# 한국노동패널(KLIPS)의 표본이탈 분석
## -가구소득을 중심으로-

이 상 호*

## Ⅰ. 들어가며

패널자료에 대한 이용가능성의 증대는 지난 30년간 응용사회과학 연구에 있어서 가장 중요한 발전이었다고 일컬어진다. 패널자료는 다양한 이슈들을 광범위하게 다루면서도 동일한 표본을 반복조사한다는 점에서 횡단면 자료와 시계열 자료의 장점을 동시에 갖고 있기 때문이다. 즉, 시간의 경과나 환경의 변화에 따른 동태적 변화 및 상태간 변이과정을 추적할 수 있을 뿐만 아니라, '미관측 이질성'(Unobserved Heterogeneity)을 통제하므로 보다 엄밀한 조건에서 사회경제적 변수들의 효과를 측정할 수 있다. 이러한 장점 때문에 서구에서 이미 수십년 전부터 패널조사가 진행되어 오고 있다. 예를 들어 미국의 NLS(1966~)와 PSID(1968~), 영국의 BHPS(1991~), 독일의 GSOEP(1984~) 등은 세계적 수준의 패널로 손꼽히고 있다. 우리나라에서도 1990년대 이후 패널조사의 필요성이 제기됨에 따라 1994년에 국내 최초로 대우패널이 시작되었으나 1998년에 종료되었으며, 한국노동패널(1998~)이 그 뒤를 이어서 진행되고 있다. 특히 최근 2~3년 동안은 중앙고용정보원의 청년패널을 포함하여 여러 가지 형태의 패널조사들이 공공연구기관이나 대학에서 시도되고 있다. 그러나 패널조사는 횟수가 반복됨에 따라 응답에 대한 피로감이 누적되면서 응답기피 현상이 나타날 수 있다. 때문에 막대한 예산과 인원이 투입되는 패널조사의 성공을 위한 전제조건으로 가장 중요시되는 문제가 바로 '어떻게 표본이탈을 최소화 할 것인가'라는 과제이다. 조사횟수가 늘어남에 따라 증가하는 표본이탈은 조사의 장기적인 지속가능성을 어렵게 할 뿐만 아니라, 만일 그러한 이탈이 특정집단에 집중되어 있거나 체계적인 패턴을 가지고 있을 때에는 조사의 대표성을 훼손할 수도 있다.

* 한국노동연구원 노동패널팀 연구원(shlee@kli.re.kr).

그렇다면 '한국노동패널(Korean Labor and Income Panel Study : 이하 KLIPS)'에도 자료의 신뢰성과 대표성을 저해하는 '체계적인 표본이탈 편의'(Systemic Non-Random Attrition Bias)가 존재하는가? 만일 존재한다면 표본이탈의 결정요인은 무엇인가? 이러한 표본이탈이 주요 변수에는 어떤 영향을 미치는가? 통계적으로는 이를 어떻게 검증하고 바로잡을 것인가? 이러한 질문에 대한 답을 찾고자 함이 이 글의 목적이다. 특히 KLIPS에서 조사되는 핵심적인 변수들 중에서 그 동안 다수 연구들에서 문제 제기가 이루어졌던 가구소득 추정에서의 표본이탈 효과를 검증하는데 초점을 맞출 것이다.

이를 위해 우선 제Ⅱ장에서는 패널의 표본이탈과 관련된 이론적·실증적 논의를 검토하고, 제Ⅲ장에서는 본 논문에서 다루게 될 분석모형을 구성할 것이다. 제Ⅳ장에서는 표본이탈의 패턴 및 실증분석 결과를 제시한 후, 마지막으로 제Ⅴ장에서 결론을 맺는다.

## Ⅱ. 표본이탈에 대한 이론적 실증적 논의

### 1. 표본이탈 모형

일반적으로 패널자료를 이용하여 노동시장 이슈를 다루고자 할 때 문제가 될 수 있는 내생적 선택의 문제는 크게 세 가지 정도로 구분할 수 있다(Jenkins, 2002). 첫번째는 조사 설계에 따른 효과이다. 예컨대, PSID의 SEO 표본의 경우 빈곤선의 1.5배에 해당하는 계층을 따로 추출하여 조사에 반영하였다. 또한 샘플링 에러에 의해 최초의 조사시점에서 확률표본이 이루어지지 않을 수도 있다. 두번째는 항목무응답 응답구조에 의한 무응답이 발생할 수 있다. 세번째는 표본이탈이다.

이 중에서 첫번째 문제는 적절한 모집단의 특성을 추출할 수 있을 경우 가중치를 통해서 해결이 가능하다. 두번째 문제는 횡단면 조사에서도 얼마든지 나타날 수 있는 문제이며, 항목무응답의 경우 '보정'(imputation)기법과 같은 해법들을 사용할 수 있다. 또한 경제적 지위에 의한 체계적 무응답의 경우에도 '해킷'(Heckit)을 활용한 해결이 가능하다(Heckman, 1979).

세번째 문제는 패널자료에서만 나타나는 특수한 상황이다. 한번의 표본추출로 1회의 조사에 그치는 횡단면 조사와는 달리 병환이나 경제적 어려움, 가족 내에서의 문제 등으로 인해 일시적 혹은 영구적으로 조사에서 빠져나갈 수 있기 때문이다. 만일 이러한 이탈이 체계적인 패턴을 가지고 일부집단, 특히 사회경제적 지위(소득, 취업형태, 종사상 지위 등)가 높거나 낮은 집단에 집중되어 있다면 경제모형을 분석함에 있어서 더욱 문제

가 될 수 있다. 이러한 비무작위적 이탈이 외생변수에만 영향을 미친다면 설명변수를 통제함으로써 해결가능하지만, 내생변수까지 영향을 받는다면 '일치 추정량'(Consistent Estimates)을 얻을 수 없을 것이기 때문이다.

패널자료의 표본이탈 모형에 대한 이론화는 Hausman and Wise(1979)가 최초로 시도하였다. 이들은 임의효과 모형(Random Effect Model) 상황에서의 표본이탈 모형을 제안하였다. Ridder(1992)는 두 개 차수에 제한된 'Hausman-Wise 모형'을 보다 일반화하였으며, 이후 여러 학자들에 의해 모형의 개선과 확장이 이루어졌다(Beketti et al, 1988; Wooldridge, 1995; Zabel, 1998; Ryu, 2001).

여기서는 Wooldridge(1995, 2002)가 제안한 '고정효과' 상황에서의 표본이탈 모형을 살펴보자. 우선 모든 조사차수에 대해 모든 관측치가 존재하는 패널자료(즉, Balanced Panel)가 있다고 가정하자. 임의의 종속변수에 관해 식 (1.1)과 같은 선형회귀 모형을 고려한다면, 전체 t기의 조사차수에 대해 i명의 개인으로 구성되는 i × t개의 관측치를 갖게 된다. 만일 미관측된 개인효과 $c_i$ ,를 고정된 것으로 가정했을 때, $x_{it}$ 가 full rank를 갖는 외생변수 벡터이고, $E(u_i u_i^{'} \mid c_i, x_i) = \sigma^2 I_T$ 를 만족한다면 식 (1.3)에 나타난 바와 같이 집단내 변환(Within Transformation)을 거친 $\hat{\beta}_{FE}$ 가 일치 추정량이 된다.

$$
\begin{aligned}
y_{it} &= \mathbf{x_{it}}\beta + c_i + u_{it}, \quad t = 1, \Lambda\ , T, \qquad u_{it} \sim N(0, \sigma^2) \qquad (1.1)\\
\bar{y}_i &= \bar{\mathbf{x}}_{\mathbf{i}}\beta + c_i + \bar{u}_i \\
y_{it} - \bar{y}_i &= (\mathbf{x_{it}} - \bar{\mathbf{x}}_{\mathbf{i}})\beta + u_{it} - \bar{u}_i
\end{aligned}
$$

$$
\ddot{y}_{it} = \ddot{\mathbf{x}}_{\mathbf{it}}\beta + \ddot{u}_{it}, \quad t = 1, \Lambda\ , T \qquad (1.2)
$$

$$
\hat{\beta}_{FE} = \left( \sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\mathbf{x}}_{\mathbf{it}}^{'} \ddot{\mathbf{x}}_{\mathbf{it}} \right)^{-1} \left( \sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\mathbf{x}}_{\mathbf{it}}^{'} \ddot{y}_{it} \right) \qquad (1.3)
$$

$$
\begin{aligned}
\hat{\beta} &= \left( N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{\mathbf{it}}^{'} \ddot{\mathbf{x}}_{\mathbf{it}} \right)^{-1} \left( N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{\mathbf{it}}^{'} \ddot{y}_{it} \right) \\
&= \beta + \left( N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{\mathbf{it}}^{'} \ddot{\mathbf{x}}_{\mathbf{it}} \right)^{-1} \left( N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{\mathbf{it}}^{'} u_{it} \right) \qquad (1.4)
\end{aligned}
$$

$$
where \quad \ddot{\mathbf{x}}_{\mathbf{it}} \equiv \mathbf{x_{it}} - T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{x_{ir}}, \quad \ddot{y}_{\mathbf{it}} \equiv y_{\mathbf{it}} - T_i^{-1}\sum_{r=1}^{T} s_{ir} y_{\mathbf{ir}}, \qquad T_i \equiv \sum_{t=1}^{T} s_{it}
$$

$$s_{it} = w_{it}\gamma + v_{it}, \quad t = 2,\Lambda\ ,T \tag{1.5}$$

$$E(u_{it} \mid c_{it}, x_i, s_i) = 0, \qquad t = 1,2,\Lambda\ ,T$$
$$E(u_{it} \mid c_{it}, x_i, s_i) = E(u_{it} \mid v_{it}) = \rho\sigma \frac{\boldsymbol{\varphi(\mathbf{w}_{it}\gamma)}}{\boldsymbol{\Phi[\mathbf{w}_{it}\gamma]}} = \rho\hat{\lambda}_{it}, \qquad t = 1,2,\Lambda\ ,T \tag{1.6}$$

$$E(\text{■}_{it} \mid \text{■}_{it}, s_{it} = 1) = \text{■}_{it}\boldsymbol{\beta} + \rho\text{■}_{it} + error \tag{1.7}$$

이제 표본이탈로 인한 '비균형패널'(Unbalanced Panel)을 고려해 보자. 개별 관측치는 $s_{it} = 1$일 때만 관찰되며 이 때의 추정량 $\hat{\beta}$은 식 (1.4)와 같은 형태를 가지게 된다.

응답/비응답을 나타내는 확률은 각각 식 (1.5)와 같은 프로빗 모형으로 구성할 수 있다. 이제 이 식을 추정하여 얻어진 확률밀도함수 $\boldsymbol{\varphi(\mathbf{w}_{it}\gamma)}$를 분포함수인 $\boldsymbol{\Phi[\mathbf{w}_{it}\gamma]}$로 나누어서 얻어진 Inverse Mills Ratio를 식 (1.1)에 대입하여 다시 within transformation을 취하면 마침내 최종적인 추정식 (1.7)이 얻어진다. 이 식을 추정하여 'Ho: ρ=0', 다시 말해 '체계적인 표본이탈이 존재하지 않는다'에 대한 귀무가설을 테스트하면 된다.

## 2. 실증연구

패널자료에서 표본이탈에 대한 연구는 1998년 「Journal of Human Resources」에서 특집으로 다루었을 정도로 중요한 이슈가 되었다. 원래 PSID와 같은 패널자료가 소득불평등이나 빈곤을 완화시키고자 하는 정책적 목표와 긴밀하게 연관되어 있었기 때문에 표본이탈로 인한 핵심변수의 영향 여부와 정도를 측정하고 보정하는 것은 자료 자체의 신뢰성과 직결되는 문제였다. 이들 연구에서 다루고 있는 표본이탈 요인은 크게 세 가지 정도로 나눌 수 있다. 첫번째, 대부분의 연구들은 주로 성별, 인종, 혼인상태, 교육년수 같은 인구통계학적 특성 및 가구소득이나 개인의 노동시장에서의 지위가 이탈에 어떤 영향을 미치는지를 다루고 있다. 두번째, 어떤 특정한 상태 그 자체보다는 이혼이나 실업 등과 같은 사회경제적 충격이 응답 여부에 어떤 영향을 미치는가에 대한 관심도 증가하고 있다. 세번째, 최근에는 응답자의 특성뿐만 아니라 면접원의 특성이나 '조사시스템'(survey mechanism)이 어떤 영향을 미치는지에 대한 연구들이 나오고 있다.

최초의 연구로는 '부의 소득세' 효과를 실험하는 GIME(Gary Income Maintence Experiment) 자료를 이용하여 표본이탈 모형을 실증하였던 Hausman and Wise(1979)를 들 수 있다. 이들이 설정한 이탈모형에는 실험/통제집단 여부, 교육년수, 근속년수, 가구소

득, 노조유무, 건강상태와 같은 설명변수들이 사용되었는데, 특히 소득이 높은 응답자일수록 실험집단 내에서 이탈확률이 높아지는 것으로 관찰되었다.

Becketti et al.(1988)은 PSID 14년간(1968～1981년)의 자료에 대해 비례위험률 모형(Proportional hazard model)을 이용하여 이탈분석을 실시하였다. 이들은 빈곤층만을 별도로 추출한 SEO 소득을 통제한 후에도 SRC 표본보다 이탈률이 더 높다는 사실을 발견했다.

Fitzgerald et al(1998)은 PSID(Panel Study of Incom Dynamics) 19년간(1968～1986년)의 자료를 이용한 분석에서 미혼자, 고령자, 비백인 등의 인구학적 특성을 가진 경우와 자가소유가 아니며 근로소득의 변동폭이 큰 경우 표본이탈이 높다는 분석결과를 제시했다. 그러나 앞의 연구결과와는 달리 고소득가구인 경우 표본이탈이 높은 것으로 나타났다.

Lillard and Panis(1998)의 경우에도 PSID 1968~1988년까지의 자료를 패널화하여 분석한 결과 백인남성인 경우와 기혼부부, 특히 결혼한지 오래된 부부일수록 응답률이 높다는 것을 발견하였다. 또한 이혼이나 별거, 배우자의 사망과 같은 혼인상태의 변화가 이탈할 위험을 높인다고 주장하였다.

주로 인구학적 특성 혹은 사회경제적 배경에 초점을 맞춘 이상의 연구들과는 달리 Zabel(1998)은 PSID와 SIPP(Survey of Income and Program Participation)에 대한 비교연구에서 인구통계학적 변수들의 영향은 그리 높지 않으며 오히려 면접원과 면접과정이 표본이탈에 많은 영향을 미친다고 주장하였다. 특히 인터뷰 시간이 짧을수록, 동일한 면접원이 계속 조사할수록 표본이탈이 감소한다는 것을 발견하였다.

Hill and Willis(2001)의 경우에도 미국의 HRC(Health and Retirement Study) 1~3차년도 자료를 이용하여 인구통계학적 특성보다는 동일한 면접원을 확보하는 것이 응답률을 높이기 위한 가장 중요한 요인임을 밝히고 있다. 더불어 응답자의 조사에 대한 몰입정도도 중요한 요인임을 지적하였다.

국내의 연구로는 경제활동인구조사를 패널화한 자료(1993~1997년)와 대우패널(1993~1998년)의 표본이탈률을 분석한 김대일 외(2000)의 연구가 있다. 이들은 두 가지 자료의 공통적인 분석결과로 남성이 여성보다, 고연령층이 저연령층보다, 이혼하였거나 미혼인 경우가 배우자가 있는 경우보다 이탈률이 높다는 분석결과를 제시하였다. 또한 교육수준별로는 고학력자가, 고용형태별로는 임금근로자가, 경제활동상태별로는 실업자・미취업자의 이탈률이 높은 반면, 가구주인 경우에는 이탈률이 낮은 것으로 보고하였다.

최근에는 KLIPS의 가중치가 표본의 대표성을 적절하게 보완해 주는지를 검토하고 있다. 김영원 외(2005), 김규성 외(2005)는 KLIPS 1~6차년도 자료를 이용하여 표본이탈의 일반적 패턴을 분석하고, 기존의 가중치에 대한 '보정'(Calibration)을 제안하고 있다.

# Ⅲ. KLIPS의 조사체계와 표본이탈의 일반적 특성

한국노동패널의 모집단은 제주도를 제외한 전국의 도시지역 가구이다. 표본틀은 1995년 인구센서스조사를 모집단으로 표집된 「고용구조특별조사」(1997년)를 이용하였다. 표본추출은 1단계에서 조사구를 선정하고 2단계에서 가구를 선정하는 '2단층화집락계통추출법'(two stage stratified cluster sampling)으로 이루어졌다. 이러한 추출법에 따라 접촉한 가구의 조사성공률은 75.5%였으며, 나머지 25.4%는 대체표본을 추출하여 5000가구를 확정지었다.

조사는 매년 4~9월 사이에 실시되었으며, '면접타계식'(face-to-face)을 원칙으로 하되 개인조사에 한해서 접촉이 어려운 경우 '유치조사'나 '전화조사'를 허용하였다. 면접조사 비중은 1차년도에 64.4% 수준이었으나 6차년도에는 86.3%까지 증가하였으며(표 1), 6차년도까지 순수한 대리응답 비중은 11.2% 정도에 그치는 수준이었다(표 2).

KLIPS는 조사성공률을 높이기 위해서 다양한 응답자 관리방법을 사용하고 있다. 우선 실사 전에 노동부 공문 및 조사의 필요성을 기술한 편지를 발송하여 조사목적에 대

**〈표 1〉 주요 패널의 원표본 조사성공률**

| | PSID | GSOEP | BHPS | KHPS | KLIPS |
|---|---|---|---|---|---|
| 2차조사 | 89%(1969) | 90%(1985) | 88%(1991) | 79%(1994) | 88%(1999) |
| 3차조사 | 86%(1970) | 86%(1986) | 83%(1992) | 66%(1995) | 81%(2000) |
| 4차조사 | 84%(1971) | 85%(1987) | 79%(1993) | 59%(1996) | 77%(2001) |
| 5차조사 | 81%(1972) | 81%(1988) | 75%(1993) | 56%(1997) | 76%(2002) |
| 6차조사 | 79%(1973) | 79%(1989) | 74%(1994) | 44%(1998) | 77%(2003) |
| 7차조사 | 76%(1974) | 78%(1988) | 71%(1995) | End | 78%(2004) |

**〈표 2〉 6차년도 면접원의 특성**

| 지 역 | 사례수 | 교육수준 | 연 령 | 노동패널 경력 |
|---|---|---|---|---|
| 수 도 권 | 19 | 1.6 | 42.3 | 3.2 |
| 대구경북 | 15 | 1.3 | 36.9 | 3.9 |
| 대전충청 | 8 | 1.3 | 44.6 | 3.4 |
| 광주전라 | 10 | 1.4 | 42.9 | 3.8 |
| 부산경남 | 15 | 1.3 | 41.5 | 3.8 |
| 강원제주 | 5 | 1.4 | 39.4 | 1.6 |
| 전 체 | 72 | 1.4 | 41.2 | 3.5 |

주 : 교육수준에서 고졸은 '1', 대졸은 '2'임.

해 응답자들이 충분히 이해할 수 있도록 한다. 또한 가구 전체가 이사하거나 일부 가구원이 분가할 경우 변동사실을 알려줄 경우에 대한 보상도 실시하고 있다. 조사가 완료되면 소정의 선물을 지급한다. 그 외에도 가구원의 생일, 결혼기념일, 회갑 등이 있을 시 축하편지를 발송하고 수집된 자료가 학술 및 정책적으로 유용하게 사용되고 있음을 알려주는 소식지도 발송한다. 이러한 노력에도 불구하고 조사초기였던 2차년도와 3차년도의 '조사성공률'(retention rate)은 각각 88%와 81%로 표본이탈이 예상보다 높게 나타났다(표 1). 이 때문에 4차년도 조사에서부터는 각 가구에 대해 3만원의 응답사례금을 지급하여 경제적 보상을 통한 응답률 유지에도 노력하였다.

응답자 관리뿐만 아니라 면접원 관리시스템도 체계적으로 운영하고 있다. 패널조사는 가구의 소득 및 경제활동과 관련된 매우 민감한 정보들을 필요로 하기 때문에 경력이 많은 면접원을 가급적 동일한 가구에 접촉시킴으로써 응답자와의 신뢰관계를 구축하는 것이 중요하다. <표 2>에서 볼 수 있듯이, 6차년도 조사에 투입된 73명의 면접원 중 1인당 평균 조사가구수는 65.9가구(최대 406가구)였다. 면접원들의 평균 조사년수는 3.5년으로 이 중에서 강원과 제주를 제외한 수도권의 노동패널 조사경력이 3.2년으로 상대적으로 다른 지역에 비해 낮은 것으로 나타났다.

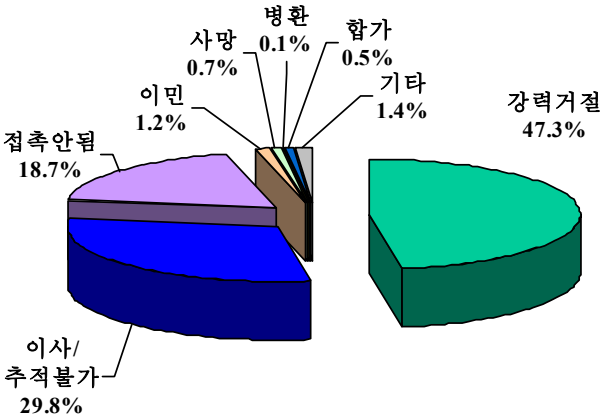
<표 3>은 전체 6개년도에 대한 가구의 연도별 무응답 패턴이 제시되어 있다. 이 표의 마지막 칼럼에서 "1"은 '응답'을, "0"은 '무응답'을 의미한다. 즉, "111110"은 5차년도까지는 응답했다가 6차년도에 이탈했음을 의미하는 것이다. 분석결과, 1차년도 원표본인 5,000가구 중에서 6차년도까지 계속응답한 가구가 61.7%를 차지하며, 이탈 후 현재까지 복귀하지 않은 가구가 5.5%, 나머지 32.8% 가구는 이탈한 후에 다시 복귀하거나 재이탈하는 가구로 나타났다. 여기서 특징적인 것은 표본이탈의 패턴이 한번 이탈하면 그 상태가 지속적인 것이 아니라, '이탈'과 '복귀'를 반복하는 매우 복잡한 양상으로 전개되고 있다는 점이다. 즉, 본격적인 이탈패턴을 분석하기 위해서는 이러한 양상이 고려되어야 함을 시사한다고 볼 수 있다.

마지막으로 이탈 사유를 살펴보면, 6차년도 조사의 경우 응답자의 강력거절이 47.3%를 차지하였고, 그 다음이 이사 등으로 인해 추적할 수 없는 경우가 29.8%를 차지하였다. 이에 반해 사망, 이민 등과 같이 표본의 자연감소가 차지하는 비중은 매우 낮은 것으로 파악되었다(그림 3).

〈표 3〉 표본이탈의 일반적인 패턴

| 빈 도 | 비 중 | 누적비중 | 응답패턴 |
|---|---|---|---|
| 3,087 | 61.7 | 61.7 | 111111 |
| 188 | 3.8 | 75.0 | 111110 |
| 142 | 2.8 | 81.4 | 111100 |
| 182 | 3.6 | 78.6 | 111000 |
| 200 | 4.0 | 71.2 | 110000 |
| 273 | 5.5 | 67.2 | 100000 |
| 127 | 2.5 | 84.0 | 110111 |
| 100 | 2.0 | 86.0 | 111011 |
| 88 | 1.8 | 87.7 | 111101 |
| 82 | 1.6 | 89.4 | 111001 |
| 531 | 10.6 | 100.0 | Other Pattern |
| 5,000 | 100.0 | | 1=응답, 0=무응답 |

〔그림 1〕 6차년도 비성공 사유




# Ⅳ. 분석결과

여기서는 과연 KLIPS에도 '비무작위적인 이탈'이 나타나는지, 나타난다면 어떤 요인들이 영향을 미치는지, 이러한 이탈요인을 통제하여 목적변수에 대해 보다 일관된 추정치를 얻을 수 있는지를 실증할 것이다. 특히 소득불평등 및 빈곤관련 연구들에서 KLIPS에 저소득계층에 대한 과대추정(over estimation)의 문제가 제기되고 있다. 이들은 패널에

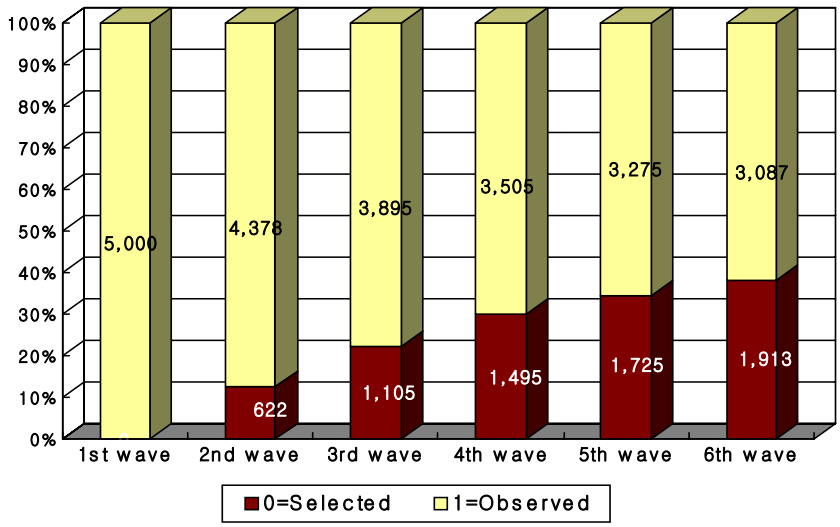
대한 빈곤연구의 경우 표본이탈로 인한 편의(bias)의 발생가능성 때문에 신중을 기할 것을 지적한다(유경준 외, 2002: 126). 따라서 본 연구에서는 가구소득을 목적변수로 설정하였을 때의 표본이탈 효과를 검증함으로써 이러한 주장의 타당성 여부도 함께 살펴볼 것이다.

분석방법은 제Ⅱ장 제1항에서 소개한 Wooldridge(1995)의 모형과 절차에 따랐다. 우선 1단계로 응답여부에 대한 프로빗 모형을 추정한 후, 2단계에서는 프로빗 모형에서 얻어진 IMR(Inverse Mills Ratio)를 회귀모형에 설명변수로 추가하여 고정효과모형을 추정한다. 마지막으로 $\hat{\lambda}_{it}$의 계수 $\rho$가 통계적으로 유의미한지 여부를 검증하여 이탈의 체계성 여부를 검증한다. 분석을 위하여 본 모형에서 다음과 같은 추가적인 전제조건을 가정하였다.

① 1기에서는 표본이 모집단으로부터 확률표집으로 추출되었다.
② 표본이탈은 2기 이후부터 발생하며, 일단 한번 이탈하면 다시 복귀할 수 없다.
③ 1기에서는 모든 (xit, yit)변수들이 관측되었다. 즉, '항목 무응답'이 존재하지 않는다.

이러한 가정에 따라 이탈모형에서 종속변수로 사용될 이탈변수는 [그림 2]와 같은 패턴을 갖게 된다. 식 (1.5)에서 제시된 표본이탈 모형에 따라 설명변수들은 '시간에 일정한 변수(time-constant variables)'와 '시간에 따라 바뀌는 변수(time-varing variables)'로 구성되었다. 전자로는 가구주의 성별과 교육수준 더미가 사용되었으며, 후자에는 (t-1)기의

〔그림 2〕 이탈변수의 연도별 패턴



〈표 4〉 1차년도 핵심 변수의 요약통계량

| | 2차년도 이후 이탈 | | 모든 연도 관측 | |
|---|---|---|---|---|
| | 평 균 | 표준편차 | 평 균 | 표준편차 |
| Inc (연간 가구총소득) | 1879.84 | 1862.62 | 1759.10 | 1887.19 |
| Sex (남자=1, 여자=2) | 1.15 | 0.36 | 1.14 | 0.34 |
| Age (만나이) | 44.15 | 12.56 | 47.75 | 13.08 |
| Edu1 (중졸 이하 =1) | 0.28 | 0.45 | 0.42 | 0.49 |
| Edu2 (고졸=1) | 0.39 | 0.49 | 0.37 | 0.48 |
| Edu3 (대학 이상=1) | 0.32 | 0.47 | 0.21 | 0.41 |
| Econ1 (임금근로자=1) | 0.48 | 0.50 | 0.43 | 0.49 |
| Econ2 (자영업자=1) | 0.28 | 0.45 | 0.30 | 0.46 |
| Econ4 (미취업자=1) | 0.24 | 0.43 | 0.28 | 0.45 |
| Mar1 (미혼=1) | 0.10 | 0.30 | 0.05 | 0.21 |
| Mar2 (기혼 유배우자=1) | 0.85 | 0.36 | 0.91 | 0.29 |
| Mar3 (기혼 무배우자=1) | 0.05 | 0.22 | 0.04 | 0.20 |
| Myhome (자가=1, 전세 등=0) | 0.49 | 0.50 | 0.60 | 0.49 |
| Con (월평균 소비지출액) | 105.96 | 74.73 | 97.12 | 64.60 |
| Fnumb (가구원수) | 3.43 | 1.36 | 3.55 | 1.36 |
| 관측치수 | 1913 | | 3087 | |

〈표 5〉 1단계 이탈모형에 대한 분석결과

| | 계 수 | 표준오차 | z |
|---|---|---|---|
| 성별-t-1기 | -0.063 | 0.043 | -1.5 |
| 연령-t-1기 | 0.009 | 0.001 | 7.2 |
| 기혼 유배우더미-t-1기 | 0.250 | 0.058 | 4.3 |
| 기혼 무배우더미-t-1기 | 0.156 | 0.070 | 2.3 |
| 미취업자더미-t-1기 | -0.157 | 0.034 | -4.7 |
| 소비-t-1기 | -0.097 | 0.022 | -4.5 |
| 자가여부-t-1기 | 0.117 | 0.033 | 3.5 |
| 지역더미1(서울)-t-1기 | -0.243 | 0.032 | -7.6 |
| 지역더미3(대구)-t-1기 | -0.368 | 0.052 | -7.2 |
| 지역더미5(인천)-t-1기 | -0.202 | 0.054 | -3.7 |
| 지역더미8(경기)-t-1기 | -0.191 | 0.036 | -5.3 |
| 지역더미9(강원)-t-1기 | -0.416 | 0.076 | -5.5 |
| 연도더미(3차년도) | 0.124 | 0.038 | 3.2 |
| 연도더미(4차년도) | 0.165 | 0.041 | 4.0 |
| 연도더미(5차년도) | 0.305 | 0.042 | 7.3 |
| 연도더미(6차년도) | 0.405 | 0.065 | 6.3 |
| 상수 | 1.144 | 0.130 | 8.8 |
| 관측치수 | 19,488 | Pseudo R2 | 0.036 |

연령, 혼인상태, 로그 월평균 소비, 주거형태, 현재 거주지역 등이 사용되었다. 마지막으로 Becketti et al.(1988)에서 언급된 '기간의존성'(duration dependency)을 측정하기 위해 연도더미를 추가하였다.

변수들의 전반적인 요약통계량은 <표 4>에 제시된 바와 같다. 첫번째 칼럼에 나열된 변수들은 모두 1차년도 당시의 특성을 나타내는 변수들이며, 두번째와 세번째 칼럼에 제시된 값들은 계속응답과 중도이탈 여부를 기준으로 구분된 값들이다. 여기서 계속응답자는 연령이 많을수록 비중이 높고, 실업자일수록 더 많이 응답하며 미혼일수록 더 많이 이탈하는 것으로 나타나고 있다. 또한 소득 및 소비수준이 높을수록 이탈자의 비중이 높은 것으로 나타나고 있다.

다음으로 변수들의 인과관계를 본격적으로 파악하기 위해 1단계 프로빗 모형을 추정하였는데, 그 결과는 앞의 요약통계량과는 상이한 패턴으로 나타나고 있다.

첫째, 이탈모형의 'Pseudo R2'가 0.036으로 매우 낮다. 이러한 특징은 PSID를 이용한 Fitzgerald et al(1998)을 비롯한 여러 분석(프로빗 모형의 Pseudo R2가 0.028~0.071 수준으로 나타남) 결과에서도 나타난 바 있다. 예측컨대 이는 일반적인 표본이탈 분석에서 사용하는 모형에 면접원 효과, 이탈직전의 쇼크(이혼, 실업, 이사, 질병 등)과 같이 관측이나 수치화가 어려운 다른 중요한 요인들이 설명변수에서 누락되었기 때문인 것으로 보인다.

둘째, 가구주의 특성은 주로 연령이 높을수록 높은 응답률을 보였다. 혼인상태의 경우 '미혼자<기혼무배우<기혼유배우'의 순으로 응답률이 높은 것으로 나타났다. 그러나 성별에 따른 응답률 차이는 통계적으로 유의미하지 않았다.

셋째, 가구주 혹은 가구의 사회경제적 상태는 응답여부에 상반되는 영향을 미치는 것으로 나타났다. 우선 가구주가 실업자인 경우 취업자인 경우보다 낮은 응답률을 보인데 반해 가구의 소비수준과는 양의 상관관계, 즉 소비수준이 높을수록 낮은 응답률을 보였다. 가구의 거주형태가 '자가'인 경우에도 응답률이 높은 것으로 나타났다. 이러한 결과를 보건데, 응답률은 양극화 현상을 보이고 있음을 알 수 있다. 즉, 고소득계층일수록 응답을 기피하는 한편, 가구주가 실업상태에 놓인 하위계층 역시 응답을 기피하는 현상이 상대적으로 두드러진다는 점이다.

마지막으로 1단계 프로빗 모형에서 얻어진 IMR을 2단계 회귀모형에 대입한 후 '고정효과'를 가정한 가구소득 모형을 추정하였다. <표 6>에는 IMR로 '수정'(correction)한 모형과 그렇지 않은 모형 각각의 추정결과가 제시되어 있다. 우선, 시간에 고정된 변수들은 '집단내 변환'을 거치면서'개인효과'와 함께 '소거'(drop out)되었음을 알 수 있다. IMR의 계수는 '-1.648'로 t-test 결과 통계적으로 유의한 것으로 나타났다. 또한 응답확률이 낮을수록 가구소득이 높다는 요약통계량의 결과와도 일치하는 것으로 나타났다.

따라서 한국노동패널에서 표본이탈은 가구소득을 추정하는데 있어서 체계적인 영향을 미치는 요인이 됨을 알 수 있다. 그럼에도 불구하고 이러한 이탈요인이 다른 설명변수들의 계수에 심각한 영향을 미칠 정도는 아닌 것으로 보인다. 이탈요인을 고려하더라도 다른 설명변수들의 추정치 부호가 바뀌는 등의 큰 변동을 보이고 있지는 않기 때문이다.

**〈표 6〉 2단계 무작위 이탈에 대한 검증결과**

| | 수정모형 (Corrected Model) | | | 고정효과모형 (Fixed Effect) | | |
|---|---|---|---|---|---|---|
| | 계 수 | 표준오차 | t 값 | 계 수 | 표준오차 | t 값 |
| Lncon (로그 월평균 소비) | 0.53 | 0.015 | 36.23 | 0.537 | 0.015 | 36.79 |
| Sex (성별) | - | - | - | - | - | - |
| Age (연령) | 0.003 | 0.005 | 0.69 | 0.023 | 0.004 | 5.32 |
| Age2 (연령제곱) | 0 | 0 | -1.18 | 0 | 0 | -2.87 |
| Edu2 (고졸) | - | - | - | - | - | - |
| Edu3 (대학 이상) | - | - | - | - | - | - |
| Mar2 (기혼 유배우자) | -0.074 | 0.05 | -1.5 | -0.055 | 0.049 | -1.11 |
| Mar3 (기혼 무배우자) | -0.06 | 0.049 | -1.21 | -0.002 | 0.049 | -0.05 |
| Fnumsq (가구원수 제곱근) | 0.279 | 0.041 | 6.88 | 0.279 | 0.04 | 6.88 |
| Myhome (자가 더미) | 0 | 0.014 | -0.02 | -0.012 | 0.014 | -0.84 |
| Econ2 (자영업자 더미) | 0.038 | 0.025 | 1.52 | 0.044 | 0.025 | 1.75 |
| Econ4 (미취업자 더미) | -0.054 | 0.022 | -2.47 | -0.057 | 0.022 | -2.59 |
| Mills(Inverse Mills Ratio) | -1.648 | 0.218 | -7.57 | | | |
| 관측치수(i*T) | | 17,583 | | | 17,639 | |
| 관측치수(i) | | 4,351 | | | 4,357 | |

# V. 결론

패널자료에서 이탈패턴을 분석해야 하는 이유는 사후적으로 자료의 대표성을 확보하는 문제 이상을 의미한다. 관측가능한 주요 변수들을 기초로 모집단을 복원하려는 시도, 즉 가중치를 통한 해법은 한계가 있으며 일차적으로는 조사과정 자체에서 양적·질적으로 최소화하려는 노력이 원래의 표본특성을 가장 잘 유지할 수 있기 때문이다. 그런 의미에서 이 글에서 사용한 분석방법 역시도 많은 한계와 과제를 갖는다고 볼 수 있다.

우선 한국노동패널에는 체계적인 비무작위적인 이탈이 나타났지만, 핵심적인 경제변

수를 분석하는데 있어서 심각한 영향을 미칠 정도는 아니었다. 이탈의 특성별로는 고소득자일수록 이탈가능성이 더 높지만, 가구주가 실업자일 때에도 역시 이탈가능성이 높은 것으로 나타나 이탈에 있어서 양극화 현상이 나타났다. 그러나 전체적인 이탈모형의 설명력은 그리 높지 않은데, 여기에는 두 가지 가능성이 있다.

첫째, 분석모형 내에 면접원 효과나 조사시스템의 구조를 나타낼 수 있는 변수들이 포함되지 못했기 때문일 수 있다. 이 경우 조사주체는 수량화가 가능한 정보들을 발굴하고 이를 체계적으로 누적할 필요가 있다. 예컨대, 면접원의 특성, 응답자의 응답몰입도, 성실도 등은 현재에도 서구 패널에서 많이 사용되고 있는 지표들이다. 둘째, 가정의 현실성에 따른 결과일 수 있다. 예컨대, Ryu(2001)는 프로빗 모형을 이용한 이탈확률 계산은 전 시점의 응답확률을 반영하고 있지 못하기 때문에 'group duration model'을 이용한 분석방법을 제안한 바 있다.

이러한 한계에도 불구하고 일반 연구자들이 패널자료를 사용하는 데에는 그리 큰 문제가 없을 것으로 판단된다. 현재 KLIPS에서 제공되는 가중치는 이러한 표본이탈을 감안하여 개발되었기 때문에 이 글에서 나타난 정도의 편의는 어느 정도 수정되었을 것으로 보인다. 다만, 이 경우에도 가중치에 대한 보정작업의 필요성은 여전히 과제로 남아 있다[1]. 따라서 「경제활동인구조사」나 「도시가계조사」 등과 같이 국가에서 공식적으로 사용하고 있는 통계와의 비교연구 등이 더욱 활발하게 진행되어야 할 것이다. 또한 분석대상이 가구가 아닌 개인일 경우, 혹은 이 글에서 간과한 다른 핵심변수들에는 어떤 결과가 나타날 것인지에 대해서도 추가적인 연구가 필요할 것이다.

## <참고문헌>

강석훈, 「KLIPS의 가중치 부여방안 연구」, 『한국노동패널연구』, 한국노동연구원, 2003.

김규성 외, 「패널조사에서 가중치 부여 방법 및 효과에 관한 연구」, 『제6회 한국노동패널 학술대회 발표논문집』, 한국노동연구원, 2005.

김대일·남재량·류근관, 「한국노동패널 표본의 대표성과 패널조사 표본이탈자의 특성연구」, 『노동경제논집』, Vol.23, 한국노동경제학회, 2000.

1) 김영원 외(2005), 김규성(2005) 등은 한국노동패널의 가중치가 일부 변수에 대해 모평균을 과대 혹은 과소 추정할 수 있는 가능성 및 가중치와 관심변수간의 상관관계가 크면 분산이 증가할 수 있기 때문에 '보정'(Calibration)기법을 이용한 가중치의 수정이 필요함을 제안한 바 있다. 그러나 이들의 경우에도 가구의 소득에 대해서는 기존의 가중치가 모평균을 불편추정하고 있음을 확인하였다.

김영원 외, 「한국노동패널 표본의 대표성과 가중치 보정방법」, 『제6회 한국노동패널 학술대회 발표논문집』, 한국노동연구원, 2005.

남재량 외, 『제6차(2003)년도 한국 가구와 개인의 경제활동』, 한국노동연구원, 2005.

유경준·김대일, 『외환위기 이후 소득분배구조 변화와 재분배정책 효과 분석』, 한국개발연구원, 2002.

Becketti, Sean, William Gould, Lee Lillard and Finis Welch, "The Panel Study of Income Dynamics after Fourteen Years:An Evaluation", *Journal of Labor Economics*, 6(4), 1988, pp.472~492.

Hausman, J. and D. Wise, "Attrition Bias in Experimental and Panel Data", *Econometrica*, 47, 1979, pp.455~473.

Ridder, Geert, "An Empirical Evaluation of Some Models for Non-Random Attrition in Panel Data", *Structural Change and Economic Dynamics*, Vol.3., No.2, 1992.

Wooldridge, J., "Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions", *Journal of Econometrics*, 68, 1995, pp.115~132.

Ryu, Keunkwan, "New Approach to Attrition Problem in Longitudinal Studies," in C. Hsiao, K. Morimune and J. Powell, eds., *Nonlinear Statistical Modeling*, Cambridge Univ. Press, ch.4, 2001.

Hill, David H. and J. Willis, Robert, "Reducing Panel Attrition," *The Journal of Human Resources*, 36(3), 2001.

Lillard, Lee A. and W. A. Panis, Constantijn, "Panel Attrition from the PSID," *The Journal of Human Resources*, 33(2), 1998.

Zabel, Jeffrey E., "An Analysis of Attrition in the PSID and the Survey of Income and Program Participation," *The Journal of Human Resources*, 33(2), 1998.

Fitzgerald, John, Gottschalk, Peter and Moffitt, Robert, "An Analysis of Sample Attrition in Panel Data", *The Journal of Human Resources*, 33(2), 1998, pp.251~299.

〈부표 1〉 각 연도별 조사방식(면접, 유치, 전화조사) 분포

(단위 : 명, %)

| | 1차년도 | 2차년도 | 3차년도 | 4차년도 | 5차년도 | 6차년도 |
|---|---|---|---|---|---|---|
| 개인응답자수 | 13,321 | 12,042 | 11,206 | 11,051 | 10,966 | 11,543 |
| 전 체 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| (1) 면접 | 64.4 | 77.7 | 83.0 | 83.7 | 83.8 | 86.3 |
| (2) 유치 | 21.7 | 7.6 | 5.1 | 3.3 | 2.6 | 1.7 |
| (3) 전화 | 1.5 | 7.4 | 5.3 | 4.6 | 3.9 | 3.8 |
| (4) 면접+전화 | 2.6 | 4.7 | 3.8 | 4.9 | 6.4 | 5.7 |
| (5) 유치+전화 | 3.7 | 2.5 | 1.4 | 1.2 | 1.0 | 0.9 |
| (6) 면접+유치 | 2.6 | 0.0 | 0.9 | 1.8 | 1.0 | 1.3 |
| (7) 면접+유치+전화 | 0.0 | 0.0 | 0.2 | 0.5 | 1.3 | 0.4 |

〈부표 2〉 각 연도별 직접응답 비중의 분포

(단위 : 명, %)

| | 1차년도 | 2차년도 | 3차년도 | 4차년도 | 5차년도 | 6차년도 |
|---|---|---|---|---|---|---|
| 개인응답자수 | 13,321 | 12,042 | 11,206 | 11,051 | 10,966 | 11,543 |
| 전 체 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| (1) 본인 | 74.0 | 88.6 | 88.3 | 83.3 | 83.8 | 83.1 |
| (2) 대리인 | 19.7 | 11.3 | 8.1 | 11.0 | 9.8 | 11.2 |
| (3) 본인+대리인 | 0.4 | 0.0 | 3.6 | 5.5 | 6.4 | 5.8 |