# 패널조사 응답지속성에 관한 연구
# : 한국노동패널조사를 중심으로

이 경 희* · 민 인 식**

여러 차수에 걸쳐 추적조사가 이루어지는 패널조사의 특성상, 중도에 조사대상자의 응답 거부·중단이 발생할 수 있고, 이는 샘플 수 감소뿐만 아니라 표본의 대표성 훼손으로 이어질 수 있다. 또한 이러한 표본이탈 또는 탈락(attrition) 문제는 패널기간이 길어질수록 심해질 가능성이 크며 추정결과의 편의도 유발할 수 있다. 이러한 문제에 대한 가장 근본적인 해결방법은 패널조사의 표본이탈률 자체를 줄이는 것일 것인데, 이를 위해서는 표본이탈의 과정과 패턴에 대한 정확한 이해가 필요하다. 이에 본 연구는 우리나라의 대표적 패널조사인 『한국노동패널조사』 1차~17차년도 자료를 이용하여 패널탈락이 어떠한 요인에 의해 발생하는지, 각 요인의 영향력은 어느 정도인지 분석하고 이를 이용하여 패널탈락 위험군 및 탈락시점에 대해 예측한다. 분석방법으로는 패널지속기간(duration)이 증가함에 따라 패널탈락 위험률(hazard rate)이 일정하게 변한다기보다는 어느 특정시점까지는 높아지다가 그 이후에는 다시 낮아진다고 가정하는 로그로지스틱 생존분석 모형을 사용한다. 분석 결과, 여성 가구주인 경우, 비광역시에 거주하거나 가구주가 임금근로자인 경우, 가구주 연령이나 교육수준, 가구소득이 높을수록, 그리고 가구원 수가 적을수록 패널지속 기간이 길어지는, 즉 패널탈락 위험이 낮아지는 것으로 나타났다. 이를 바탕으로 한 예측 결과는, 각 가구의 특성에 따라 패널탈락 예상 시점이 달라지며, 특히 가구주가 남성이거나 학력이 고졸미만, 비임금근로자 또는 비취업자인 경우, 광역시에 거주하거나 가구주 연령이 45세 미만인 가구의 경우, 상대적 위험군으로 분류가능함을 보여준다. 또한 『한국노동패널조사』와 같은 장기 패널조사의 경우 대표성과 신뢰성, 지속성을 확보하기 위해서는 이러한 패널탈락 위험군에 대한 사전관리가 필요함을 시사한다.

## Ⅰ. 서 론

미시자료를 이용한 실증분석을 해본 경험이 있는 사회과학 연구자들은 아마 진정한 효과(true effect)나 인과관계를 정확하게 밝혀내는 것이 쉬운 작업이 아님을 잘 알 것이다. 다른 모든 조건을 통제할 수 있는 실험상황과는 달리, 복잡다단한 관계들이 얽혀 있는 현실세계의 현상을 분석하기 위해서는 가능한 한 많은 요인들을 되도록 여러 번에 걸쳐 관찰, 측정하고 이를 분석에 이용해야 한다. 그러나 현실적으로 모든 요인들이 관측가능하지도 않을뿐더러 관측가능하다 하더라도 전부를 측정, 조사할 수는 없기 때문에, 많은 연구자들은 관측되지 않은 요인들의 효과를 방법론적으로라도 통제할 수 있는 패널 자료를 선호한다.

---

* 한국노동연구원 연구위원 (kheelee@kli.re.kr)
** 경희대학교 경제학과 교수 (imin@khu.ac.kr)

이러한 수요가 반영되어서인지 국내외를 막론하고 수많은 패널조사들이 만들어지고 패널 데이터가 축적되고 있다. 그런데 여러 차수에 걸쳐 추적조사가 이루어지는 패널조사의 특성상, 중도에 조사대상자의 응답 거부·중단이 발생할 수 있고, 이는 샘플 수 감소 및 표본의 대표성 상실로 이어질 수 있다. 또한 Hausman & Wise(1979)를 비롯한 여러 연구들에 의하면, 이러한 표본이탈/탈락(attrition)은 추정결과의 편의(bias)를 유발할 수 있다. 표본이탈로 인한 추정결과의 편의를 보정하기 위한 다양한 방법론들이 개발, 사용되고 있지만, 가장 근본적인 해결책은 아마도 패널조사의 표본이탈률 자체를 줄이는 것일 것이며, 이를 위해서는 표본이탈의 과정과 패턴에 대한 정확한 이해가 필요하다.

이에 본 연구에서는 우리나라의 대표적 패널조사인 『한국노동패널조사』(KLIPS)에서 설문응답 지속성(지속기간)에 영향을 미치는 요인들이 무엇인지, 영향력은 어느 정도인지 분석하고 이를 이용하여 표본이탈 위험군 및 이탈시점에 대해 예측함으로써, 표본이탈률을 낮추고 패널 데이터의 공신력과 지속성을 높이는데 도움이 되는 방향을 알아보고자 한다. 이를 위해 본 연구에서는 1~17차년도 KLIPS 자료와 로그로지스틱 생존분석 모형을 이용하며 median duration 예측 값을 추정한다. 본 연구의 나머지 부분은 다음과 같이 구성되어 있다. 이어지는 다음 장에서는 패널탈락을 분석한 선행연구들을 소개하고, 제3장에서는 분석 데이터 및 기초통계, 패널탈락 여부에 따른 각 설명변수의 차이 검증 결과를 제시한다. 제4장에서는 계량분석 방법론과 분석결과에 대해 설명하고, 마지막으로 제5장에서는 분석결과를 요약하고 시사점을 제시한다.

## II. 선행연구 검토

패널조사를 유지하는 동안 조사대상의 탈락을 최소화 또는 방지하는 것은 패널 데이터의 신뢰성과 유의성을 확보하기 위해 매우 중요한 의미를 갖는다. 때문에 패널탈락률 감소방안 마련을 위한 연구는 다양한 접근방식으로 이루어져왔다. 일반적으로 패널탈락 연구는 크게 두 가지 흐름, 표본 개인적 특성에 따른 영향과 조사의 환경적 요인에 따른 영향을 분석한 연구로 나뉜다.[1] 표본의 개인적 요인은 표본의 성별이나 연령, 경제활동상태, 결혼상태 등 조사대상의 인구통계학적 특성을 의미하며, 조사의 환경적 요인은 조사 진행기간, 면접원의 특성 및 직업숙련도 등의 외부적 요인을 포괄한다.

표본의 개인적 특성에 초점을 두고 선행된 연구로는 Fitzgerald, Gottschalk & Moffitt(1998)가 미국의 PSID(the Panel Study of Income Dynamics) 자료를 이용한 연구가 있는데, 여기에서는 탈락표본들의 인구통계학적 특성을 분석하여, 소득이나 교육수준이 상대적으로 낮거나 가정의 안정도 및 거주 지역적 유동성과 같은 표본의 환경이 불안정하거나 가변성이 높을수록 탈락경향성이 높은 점을 발견하였다. Lillard & Constatijn(1998) 역시 패

1) 이외에도 최근 들어서는, 패널탈락의 결정요인이 주된 관심은 아니지만, 다양한 분석에서 패널 데이터의 표본 이탈로 인한 편의(attrition bias) 문제를 해결하기 위해 패널탈락 결정식 등을 도구적 수단으로 이용하는 연구들이 이루어지고 있다(예. Cheng and Trivedi(2015) 등)

널탈락에 있어 표본의 개인적인 특성에 초점을 두어 연구하고 결혼상태 등에 변화가 생길 경우 탈락률에 영향을 준다고 주장하였다. 우리나라에서도 패널자료가 만들어지기 시작한 이후 2000년대 들어 패널탈락에 관한 연구가 점차적으로 이루어지고 있다. 김대일·남재량·류근관(2000)은 한국노동패널(KLIPS) 표본의 대표성 검증 및 표본 이탈자 특성연구를 통해 표본의 성별이나 연령, 결혼유무, 교육수준, 경제활동상태 등의 차이에 따라 이탈가능성에 차이를 보인다고 지적하였다. 구체적으로 여성보다는 남성의 경우, 학력이 높을수록, 연령이 낮을수록, 미혼 또는 이혼이나 별거로 배우자와 함께하지 않는 경우일수록, 고용상태가 불안정할수록 패널탈락률이 상대적으로 높게 나타났다. 이는 교육수준과 혼인여부, 취업상태에 따라 상이한 패널탈락 경향성을 보인다는 점을 지적한 이상호(2005)의 연구결과와도 유사하다. 이상호(2005)에 따르면, 가구주의 나이가 많을수록, 거주형태가 자가인 경우 탈락률이 낮은 경향이 있는 반면, 소득수준이 높거나 결혼상태가 미혼인 경우 탈락률은 높아졌다.

한편, 표본의 개인적 특성보다는 면접원의 특성에 초점을 두어 패널조사 응답결과를 살펴본 연구결과도 있다. 신선옥(2008)은 면접원의 경력여부와 교육수준에 따라서 표본대상들이 얼마나 조사에 협조하는지 여부가 달라질 수 있다고 밝혔다. Zabel(1998) 역시 조사대상자들의 인구통계학적 요인들보다 면접원이나 면접 과정이 패널탈락에 더 크게 영향을 미칠 수 있다고 지적하였다. 기존의 패널탈락에 관한 선행연구는 조사의 환경적 요인보다 표본의 개인적 특성에 초점을 둔 것이 많지만 이러한 선행연구들 역시 면접원 특성이나 조사의 특성과 같은 외부 환경적 요인 변수도 분석에 반영할 필요가 있다고 지적한 바 있다.(이상호, 2005)

이상협 외(2010)는 분석 시 표본의 개인적 특성과 조사의 환경적 요인 두 가지를 모두 고려하였다. 그는 1998년부터 2008년까지 11년간 조사된 KLIPS의 가구주 관련 자료를 이용하여 연령이나 거주지역, 학력, 결혼상태 및 입주형태 등 표본개인의 인구통계학적 특성과 조사 진행기간이나 방식 등 외부적 요인이 패널탈락에 어떠한 영향을 끼치는지를 생명표방법과 Cox 비례위험모형을 통해 분석하였다. 분석을 통해 확인된 패널탈락의 특징은 다음과 같다. 연령의 경우, 60세 이상의 표본이탈이 가장 적고 30세 이하의 표본이탈이 가장 많았으나 조사기간이 길어질수록 고령으로 인한 건강상의 문제가 영향을 미쳐 60세 이상의 패널탈락률이 증가하며 두 집단 간 차이가 감소함을 보이고 있다. 거주지역의 경우, 서울>경기도>중부지방>남부지방 순으로 패널탈락률이 높은 것으로 나타났다. 학력의 경우, 초중고 교육수준에서는 미미하였으나 전문대 이상 교육수준에서는 탈락률이 높아져 교육수준이 높아질수록 패널탈락률이 커짐을 보였다. 결혼상태에 있어서는 기혼보다는 미혼의 경우가, 거주형태에 있어서는 자가 거주자보다 세입자인 경우가 패널탈락률이 높은 것으로 나타나 환경변동성이 높을수록 탈락률이 높아질 수 있음을 보여주었다. 조사방법의 경우, 면접, 유치, 전화·혼합이 변수 값으로 설정되었으며 패널탈락의 위험은 유치>전화·혼합>면접 순으로 나타났다.

이상의 선행연구들의 결과를 종합해보면 패널조사의 표본탈락률은 표본의 인구통계학적 특성이라는 내부적 요인과 조사환경에 따른 외부적 요인 모두에서 영향을 받는 것으로 나타나고 있다. 특히 연령이 낮을수록, 학력이 높을수록, 결혼상태나 고용상태가 불안정할수록 패

널탈락률이 높다는 것이 공통적으로 확인되고 있다. 본 연구에서는 이러한 선행연구들을 참고하여 주로 표본의 인구통계학적 특성에 주안점을 두고 총 17차의 최신 KLIPS 자료를 토대로 패널조사 응답지속성(패널탈락)에 영향을 미치는 요인들이 무엇인지 파악하고자 한다. 또한 기존 선행연구들과는 달리 설명변수들의 영향력 추정뿐만 아니라 이에 기초한 탈락시점 예측을 통해 보다 실질적으로 패널유지율을 제고할 수 있는 방안에 대한 시사점을 모색해보고자 한다.

# III. 연구데이터 및 기초통계 분석

## 1. 연구데이터

본 연구에 사용된 자료는 한국노동연구원(Koreal Labor Institute: KLI)에서 1998년부터 조사하기 시작한 한국노동패널(Korean Labor and Income Panel Survey: KLIPS) 자료이다. 2015년 현재 KLIPS는 17차(2014년) 조사가 완료되어 (학술대회용으로) 공개되어 있다. 1998년 제주도를 제외한 전국에서 추출된 5,000개 가구표본과 15세 이상 생산가능인구에 속하는 13,317명의 가구원으로 시작하였다. 17년 간 진행되어온 패널조사로 장기간에 걸쳐 성공적으로 구축되고 있기 때문에 각종 노동정책의 수립 및 평가 그리고 노동관련 연구의 활성화를 위한 필수적인 자료로 자리매김하고 있다.

다만, 동일한 가구와 가구원을 조사해야 하는 패널조사의 특성 때문에 원 가구(initial households)를 지속적으로 패널조사에서 포함하는 것이 쉬운 일은 아니다. 조사차수가 증가할수록 원 가구가 지속적으로 줄어드는 것은 문제가 된다. 이러한 패널탈락(panel attrition)이 어느 시점에서 주로 발생하고 어떠한 요인에 의해 발생하는지 분석하는 것이 본 연구의 주요 내용이다. 패널탈락을 연구하는 방법론으로 생존분석(survival analysis)과 로짓분석(logit analysis)을 선택할 수 있다. 생존분석에서 관심변수는 패널탈락까지 걸린 시간(사건발생 시간)이고 로짓분석에서는 시간과 무관하게 패널탈락 발생여부가 관심변수이다. 본 연구에서는 사건발생까지 걸린 시간을 관심변수로 설정한 생존분석 방법론을 선택한다.

KLIPS 1차(1998년) ~ 17차(2014년) 조사에서 패널탈락 시점, 즉 조사지속 기간을 연구의 관심변수로 설정한다. 그러나 패널탈락 여부를 살펴보면 최초 탈락가구가 지속적으로 탈락하는 것이 아니고 추후 다시 패널에 복귀하는 경우도 있다. 이런 가구의 경우에는 최초 탈락시점까지만 연구의 관심대상으로 삼는다. 즉 모든 조사대상 가구에는 사건발생까지 걸린 시간(이하 duration 변수)이 한 번씩만 존재하도록 연구데이터를 구축하였다. 생존분석에서는 사건이 발생한 가구도 있지만 마지막 조사인 17차 조사까지 사건이 발생하지 않은 (right-censored) 가구도 존재한다. right-censored 가구인 경우에는 조사시작 시점부터 17차 조사시점까지의 기간을 관심변수로 계산하였다.

관심변수인 $T$ (사건발생까지 시간)는 연속형 변수로 간주한다. 가령 1차 조사에서 시작한

가구가 3차 조사시점에서 탈락했다면 $T=3$이 된다. 3차 조사시점 직전($T-\triangle t$)까지는 탈락한 것으로 간주되지 않고 3차 조사시점에서 패널탈락으로 결정되었다고 간주한다. 1차 조사에서 시작해서 17차까지 계속 조사된 가구는 right-censored 가구로서 $T=17$ 을 갖게 된다. 또한 17차 조사에서 신규로 들어온 가구는 17차 조사 한 번만 조사에 답하고 아직 탈락하지 않았기 때문에 $T=1$ 이 된다. 표 1에서는 1차 ~ 17차 조사의 가구 표본 수와 응답률을 정리해서 제시하고 있다.

<표 1> KLIPS 1차 ~ 17차 조사표본과 응답률

| 조사차수 | 응답=0 | 응답=1 | 조사대상 표본 수 | 응답률 (%) |
|---|---|---|---|---|
| 1차 | 0 | 5,000 | 5,000 | 100 |
| 2차 | 622 | 4,507 | 5,129 | 87.7 |
| 3차 | 988 | 4,266 | 5,254 | 81.2 |
| 4차 | 1,189 | 4,248 | 5,437 | 78.1 |
| 5차 | 1,286 | 4,298 | 5,584 | 76.9 |
| 6차 | 1,246 | 4,592 | 5,838 | 78.6 |
| 7차 | 1,296 | 4,761 | 6,057 | 78.6 |
| 8차 | 1,402 | 4,849 | 6,251 | 77.5 |
| 9차 | 1,448 | 5,001 | 6,449 | 77.5 |
| 10차 | 1,558 | 5,069 | 6,627 | 76.4 |
| 11차 | 1,685 | 5,116 | 6,801 | 75.2 |
| 12차 | 1,835 | 6,721 | 8,556 | 78.5 |
| 13차 | 2,153 | 6,683 | 8,836 | 75.6 |
| 14차 | 2,339 | 6,686 | 9,025 | 74.0 |
| 15차 | 2,437 | 6,753 | 9,190 | 73.4 |
| 16차 | 2,604 | 6,785 | 9,389 | 72.2 |
| 17차 | 2,699 | 6,838 | 9,537 | 71.7 |

주) 12차(2009년) 이후 수치는 12차년도 당시 응답가구를 원 표본으로 하는 통합표본만을 대상으로 한 것이 아니라 모든 조사대상 가구(12차 년도에 응답하지 않아 통합표본에 속하지 않는 가구도 포함)를 기준으로 작성함.

5,000가구로 시작한 1차 조사를 시작으로 2차 조사부터 표본탈락가구가 발생하기 시작한다. 2차 조사의 응답률은 87.7%이다. 특정 시점에서 탈락하면 다시 복귀하는 것보다는 지속적으로 탈락할 가능성이 크기 때문에 응답률은 지속적으로 감소하는 것으로 보인다. 2014년의 17차 조사에서는 9,537 가구 중 6,838가구가 조사에 응해 약 71%의 응답률을 보이고 있다.

## 2. 기초통계 분석

본 소절에서는 관심변수인 사건발생까지 걸린 시간(duration)에 대한 기초통계와 모형의 설명변수에 대해서 요약통계량을 제시한다. 먼저 1차~17차에 포함되어 있는 9,537가구 중 패널탈락 시점 또는 right-censored 시점을 계산하면 표 2와 같다.[2)]

<표 2> duration 변수의 빈도표

| duration | 패널탈락 | right-censored | 합계 | 비율 (%) |
|---|---|---|---|---|
| 1 | 0 | 143 | 143 | 1.5 |
| 2 | 1,144 | 161 | 1,305 | 13.8 |
| 3 | 740 | 120 | 860 | 9.1 |
| 4 | 576 | 138 | 714 | 7.5 |
| 5 | 386 | 195 | 580 | 6.1 |
| 6 | 277 | 1,303 | 1,580 | 16.7 |
| 7 | 190 | 109 | 299 | 3.1 |
| 8 | 148 | 107 | 255 | 2.7 |
| 9 | 99 | 117 | 216 | 2.2 |
| 10 | 117 | 105 | 222 | 2.3 |
| 11 | 111 | 119 | 230 | 2.4 |
| 12 | 111 | 125 | 236 | 2.5 |
| 13 | 93 | 73 | 166 | 1.7 |
| 14 | 73 | 78 | 151 | 1.6 |
| 15 | 52 | 48 | 100 | 1.0 |
| 16 | 47 | 47 | 94 | 1.0 |
| 17 | 30 | 2,235 | 2,265 | 24.0 |
| 합계 | 4,193 | 5,223 | 9,416 | 100 |

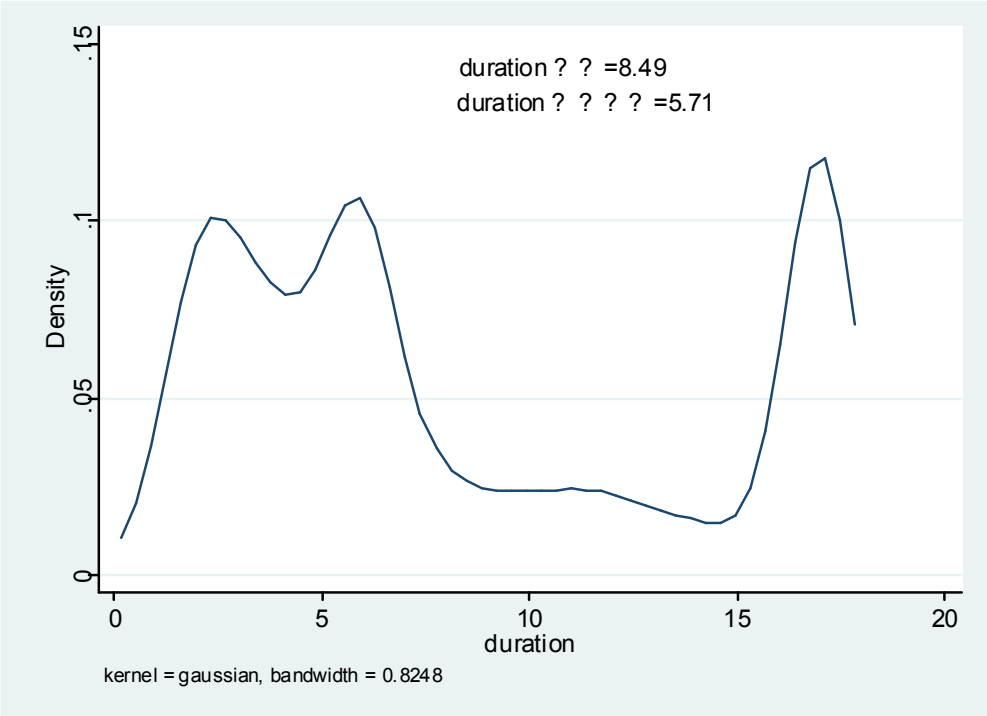
duration=1인 경우는 17차 조사에 신규로 들어와서 17차 조사에 응답한 가구에 해당한다. 모두 right-censored 표본이라고 말할 수 있다. duration=2가 꽤 높은 비율을 차지한다. 패널에 신규로 들어와서 1번 조사된 후 그 다음 해 조사에서 탈락한 경우이다. duration=2인 가구 중 1,144가구가 패널탈락 가구에 해당한다. duration=6이 16.7%로 높은 이유는 KLIPS 12차 조사에서 그 동안 탈락된 가구를 보완하기 위해 새로운 가구표본을 대폭 포함하였다. 따라서 12차에서 새로 시작된 가구들이 17차까지 지속적으로 조사된 가구, 즉 duration=6에서

2) 총 9,537 가구 중 9,416가구만을 대상으로 duration을 계산하였다. 제외된 가구는 한 번도 노동패널 조사에 응답한 적이 없고 단지 가구번호만 존재한다.

right-censored 가구가 많이 존재하기 때문이다. 1차 조사에 포함된 5,000가구 중 '17차까지 탈락하지 않고 조사된 표본+17차에서 탈락한 가구'는 2,265가구이다. 따라서 1차 원 가구의 17차까지의 유지율은 (2,235/5,000)×100=44.7%이다.[3] 노동패널 조사팀에서는 조사에 성공하지 못한 이유를 4가지 이유로 구분하고 있다.[4] 조사를 성공하지 못한 이유는 강력거절, 이사로 인한 추적실패, 접촉불가, 기타 사유로 구분한다. 탈락한 가구의 50.4%는 강력거절, 20.2%는 이사로 인한 추적실패, 20.3%는 접촉불가 사유임을 확인할 수 있다.

그림 1에서는 duration 변수의 분포를 비모수적 밀도함수(non-parametric density function)로 표현하고 있다. duration 변수 평균은 8.5년이다. 그러나 right-censored 가구가 포함되어 있기 때문에 패널탈락까지 걸린 평균 시간이라고 말할 수 없다. duration=2년, 6년 그리고 17년일 때 가장 높은 확률밀도를 보여주고 있다.[5] duration 변수의 분포는 정규분포와 전혀 다른 분포임을 예상할 수 있다.

<그림 1> duration 변수의 density function



경 변수로 구분할 수 있다. 가구주의 특성은 나이, 성별, 거주지역, 교육수준, 혼인상태, 취업형태, 가구원 수 변수를 포함하였고 경제적 환경 변수는 주거점유형태와 가구소득 변수를 선택한다. 특히 주거점유형태는 자가인 경우와 전월세인 경우로 구분하고 자가인 경우에 비해 전월세 거주자가 이사 가능성이 크고 따라서 패널추적이 실패할 가능성이 클 것으로 예상되기 때문에 설명변수로 포함한다. 대부분의 설명변수들이 조사시점에 따라 변하기 때문에(time-varying) 어떤 시점의 설명변수 값을 포함하느냐가 문제이다. 가령 1차 원 가구 중 6차에서 탈락한 가구는 duration=6가 된다. 그러나 탈락한 시점에서 설명변수 값이 존재하지 않기 때문에 $t=5$시점에서 설명변수 값을 사용할 수밖에 없다. 즉 탈락이 발생하기 직전 가구특성 변수를 이용한다. 그러나 right-censored 가구는 마지막 조사시점의 가구특성 변수 값을 선택할 수 있다. 표 3에서는 생존분석 모형에서 설명변수로 사용되는 변수의 정의와 기초통계량을 제시한다.

<표 3> 설명변수 정의와 기초통계량

| 변수이름 | 정의 | 평균 | 표준편차 |
|---|---|---|---|
| $gender$ | 가구주 성별 (1=남자 2=여자) | 1.24 | 0.42 |
| $region$ | 거주지역 (1=광역시 2=비광역시) | 1.51 | 0.49 |
| $married$ | 가구주 혼인상태<br>(1=미혼 2=기혼유배우 3=기혼무배우) | 2.08 | 0.57 |
| $edu$ | 가구주 교육수준<br>(1=고졸미만 2=고졸 3=전문대졸 이상) | 2.04 | 0.80 |
| $employed$ | 가구주 취업형태<br>(1=임금근로자 2=자영업/무급가족종사자<br>3=무직/비경제활동) | 1.76 | 0.84 |
| $hsize$ | 가구원 수 | 2.79 | 1.37 |
| $age$ | 가구주 나이 | 50.2 | 16.4 |
| $lincome$ | 가구소득(단위: 만원)의 로그값 | 7.74 | 1.00 |
| $owner$ | 주거 점유형태<br>(1=자가 0=전/월세) | 0.53 | 0.49 |

패널탈락 여부와 설명변수의 관계를 이변량 분석(bivariate analysis)을 통해 예상하고자 한다. 먼저 패널탈락 여부에 따라 연속형 변수인 가구소득, 가주주 나이, 가구원 수 변수의 평균 차이가 통계적으로 유의한지 t-검정 결과를 표 4에서 제시한다. 가구원 수가 많을수록 패널탈락 확률이 높고 가구주 나이가 낮을수록 패널탈락 가능성이 높아진다. 가구소득은 낮을수록 패널탈락 확률이 높아진다. 또한 평균차이는 1% 유의수준에서 유의함을 확인할 수 있다. 이러한 결과는 기존 선행연구들(예. Fitzgerald, Gottschalk & Moffitt(1998), 김대일・남재량・류근관(2000), 이상협 외(2010) 등)과 일치한다.

<표 4> 패널탈락 여부에 따른 t-검정

| | 패널탈락 | 패널지속 | t-value | p-value |
|---|---|---|---|---|
| $hsize$ | 2.94 | 2.66 | -9.54 | 0.000*** |
| $lincome$ | 7.45 | 7.96 | -24.04 | 0.000*** |
| $age$ | 45.4 | 54.2 | 26.11 | 0.000*** |

주) ***, **, *는 각각 1%, 5%, 10% 유의수준에서 유의함을 나타냄.

<표 5> 패널탈락 여부에 따른 카이제곱 검정

| 설명변수 | 범주 | 패널탈락 (%) | 패널지속 (%) | 카이제곱 검정 (p-value) |
|---|---|---|---|---|
| $gender$ | 남자 | 45.9 | 54.0 | 7.46 (0.00)*** |
| | 여자 | 42.5 | 57.4 | |
| $region$ | 광역시 | 50.1 | 49.8 | 86.0 (0.00)*** |
| | 비광역시 | 40.2 | 59.5 | |
| $edu$ | 고졸미만 | 40.2 | 59.8 | 41.9 (0.00)*** |
| | 고졸 | 48,4 | 51.5 | |
| | 전문대졸 이상 | 46.0 | 53.9 | |
| $employed$ | 임금근로자 | 45.2 | 54.7 | 1.64 (0.44) |
| | 자영업/무급가족 종사자 | 43.9 | 56.0 | |
| | 무직/비경활 | 45.7 | 54.2 | |
| $owner$ | 자가 | 38.2 | 61.7 | 195.1 (0.00)*** |
| | 전월세 | 52.9 | 47.7 | |
| $married$ | 미혼 | 56.7 | 43.2 | 90.3 (0.00)*** |
| | 기혼유배우 | 44.8 | 55.1 | |
| | 기혼무배우 | 39.0 | 60.9 | |

주) ***, **, *는 각각 1%, 5%, 10% 유의수준에서 유의함을 나타냄.

범주형 설명변수와 패널탈락 여부와의 관계는 카이제곱 검정을 통해 판단할 수 있다. 표 5에서는 패널탈락 여부와 범주형 변수의 이원 빈도표와 카이제곱 검정 결과를 제시한다. 남자 가구주인 경우 패널탈락 확률이 45.9%인데 여자 가구주의 탈락확률은 42.5%이다. 남자 가구주의 패널탈락 가능성이 더 높은 것으로 나타났는데, 이러한 차이는 통계적으로도 유의하다. 거주지역에 따른 패널탈락 확률 차이는 좀 더 분명하다. 광역시에 거주하는 가구주일수록 탈락확률이 높으며 이러한 차이는 1% 유의수준에서 유의하다. 가구주 교육수준에 따른 패널탈

락 확률차이도 존재한다. 고졸과 대졸인 경우 탈락확률이 높고 고졸미만인 경우 지속확률이 더 높다. 가구주 취업형태와 패널탈락 여부는 통계적으로 유의하지 않다. 주거점유형태는 패널탈락 여부와 가장 분명하게 관계가 있는 것으로 보인다. 자가인 경우 탈락확률은 38.2%이지만 전월세 가구는 52.9%로 탈락확률이 매우 높아진다. 전월세 가구의 이사 가능성이 높기 때문에 패널추적 조사실패로 이어질 가능성이 크기 때문인 것으로 보인다. 혼인상태 또한 패널탈락 가능성과 유의한 연관성을 가진다. 가구주가 기혼자인 경우(기혼유배우 44.8%, 기혼무배우 39%)에 비해 미혼인 경우(56.7%) 패널탈락 확률이 현저히 높게 나타난다.

표 4와 표 5에서는 단지 패널탈락 여부와 설명변수의 관계를 살펴보고 있고 duration과는 어떠한 관계인지 전혀 판단할 수 없다. 따라서 연구의 관심주제인 생존분석을 통해 설명변수들이 패널탈락까지 걸린 시간에 어떠한 영향을 미치는지 분석하고자 한다.

# Ⅳ. 계량방법론 및 실증분석

## 1. 계량방법론

관심변수가 사건발생까지 걸린 시간인 경우 생존분석 모형을 설정할 수 있다. duration 변수는 항상 0보다 큰 값을 가지고 정규분포(normal distribution)와는 전혀 다른 분포일 가능성이 크다. 생존분석 모형에서는 duration 자체를 추정할 수도 있지만 hazard rate을 추정하는 것에 관심을 가질 수도 있다. $t$ 시점의 hazard rate인 $\lambda(t)$는 다음과 같이 조건부 확률(conditional probability)의 극한(limit)으로 정의한다.[6)]

$$\lambda(t) = \lim_{\triangle t \to 0} \frac{\Pr(t \leq T < t + \triangle t | T \geq t)}{\triangle t} \qquad \text{(식 1)}$$

식 1을 해석하면 다음과 같다. 연속형 시간 확률변수인 $T$가 사건발생 시점이고 $t$시점 이전까지 사건이 발생하지 않았다는 조건이 주어졌을 때 $t$ 시점이 되는 순간에 사건이 발생할 확률이 된다. 본 연구에서 패널탈락 사건은 탈락 조사년도에 면접원이 조사하러 가기 전까지는 패널이 지속된다고 가정한다.

hazard rate은 다음과 같이 생존함수(survival function)와의 관계로 표현할 수 있다.

$$\lambda(t) = \frac{f(t)}{S(t)} \qquad \text{(식 2)}$$

위 식에서 $S(t)$는 $\Pr(T \geq t)$로 정의되며 생존함수이다. 또한 $f(t)$은 $1 - S(t)$의 1차 미분함

6) 사건발생 시간 $T$가 연속형 확률변수인 경우에 해당한다.

수로 정의된다.

hazard rate 추정모형은 비모수적(non-parametric), 준모수적(semi-parametric) 그리고 모수적 접근법(parametric approach)을 사용할 수 있다. 본 연구에서는 모수적 접근법을 선택한다. 모수적 접근법은 hazard rate 자체를 추정하는 proportional hazard (PH) metric이 있고 duration 변수 자체를 종속변수로 하는 accelerated failure time (AFT) metric으로 나눌 수 있다. AFT metric은 다음과 같이 duration 변수의 로그값을 종속변수로 설정한다.

$$\log(t_j) = X_j\beta + e_j \qquad \text{(식 3)}$$

위 식에서 $t_j$는 항상 0보다 큰 값이어야 한다. 모수적 접근법에서는 오차항 $e_j$에 대해 0보다 큰 값을 갖는 확률분포를 가정한다. 주로 사용하는 분포함수는 지수분포(exponential distribution), 웨이블분포(Weibull distribution) 그리고 로그로지스틱 분포(loglogistic distribution) 등이다. 지수분포 가정에서는 hazard rate이 duration($t_j$)과 무관하게 일정하다. 즉 constant hazard rate이 도출된다. 웨이블분포 가정 하에서는 hazard rate이 duration이 증가함에 따라 일정한 방향으로 감소(decreasing) 또는 증가(increasing) 형태라고 가정한다. 반면 로그로지스틱 분포 가정에서는 비단조적(non-monotone) 형태로 hazard rate이 변한다고 가정한다. 본 연구에서는 duration이 증가함에 따라 패널탈락 hazard rate이 일정하게 변한다기보다는 어느 특정시점에서 가장 높아지고 그 이후에는 다시 낮아진다고 가정하고자 한다. 패널조사에 참여하는 가구들은 매년 조사될 것으로 예상하고 있기 때문에 패널지속성을 유지하고자 할 것이다. 그러나 이사나 경제적 상황 변화에 따른 패널탈락이 어느 시점부터 나타나기 시작하고 특정 시점에서 탈락가능성이 높아질 것으로 예상한다. 이러한 hazard rate의 패턴을 반영하기 위해 로그로지스틱 분포를 가정한다.

로그로지스틱 모형에서 hazard rate은 다음과 같이 계산된다.

$$\lambda(t,X) = \frac{\mu^{1/\gamma}\ t^{(1/\gamma-1)}}{\gamma[1+(\mu t)^{1/\gamma}]} \qquad \text{(식 4)}$$

위 식에서 $\mu = \exp(-X\beta)$로 정의한다. shape parameter인 $\gamma$에 따라 hazard rate 형태가 결정된다. $\gamma < 1$이면 hazard rate이 처음에 증가하다가 어느 최대점을 지나면 그 이후에는 감소하게 된다. $\gamma \geq 1$이면 hazard rate은 꾸준히 감소한다. 식 2에서 설명하였듯이 hazard rate인 $\lambda(t)$를 구할 수 있다면 생존함수인 $S(t)$ 역시 계산할 수 있다.[7] $\gamma < 1$인 경우 duration 변수의 기댓값과 중앙값은 다음과 같이 쓸 수 있다.

$$E(t) = \int_0^\infty t f(t)dt = \int_0^\infty S(t)\,dt = \frac{1}{\lambda}\frac{\gamma\pi}{\sin(\gamma\pi)} \qquad \text{(식 5)}$$

---

7) $S(t) = 1+(\lambda t)^{1/\gamma-1}$ 이 된다.

$$median(t) = \{t : S(t) = 1/2\} \quad (식\ 6)$$

기울기 모수인 $\beta$와 부가모수(ancillary parameter)인 $\gamma$를 추정하기 위해서는 다음과 같이 로그우도함수를 쓸 수 있다. duration=6인 경우 $t = 6$ 시점에서 패널에서 탈락한 가구도 있고 right-censored 가구도 있기 때문에 두 가지 타입의 가구는 로그우도함수에서 기여분(contribution)이 서로 다르다. $t_j$에서 패널탈락 가구의 우도기여는

$$L_j = f(t_j) = S(t_j)\lambda(t_j) \quad (식\ 7)$$

이며, duration=$t_j$에서 right-censored 경우에 우도기여는

$$L_j = S(t_j) \quad (식\ 8)$$

이다. 전체 관측치에 대한 우도함수는 식 7과 식 8의 우도기여를 하나의 식으로 표현하여 다음과 같이 나타낼 수 있다.

$$L = \Pi_{j=1}^{n} L_j = \Pi\ \lambda(t_j)^{d_j} S(t_j) \quad (식\ 9)$$

위 식에서 $d_j$는 가구 타입을 나타내는 더미변수로서, right-censored 가구인 경우 $d_j = 0$, 패널탈락 가구인 경우 $d_j = 1$의 값을 갖는다. 식 9에 로그를 취한 로그우도함수를 최대화하여 $\beta$와 $\gamma$를 추정할 수 있다. 추정치 $\hat{\beta}$와 $\hat{\gamma}$을 식 4에 대입하면 hazard rate을 얻을 수 있다.

## 2. 실증분석 결과

본 소절에서는 로그로지스틱 생존분석 모형의 추정결과를 제시하고 그 결과를 해석한다. duration 변수의 로그값을 종속변수로 AFT metric에서 모수를 추정한다. 모형 1에서는 설명변수를 전혀 포함하지 않고 단지 로그로지스틱 분포모수인 $\gamma$를 추정한다. 모형 2에서는 표 3의 설명변수를 포함하여 모형을 추정한다. 모형 1과 모형 2의 추정결과는 표 6에 제시되어 있다.

$$모형\ 1:\ \log(t_j) = \alpha + e_j \quad (식\ 10)$$
$$모형\ 2:\ \log(t_j) = \alpha + \beta X_j + e_j \quad (식\ 11)$$

모형 1의 $\hat{\gamma}$ 추정치가 1보다 작기 때문에 duration이 증가함에 따라 hazard rate이 증가하다가 감소하는 패턴임을 알 수 있다. 설명변수를 포함한 모형 2에서 $\hat{\gamma} = 0.523 < 1$ 이며 역시

일관된 결과로 추정된다. AFT metric에서 $\hat{\beta}$에 대한 해석은 $\frac{\partial log(t_j)}{\partial X}$ 이 된다. $\hat{\beta}>0$인 경우, $X$변수가 증가하면 사건발생까지 걸린 시간, 즉 패널지속 기간이 길어진다. 남자 가구주에 비해 여자 가구주인 경우 패널지속 기간이 길어지며, 가구주 교육수준이 높을수록 패널지속 기간이 길어진다. 광역시에 비해 비광역시에 거주하는 가구일수록 패널지속 기간이 길어진다. 광역시 거주가구일수록 이사 가능성이 더 높기 때문으로 예상되는 결과이다. 한편, 점유형태 변수는 자가일수록 패널지속 기간이 길어지지만 통계적으로 유의하지는 않다.[8] 가구주 혼인상태 변수 또한 통계적으로 유의하지 않다. 반면, 가구주의 취업상태 변수는 통계적으로 매우 유의한 것으로 나타났는데, 임금근로자인 가구주의 패널지속 기간이 자영업이나 무직/비경활인 경우에 비해 길어짐을 확인할 수 있다. 소득 변수는 예상대로 소득이 높을수록 패널지속 기간이 길어지며 가구원 수는 많을수록 패널지속 기간이 짧아진다. 마지막으로 가구주 연령변수를 살펴보면 연령이 높을수록 패널지속 기간이 길어진다.

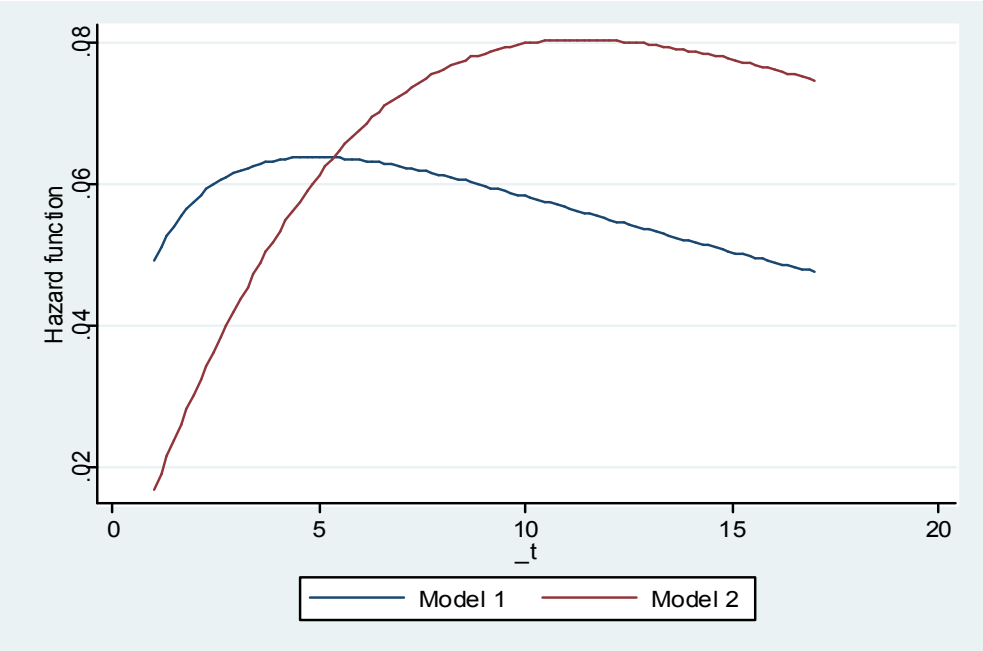
<표 6> 로그로지스틱 모형 추정결과

| 변수 | 모형 1 | 모형 2 |
|---|---|---|
| $gender2$ (여자) | | 0.066 (0.036)* |
| $edu2$ (고졸) | | 0.198 (0.035)*** |
| $edu3$ (전문대졸 이상) | | 0.157 (0.038)*** |
| $region2$ (비광역시) | | 0.135 (0.023)*** |
| $owner1$ (자가) | | 0.018 (0.025) |
| $married2$ (기혼유배우) | | -0.004 (0.049) |
| $married3$ (기혼무배우) | | -0.032 (0.052) |
| $employed2$ (자영업) | | -0.122 (0.029)*** |
| $employed3$ (무직/비경활) | | -0.141 (0.033)*** |
| $lincome$ (로그 소득) | | 0.568 (0.014)*** |
| $hsize$ (가구원 수) | | -0.151 (0.012)*** |
| $age$ (가구주 나이) | | 0.046 (0.001)*** |
| $\alpha$ (상수항) | 2.475 (0.016)*** | -4.005 (0.123)*** |
| $\gamma$ (shape 모수) | 0.762 (0.009)*** | 0.523 (0.007)*** |
| $\log L$ | -9968.1 | -7586.7 |

8) 이사가능성은 거주지역, 소득, 가구주 직업, 가구주 나이 변수들에 의해 통제될 가능성이 크기 때문에 점유형태 변수 자체는 통계적으로 유의하지 않은 것으로 추론할 수 있다.

그림 2에서는 모형 1과 모형 2에서 hazard rate의 예측값을 비교하고 있다. 표 6의 추정결과에서 확인할 수 있듯이 $\hat{\gamma} < 1$ 이므로 hazard rate이 증가하다가 감소하는 패턴으로 나타난다. 모형 2에서는 설명변수 값이 평균에 있다고 가정하고 hazard rate를 계산한다. 설명변수를 통제하지 않은 모형 1에서는 duration=5년일 때 hazard rate이 가장 높다. 그러나 설명변수를 통제한 모형 2에서는 duration=10년일 때 hazard rate이 가장 높아지고 그 이후에는 완만하게 hazard rate이 낮아진다. 패널지속 기간이 길어질수록 모형 2에서 패널탈락의 hazard rate이 모형 1의 경우에 비해 훨씬 높아진다는 것을 알 수 있다. 그림 2를 통해서 패널조사에서 패널탈락 가능성은 단순히 시간을 따라서 발생한다기보다는 해당 가구의 특성에 의해 hazard rate이 많이 달라진다고 해석할 수 있다.

<그림 2> hazard rate 그래프



<그림 3> hazard rate 그래프 : 광역시 vs. 비광역시

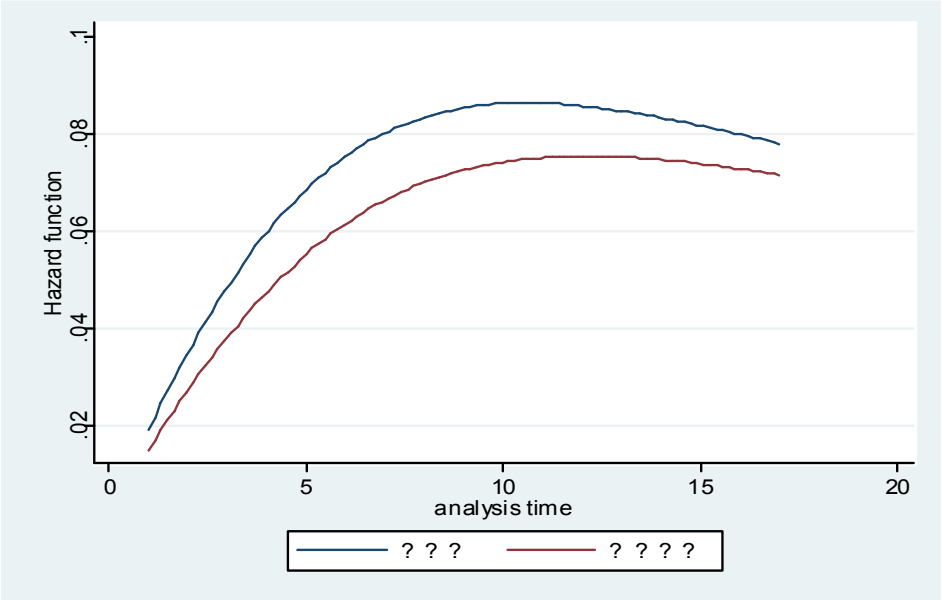


표 6의 추정결과를 이용하여 가구 $j$의 패널탈락 시점을 예측해 볼 수 있다. 본 연구에서는 패널탈락 예측시점을 median duration으로 대신한다.[9] median duration 예측값은 식 6을 이용하여 계산할 수 있다. 표 7에서는 설명변수 값이 일정하게 주어졌을 때 median duration을 정리하여 보여준다.[10] 남자 가구주에 비해 여자 가구주의 패널지속 기간이 0.8년 더 길고, 고졸미만 가구주에 비해 고졸/전문대졸 이상 가구주의 경우 패널지속 기간이 2년 내외 더 길 것으로 예상된다. 또한 광역시 거주가구에 비해 비광역시 거주가구가 패널지속 기간이 1.7년 더 길 것으로 예측된다. 가구주의 직업이 임금근로자인 경우 자영업/무직/비경활 가구주인 경우에 비해 패널지속 기간이 1.5년~1.7년 정도 길 것으로 기대된다. 가구주 연령이 45세이고 다른 설명변수가 평균에 있는 가구의 패널탈락 시점은 9.13년, 가구소득이 3,500만원이고 다른 설명변수가 평균에 있는 가구의 패널탈락 시점은 15년이라고 예측할 수 있다.

<표 7> median duration 예측값

| 설명변수 | 범주 | median duration (년) |
|---|---|---|
| $gender$ | 남자 | 11.7 |
| | 여자 | 12.5 |
| $edu$ | 고졸미만 | 10.5 |
| | 고졸 | 12.8 |
| | 전문대졸 이상 | 12.3 |
| $region$ | 광역시 | 11.0 |
| | 비광역시 | 12.7 |
| $owner$ | 전월세 | 11.7 |
| | 자가 | 11.9 |
| $married$ | 미혼 | 12.0 |
| | 기혼유배우 | 11.9 |
| | 기혼무배우 | 11.6 |
| $employed$ | 임금근로자 | 12.7 |
| | 자영업 | 11.2 |
| | 무직/비경활 | 11.0 |
| $lincome$ | 3500만원 | 15.0 |
| $age$ | 45세 | 9.13 |

9) 생존확률이 50%인 시점, 즉 패널탈락 가능성이 50%가 되는 시점을 패널탈락 예측시점으로 간주하고 분석한다.

10) 지정된 설명변수 이외의 다른 설명변수들의 값은 평균이라고 가정하고 계산한다. 따라서 다른 설명변수들의 값이 달라지면 median duration, 즉 패널탈락 예측시점 또한 달라진다.

# Ⅴ. 결론 및 시사점

본 연구에서는 여러 번에 걸쳐 표본조사가 이루어지는 패널조사의 고질적인 문제인 패널탈락의 문제에 대해 우리나라의 대표적인 장기 패널조사인 KLIPS(1~17차) 자료를 이용하여 살펴보았다. 로그로지스틱 생존분석 모형 추정 결과, 기존의 선행연구들과 마찬가지로 표본의 개별특성들이 패널지속성을 결정하는 중요한 요인인 것으로 확인되었다. 구체적으로, 가구주의 성별, 연령, 학력, 거주지역, 취업상태, 가구소득, 가구원 수 등이 KLIPS 설문응답 지속기간에 유의한 영향을 미치는 것으로 나타났다. 남자 가구주에 비해 여자 가구주인 경우, 가구주 연령이나 교육수준, 가구소득이 높을수록, 광역시에 비해 비광역시에 거주하는 가구의 경우, 가구주가 자영업이나 무직/비경활인 경우에 비해 임금근로자인 경우, 그리고 가구원 수가 적을수록 패널지속 기간이 길어지는 양상을 보였다. 한편, 점유형태 변수(자가 보유여부)와 혼인상태 변수는 통계적으로 유의하지 않았다. 패널지속성에 영향을 미칠 것으로 생각되는 이사가능성은 거주지역, 소득, 가구주 직업, 가구주 나이와 같은 변수들에 의해 통제될 가능성이 크기 때문에 점유형태 변수 자체는 통계적 유의성을 확보하지 못한 것으로 보인다.

더 나아가 본 연구에서는 상기한 로그로지스틱 생존분석 모형 추정결과를 이용하여 생존확률이 50%, 즉 패널탈락 가능성이 50%인 시점인 median duration을 추정하고 이를 통해 특정 가구의 패널탈락 시점을 예측해보았다. 지정된 변수 이외의 다른 설명변수들은 평균값을 가진다고 가정한 상태에서 볼 때, 남자 가구주(11.7년)에 비해 여자 가구주(12.5년)의 경우, 고졸미만 가구주(10.5년)에 비해 고졸/전문대졸 이상 가구주(12.8/12.3년)의 경우, 패널지속 기간이 각각 0.8년과 2년 내외 더 길 것으로 예상된다. 또한 광역시 거주가구(11년)에 비해 비광역시 거주가구(12.7년)의 경우, 가구주의 직업이 자영업/무직/비경활(11.0/11.2년)인 경우에 비해 임금근로자(12.7년)인 경우, 패널지속 기간이 1.5년~1.7년 정도 더 길 것으로 예측된다. 가구주 연령이 45세이고 다른 설명변수가 평균에 있는 가구의 패널탈락 시점은 9.13년, 가구소득이 3,500만원이고 다른 설명변수가 평균에 있는 가구의 패널탈락 시점은 15년일 것으로 기대된다. 이러한 예측 결과를 보면, 각 가구의 특성에 따라 패널탈락 예상 시점이 달라진다는 점을 알 수 있으며, 특히 가구주가 남성이거나 학력이 고졸미만, 비임금근로자 또는 비취업자인 경우, 광역시에 거주하거나 가구주 연령이 45세 미만인 가구의 경우 패널조사에 응답을 지속하는 기간이 10~11년을 넘지 못하여, 상대적으로 응답중단 가능성이 높은 위험군으로 분류가능할 것으로 보인다. 이와 같은 패널탈락 위험군에 대해서는 각 가구의 특성에 따른 패널탈락 예측 시점이 도래하기 전에, 선물 또는 금전적인 인센티브, 감사 편지・전화, 면접원 교체・업그레이드 등의 사전관리를 시도한다면 패널조사에서 이탈할 위험률을 낮추는데 도움이 될 것으로 생각된다. 또한 패널탈락 예측 시점을 이용한 이러한 사전관리 방법의 실효성에 대한 지속적인 모니터링 및 분석 노력도 KLIPS의 응답지속성을 제고하는데 도움이 될 것으로 보인다.

# 참 고 문 헌


김대일・남재량・류근관(2000), 「한국노동패널 표본의 대표성과 패널조사 표본이탈자의 특성연구」, 『노동경제논집』, 23(S), pp. 1-33.

이상호(2005), 「한국노동패널(KLIPS)의 표본이탈 분석-가구소득을 중심으로」, 『노동리뷰』, 통권 제11호, pp. 66-80.

이상협・박찬용・정성석・최혜미(2011), 「한국노동패널 탈락 분석」, 『한국데이터정보과학회지』, 22(1), pp. 1-8.

신선옥(2008), 「한국노동패널조사의 응답자 태도에 면접원이 미치는 효과」, 『노동리뷰』, 통권 제37호, pp. 74-82.

Cheng, Terence C. and Trivedi, Pravin K.(2015), “Attrition Bias in Panel Data: A Sheep in Wolf's Clothing? A Case Study Based on the Mabel Survey”, *Health Economics*, 24(9), pp. 1101 - 1117.

Lillard, Lee A. and W.A. Panis, Constatijn(1998), “Panel Attrition from the Panel Study of Income Dynamics: Household Income, Marital Status, and Mortality,” *The Journal of Human Resources*, 33(2), pp. 437-457.

Fitzgerald, John, Gottschalk, Peter and Moffitt, Robert(1998), “An Analysis of Sample Attrition in Panel Data: The Michigan Panal Study of Income Dynamics,” *The Journal of Human Resources*, 33(2), pp. 251-299.

Hausman, Jerry A. and Wise, David A.(1979), “Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment,” *Econometrica*, 47(2), pp. 455-473.

Zabel, Jeffrey E.(1998), “An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior,” *The Journal of Human Resources*, 33(2), pp. 479-506.