한 무응답에 의한 결측자료는 조사 자료의 신뢰성에 영향을 끼치고, 통계적인 방법으로 완전자료를 구성하

[1]Corresponding author: Center for Employment Survey and Analysis, Korea Employment Information Service, 77-11, Mullae-dong 3-ga, Yeongdeungpo-gu, Seoul 150-093, Korea. E-mail: ksnam62@keis.or.kr

여 정형화된 통계분석을 수행하게 된다. 이와 같이 패널조사와 같은 종단면 연구에서는 웨이브무응답이 발생하며, 무응답을 처리하기 위해 대체방법이나 가중값 조정 등의 방법을 사용하여 무응답으로 인한 편향을 최소화 할 필요가 있다.

선행연구로서 종단면 연구에서 결측치를 다루는 여러가지 연구들이 발표되었고 (Demirtas, 2004; Molenberghs와 Verbeke, 2004; Hogan과 Larid, 1997; Gorbein 등, 1992), Little (1995)은 결측치 연구에 대한 중요한 통계적인 체계를 수립하였고, Hogan 등 (2004)은 결측치 연구에 대한 다양한 적용을 소개하였다.

본 연구에서는 14차까지 진행되어온 한국노동패널(Korean Labor and Income Panel Study; KLIPS)을 이용하여 10년 이상 진행되어 온 이후의 패널의 탈락 현상을 몇 개의 패턴으로 나누어 분석하고자 한다. 조사횟수가 늘어남에 따라 증가하는 표본탈락은 조사의 장기적인 지속가능성을 어렵게 할 뿐만 아니라, 만일 그러한 탈락이 특정집단에 집중되어 있거나 체계적인 패턴을 가지고 있을 때에는 조사의 대표성을 훼손할 수도 있기 때문이다. 또한 본 연구에서의 패널탈락의 분석대상은 9차에서 12차까지를 대상으로 한다. 이는 패널조사에서 패널 구축 이후 초기에는 패널 탈락도 많이 발생하지만, 이에 대한 연구도 비교적 많이 존재한다. 그러나 10여년 정도 패널이 진행되면 패널 대상자도 본인 스스로 패널이라는 사실을 각인하고, 조사에 비교적 협조적이며, 패널탈락도 초기와는 다른 양상일 수 있기 때문이다.

본 연구에서 사용한 분석내용은 각 차수의 무응답 여부에 따른 각 그룹 간에 패턴 여부와 그룹 간에 관련성이 있는 지를 살펴보고, 이전 년도의 독립변수 요인에 따라 다음연도의 응답여부 간에 관련성이 있는 지를 살펴본다. 그리고 로짓분석을 이용하여 다른 독립변수의 영향을 제거하고 이전 년도의 독립변수 요인에 따라 다음연도의 응답여부 간에 관련성이 있는 지를 살펴보고, 마지막으로 데이터마이닝의 의사결정나무분석을 이용하여 이전 년도의 응답결과 자료를 활용하여 패널 탈락여부에 어떠한 변수들이 영향을 주는 지를 분석하고자 한다.

본 연구의 구성은 1장에서 서론을 다루고, 2장에서는 패널의 탈락과 관련된 이론적 배경을 살펴보고, 3장에서는 패널탈락과 관련된 실증분석을 통하여 탈락 패턴을 분석하고, 마지막으로 4장에서는 결론을 다루고자 한다.

2. 이론적 배경

Lee (2005)는 해외의 패널탈락 선행 연구를 정리한 결과, 패널탈락 요인은 인구통계학적 특성 및 개인의 노동시장내 지위, 사회경제적 충격, 조사관계자의 특성 및 조사시스템 등 크게 세 가지로 분류된다고 하였다. 그리고 그의 연구에서 노동패널 6차년도까지의 자료를 이용하여 체계적인 비무작위적인 이탈이 나타났지만, 핵심적인 경제변수를 분석하는데 있어서 심각한 영향을 미칠 정도는 아니었고, 이탈의 특성별로는 고소득자일수록 이탈가능성이 더 높지만, 가구주가 실업자일 때에도 역시 이탈가능성이 높은 것으로 나타나 이탈에 있어서 양극화 현상이 나타나는 것으로 연구되었다.

KLIPS를 이용한 패널탈락 관련 연구로는 Kim 등 (2005a), Kim 등 (2005b)에서 KLIPS의 가중치가 표본의 대표성을 적절하게 보완해 주는지를 검토하고 KLIPS 1∼6차년도 자료를 이용하여 표본이탈의 일반적 패턴을 분석하고, 기존의 가중치에 대한 보정(Calibration)을 제안하고 있다.

Chun 등 (2009)은 한국고용정보원의 대졸자직업이동 경로조사를 이용하여 패널탈락은 인구통계적 효과는 통계적으로 유의미한 효과는 있지만 전체의 설명력에서는 미미하며, 조사시스템의 효과, 응답자 사례, 사회·경제적 분위기 등과 동일면접원 여부가 패널탈락에 영향을 많이 끼치는 것으로 연구되었다.

Son과 Shin (2009)은 한국복지패널의 1∼3차 자료를 이용한 분석에서 저소득층가구와 일반가구의 결측

Table 3.1. An analysis of the target and distribution

9차 응답여부	10차 응답여부	11차 응답여부	12차 응답여부		전체
			응답안함	응답함	
응답함 (3,820명)	응답안함 (163명)	응답안함 (116명)	그룹 1000 (98명)	그룹 1001 (18명)	116
		응답함 (47명)	그룹 1010 (5명)	그룹 1011 (42명)	47
		소계	103명	60명	163
	응답함 (3,657명)	응답안함 (158명)	그룹 1100 (109명)	그룹 1101 (49명)	158
		응답함 (3,499명)	그룹 1110 (156명)	그룹 1111 (3,343명)	3,499
		소계	265명	3,392명	3,657

패턴의 분석에서는 일반가구의 패널 이탈율이 저소득가구에 비해 높게 나타고 있으며, 지역별로도 일반가구와 저소득 가구의 이탈율이 상이하게 나타나고, 이러한 결과는 저소득가구에 대해서도 마찬가지의 결과를 도출하였다.

Lee 등 (2011)은 1차에서 11차까지의 자료를 이용하여 패널탈락 시간에 대한 정보를 이용하여 생명표 방법과 Cox 비례위험모형으로 패널탈락에 영향을 주는 요인 분석에서 고학력, 미혼, 미취업자의 경우 패널탈락위험이 높은 것으로 연구되었다.

또한 무응답이 많으면 자료 분석에도 문제가 있지만 패널 탈락과 연결될 수 있다는 점에서 유사한 점이 있으며, 이에 대한 연구로서 유사한 탈락에 대한 패널의 무응답과 관련한 연구로서 Son 등 (2011)은 추적조사를 통하여 무응답패턴을 연구하여 본조사와 추적조사간의 차이가 없는 것으로 분석되었고, Oh와 Chun (2009)은 임금자료에 대해 무응답대체를 연구하여 자료기반 회귀대체와 비교하면, 모형기반 회귀대체의 성능은 거의 유사하지만, 모형을 세우고 대체를 적용하는 데 있어 더 효율적으로 연구되었다.

3. 패널탈락분석

본 연구에서의 패널탈락의 분석대상은 9차에서 12차까지를 대상으로 한다. 이는 패널조사에서 패널 구축 이후 초기에는 패널 탈락도 많이 발생하지만, 이에 대한 연구도 비교적 많이 존재한다. 그러나 10여년 정도 패널이 진행되면 패널 대상자도 본인 스스로 패널이라는 사실을 각인하고, 조사에 비교적 협조적이며, 패널탈락도 초기와는 다른 양상일 수 있기 때문이다.

Table 3.1을 보면, 초기 패널 구축 가구(5,000 가구)에서 분가 가구나 새로 패널에 진입하는 가구는 분석에서 제외하고, 10여년이 지난 처음의 5,000가구에 대해서만 분석에 포함하며, 9차에서 조사에 응답하지 않은 1,343가구도 9차의 정보를 알 수 없기에 분석에서 제외하였다. 패널의 패턴을 정의하기위해 9차에 응답한 가구에 대해 10차, 11차, 12차의 응답여부에 따라 응답을 하였으며 1, 응답을 하지 않았다면(패널탈락) 0을 부여하여 4자리 숫자의 8가지 응답패턴을 부여하였다.

본 연구에서 사용한 분석내용은 첫째, 위에서 제시한 8개 그룹 간에 패턴이 존재하는가? 즉, 9차의 독립변수(응답결과)들과 8개의 그룹 간에 관련성(10, 11, 12차 응답여부를 예측)이 있는지를 살펴보고, 둘째, 8개의 그룹에 해당하는 표본의 수가 적을 수 있기에 10차의 응답결과들과 4개의 그룹 간에 관련성(11차와 12차의 응답여부를 예측)이 있는 지를 살펴보고, 셋째, 이전 년도의 독립변수 요인에 따라 다음연도의 응답여부 간에 관련성이 있는 지를 살펴본다. 넷째는 로짓분석을 이용하여 다른 독립변수의

영향을 제거하고 이전 년도의 독립변수 요인에 따라 다음연도의 응답여부 간에 관련성이 있는 지를 살펴보고, 마지막으로 의사결정나무분석을 이용하여 이전 년도의 결과를 이용하여 다음 년도의 응답여부를 분석하고자 한다.

3.1. 결측패턴 유형 분류 및 단일변수를 이용한 패널탈락 분석

9차부터 12차 까지의 4차년도 자료를 활용한 결측패턴은 모두 8가지 유형(1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111)인데, 각 결측유형별 케이스가 적어서 통계적 유의성을 판단하기에 어려움이 있을 뿐만 아니라, 유형간에 특징적인 차이점을 발견하기 어려웠다. 이에 따라 8가지 유형을 네 가지 유형(1: 계속유지, 2: 패널복귀, 3: 단기탈락, 4: 장기탈락)으로 조정하였다. '계속유지' 유형은 9차부터 12차 까지 모두 응답한 경우로써, 8가지 유형의 '1111'이 여기에 해당한다. '패널복귀' 유형은 9차 조사 이후에 탈락했다가 12차 조사에 복귀한 경우로써, 8가지 유형의 '1001', '1101', '1011' 등이 여기에 해당한다. '단기탈락' 유형은 9차 조사 성공이후에 12차 조사에서 탈락한 패널 중 10차 또는 11차에서 1회 이상 응답한 경우로써, 8가지 유형의 '1100', '1010', '1110' 등이 여기에 해당한다. 마지막으로 '장기탈락' 유형은 9차 조사 성공 이후에 단 한번도 조사에 응하지 않은 경우로써, 8가지 유형의 '1000'이 여기에 해당한다.

각 차수별로 이전 년도의 독립변수 요인에 따라 다음연도의 응답여부 간에 관련성이 있는 지를 살펴본 것이 Table 3.2이다.

먼저 가구주 학력은 10차($p = 0.020$)와 12차($p = 0.000$)에서 유의미하게 관련성이 있고, 가구주의 종사상지위는 10차($p = 0.033$)에서, 가구주의 정규직 여부는 12차($p = 0.062$)에서, 입주형태는 10차($p = 0.000$)와 12차($p = 0.099$), 분가여부는 12차($p = 0.001$), 이사여부는 10차($p = 0.001$), 11차($p = 0.000$), 12차($p = 0.002$) 모두에서 유의미한 관련성이 있는 것으로 분석되었다.

가구주의 학력에서는 학력이 낮을수록 응답비율이 10차와 12차에서 높고, 종사상 지위는 10차에서 상용직의 경우에 응답하지 않을 비율이 상대적으로 높고, 가구주의 정규직 여부에서는 12차에서 비정규직에서 상대적으로 응답할 비율이 높고, 입주형태에서는 10차와 12차에서 자기집의 경우에 응답할 확률이 높게 분석되었다. 분가여부에서는 12차에서 분가가 없는 경우에 응답할 비율이 높게 분석되었고, 이사여부에서만 모든 차수에서 이사 경험이 없는 경우네 응답할 비율이 상대적으로 높은 것으로 분석되었다. Table 3.3에서와 같이 연속형 변수 독립변수에서는 가구주의 만나이에서 모든 차수($p = 0.029, p = 0.022, p = 0.000$)에서 응답한 그룹에서 고연령층으로 분석되었고, 총근로소득은 12차에서만 응답하지 않은 그룹이 소득이 높고($p = 0.007$), 면접원의 총방문횟수는 응답하지 않은 그룹의 평균 방문횟수가 많은 것으로 모든 차수에서 분석되었다.

3.2. 로짓분석을 이용한 패널탈락 분석

앞 장에서 실시한 단일변수에 대한 검정 결과에서 사용한 독립변수 가운데, 가구주의 종사상 지위와 정규직 여부 변수는 취업자만 응답할 수 있기 때문에 모형에 포함되는 케이스가 작아지는 문제가 있다. 따라서 로짓분석을 실시한 분석모형에서는 제외하였다. 또한 월평균 저축액 변수는 무응답이 많기 때문에 역시 활용하기가 어려워 저축여부 변수를 대신 사용하였다.

로짓모형은 종속변수가 이진 범주를 가진 경우에 사용하는 회귀분석 방법인데, 본 연구에서는 패널탈락여부(탈락: 0, 유지: 1)를 종속변수로 하여 분석을 실시하였다. 또한 독립변수로 활용한 변수는 앞 장에서 사용한 변수를 중심으로 모형에 포함시켰다. 자료의 종류가 범주형인 경우에는 가변수(dummy variable)로 변환하여 사용하였는데, 범주가 2개인 이진 자료의 경우에는 성별(여성: 0, 남성: 1), 저축

Table 3.2. Test for discrete variables

가구주의 성별									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
남성	128	2,940	3,068	145	2,695	2,840	125	2,536	2,661
	4.17%	95.83%	100%	5.11%	94.89%	100%	4.70%	95.30%	100%
여성	32	664	696	20	412	432	23	388	411
	4.60%	95.40%	100%	4.63%	95.37%	100%	5.60%	94.40%	100%
소계	160	3,604	3,764	165	3,107	3,272	148	2,924	3,072
	4.25%	95.75%	100%	5.04%	94.96%	100%	4.82%	95.18%	100%
$\chi^2(p)$	.252(.615)			.177(.674)			.627(.428)		
가구주의 학력									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
무학	51	1,541	1,592	60	1,247	1,307	41	1,178	1,219
	3.20%	96.80%	100%	4.59%	95.41%	100%	3.36%	96.64%	100%
초졸	62	1,234	1,296	55	1,106	1,161	50	1,042	1,092
	4.78%	95.22%	100%	4.74%	95.26%	100%	4.58%	95.42%	100%
중졸	47	829	876	50	754	804	57	704	761
	5.37%	94.63%	100%	6.22%	93.78%	100%	7.49%	92.51%	100%
고졸	160	3,604	3,764	165	3,107	3,272	148	2,924	3,072
	4.25%	95.75%	100%	5.04%	94.96%	100%	4.82%	95.18%	100%
$\chi^2(p)$	7.869(.020)			3.107(.212)			17.611(.000)		
가구주 취업여부									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
취업자	110	2,470	2,580	114	2,179	2,293	104	2,054	2,158
	4.26%	95.74%	100%	4.97%	95.03%	100%	4.82%	95.18%	100%
미취업자	50	1,134	1,184	51	928	979	44	870	914
	4.22%	95.78%	100%	5.21%	94.79%	100%	4.81%	95.19%	100%
$\chi^2(p)$	.003(.954)			.081(.776)			.000(.995)		
가구주의 종사상 지위									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
상용직	59	1,067	1,126	50	936	986	50	866	916
	5.24%	94.76%	100%	5.07%	94.93%	100%	5.46%	94.54%	100%
임시직	9	113	122	3	96	99	5	113	118
	7.38%	92.62%	100%	3.03%	96.97%	100%	4.24%	95.76%	100%
일용직	8	260	268	16	243	259	8	220	228
	2.99%	97.01%	100%	6.18%	93.82%	100%	3.51%	96.49%	100%
고용주/	32	1,102	1,034	45	881	926	40	832	872
자영업	3.09%	96.91%	100%	4.86%	95.14%	100%	4.59%	95.41%	100%
무급가족	2	28	30	0	23	23	1	23	24
종사자	6.67%	93.33%	100%	0.00%	100%	100%	4.17%	95.83%	100%
$\chi^2(p)$	10.484(.033)			2.836(.586)			1.881(.758)		

(continue)

가구주의 정규직/비정규직 여부									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
정규직	52	969	1,021	48	854	902	48	776	824
	5.09%	94.91%	100%	5.32%	94.68%	100%	5.83%	94.17%	100%
비정규직	22	466	488	21	421	442	15	423	438
	4.51%	95.49%	100%	4.75%	95.25%	100%	3.42%	96.58%	100%
소계	74	1,435	1,509	69	1,275	1,344	63	1,199	1,262
	4.90%	95.10%	100%	5.13%	94.87%	100%	4.99%	95.01%	100%
$\chi^2(p)$	.242(.623)			.198(.656)			3.475(.062)		
가구내 근로소득자 유무									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
있음	137	3,067	3,204	147	2,975	3,122	158	2,916	3,074
	4.28%	95.72%	100%	4.71%	95.29%	100%	5.14%	94.86%	100%
없음	26	590	616	35	618	653	35	598	633
	4.22%	95.78%	100%	5.36%	94.64%	100%	5.53%	94.47%	100%
소계	163	3,657	3,820	182	3,953	3,775	193	3,514	3,707
	4.27%	95.73%	100%	4.82%	95.18%	100%	5.21%	94.79%	100%
$\chi^2(p)$	.004(.951)			.499(.480)			.161(.688)		
입주형태									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
자가	90	2,461	2,551	117	2,434	2,551	116	2,393	2,509
	3.53%	96.47%	100%	4.59%	95.41%	100%	4.62%	95.38%	100%
전세	31	648	679	34	588	622	43	583	626
	4.57%	95.43%	100%	5.47%	94.53%	100%	6.87%	93.13%	100%
월세	36	385	421	22	428	450	24	407	431
	8.55%	91.45%	100%	4.89%	95.11%	100%	5.57%	94.43%	100%
기타	6	156	162	9	141	150	10	133	143
	3.70%	96.30%	100%	6.00%	94.00%	100%	6.99%	93.01%	100%
$\chi^2(p)$	22.559(.000)			1.328(.722)			6.277(.099)		
분가 여부									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
없음	150	3,409	3,559	171	3,359	3,530	162	3,204	3,366
	4.21%	95.79%	100%	4.84%	95.16%	100%	4.81%	95.19%	100%
있음	13	248	261	11	234	245	31	312	343
	4.98%	95.02%	100%	4.49%	95.51%	100%	9.04%	90.96%	100%
$\chi^2(p)$	.349(.554)			.063(.802)			11.265(.001)		
이사 여부									
구분	10차응답여부			11차응답여부			12차응답여부		
	응답안함	응답함	소계	응답안함	응답함	소계	응답안함	응답함	소계
있음	33	431	464	29	408	437	34	373	407
	7.11%	92.89%	100%	6.64%	93.36%	100%	8.35%	91.65%	100%

(continue)

없음	130	3,226	3,356	153	3,185	3,338	159	3,143	3,302
	3.87%	96.13%	100%	4.58%	95.42%	100%	4.82%	95.18%	100%
소계	163	3,657	3,820	182	3,593	3,775	193	3,516	3,709
	4.27%	95.73%	100%	4.82%	95.18%	100%	5.20%	94.80%	100%
$\chi^2(p)$	10.465(.001)			3.548(.060)			9.198(.002)		

Table 3.3. Test for continuous variables

구분		10차응답여부			11차응답여부			12차응답여부		
		N	mean	sd	N	mean	sd	N	mean	sd
가구주	응답안함	160	52.73	13.27	165	53.48	12.20	148	53.10	12.28
나이	응답함	3,604	54.95	12.48	3,107	55.70	12.13	2,924	56.81	12.08
(세)	$t(p)$	−2.190(.029)			−2.291(.022)			−3.639(.000)		
가구	응답안함	162	3,536	3,756	179	3,207	2,889	191	4,328	4,418
총소득	응답함	3,640	3,359	4,161	3,574	3,557	4,188	3,492	3,585	3,579
(만)	$t(p)$	.533(.594)			−1.106(.269)			2.286(.023)		
월평균	응답안함	163	179.05	111.79	182	198.24	135.31	193	213.20	137.85
생활비	응답함	3,656	176.50	118.87	3,593	467.44	16,680	3,516	476.85	16,861
(만)	$t(p)$	.269(.788)			−.218(.828)			−.217(.828)		
총	응답안함	163	3.43	2.09	182	2.71	1.79	193	2.86	1.88
방문	응답함	3,657	3.04	1.89	3,589	2.65	1,82	3,516	2.24	1,42
횟수	$t(p)$	2.357(.020)			.415(.678)			4.550(.000)		

여부(안 함: 0, 함: 1), 취업여부(미취업: 0, 취업: 1), 근로소득자 유무(무: 0, 유: 1), 이사여부(안함: 0, 함: 1), 분가여부(없음: 0, 있음: 1) 등이 분석에 포함되었다. 범주가 3개 이상인 변수는 다음과 같이 사용하였다. 입주형태(전세여부, 월세여부, 기타여부)는 '자가'를 기준변수로 사용하였고, 학력(중졸여부, 고졸여부)은 대졸여부를 기준변수로 사용하였다.

로짓모형에는 모두 6개의 모형을 통해 패널탈락에 영향을 주는 변수를 확인하였는데, 모형 1에서 모형 3은 전년도 독립변수가 다음 조사의 탈락여부에 미치는 효과를 살펴본 것이다. 따라서 모형 1은 9차년도 독립변수가 10차년도 탈락여부, 모형 2는 10차년도가 11차년도 탈락여부, 모형 3은 11차년도가 12차년도 탈락여부에 미치는 효과를 살펴보았다. 모형 4에서 모형 6은 2년 이상의 간격을 통해 탈락여부를 분석한 것이다. 모형 4는 9차년도 독립변수가 11차년도 탈락여부, 모형 5는 10차년 독립변수가 12차년도 탈락여부, 모형 6은 9차년도 독립변수가 12차년도 탈락여부에 미치는 효과를 분석한 것이다. 모형에 포함된 독립변수 중에서 패널의 개인 배경 변수라고 볼 수 있는 것들로는 가구주의 성별, 나이, 학력, 취업여부, 근로소득자 유무 등이며, 패널의 가구배경 변수로 볼 수 있는 것들로는 월평균생활비, 입주형태, 이사여부, 분가여부, 가구총소득, 저축여부 등이다.

모형 1에서 학력이 중졸인 경우, 대졸에 비해 응답확률이 높고, 입주형태가 월세인 경우에는 자가에 비해 패널 탈락 확률이 높고, 이사를 하는 경우, 총방문횟수가 많을수록 패널 탈락 비율이 높게 나타났다. 모형 2에서는 나이만 유의한 영향을 주는 것으로 나타났는데, 고연령층일수록 패널 유지 확률이 높게 나타났다. 모형 3에서는 나이가 많을수록, 대졸에 비해 중졸일수록, 근로소득자일수록, 방문횟수가 적을수록, 분가 가구 있을수록, 가구총소득이 높을수록 패널 유지 확률이 높은 것으로 나타났다. 모형 3은 다른 모형과 다른 변수들이 유의한 영향을 주는 것으로 나타났는데, 특히 가구총소득이 높을수록 패널 유지가 잘 된다는 것은 다른 패널들의 결과와 배치되는 것이다.

모형 4에서 총방문횟수가 증가할수록, 입주형태가 자가보다는 월세일수록, 이사를 할수록, 비저축자일수록 패널 탈락 확률이 높은 것으로 나타났다. 모형 1의 경우와 유사한 결과인데, 모형 1에서 이사여부는 유의수준 0.1 수준에서는 통계적으로 유의한 영향을 주는 것으로 파악된다. 또한 모형 5에서는 나이가 적을수록, 학력이 높은 경우, 총방문횟수가 많을수록 패널 탈락 확률이 높은 것으로 나타났다. 이는 모형 2의 결과와 유사하다. 한편 모형 6에서는 가구주가 여성일수록, 저연령, 고학력, 근로소득자가 아닐수록, 방문횟수가 많을수록, 이사를 할수록, 비저축자 일수록 패널 탈락 확률이 높은 것으로 나타났다. 특히 가구주가 남성일수록 패널유지가 잘되는 것으로 나타났는데, 이는 가구주가 여성인 경우에는 직장생활과 가사일을 병행하는 경우가 많아 조사가 어려운 경우가 있고, 경제적 문제, 집안에 남자의 부재 등 상대적으로 조사가 어렵기 때문인 것으로 추측된다.

위에서 살펴본 바와 같이 1년 전의 개인적, 가구적 환경들이 다음해의 조사여부에 미치는 원인은 조사년도마다 다른 양상을 보이고 있다. 하지만 10차년도 독립변수들이 포함된 모형 2와 모형 5는 다른 모형에 비해 유의한 변수들이 없는 것으로 나타났는데, 이는 10차년도의 외부적 요인 또는 조사시스템의 문제가 있는지를 검토해 볼 필요가 있음을 보여주고 있다. 2007년에 실시된 10차 조사는 CAPI(Computer Assisted Personal Interview) 조사 방식을 시범적으로 운용한 해이다. 물론 대전과 충청지역에만 제한적으로 적용하였지만, 현재까지 패널유지에 어떤 영향을 미쳤는지에 대한 연구결과가 없는 상태이다. Seong과 Choi (2011)는 10차 조사의 CAPI와 PAPI(Paper and Pencil Interview) 방식 간에 통계적 유의성이 없기 때문에 모드효과(mode effect)가 없다고 설명하지만, 이는 조사된 자료의 분포 및 평균차이 등과 같은 통계적 유의성에 초점을 맞춘 것이기 때문에 자료의 연속성 등의 측면과는 다른 문제이다.

이사여부가 유의한 영향을 주는 모형에서는 입주형태가 월세인 경우가 자가인 경우에 비해 패널타락이 많이 일어나는 것을 볼 수 있다. 또한 전년도 독립변수를 활용한 모형 1과 모형 3 보다는 2년 전 독립변수를 활용한 모형 4와 3년 전 독립변수를 활용한 모형 6에서 더 유의한 영향을 주는 것으로 나타났다. 보통 입주형태가 자가가 아닌 경우, 2년 단위의 계약이 이루어진다는 측면에서, 그리고 전년도에 이사를 했다면 올해 보다는 이사 후 2년이 경과한 다음 년도에 다시 이사할 가능성이 높다는 측면에서 패널탈락이 이루어질 수 있다는 점을 주의 깊게 살펴보아야 할 것이다.

앞에서 제시한 모형들에서 가구총소득은 유의한 영향을 주지 않지만 저축여부는 몇 개의 모형에서 유의한 영향을 주는 것으로 나타났다. 만일 10차 조사의 결과가 다른 차수의 조사와의 연계성이 담보된다면, 10차 조사 당시의 사회 특수성에 관심을 가져야 한다. 따라서 다른 차수의 경우에도 저축여부가 패널유지에 유의한 영향을 주는 지를 파악할 필요가 있다.

10차 조사에서 수집한 자료를 독립변수로 활용한 모형 2와 모형 5에서, 유의한 영향을 주는 변수가 다른 모형과 다르다는 사실 뿐만 아니라 일부 변수의 경우에는 방향성이 다른 모형과 다르게 나타나는 사실에도 주목할 필요가 있다. 이는 분명히 다른 차수의 조사 결과와는 다른 내재적 요인들이 작용하는 것이기 때문에 향후 연구에서 면밀히 검토해야 할 것이다.

나이가 많을수록, 학력이 낮을수록 패널탈락이 덜 발생한다는 분석결과는 기존에도 발표되었던 내용이다. 총방문횟수가 증가할수록 패널탈락이 많이 발생한다는 사실은 해석에 주의할 필요가 있다. 패널탈락자들은 조사가 잘 되지 않기 때문에 많은 방문을 할 것이다. 이로 인해 방문횟수가 증가하는 것이다. 즉, 방문횟수가 탈락 여부에 영향을 미치는 인과관계로 해석하기에는 어려움이 있다. 하지만 이 결과에서 확인할 수 있는 중요한 내용은 전년도 또는 2~3년전의 방문횟수가 몇 해가 지나도 영향을 준다는 사실이다. 결국 많은 방문을 하는 패널가구는 몇 해가 지나도 응답을 하지 않는다는 것인데, 조사 실패가 장기화되는 가구는 단순히 방문을 많이 한다고 해서 패널복귀가 이루어지지는 않는다는 것을 보여주는 것이다.

Table 3.4. Logit analysis on whether the panel attrition

구분	model1(9차 → 10차)			model2(10차 → 11차)			model3(11차 → 12차)		
	B	S.E	Exp(B)	B	S.E	Exp(B)	B	S.E	Exp(B)
성별	.210	.229	1.234	−.057	.267	.944	.333	.274	1.395
나이	.006	.009	1.006	.021*	.010	1.021	.034**	.011	1.034
학력(중졸)	.618*	.254	1.855	.114	.245	1.120	.642*	.261	1.901
학력(고졸)	.189	.211	1.208	.220	.223	1.247	.307	.217	1.360
취업여부	.013	.240	1.013	.215	.250	1.240	.068	.284	1.070
근로소득자	.067	.308	1.069	.115	.323	1.122	.701*	.355	2.016
월평균생활비	.000	.001	1.000	.000	.000	1.000	.000	.000	1.000
총방문횟수	−.089*	.040	.915	−.044	.044	.957	−.223***	.050	.800
입주형태(전세)	−.109	.225	.896	.136	.239	1.146	−.202	.234	.817
입주형태(월세)	−.749***	.230	.473	.165	.279	1.179	.409	.330	1.505
기타	.003	.441	1.003	.426	.527	1.532	−.483	.402	.617
이사여부	−.412†	.217	.662	−.138	.270	.871	−.128	.276	.880
분가여부	.244	.303	1.276	−.032	.357	.968	.889***	.274	2.432
가구총소득	.000	.000	1.000	.000†	.000	1.000	.000*	.000	1.000
저축여부	.401*	.200	1.493	−.146	.220	.864	.478*	.212	1.613
상수항	2.299**	.819	9.964	1.957*	.944	7.075	−.814	.917	.443
χ^2	40.938***			11.978			63.747***		
$-2LL$	1274.341			1196.153			1034.622		
Nagelkerke's r^2	0.037(n = 3,739)			0.012(n = 3,234)			0.068(n = 3,037)		

구분	model4(9차 → 11차)			model5(10차 → 12차)			model6(9차 → 12차)		
	B	S.E	Exp(B)	B	S.E	Exp(B)	B	S.E	Exp(B)
성별	.148	.179	1.159	.302	.198	1.352	.345*	.154	1.411
나이	.012†	.007	1.013	.024**	.008	1.025	.019**	.006	1.020
학력(중졸)	.279	.198	1.322	.356†	.190	1.427	.383*	.171	1.467
학력(고졸)	.157	.172	1.170	.413*	.171	1.511	.346*	.148	1.414
취업여부	−.087	.196	.917	.301	.202	1.351	.011	.171	1.011
근로소득자	.361	.241	1.434	.262	.257	1.299	.496*	.210	1.642
월평균생활비	.000	.001	1.000	.000	.000	1.000	−.001	.001	.999
총방문횟수	−.069*	.032	.933	−.073*	.034	.930	−.081**	.028	.923
입주형태(전세)	−.118	.176	.889	−.058	.183	.944	−.172	.149	.842
입주형태(월세)	−.598***	.187	.550	.016	.221	1.017	−.309†	.079	.734
기타	−.102	.322	.903	−.174	.338	.840	−.244	.266	.783
이사여부	−.474**	.172	.622	−.214	.205	.807	−.517***	.151	.596
분가여부	.199	.243	1.221	−.470	.336	.625	.152	.218	1.164
가구총소득	.000	.000	1.000	.000	.000	1.000	.000	.000	1.000
저축여부	.384*	.159	1.467	−.097	.177	.908	.372**	.141	1.451
상수항	1.302*	.635	3.678	.761	.726	2.141	.173	.548	1.188
χ^2	49.888***			36.841**			79.094***		
$-2LL$	1863.697			1693.410			2294.689		
Nagelkerke's r^2	0.034(n = 3,739)			0.012(n = 3,234)			0.068(n = 3,739)		

† : pr < 0.1, ∗ : pr < 0.05, ∗∗ : pr < 0.01, ∗ ∗ ∗ : pr < 0.001

3.3. 의사결정나무를 이용한 패턴분류 분석

본 연구에서는 의사결정나무기법 중에서 CHAID(Chi square AID; Kass, 1980)를 이용하였는데, 목표변수(target variable)는 네 가지 패널유형이다. CHAID와 같은 의사결정나무(decision tree)는 로짓모형과 다른 특성이 있는데, 목표변수가 아닌 입력변수(회귀모형에서의 독립변수)에 무응답이 있어도 분류를 실시한다는 것이다. 이를 통해 회귀모형과의 비교가 가능한 특징이 있다. 이 때 사용한 입력변수(input variable)는 앞의 로짓모형에서 사용한 변수를 사용하되 저축여부 대신에 저축액 등과 같은 사용가능한 연속형 자료를 사용하였다. 연속형 변수의 분리점은 모형이 스스로 정할 수 있는 옵션을 채택했다. 본 연구에서 작성한 최종모형은 Figure 3.1과 같이 15개의 잎을 가진 의사결정나무이다. 의사결정나무는 각 모형마다 분리기준으로 사용하는 통계량이 다른데, CHAID는 χ^2 통계량을 기준으로 한다. 뿌리마디에서 제일 먼저 분리기준으로 선택된 변수는 입주형태이고, 두 번째 마디에서부터는 다른 변수가 선택되었는데, 입주형태가 '자가'인 경우에는 '이사여부', 입주형태가 '전세'와 '기타'인 경우에는 '성별', 입주형태가 '월세'인 경우에는 '이사여부' 변수가 분리기준으로 선택되었다. 이를 통해 확인할 수 있는 의사결정규칙들이 발견되는데, 입주형태가 '자가(1)'이고 이사를 '하지않은 경우(2)'이고, 월평균생활비가 189만원 이하인 경우, 91.55%에 해당하는 1,192명이 패널유지가 계속 유지된다는 것을 알 수 있다. 동일 조건에서 월평균생활비가 189만원 초과이면서 250만원 이하인 경우에는 87.02%로 나타났다. 동일 조건에서 월평균 생활비가 250만원 초과인 경우, 다시 가구총소득이 2,630만원 이하인 경우에는 총 16명 중 6명에 해당하는 37.5%가 단기 탈락인 것으로 나타났다. 또한 아래쪽 자식마디에서 발견된 월평균생활비와 가구총소득은 그 부모마디에 속한 변수에 따라 기준과 해석이 달라지는 것으로 나타났다. 따라서 소득과 소비 관련 변수는 회귀분석에서처럼 어떤 변수와 종속변수 사이의 정(+)의 관계 또는 부(−)의 관계로 해석하지 말고 다른 변수와 동시에 고려한 해석이 필요함을 알 수 있었다.

4. 결론

응답대상이 가구인 가구패널조사에서 가구원과 같은 개인의 탈락여부는 전적으로 가구주의 탈락여부에 종속되어 있다. 따라서 노동패널의 탈락여부는 가구주의 응답여부가 중요한데, 특정년도의 탈락여부를 판단하기 위해 전년도 조사에서 얻어진 가구주의 개인특성변인과 가구특성변인들을 통해 확인할 수 밖에 없다. 패널조사는 전년도와 동일한 면접원이 응답자를 조사할 수 있도록 조치를 취하는 경우가 많다. 이는 응답자 패널과 면접원 사이의 신뢰대(rapport) 형성을 통해 조사 성공률을 높이고자 하는 목적 때문이다. 또한 면접원이 보유한 면접기술은 응답자에게 적지 않은 영향을 끼치게 된다는 사실은 널리 알려져 있다. 또한 조사 방식의 변화여부, 패널유지활동과 같은 조사시스템이 패널에 주는 영향에 대해서는 이 연구에서 다루지 않았다. 선행 연구에서도 조사 시스템이 패널유지에 미치는 영향이 클 것으로 보고 있지만, 현재의 조사 현실상 조사시스템이 패널 개인에 미치는 영향에 대해 분석하기는 쉽지 않기 때문이다.

본 연구에서 활용하고 있는 자료는 2006년부터 2009년까지 실시한 9차∼12차 조사 자료이다. 패널조사의 경우에는 웨이브 초반에 많은 패널들이 탈락한다. 따라서 패널이 어느 정도 안정상태의 단계에 접어들게 되면, 오랜 기간 남아있는 패널은 유지 상태를 존속하려는 경향이 강한 편으로 생각할 수 있다. 그러나 한국노동패널 조사방식이, 9차 조사까지는 PAPI, 10차 조사는 PAPI와 일부 CAPI, 11차와 12차 조사는 CAPI 조사 방식을 적용하였다. 물론 각 조사년도마다 일부 불가피한 경우에는 PAPI 조사를 진행하는 경우도 있다. 하지만 종이로 되어 있는 설문지 조사 방식에서 노트북을 활용한 CAPI 조사를 전면적으로 적용하게 될 경우, 예상치 못한 문제가 발생할 가능성이 있다. 물론 10차 조사에서 일부 지역에 대한 시범 운용을 통해 점검을 실시했고, 11차 조사에 전면 도입하는 방식을 적용하였으며, 전년도

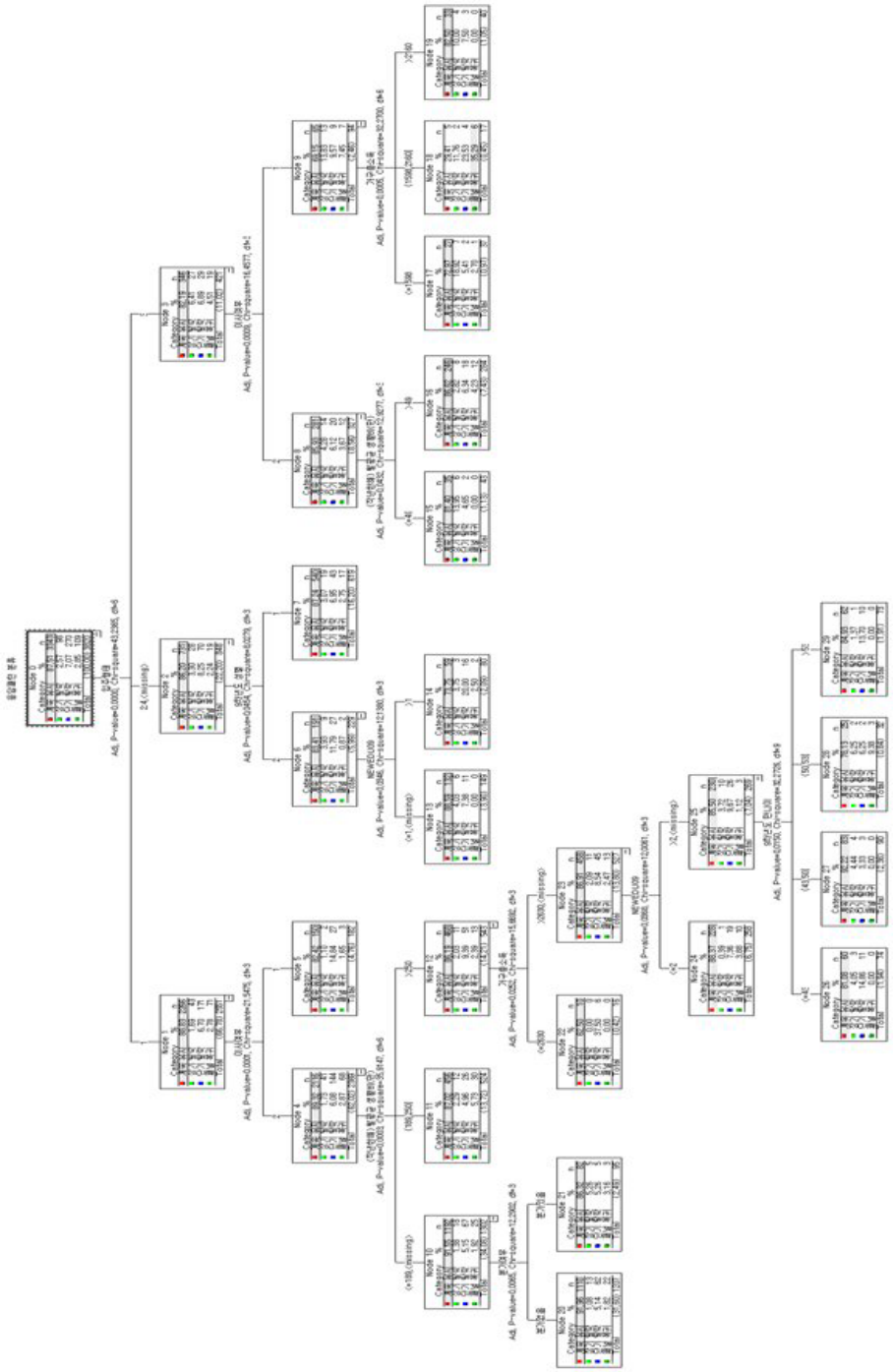

Figure 3.1. Decision tree model

조사 정보를 CAPI 시스템에 탑재하여 자료의 연속성 및 신뢰성을 보장하고자 노력하였다. 그럼에도 불구하고 조사 방식의 전환이 어떤 영향을 주었는지에 대한 면밀한 검토가 진행되어야 할 필요가 있음을 이 연구에서 찾을 수 있었는데, 단일 변수를 통한 자료의 분포 및 연속성 등이 문제가 없다고 하더라도 패널유지 측면에서 각각의 변수들이 모형에 포함되었을 경우, 심각한 불균형이 있음을 알게 되었다. 노동패널조사는 현재 15차년도 조사를 진행하고 있기 때문에 조사 진행이 오래되지 않은 패널의 패널유지와는 전혀 다른 양상이 전개될 것이다. 패널이 안정적으로 정착된 상황에서 패널유지는 개인의 심경변화와 같은 일시적 반응 보다는 가구 및 개인의 급격한 상태 변화와 같은 특별한 경우에 발생할 것이다.

로짓모형과 의사결정나무를 통해 패널복귀와 탈락에 대한 유형이 어떤 변수에 의해 영향을 받는지를 분석하였다. 물론 종속여부가 다르고 분석에 활용한 변수가 몇 차년도 자료였는지에 따라 해석이 다르지만, 다른 패널탈락분석에서 확인되었던 내용들도 여전히 영향력을 주고 있음을 알 수 있었다. 예를 들어 나이가 적고 학력이 높고 소득이 많으면 이탈이 많이 발생하는 내용에 대해서는 동일한 결과를 얻었다. 다른 국내연구의 패널 이탈분석에는 국내패널이 장기화되지 못한 탓에 패널초기 시점에 초점이 맞추어진 반면, 본 연구에서는 패널이 어느 정도 안정화단계에 접어든 9차년도 이후 자료를 분석했음에도 여전히 유의한 개인적 영향 요인이 존재한다는 사실을 밝혀냈다. 또한 가구 대상 패널에서 이사여부와 입주형태는 중요한 영향요인임을 알 수 있었다. 한편 로짓모형에서는 저축여부, 의사결정나무에서는 월평균 생활비와 가구총소득과 같이 소득 또는 소비와 관련된 변수가 유의한 영향을 주는 것으로 나타났는데, 이에 대해서는 좀 더 면밀한 검토를 통해 확인을 해야 할 것으로 보인다. 왜냐하면 의사결정나무에서 가구소득과 월평균 생활비의 경우, 다른 변수들과의 관계에 따라 다른 유지 패턴이 나타나기 때문이다. 따라서 단순히 소득과 소비의 높고 낮음에 따라 패널이 유지되는지 또는 탈락하는지를 판단해서는 안 될 것이다.

웨이브 초기 패널 탈락은 조사시스템에 의해 많은 영향을 받을 수 있다. 특히 동일 면접원 여부, 패널유지활동, 조사방식 등 조사시스템에 의해 많은 영향을 받을 수 있다. 국내의 대부분의 패널조사는 웨이브가 짧아서 어떤 배경적 변인보다도 조사시스템에 의한 영향을 예상할 수 밖에 없었다. 더군다나 대상이 특화되어 있는 패널조사에서는 조사시스템의 특성이 영향을 줄 수 밖에 없게 된다. 예를 들어 중고령자 대상 패널에서는 동일면접원 투입, 면접원과 응답자의 rapport 형성은 패널의 유지 및 탈락에 많은 영향을 줄 수 있다. 조사방식은 패널 유지 보다는 자료의 연계성 및 신뢰성에 많은 영향을 준다. CAPI 방식의 도입으로 이전 차수의 조사결과와의 연계검토, 조사 분기 문항의 적정성을 통한 불필요한 내용 검토 불필요, 적정 수준의 로직 적용을 통한 이상값 검출 등에 효과적이다.

노동패널은 국내에서 가장 오래된 패널조사로서 웨이브가 길기 때문에 10차 이후에 유지되는 패널의 조사에 대한 충성도는 높은 수준이다. 따라서 웨이브가 짧은 패널조사에서의 패널 탈락과는 다른 요인들이 영향을 줄 것이다. 또한 조사 초기의 결측패턴과 안정화 단계의 결측패턴은 다른 함의가 있을 것이므로 어떤 외부적인 효과가 탈락 및 패턴에 영향을 주는지를 판단하는 것은 중요한 연구로 볼 수 있다.

거주기간과 이사횟수 등을 고려한 모형을 통해 분석한다면 좀 더 설명력 있는 모형을 산출 할 수 있을 것이다. 향후 연구에서는 이 변수들을 투입해 로짓모형과 의사결정나무 분석을 실시할 것이며, 네 가지 패턴을 종속변수로 하여 로지스틱 회귀분석이나 수량화 방법을 통해 각 패턴에 영향을 주는 변수가 무엇인지를 확인하고자 한다.

References

Chun, Y. M., Yoon, J. H. and Oh, M. H. (2009). An analysis of panel attrition in GOMS(Graduates Occupational Survey), *The Korean Journal of Applied Statistics*, **22**, 951–993.

Demirtas, H. (2004). Modeling incomplete longitudinal data, *Journal of Modern Applied Statistical methods*, **3**, 305–321.

Gorbein, J. A., Lazaro, G. G. and Little, R. J. A. (1992). Incomplete data in repeated measures analysis, *Statistical Methods in Medical Research*, **1**, 275–295.

Hogan, J. W. and Laird, N. M. (1997). Model-based approaches to analyzing incomplete longitudinal and failure-time data, *Statistics in Medicine*, **16**, 259–272.

Hogan, J. W., Roy, J. and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies, *Statistics in Medicine*, **23**, 1455–1497.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of data, *Applied Statistics*, **29**, 119–127.

Kim, K. S., Hang, Y. E. and Park, J. W. (2005a). The study of panel survey in weighting methods and effects, In *Proceedings of the 6th Korea Labor and Income Panel Conference*, Korea Labor Institute.

Kim, Y. W., Kim, J. K., Lee, K. J. and Jo, Y. M. (2005b). Sample of the performance and weighting correction method for KLIPS, *Proceedings of the 6th Korea Labor and Income Panel Conference*, Korea Labor Institute.

Little, R. J. A. (1995). Modeling the drop-out mechanism in longitudinal studies, *Journal of the American Statistical Association*, **90**, 1112–1121.

Lee, S. H. (2005). The analysis of sample deviation for KLIPS, *Monthly Labor Review*, **11**, 66–79.

Lee, S. H., Park, C. Y., Chung, S. S. and Choi, H. M. (2011). Panel attrition factors in Korean labor and income panel study, *Journal of Korea Data Information Science Society*, **22**, 1–8.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer Series in Statistics, Springer-Verlag, New-York, 683.

Oh, M. H. and Chun, Y. M. (2009). A study on imputation method for wage data in OES(Occupational Employment Statistics), *Journal of the Korean Data Analysis Society*, **11**, 1399–1410.

Seong, J. M. and Choi, H. M. (2011). Using computers to investigate how the impact of data quality - CAPI experiment KLIPS, *Proceedings of the Fall Conference of the Korean Statistical Society*.

Son, C. K., Hong, K. H. and Lee, G. S. (2011). Nonresponse pattern analysis from follow-up survey - Focus on smoking status survey-, *Journal of the Korean Data Analysis Society*, **13**, 2977–2985.

Son, C. K. and Shin, J. D. (2009). The study on missing patterns of low-income households for Korea welfare panel study, *The 2nd Korea Welfare Panel Conference*, The Korea Institute for Health and Social Affairs.