Asia-Pacific Journal of Business (아태비즈니스연구)
Vol. 11, No. 4, December 2020 (pp.291-301)
https://doi.org/10.32599/apjb.11.4.202012.291.

APJB
ISSN 2233-5900 (Print)
ISSN 2384-3934 (Online)

한국복지패널 마모패턴 특성 및 패널 이탈 모형 추정 연구*

박승환
강원대학교 정보통계학과 조교수

A Study on the Patterns of Panel Attrition in the Korea Welfare Panel Study

Seung-Hwan Park[a]

[a]Department of Information Statistics, Kangwon National University, South Korea

Received 30 November 2020, Revised 16 December 2020, Accepted 21 December 2020

Abstract

Purpose - The purpose of this study was to investigate several household characteristics related to panel attrition, examining how they may have conditioned the panel data in the Korea Welfare Panel Study (KOWEPS).
Design/methodology/approach - We studied the cause of the differences in household income between the original panel and the new panel in KOWEPS.
Findings - To summarize our findings, whereas it is highly likely that a low-income household or a household without health insurance will remain in the panel, it is highly likely that a high-income household or a household of more than three members will be taken off the panel.
Research implications or Originality - The proportion of low-income household tends to decrease over the years, which appears to result from an overall increase in household income. Such changes are reflected in the pattern in which older panels have higher estimates of household income than newer panels.

Keywords: Complex Sampling Design, Design Effect, Panel Attrition, Panel Conditioning
JEL Classifications: C13, C30, C83

Ⅰ. 서론

한국복지패널조사(Korea welfare panel study, KOWEPS)는 외환위기 이후 빈곤층, 근로빈곤층, 차상위층의 가구형태, 소득수준, 취업상태가 급격히 변화하고 있는 상황에서 이러한 계층의 규모 및 생활실태 변화를 동태적으로 파악함으로써 정책 형성에 기여하기 위한 조사이다.

한국복지패널 조사는 2006년에 제1차 조사를 시작하여 2018년에 제13차 조사를 완료하였다. 2006년 제1차 조사에서 7,072가구가 조사되었고 제6차인 2011년 조사까지 5,000가구 이상의 원패널 규모를

* 이 논문은 2018년도 강원대학교 대학회계 학술연구조성비로 연구하였음.
* 이 논문은 오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019), 2019년 한국복지패널 심층분석 보고서. 한국보건사회연구원의 주요 내용을 중심으로 수정하여 작성하였음.
[a] First Author, E-mail: stat.shpark@kangwon.ac.kr

유지하고, 원패널 유지율은 약 75% 이상을 유지하였다. 그러나 조사거부 및 자연손실 등으로 원표본이 지속적으로 감소하는 문제와 표본 탈락으로 인한 패널 표본가구 분포 상의 문제점이 제기되고, 이러한 문제점을 개선하여 최초 구축 당시 표본규모와 대표성을 유지할 수 있도록 하기 위하여 제7차년도인 2012년에 1,800가구의 신규패널을 추가하였다. 2018년 제13차 조사에서 원표본 가구수는 4,266로 원표본 유지율은 60.3%이며 신규패널의 원표본 가구수는 1,392이다

패널조사에서는 동일한 가구를 대상으로 계속 조사를 수행하므로 조사가 거듭될수록 패널마모가 발생하는 것은 불가피하다. 패널마모는 이사, 가구 분화 혹은 가구 확장 등의 패널가구의 변화 때문에 발생하기도 하고, 혹은 장기간 조사 응답에 의한 피로감 때문에 생기기도 한다. 어떤 이유이든 원표본수의 감소는 패널표본의 대표성을 저하시키고 또한 표본수 감소로 인한 통계결과의 신뢰도 저하를 가져온다. 이러한 점은 패널조사가 품고 있는 피하기 어려운 본질적인 단점이기도 하다.

이러한 배경에서 본 연구의 목적은 차수 별 가중치 변화 분석, 마모패턴의 특성 분석, 패널 이탈 확률 모형 추정에 관한 연구이다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 한국복지패널의 가중치 산정 방식 가중치 분포를 분석하고 패널 마모패턴 특성을 분석 한다. Ⅲ장에서는 비랜덤결측(MNAR: Missing Not at Random)가정 하에서 패널 이탈 확률 모형을 추정할 수 있는 방법론을 제안한다. Ⅳ장에서는 Ⅲ장에서 제안한 비랜덤결측 가정 하에서 패널 이탈 확률 모형을 13차 복지패널을 대상으로 적용하고 그 결과를 해석한다. Ⅴ장에서는 본 연구의 분석내용을 정리하고 그 시사점 및 의의를 기술한다.

Ⅱ. 한국복지패널 가중치 및 패널 마모패턴 분석

1. 한국복지패널 개요

한국복지패널의 조사목적은 외환위기 이후 빈곤층, 근로빈곤층(working poor), 차상위층(near poor)의 가구형태, 소득수준, 취업상태가 급격히 변화하고 있는 상황에서 이들 계층의 규모 및 생활실태 변화를 동태적으로 파악함으로써 정책형성에 기여함과 동시에 정책지원에 따른 효과성을 제고 하고자 함이다.

한국복지패널의 목표모집단은 13차 조사기준 2017년 현재 전국에 거주하는 가구이며 조사 모집단은 2005년 인구센서의 90%조사구로 섬과 특수시설 조사구는 제외하였다.

한국보건사회연구원은 한국복지패널 표본으로 '2005년 인구주택총조사' 자료로부터 확률비례추출 한 '2006년 국민생활실태조사' 최종 조사완료가구의 소득 자료를 기준으로 일반가구와 저소득층 가구를 구분하여 두 층으로 부터 각각 3,500가구씩 총 7,000가구를 선정하였다. 최종 패널가구로 구축된 표본가구는 7,072가구였으며, 표본추출과정에서 저소득층가구는 향후 패널 소실과 통계적 유의미성을 고려하여 과대표집하였다.

표본추출은 조사구 규모에 따른 층화 확률 비례추출을 사용하였다. 먼저 인구센서스 90% 자료로부터 517개 조사구를 표본으로 추출하여 가구의 소득 및 가구원의 경제활동 상태 등을 조사하였다. 1단계 조사 자료로부터 일반 가구와 저소득층 가구를 각각 3,500가구씩 총 7,000가구를 표본가구로 추출하였다. 소득 규모별로 구분된 2개의 층에서 지역별, 조사구별로 확률 비례 계통 추출에 따라 일반 가구와 저소득 가구를 표본으로 추출하였다. 상대적으로 규모가 작은 저소득층 가구에 대해서는 추출률을 상향 조정하여 일반 가구와 동일한 수준인 3500가구를 표본 가구로 선정하였다.

한국복지패널의 6차년도 조사 이후에 원표본 가구 유지율이 감소하는 상황에서 신규 표본가구의 추가 필요성이 제기되었으며, 저소득층 가구 및 가구원의 분포가 어느 정도 치우침 현상이 발생하였고, 지역별 표본 규모는 잦은 이주와 탈락 등의 사유로 변동이 발생하였다. 따라서 한국복지패널 7차년도 조사에는 1차년도 표본규모를 유지하고자 약 1,800가구를 추가하여 신규패널을 구축하였으며, 표본추출은 1차년도 표본추출방식과 동일한 방식으로 추출하였다. 1차부터 13차까지 복지패널 표본가구 현황은 〈Table 1〉에 나타나 있다.

Table 1. Household Samples by Year in KOWEPS

패널구분	기존패널		신규패널		전체패널	
가구구분	원가구	전체가구	원가구	전체가구	원가구	전체가구
1차	6,928	7,072	-	-	6,928	7,072
2차	6,401	6,580	-	-	6,401	6,580
3차	6,017	6,314	-	-	6,017	6,314
4차	5,812	6,207	-	-	5,812	6,207
5차	5,561	6,034	-	-	5,561	6,034
6차	5,272	5,735	-	-	5,272	5,735
7차	5,141	5,732	-	-	5,141	5,732
8차	4,995	5,619	1,687	1,693	6,682	7,312
9차	4,804	5,438	1,591	1,610	6,395	7,048
10차	4,680	5,343	1,532	1,571	6,212	6,914
11차	4,474	5,189	1,477	1,534	5,951	6,723
12차	4,309	5,081	1,425	1,500	5,734	6,581
13차	4,170	4,997	1,389	1,477	5,559	6,474

Soucre: 오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019)

2. 한국복지패널 가중치 산출 방식 및 분포 현황

한국복지패널 가중치 산출은 가구가중치와 개인가중치를 구분하여 산출한다. 먼저 개인가중치 산출에 있어 종단 가중치와 횡단 가중치를 각각 산출토록 한다. 해당 차수의 가중치는 이전 차수의 가중치를 기준으로 하여 가구 가중치와 개인 가중치를 산정하는 방식이다. 가구 가중치 산정에 있어 조사 차수가 증가함에 따라 가구 개념이 설계 당시의 가구의 상태의 정의와 달라질 수 있고 가구의 생성과 소멸이 반복적으로 이루어지기 때문에 2차 조사 이후의 가중치는 개인 가중치를 중심으로 산정한다.

개인 가중치 산출은 이전 차수의 종단 가중치를 바탕으로 산출한다. 탈락 가구원이 신규로 진입한 경우에는 해당 차수의 가구평균 가중치 혹은 전체 평균 가중치를 부여하도록 한다. 탈락가구원이 아닌 분가 후 결혼 등의 사유로 새로 패널에 들어온 신규 가구원의 해당 차수 가중치는 0으로 부여한다.

개인 종단면 가중치는 먼저 로지스틱회귀를 이용하여 응답확률을 추정하여 무응답 보정을 실시한다. 로지스틱 모형의 설명 변수로는 성, 연령, 지역, 경제활동 상태 변수를 사용하였다. 앞에서 언급한 개인별 변동 상황에 따라 가중치를 조정한다. 마지막으로 성, 연령, 지역 등 응답자의 인구학적 특성에 따라 모집단 정보를 이용하여 사후 조정을 실시한다.

개인 횡단면 가중치는 앞서 계산한 해당 차수의 종단면 가중치를 기본으로 하여 무응답 보정 과정을 거친다. 개인별 변동 상황에 따른 0으로 부여된 가중치에는 가구별 평균 가중치를 부여한다. 마지막으로 성, 연령, 지역 등 응답자의 인구학적 특성에 따라 모집단 정보를 이용하여 사후 조정을 실시한다.

종단 가중치와 횡단 가중치 모두 지나치게 큰 가중치 값들은 추정량의 분산을 과대하게 만들 수 있어 추정의 정확도에 문제를 만들 수 있다. 따라서 상위 1%의 극단 가중치들은 절단하고 성*연령*지역 내 관측치에 동일하게 배분하였다. 가중치에 대한 특이치 조정은 추정량의 편향(Bias)이 증가할 수 있지만, 추정량의 분산 감소를 통해서 추정량의 평균제곱오차(Mean Squared Err or: MSE)를 줄임으로써 전체적으로 추정의 정확도를 높일 수 있다.

7차 조사에 새롭게 추가된 신규 패널은 '2011년 복지욕구실태조사'의 최종 가중치를 기본 가중치로 고려하여 1800가구의 추출 확률의 역수를 곱하여 신규 패널의 설계 가중치를 조정하였다. 7차 조사의 종단가중치는 0으로 부여하고 횡단 가중치의 경우 기존 패널과 신규패널을 합하여 사후 조정을 통하여 새롭게 가중치를 부여한다.

1차부터 13차까지 한국복지패널 기존패널과 신규패널의 횡단 가중치에 대하여 가중치 분포를 분석한 결과는 〈Table 2〉와 〈Table 3〉에 나타나 있다. 횡단 가중치의 최댓값, 75%, 50%, 25% 백분위수, 최솟값, 평균, CV를 구하여 가중치 분포를 분석하였다.

기존패널의 횡단가중치는 표본수가 차수가 지남에 따라 감소함에 따라 가중치 값이 전체적으로 증가하는 것을 알 수 있다. 전반적인 값이 증가했을 뿐만 아니라, 4분위수 범위를 보게 되면 차수가 지남에 따라 가중치의 산포도 증가했음을 알 수 있다. CV는 1차에 150% 정도였으나 3차 이후로 는 120%정도를 유지하고 있다.

신규 패널의 횡단 가중치 값은 같은 차수의 기존 패널 가중치 값에 비하여 약 절반 정도이다. 신규 패널의 가중치는 전체 가구, 일반 가구, 저소득가구에서 서로 비슷한 값을 지닌다. 차수가 지남에 따라 일반, 저소득 모두 평균 가중치 값이 증가함을 알 수 있고 가중치의 분산 역시 증가함을 알 수 있다.

Table 2. Distribution of Survey Weights for Old Panel in KOWEPS

	최대	75%	50%	25%	최소	평균	CV
1차	9,321	3,090	1,986	1,138	160	2,247	154
2차	15,053	3,340	2,111	1,264	131	2,456	151
3차	28,136	3,578	2,078	1,140	95	2,600	131
4차	20,714	3,730	2,136	1,133	48	2,686	128
5차	21,593	3,885	2,193	1,150	83	2,804	126
6차	21,633	4,164	2,281	1,178	75	2,991	122
7차	34,994	3,709	2,116	1,116	49	2,741	118
8차	10,399	4,022	2,266	1,177	87	2,884	129
9차	11,282	4,216	2,400	1,221	66	3,021	128
10차	11,795	4,283	2,464	1,245	71	3,125	125
11차	12,171	4,548	2,524	1,250	65	3,260	125
12차	12,643	4,766	2,604	1,293	69	3,368	125
13차	13,300	4,829	2,649	1,294	70	3,464	122

Soucre: 오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019)

Table 3. Distribution of Survey Weights for New Panel in KOWEPS

	최대	75%	50%	25%	최소	평균	CV
1차	9,321	3,090	1,986	1,138	160	2,247	154
2차	15,053	3,340	2,111	1,264	131	2,456	151
3차	28,136	3,578	2,078	1,140	95	2,600	131
4차	20,714	3,730	2,136	1,133	48	2,686	128
5차	21,593	3,885	2,193	1,150	83	2,804	126
6차	21,633	4,164	2,281	1,178	75	2,991	122
7차	34,994	3,709	2,116	1,116	49	2,741	118
8차	10,399	4,022	2,266	1,177	87	2,884	129
9차	11,282	4,216	2,400	1,221	66	3,021	128
10차	11,795	4,283	2,464	1,245	71	3,125	125
11차	12,171	4,548	2,524	1,250	65	3,260	125
12차	12,643	4,766	2,604	1,293	69	3,368	125
13차	13,300	4,829	2,649	1,294	70	3,464	122

Soucre: 오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019)

3. 마모패턴 특성 분석

1차부터 13차까지 한국복지패널의 마모패턴 특성을 분석한다. 먼저 표본가구의 마모패턴을 나타내는 변수를 생성한다. 기존패널의 경우 총 13차에 걸쳐 조사가 이루어졌으므로 자릿수는 13자리로 하고, 각 자리는 조사연도를 나타내는 것으로 한다. 각 연도에서 패널은 응답을 하여 패널로 유지가 되거나 아니면 응답 거절 등으로 패널에서 제외된다. 패널유지인 경우에 '1'의 값을 지정하고, 만일 패널마모인 경우는 '0'의 값을 지정한다. 신규패널 가구는 7차부터 조사에 포함되고 분가가구는 분가 이후에 패널에 포함되기 때문에 신규패널 가구와 분가가구는 패널에 포함되지 않는 연도가 존재한다. 이러한 경우에는 '.'의 값을 지정한다.

기존 패널과 신규 패널에서 패널 가구를 원가구, 분가가구, 기타가구로 구분한다. 원가구는 13차 조사에서 계속 응답을 한 가구와 계속 응답을 하다가 일정시점부터 응답을 하지 않아 패널 마모가 일어난 가구이다. 분가가구는 1차 이후에 패널에 진입한 가구이며, 기타가구는 1차 혹은 그 후 차수까지 응답을 연속으로 하고 그 후 응답을 거절하였다가 다시 응답한 마모패턴이 불규칙한 가구이다.

13차 조사 기존 패널은 1차부터 13차까지 모두 응답한 가구 4,170가구와 분가가구와 기타가구 827가구(분가가구 731가구, 기타가구 96가구)를 합쳐 총 4,997가구로 이루어져 있다. 13차 조사 신규 패널은 7차부터 13차까지 모두 응답한 가구는 1,389가구와 분가가구와 기타가구 88가구(분가가구 85가구, 기타가구 3가구)를 합쳐 총 1,477가구가 조사 완료되었음을 알 수 있다. 13차까지 전체가구 중 원가구의 비율은 약 94% 이상으로 상당히 높은 편이다. 기존 패널과 신규패널의 마모패턴 현황은 〈Table 4〉과 〈Table 5〉에 각각 나타나 있다.

Table 4. Panel Attrition Pattern of Old Panel in KOWEPS

가구구분	범주	마모패턴	가구수
원가구	p1	"1111111111111"	4170
	p2	"1111111111110"	139
	p3	"1111111111100"	165
	p4	"1111111111000"	206
	p5	"1111111110000"	124
	p6	"1111111100000"	191
	p7	"1111111000000"	146
	p8	"1111110000000"	131
	p9	"1111100000000"	289
	p10	"1111000000000"	251
	p11	"1110000000000"	205
	p12	"1100000000000"	384
	p13	"1000000000000"	527

Source: 오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019)

Table 5. Pannel Attrition Pattern of Old Panel in KOWEPS

가구구분	범주	마모패턴	가구 수
원가구	np1	"1111111"	1389
	np2	"1111110"	36
	np3	"1111100"	52
	np4	"1111000"	55
	np5	"1110000"	59
	np6	"1100000"	96
	np7	"1000000"	110

Source: 오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019)

가구소득과 패널 이탈과의 관계 분석을 위하여 저소득 가구의 비율을 패널 마모패턴별로 비교 분석한다. 일반 가구(1)와 저소득 가구(2)로 구분되어 있는 균등화 소득 에 따른 가구 구분 변수를 이용해 저소득 가구를 정의한다. 〈Table 6〉과 〈Table 7〉은 기존 패널 및 신규 패널에 대해 마모패턴별 저소득 가구 비율을 나타낸 것이다. 13차까지 모두 응답한 패널 가구의 저소득 가구 비율이 상대적으로 일찍 탈락한 패널에 비하여 높게 나타난다. 이는 기존 패널과 신규 패널 모두 같은 경향으로 나타나고 있다. 또한 신규 패널의 저소득 가구 비율이 전체적으로 기존 패널의 저소득 가구 비율보다 높게 나타난다. 따라서 패널 이탈에 있어 소득이 높은 가구가 소득이 낮은 가구에 비하여 탈락 확률이 높을 것이라고 예상된다. 이러한 관계를 패널 이탈 확률 모형을 통해 추정하고자 한다.

Table 6. The Proportion of Low-income Household by Panel Attrition Pattern of Old Panel in KOWEPS

마모패턴	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차	11차	12차	13차
p1	0.487	0.421	0.433	0.423	0.412	0.424	0.430	0.409	0.418	0.411	0.410	0.418	0.423
p2	0.504	0.475	0.468	0.482	0.446	0.424	0.446	0.410	0.432	0.432	0.439	0.489	
p3	0.521	0.491	0.485	0.448	0.461	0.467	0.461	0.467	0.479	0.430	0.448		
p4	0.466	0.442	0.437	0.432	0.447	0.417	0.413	0.388	0.374	0.388			
p5	0.516	0.444	0.492	0.468	0.460	0.452	0.500	0.444	0.435				
p6	0.539	0.455	0.492	0.476	0.455	0.476	0.445	0.461					
p7	0.568	0.500	0.548	0.527	0.507	0.514	0.479						
p8	0.588	0.542	0.542	0.527	0.527	0.511							
p9	0.474	0.377	0.408	0.426	0.401								
p10	0.446	0.386	0.402	0.398									
p11	0.415	0.346	0.346										
p12	0.318	0.260											
p13	0.271												
전체	0.464	0.414	0.431	0.420	0.408	0.416	0.410	0.389	0.392	0.384	0.381	0.382	0.381

Source: 오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019)

Table 7. The Proportion of Low-income Household by Panel Attrition Pattern of New Panel in KOWEPS

마모패턴	7차	8차	9차	10차	11차	12차	13차
np1	0.544	0.528	0.549	0.544	0.542	0.535	0.550
np2	0.528	0.611	0.556	0.500	0.556	0.444	
np3	0.596	0.558	0.577	0.635	0.577		
np4	0.582	0.582	0.564	0.545			
np5	0.441	0.424	0.475				
np6	0.427	0.406					
np7	0.364						
전체	0.526	0.522	0.546	0.539	0.535	0.517	0.534

Source: 오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019)

Ⅲ. 비랜덤결측 가정 하에서 패널 이탈 확률 모형 추정

1. 결측자료 모형

결측자료 모형은 응답확률과 조사변수 간의 종속적 관계를 기반으로 세 가지로 분류할 수 있다(Little and Rubin, 2002). 논의를 위해 먼저 조사변수와 보조변수를 각각 Y와 X로 표기하고 개별개체의 응답 지시자를 R로 정의한다. R은 응답하였으면 1 응답하지 않았다면 0의 값을 갖는다. 즉 $R=0$인 경우 Y는 관측 되지 않고 $R=1$인 경우만 Y는 관측된다. X는 R과 상관없이 항상 관측된다고 가정한다.

응답확률이 조사변수 Y와 보조변수 X와 서로 무관한 형태인 무응답으로 완전랜덤무응답(MCAR: missing completely at random)이라 부른다. MCAR은 마치 무응답자가 전체 표본으로부터 임의로 확률추출된 것으로 자료 분석에서 무응답의 영향력이 무시될 수 있는 완전히 임의적인 상황을 뜻하며 응답 자료만 갖고 분석하여도 전체를 대표하는데 문제가 없다.

응답확률은 조사변수 Y에 좌우되지 않지만 보조변수 X의 일부 혹은 전부에 좌우되는 형태인 무응답으로 랜덤무응답(missing at random, MAR) 혹은 무시가능무응답(ignorable non-response)이라고 부른다. MAR은 보조변수 X에 의존하는 응답모형을 통해 충분히 설명할 수 있게 된다. 즉 보조변수 X가 주어지면 조사변수와 응답 지시자는 서로 독립의 관계($Y \perp R|X$)가 성립된다. 랜덤무응답 가정하에 응답확률은 다음과 같이 표현할 수 있다.

$$f(R|Y,X) = f(R|X) \tag{1}$$

응답확률이 하나 혹은 그 이상의 조사변수 Y에 좌우되는 형태의 무응답은 비랜덤무응답(not missing at random, NMAR) 혹은 무시불가무응답(non-ignorable non-response)이라 부른다. 따라서 무응답 메커니즘에 대한 조건부 확률은 더 이상 단순화되지 않게 된다.

본 연구에서는 패널 이탈 확률 모형에 대하여 세 가지 무응답 발생체계 중 랜덤무응답 가정과 비랜덤무응답 가정을 고려하고자 한다. 랜덤무응답 상황이라면 패널이탈 확률 모형의 설명 변수로 이전 차수에서 관측된 변수를 사용할 수 있기 때문에 모형에 대한 가정은 식 (1)을 통해서 이루어진다. 하지만 비랜덤무응답 가정하에서는 해당차수의 응답여부에 해당 차수의 조사 변수가 영향을 주기 때문에 무응답한 조사 변수에 대한 처리가 필요하다.

2. 비랜덤결측 모형 가정 하에서 패널 이탈 확률 모형 추정 방법

패널 이탈 확률 모형에서 다루고자 하는 자료 구조는 〈Table 8〉과 같다. 해당 차수의 패널 응답 여부를 R, $R=1$인 경우에만 응답 한 조사 자료를 Y라 한다. 보조변수 X는 응답여부 R과 상관없이 항상 응답한 변수로 이전 차수의 조사 자료를 이용하여 가정한다.

Table 8. Data Structure of Panel Attrition Probability Model

응답여부	R	Y	X
응답	1	O	O
무응답	0	X	O

즉, 이전 차수의 패널 중 다음 차수에 응답한 패널은 $R=1$로 표기되며 $R=1$인 패널에 대해서만 조사 자료 Y가 관측된다. 여기서 이전 차수의 패널 표본을 A라고 하고 다음 차수의 표본을 A_v라 하자.

여기서 우리의 관심 모수는 표본 A에서 탈락하여 다음 차수 표본A_v에 포함 되지 않을 확률, 패널 이탈 확률 모형을 결정하는 모수 θ라 하자.

$$f(R=0|Y,X)=f(R=0|Y,X;\theta) \tag{2}$$

랜덤무응답 가정하에서 패널 이탈 확률 모형은 다음과 같이 표현되어 모두 응답한 이전 차수 조사 자료 X를 설명 변수로 하여 모수 θ를 추정할 수 있다.

$$f(R=0|Y,X)=f(R=0|X;\theta) \tag{3}$$

비랜덤무응답의 경우 단순 임의표본(simple random sample)을 가정하고 관측치로부터 얻어지는 관측 가능도 함수(observed likelihood function)은 다음과 같이 표현된다.

$$L_{obs}(\theta,\alpha)=\prod_{i=1}^{n}\{f(r_i|y_i;\theta)g(y_i|x_i;\alpha)\}^{r_i}\left\{\int f(r_i|y_i;\theta)g(y_i|x_i;\alpha)dy\right\}^{1-r_i} \tag{4}$$

여기서 $g(y_i|x_i)$는 조건부 분포를 나타내는 확률 밀도 함수이며 $g(y|x)=g(y|x;\alpha)$를 따른 다는 모수적 모형 가정을 사용하였다. 또한 X는 Y에 대한 surrogate variable로 $f(R|Y,X)=f(R|Y)$를 가정하였다. 즉, 참값 Y를 아는 경우 X에 대한 추가적인 정보는 R과 Y의 조건분 분포를 예측 하는데 아무런 도움이 되지 않는다는 것이다. 이러한 경우 X를 surrogate variable이라고 한다.

이러한 우도함수를 최대화하는 최대 우도 추정량 은 EM 알고리즘을 이용하여 쉽게 계산될 수 있다. 이 경우 사용되는 EM 알고리즘 방법은 다음과 같이 기술된다.

[Step 1] validation sample로부터 모수에 대한 초기치 $\hat{\theta}^{(0)}$와 $\hat{\alpha}^{(0)}$를 계산한다.

[Step 2] 현재 step(t-step)의 모수 추정치 $\hat{\theta}^{(t)}$와 $\hat{\alpha}^{(t)}$를 이용하여, 다음의 θ와 α에 관한 mean score function을 계산한다.

$$\overline{S}_1(\theta|\hat{\theta}^{(t)},\hat{\alpha}^{(t)})=\sum_{i=1}^{n}w_i\left[r_iS(\theta;r_i,y_i)+(1-r_i)E\left\{S(\theta;r_i,Y)|r_i,x_i;\hat{\theta}^{(t)},\hat{\alpha}^{(t)}\right\}\right] \tag{5}$$

$$\overline{S}_2(\alpha|\hat{\theta}^{(t)},\hat{\alpha}^{(t)})=\sum_{i=1}^{n}w_i\left[r_iS_2(\alpha;x_i,y_i)+(1-r_i)E\left\{S_2(\alpha;x_i,Y)|r_i,x_i;\hat{\theta}^{(t)},\hat{\alpha}^{(t)}\right\}\right] \tag{6}$$

[Step 3] 다음 step의 모수 추정치를 얻기 위하여, Step2의 mean score function을 0을 만들어주는 해를 찾는다. 즉, $\hat{\theta}^{(t+1)}$와 $\hat{\alpha}^{(t+1)}$를 얻기 위하여, $\overline{S}_1(\theta|\hat{\theta}^{(t)},\hat{\alpha}^{(t)})=0$ 과 $\overline{S}_2(\alpha|\hat{\theta}^{(t)},\hat{\alpha}^{(t)})=0$ 을 만족하는 θ와 α를 찾는다.

[Step 4] Step2와 Step3의 과정을 Step3에서 얻어지는 새로운 추정치$\hat{\theta}^{(t)}$와 $\hat{\alpha}^{(t)}$가 수렴할 때 까지 반복한다.

여기서 Step 2 는 E-step, Step 3 는 M-step 이라 칭한다. E-step 을 제대로 계산하기 위해서는 Monte Carlo 근사를 실시하기도 한다. 식 (4) 와 식 (5) 에서 사용되는 w_i는 샘플링 가중치를 나타내며 score function에 대한 조건부 기댓값을 계산하기 힘든 경우 Kim (2011) 이 제안한 parametric fractional imputation 을 이용하여 θ의 최대우도 추정량을 구할 수 있다.

Ⅳ. 실증분석 결과

한국복지패널 1차부터 13차까지 가구 패널 중 12차까지 모두 응답하였던 가구 패널에 대하여 13차 조사에서 탈락할 확률에 대한 로지스틱 회귀분석을 실시한다. 패널 이탈 확률 모형에서 설명 변수로 사용한 가구 및 가구주의 인구, 사회학적 특성 변수를 재정의 하였다. 각 변수에 대한 정의와 특성은 〈Table 9〉에 나타나 있다.

Table 9. Variables Used in Model

변수	정의	특성
소득	로그 경상 소득	12차 소득은 전체 관측, 13차 소득은 응답자만 관측
성별	성별을 나타내는 가변수	1=남자, 2=여자
나이	만 나이 기준	
권역	권역 구분	1=서울, 2=광역시, 3=시 4=군, 5=도농복합
교육수준	가구주 교육 수준	1=중졸 이하, 2=고졸 이하, 3=대졸 이하, 4=대학원 이상
혼인상태	가구주 혼인상태	1=유배우자, 2=미혼, 3=사별, 이혼, 별거, 기타
가구원수	가구 내 본인을 포함한 총 가구원 수	1=1인, 2=2인, 3=3인, 4=4인, 5=5인 이상

Note: 한국보건사회연구원, 한국복지패널 1~13차 원자료[데이터파일]. 내부자료

설명변수가 p개인 로지스틱 회귀모형은 다음과 같다. 이는 성공할 확률이 실패할 확률의 몇 배인지를 나타내는 오즈에 로그를 취한 함수의 형태를 사용한 회귀모형이다.

$$\log\left(\frac{P(R=1|Y,X_1,\ldots,X_p)}{P(R=0|Y,X_1,\ldots,X_p)}\right)=\beta_0+\beta_1x_1+\cdots+\beta_px_p \tag{7}$$

여기서 응답여부 변수 R_i는 13차 패널에서 응답하였으면 1, 무응답이면 0을 갖는다. Y는 13차 소득으로 응답 여부에 따라 13차 패널에 대해서만 13차 소득을 알 수 있다. 성별, 나이 등 가구주의 인구, 사회학적 변수는 12차와 13차가 크게 다르지 않다는 가정하에 12차의 조사값을 사용하도록 한다. 무응답 모형의 가정에 대하여 랜덤무응답 가정과 비랜덤무응답 가정 두 경우를 고려한다.

먼저 패널 탈락 모형에 대하여 랜덤무응답 가정을 적용하면 무응답이 있는 13차 소득대신 모두 관측된 12차 소득을 설명 변수로 로지스틱 회귀분석을 실행할 수 있다. 응답여부와 12차 소득, 가구주 나이, 성별 등을 설명 변수로 사용한 로지스틱 회귀분석 결과는 〈Table 10〉에 나타나 있다.

Table 10. Result of Logistic Regression for Panel Attrition Under MAR Assumption

설명변수	최종 모형		
	B	S.E.	유의확률
상수항	3.370	1.238	0.007
로그경상소득(12차)	0.161	0.119	0.177
나이	-0.020	0.007	0.004
성별(여)	0.266	0.221	0.229
교육수준(고졸이하)	-0.103	0.198	0.603
교육수준(대졸이하)	-0.561	0.231	0.015
교육수준(대학원이상)	-0.808	0.427	0.059
혼인상태(미혼)	-0.264	0.345	0.444
혼인상태(사별, 이혼 등)	-0.056	0.228	0.805

랜덤무응답 가정하의 패널 이탈 확률 최종 모형에 포함 된 설명 변수는 12차 패널 로그경상소득, 가구주

나이, 성별, 교육수준, 혼인상태이다. 즉 가구의 패널 탈락 여부가 가구의 기본적인 특성 변수들에 영향을 받는 것으로 나타났으며, 나이가 많을수록 무응답 가능성이 높고, 남성 가구주가 여성 가구주에 비해 무응답 가능성이 높게 나타났다. 혼인 상태별로는 유배우자에 비하여 미혼의 무응답 확률이 높게 나타났다. 교육 수준은 중졸 이하에 비하여 교육수준이 높아질수록 패널 탈락 확률이 높게 나타났다. 랜덤무응답 가정하에서 조사변수인 소득의 영향을 살펴보면 소득이 높아질수록 탈락 확률이 낮게 나타나지만 이에 대한 통계적 유의성을 살펴보면 유의하지 않음을 알 수 있다.

비랜덤무응답 가정 하에서는 로그경상소득 변수로 13차 조사 자료를 사용하여야 한다. 13차 패널에 대해서만 소득 응답값이 있으므로 13차 탈락 패널의 소득 무응답 값에 대하여 III장에서 제안 하였던 EM 알고리즘 방법을 적용하여 패널 탈락 모형에 대한 로지스틱 회귀분석을 실행한다. 응답여부와 13차 소득, 가구주 나이, 성별 등을 설명 변수로 사용한 로지스틱 회귀분석 결과는 〈Table 11〉에 나타나 있다.

Table 11. Result of Logistic Regression for Panel Attrition Under NMAR Assumption

설명변수	최종 모형		
	B	S.E.	유의확률
상수항	11.485	1.228	0.000
로그경상소득(13차)	-0.669	0.115	0.000
나이	-0.042	0.007	0.000
성별(여)	0.128	0.225	0.569
교육수준(고졸이하)	0.140	0.199	0.483
교육수준(대졸이하)	-0.208	0.231	0.368
교육수준(대학원이상)	-0.201	0.427	0.638
혼인상태(미혼)	-1.052	0.340	0.002
혼인상태(사별, 이혼 등)	-0.434	0.230	0.059

비랜덤무응답 가정하의 패널 이탈 확률 최종 모형에 포함 된 설명 변수는 13차 패널 로그경상소득, 가구주 나이, 성별, 교육수준, 혼인상태이다. 나이가 많을수록 무응답 가능성이 높고, 남성 가구주가 여성 가구주에 비해 무응답 가능성이 높게 나타났다. 혼인 상태별로는 유배우자에 비하여 미혼의 무응답 확률이 높게 나타났다. 교육 수준은 중졸 이하에 비하여 교육수준이 높아질수록 패널 탈락 확률이 높게 나타났으나 통계적 유의성은 없는 것으로 나타났다. 랜덤무응답 가정하에서 13차 조사변수 소득의 영향을 살펴보면 소득이 높아질수록 탈락 확률이 높게 나타났으며 이에 대한 통계적 유의성도 강하게 나타남을 알 수 있다.

패널 이탈 확률 모형에 대하여 랜덤무응답 가정과 비랜덤무응답 가정하에서 로지스틱 회귀분석 결과를 살펴보았다. 소득이 패널 탈락에 미치는 영향이 두 가정의 경우 서로 상반되게 나타났다. 랜덤무응답 가정의 경우 소득의 영향의 통계적 유의성이 부족하며 비랜덤무응답 가정하에서 소득의 영향은 〈Table 6〉과 〈Table 7〉의 결과에서 확인하였던 패널 이탈에 있어 소득이 높은 가구가 소득이 낮은 가구에 비하여 탈락 확률이 높다는 결과와 일치함을 알 수 있다. 따라서 패널 마모패턴 특성 분석 등의 결과와 비교하였을 때 비랜덤무응답 가정하에서 분석하여 얻어진 소득이 높을수록 패널 탈락의 확률이 높아지는 결과가 타당함을 알 수 있다.

V. 결론

본 연구에서는 한국복지패널에서 패널마모 패턴에 다른 가구 특성 차이 분석을 통한 패널조건화 현상을 살펴보았다. 가구소득 추정에 있어서 기존패널과 신규패널의 차이 발생 여부를 살펴보고 그 원인이 패널

마모에 따른 가구 특성 차이 때문일 수 있음을 보였다.

실증분석 결과에 따르면, 한국복지패널의 패널마모는 단조마모패턴으로 발생하며, 가구의 특성과 가구소득 추정에 영향을 미친다. 패널이 오래될수록 기존패널에서는 일반가구 비율이 증가하지만 신규패널에서는 차수가 지나도 그 비율이 유사한 값으로 유지된다. 이로 인하여 기간이 긴 기존패널의 가구소득은 기간이 짧은 신규패널의 가구소득 보다 더 크게 추정되는 경향이 있다고 추론할 수 있다.

패널 이탈 확률 모형에 대하여 랜덤무응답 가정과 비랜덤무응답 가정하에서 로지스틱 회귀분석 결과를 살펴보았다. 소득이 패널 탈락에 미치는 영향이 두 가정의 경우 서로 상반되게 나타났는데 패널 마모패턴 특성 분석 등의 결과와 비교하였을 때 비랜덤무응답 가정하에서 소득이 높을수록 패널 탈락의 확률이 높아지는 결과가 타당함을 알 수 있다.

패널조사에서 패널조건화 현상은 불가피하긴 하지만 조사결과를 바탕으로 한 통계적 추론에 있어 패널조건화로 인한 왜곡이 발생하지 않아야 한다.

References

Kim, Jae-Kwang (2011), “Parametric Fractional Imputation for Missing Data Analysis”, *Biometrika*, 98, 119-132.

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

오미애, 이혜정, 신재동, 이계오, 박승환, 손창균 (2019), *2019년 한국복지패널 심층분석 보고서*, 한국보건사회연구원.