한국인구학 제29권 제3호(2006년), pp. 159~188
연구논문

횡단조사자료 종단화의 가치와 한계: 경제활동인구조사와 도시가계조사*

이 지 연** · 김 진***

본 연구의 목적은 통계청에서 생산하고 있는 가구단위 조사 중 반복횡단조사로 설계된 경제활동인구조사와 도시가계조사의 표본가구관리명부 자료를 이용하여 1998년에서 2002년까지 패널자료를 구축하고, 패널의 생존기간과 패널무응답 집단의 사회인구학적인 특성을 분석하는 것이다. 경활의 표본가구관리명부 자료를 월별 패널로 구축한 결과, 모두 응답한 가구는 46%였다. 반면에 분기별 패널로 구축된 도시가계에서 모두 응답한 가구는 34.%로 경활보다는 낮았다. 콕스비례위험모형을 통해 가구와 가구원의 사회경제적 특성이 패널의 생존기간에 미친 영향을 분석한 결과 개인차원에서는 연령, 가구차원에서는 주택소유형태에 따라 체계적인 표본소실이 발생하고 있었다. 개인의 특성별로는 여자보다는 남자가, 장년층보다는 젊은층의 패널소실 위험율이 높았다. 학력이 높을수록 패널소실확률도 함께 증가했으며, 취업자보다는 실업자의 패널소실 확률이 높았다.

핵심단어: 패널자료, 패널소실, 콕스비례위험모형, 경제활동인구조사, 도시가계조사

I. 서 론

1. 한국 패널조사의 역사

한국에서 특정한 사회경제적 현상을 시간의 흐름에 따라 반복적으로 측정하는 종단조사(longitudinal survey), 특히 동일한 대상을 반복적으로 조사하는 패널조사(panel survey)의 역사는 그리 길지 않다. 미국은 1960년대 후반부터 대표

* 이 연구는 통계청 통계개발팀에서 수행된 연구과제의 일부이다. 본 연구의 관점은 필자들의 개인적인 견해이며, 통계청의 공식입장을 대표하는 것은 아님을 밝혀둔다.
** 통계개발원 사회통계실 사무관
*** 통계청 지역통계과 사무관

적인 패널조사인 Personal Survey of Income Dynamics(PSID)나 National Longitudinal Survey(NLS)등을 시작했다. 그 이후로 패널조사를 실시하는 나라들이 속속 늘어나면서 현재 전 세계적으로 20여 개국에서 대규모 패널조사를 정기적으로 실시하고 있다. 대다수의 패널조사들이 대학이나 연구기관을 통해서 실시되고 있지만 캐나다, 프랑스, 네덜란드 등에서는 국가통계기관이 직접 패널조사를 실시하고 있다(Gong, 2004).

한국 최초의 패널조사는 대우경제연구소가 1993년부터 연단위로 실시한 '한국가구패널조사(Korean Household Panel Study)'이다. 이 조사는 가구 및 가구원의 경제활동과 이에 영향을 미치는 사회경제적 변수들을 측정하기 위해 전국의 4,547가구와 18세 이상의 가구원 10,460명을 대상으로 처음 실시되었다. 그러나 높은 패널 소실율과 외환위기 등으로 인해 1998년에는 약 54%의 가구만이 패널로 남게 되면서 6회 차 조사를 마지막으로 조사가 중단되었다(대우경제연구소, 1999).

그 이후 가장 대표적인 패널조사는 '한국노동패널조사(Korea Labor and Income Panel Study)'로서, 한국노동연구원이 1998년부터 매년 전국 5천 가구, 15세 이상 경제활동인구 13,321명을 대상으로 경제행위 유형과 가구의 경제적 상황변화를 조사하고 있다(한국노동연구원, 2005). 이 외에도 최근 2-3년 사이에 전국규모의 패널조사들이 급격히 늘고 있는데, 정부승인통계로 등록된 패널조사만도 한국직업능력개발원의 '한국교육고용패널조사(2004년)', 보건복지부의 '한국복지패널(2005년)' 등이 있다.

2. 패널조사 수요증가 원인

패널조사를 포함하는 종단조사와 일반적인 횡단조사간의 가장 큰 차이점은 시간이 자료의 수집과 분석에 있어 중요한 변수로 작용하는가에 달려 있다(Gong, 2004). 포괄적인 의미로는 시간의 흐름에 따른 변화를 측정하는 모든 조사를 종단조사라고 부를 수 있지만, Menard(2002)는 다음의 세 가지 요건을 언급했다. 첫째, 조사되는 각 변수나 항목이 적어도 두 시점 이상의 기간을 두고 수집되어야 한다. 둘째, 첫 번째 조사 시점에서 다음번 조사 시점까지 조사대상자가 동일하거나 비교 가능한 집단이어야 한다. 마지막으로, 종단조사 자료는 상이한 시점들에서 비교 가능해야만 한다.

최근 패널조사에 관한 수요가 급격히 증가하고 있는 것은 패널자료를 통해 특정정책이나 프로그램이 목표 집단(target group)에 미친 인과적인 효과를 직접

적으로 평가하거나 예측(forecasting)해 볼 수 있는 개인차원의 정보를 얻을 수 있기 때문이다. 근래 종단자료(longitudinal data) 분석 기법이 발달하고 각종 통계분석패키지들이 널리 보급되었다는 점 또한 패널조사 수요를 확대시킨 한 요인으로 작용하고 있다.

3. 통계청 가구단위조사의 종단적 특성

통계청에서는 2005년 현재 총 64종의 통계를 작성하고 있는데, 원래부터 종단조사로 설계된 조사는 없었다. 그러나 가구를 단위로 하는 표본조사 중에서 '경제활동인구조사(이하 경활)', '도시가계조사(이하 도시가계)'등의 조사는 반복횡단조사로 설계되었지만 종단자료의 속성을 가지고 있다. 경활은 전국의 약 33,000가구를 대상으로 매월 15일 현재 만15세 이상 인구의 경제활동상태를 조사한다. 도시가계는 경활 표본으로 선정된 가구 중 도시지역에 거주하는 가구원수 2인 이상의 약 5천 가구를 대상으로 가구소득 및 소비실태를 매월 조사하여 분기별로 발표하는 조사다[1].

두 조사에서 표본은 인구주택총조사를 표본프레임으로 해서 지역, 거처, 가구주의 특성을 기준으로 추출된다. 한번 선정된 표본은 차기 인구주택총조사를 통해 표본프레임이 갱신되기까지 5년간 유지된다. 즉 동일한 대상을 상대로 2002년까지는 5년 동안, 연동표본이 도입된 2003년부터는 약 3년 동안 매월 조사가 반복 실시되는 것이다. 이러한 특성들은 앞서 언급한 Menard(2002)의 종단조사의 세 가지 조건을 충족시키고 있다. 또한, 경활과 유사한 조사인 Labour Force Survey를 실시하고 있는 캐나다, 영국, 뉴질랜드 통계청에서도 조사기간이 중복되는 가구들을 종단자료 형태로 구축해서 제공하고 있다(Laux and Tonks, 1998; Kuzmicich and Wigbout, 2001; Rowe and Nguyen, 2004). Neumark and Kawaguchi(2004)에 따르면 반복횡단조사 결과를 패널자료로 구축할 경우 처음부터 패널조사로 설계된 자료들을 이용하는 것에 비해 네 가지 장점이 있다.

먼저 표본규모가 일반적인 패널조사에 비해서 상대적으로 크고, 지리적 정보의 이용가능성이 높다. 또한 패널들이 조사주제에 따라 코호트나 특정 계층으로 한정되어 있지 않고, 조사결과를 연구자들이 상대적으로 빨리 얻을 수 있다는 것이다. 그러나, 횡단조사로 설계된 자료를 패널로 구축해서 분석하게 될 경우

1) 도시가계조사는 2003년부터 가구소비실태조사와 통합되면서, 표본의 크기를 7,500가구로 확대하고 조사범위를 농촌을 포함한 전국으로 확대시키면서 '전국가계조사'로 명칭을 개정하였다.

두 가지 방법론적인 문제가 제기된다. 표본의 체계적인 탈락으로 인해 발생하는 편향과 응답오차로 인해 발생하는 편향이 그것으로(Clarke and Tate, 1999), 본 연구는 전자의 문제에 중점을 두고자 한다.

4. 연구목적

이 연구의 목적은 통계청에서 생산하고 있는 가구조사 중 반복횡단조사로 설계된 경제활동인구조사와 전국가계조사자료를 이용하여 패널자료를 구축하고, 패널의 생존기간과 패널무응답집단의 사회인구학적 특성을 분석하는 것이다. 먼저 본 연구는 패널조사가 가진 일반적 성격과 장단점을 간략히 기술하고, 통계청에서 작성된 가구단위조사가 가진 패널자료로서의 특성과 선행연구들을 살펴 볼 것이다. 다음으로 경활과 도시가계의 표본가구관리명부 자료를 이용하여 1998년에서 2002년까지 패널자료를 구축하고, 패널의 무응답율 추이와 모두 응답한 집단과 무응답집단 간의 사회인구학적 특성을 살펴볼 것이다. 또한, 가구와 가구원의 사회경제적인 특성이 패널의 생존기간에 과연 얼마만큼의 영향을 미치는지는 콕스비례위험모형을 통해 심층적으로 분석해보고자 한다. 마지막으로 본 연구는 경활 및 가계조사자료를 패널자료로 활용하고자 할 때 고려되어야 할 사안들을 제언하고자 한다.

Ⅱ. 패널자료의 장단점과 경제활동 및 도시가계패널 선행연구

1. 패널자료의 특성

패널조사는 시간의 흐름에 따른 변수들의 변화 상태를 반복적으로 측정하기 때문에 변화과정 뿐만 아니라 특정한 원인에 대한 결과가 장기간에 걸쳐 발생할 경우에도 이를 파악할 수 있다는 장점이 있다. 이러한 특성 때문에 패널조사는 사회변화 혹은 개인의 태도나 행위변화에 관한 인과관계 분석에 가장 적합한 통계자료를 제공할 수 있다(Davies, 1994). 또한 패널조사를 포함한 종단조사들은 연령효과, 코호트효과, 시기효과를 각각 분리해 낼 수 있다(Diggle et al., 2002). 예를 들어 한 조사를 통해 연령에 따라 희망자녀수가 다르다는 사실이 밝혀졌다고 하자. 하지만 그 차이가 개인의 연령이나 생애주기단계상의 위치가 다르기

때문인지(연령 효과), 베이비붐이나 베이비버스트 세대와 같이 출생시점이 달라서 생기는 상이한 사회화 경험 때문인지(코호트 효과), 경제위기와 같은 특정 사건의 영향(시기효과) 때문인지는 일회적인 측정으로 끝나는 횡단조사를 통해 파악해내기란 어렵다.

종단조사는 특정현상을 여러 번에 걸쳐서 반복적으로 조사하기 때문에 횡단조사에 비해 비용이 많이 든다는 단점이 있다. 특히, 패널조사의 경우 동일인을 지속적으로 추적하기 위해서는 더 많은 시간과 경제적인 비용이 요구된다. 일반적으로 패널조사는 조사 설계에서 공표까지의 전 과정에 시간이라는 차원이 개입되기 때문에 횡단조사에 비해 더 복잡하다. 패널조사의 계획과 관리가 어려운 근본적인 이유는 장기적으로 설정된 목표나 수요에 비해 단기간에 동원 가능한 자원들이 제한되기 때문이다. 따라서 조사 초기부터 장기간의 목표와 자원이 충분히 고려된 상태에서 진행되지 않으면 패널조사는 지속되기 어렵다.

패널조사의 가장 큰 문제점은 패널 소실(panel attrition)이다. 시간이 지날수록 패널조사는 처음 조사와 다음 차수 조사 간에 동일한 표본을 유지하기가 점차 어려워진다. 조사대상자가 응답을 거절하거나 거주지 이전 및 사망 등의 이유로 조사에서 탈락되는 등 절단사례(censored cases)들이 점차 많아지기 때문이다. 이 때 절단된 사례들이 무작위로 발생하는 것이 아니라, 일정한 소득이나 직업이 없어서 추적이 어려운 빈곤층과 같이 특정 계층이나 집단에서 주로 발생했다면, 표본에 남아있는 패널이 모집단의 대표성을 유지하기 어려울 것이다. 패널조사에서는 조사 차수가 길어질수록 더 많은 정보를 수집할 수 있고, 조사간격이 짧을수록 기억오차가 줄어드는 장점이 있지만, 응답자의 부담은 그만큼 더 증가하게 되고 패널 소실의 가능성 또한 더 커진다는 단점이 있다.

2. 경제활동 및 도시가계 패널구축: 선행연구

한국에서 패널조사는 1990년대 중반에 최초로 실시되었지만, 이보다 10여년 앞서서 사후적으로 패널자료를 구축한 연구가 시도되었다. 서로 다른 두 시점상의 경활자료를 연계시킨 최초의 사례는 류재우 · 배무기(1984)의 연구이다. 이들은 한 달 동안의 노동력상태 변화를 파악하기 위해서 1983년 4월과 5월의 2회차 경제활동인구조사 원시자료에서 조사구역, 가구번호, 성, 생년월일 등을 추출해서 동일한 응답자의 것으로 간주되는 월별 자료를 연계시켰다.

이후 장기간의 경활자료를 연계해서 본격적인 패널분석을 실시한 것은 남재량(1997)의 연구이다. 실업률 추세변화를 분석하기 위해서 이 연구는 1982년부

터 1994년까지 경활의 하반기 자료들을 이용하여 5년 단위 패널을 구축했다. 월별 원시자료에서 개인의 조사구번호, 구역, 거처, 가구원번호, 그리고 출생년월일을 연계해서 개인별로 패널 ID를 생성한 후 바로 이웃하는 두 시점끼리 비교해서 동일인을 식별해 낼 수 있었다고 한다. 패널 기간 중 1983년과 1984년 9월과 12월 자료는 개인의 생년월일 정보에 누락이 있어서 개인별 ID에 생년월일 대신 성과 연령을 포함시켰는데, 생년월일 없이도 거의 정확한 매칭이 가능했다고 한다.

이러한 방법을 통해서 이웃한 두 달의 자료들을 개인별로 매칭시킬 수 있었던 비율은 전체 13년간의 패널 기간 중 평균 91.71%였다고 한다. 최근에 올수록 개인별 ID를 통한 자료 매칭율이 높아지고 있는데, 1987년 10월 이후부터는 91%이상, 90년대에는 평균 97%정도 까지 매칭이 가능했다고 한다(남재량, 1997). 이후에 남재량 · 류근관(1999)은 패널조사로 설계된 조사 자료들은 아니지만, 사후에 동일인을 식별해 낼 수 있는 정보들을 통해서 동일인으로 추정되는 자료들을 시간의 흐름에 따라 연계시킨 것을 '이동패널(moving panel)'이라고 불렀다.

경활자료를 월별로 연계해서 구축한 이동패널을 이용하면 어떤 분석이 가능한지 잠시 살펴보자. 횡단조사로 설계된 경활자료를 이용하면 취업률과 실업률은 알 수 있지만 개인들의 노동력 상태간(취업, 실업, 비경제활동)의 변화를 파악할 수 없다. 남재량 · 류근관(1999)은 1985년부터 1997년까지 경활패널구축한 후 노동력 상태별 유동률을 계산했다 (<부표 1> 참조). 이들에 의하면, 1985년부터 1989년까지 남성의 월별평균 취업유동률은 0.056이었다. 이는 지난달에 취업되어 있던 사람들을 100명이라고 할 때 이번 달에 실업이나 비경제활동인구에서 새로 취업상태로 들어왔거나 취업상태에서 다른 노동력 상태로 빠져나간 남성은 5.6명이라는 의미이다. 여성의 유동률은 남성에 비해 높게 나타나고 있으며 실업상태로의 유출입정도를 측정하는 수치는 남성보다 훨씬 높게 나타나고 있음을 알 수 있다. 이렇듯 패널자료는 개인적인 상태의 변화과정을 측정할 수 있고, 이러한 변화에 영향을 미친 요인들 간의 인과관계 분석이 가능한 정보를 제공할 수 있다는 장점이 있다.

개인별 자료를 패널로 구축하는 방법은 특정조사 내에서 뿐만 아니라 동일한 표본프레임으로부터 추출된 표본을 사용하는 상이한 조사 간에도 가능했다. 남재량 · 이창용(2001)은 경활과 도시가계를 결합하여 1982년부터 1999년까지 5년 단위 패널을 만들어서 외환위기와 실업률 변화에 관한 연구를 수행하기도 했다.

도시가계 자료를 패널로 구축한 사례는 황덕순(2002)과 이병희·정재호(2002)의 연구가 있다. 도시가계는 가구원수 2인 이상의 비농촌지역에 거주하는 가구 약 5,000여 가구를 대상으로 조사가 이루어진 자료이다. 두 연구 모두 1998-2000년 까지 총 36개월의 도시가계 월별자료를 가구별 ID로 묶어 패널자료를 만들었다. 먼저 3개월간의 자료를 연결하여 분기패널을 만들었는데, 이 때 분기 중에 누락이 발생한 가구는 패널에서 제외시켰다. 여기서 얻은 분기별 평균 가구수는 3,839가구였다. 다음으로 연속되는 두 개의 분기패널을 연결시켜서 6개월간 누락이 한번이라도 발생한 가구는 제외시키면, 평균 3,437가구가 확보되었다. 한편, 1998년부터 3년간 모두 조사된 가구만을 연결할 경우 1,475가구의 패널자료를 구축할 수 있었다. 패널소실의 유형은 자영자가구 보다는 근로자가구와 무직업자가구에서, 그리고 소득이 낮은 계층에서 소실율이 약간 더 높게 나타났다고 한다.

3. 경제활동과 도시가계 패널 특성: 거처패널

반복횡단조사로 설계된 경활과 도시가계 조사를 패널자료로 사용하기 위해서는 두 가지 방법론적인 문제점이 사전에 고려되어야 한다. 일반적인 가구패널조사와는 달리 경활과 도시가계는 거처를 표본선정의 단위로 할 뿐, 추적조사는 실시하지 않기 때문에 '거처패널'이라고 부르는 것이 보다 정확한 표현일 것이다. 거처패널과 가구패널은 패널의 무응답 유형에서 차이가 날 수 있다. 일반적으로 패널조사 무응답 유형에는: 무응답, 부적격(non eligible), 미상(unknown)이 포함된다(Nathan, 1998). 무응답은 조사대상자가 집에 없거나, 응답을 거절한 경우이다. 부적격은 사망하거나, 이민가는 경우이다. 미상은 응답자가 이주한 장소를 파악할 수 없어서 추적이 불가능 할 때 발생한다. 경활과 도시가계조사는 가구를 추적조사하지 않기 때문에 일반 가구단위 패널에 비해서 미상으로 인한 패널의 소실이 많이 발생하고, 따라서 패널응답률이 낮아질 수밖에 없다.

Nathan(1998)은 영국의 주요패널조사에서 나타나는 소실에 관한 연구에서 패널조사 응답률은 조사완료 사례수를 조사선정 사례수로 나눈 것이라고 했다. (응답률 = 조사완료사례/(조사완료+무응답+부적격+미상)*100). 이 때 현실적으로 조사되어서는 안 될 부적격 사례를 분모에서 제외시키게 되면 응답률의 하한선이 설정된다. 만약 미상인 사례와 부적격 사례를 명확히 파악 할 수 없기 때문에 이들을 분모에서 제외시키게 되면 응답률의 실제적인 상한선을 설정할 수 있다(응답률 상한선 = 조사완료사례/(조사완료+무응답)*100).

예를 들어, 총100개의 표본 중에서 70개 표본은 조사되었고, 무응답, 부적격, 미상이 각각 10개씩 발생했다고 하자. 정확한 패널응답률은 70%이지만, 실제적으로는 부적격과 미상의 구분이 불분명하므로 응답하한선은 77%, 응답상한선은 87.5%가 되는 셈이다. Nathan의 방식을 적용해 보면 거처패널은 정확한 패널응답률 산출방식으로만 계산되어지는데 비해, 일반적인 가구패널은 실질적인 응답하한선 내지 응답상한선 방식으로도 계산될 수 있다.

두 번째는 거처가 표본선정 단위이기 때문에 봄과 가을 이사철에는 전출로 인한 무응답이 증가하면서 패널소실에 일정한 계절성이 나타날 것이라는 점이다. 전국규모의 조사들이 인구이동이 적은 달을 선정해 조사를 진행하는 이유도 이러한 계절성이 고려된 것이다. 하지만 사후적으로 패널로 구축된 경활과 도시가계자료는 이러한 계절성을 조절할 수가 없다. 따라서, 거처패널의 특성이 최종적인 추정치에 어떠한 편향을 발생시키는지는 별도의 연구를 통해서 확인해 봐야 한다는 문제점이 있다.

III. 연구방법

1. 자료

이 연구는 1998년 2002년까지 경활과 도시가계의 표본가구관리명부 자료를 사용하였다. 가구관리명부는 특정조사의 표본으로 선정된 가구의 관리를 목적으로 매월 조사원에 의해 작성된다. 연구목적인 패널표본의 대표성을 파악하기 위해서는 다음의 두 가지 정보가 필수적이다. 첫 번째는 표본가구와 가구원 각각의 사회인구학적 특성, 두 번째는 사망이나 전출, 조사표 미제출 등 표본에서 제외된 사유에 대한 기록이다. 경활과 도시가계의 원자료에는 위의 두 가지 정보가 불충분했지만, 가구관리명부에는 있었다. 또한 원자료에 비해 자료의 용량이 상대적으로 적어 분석이 용이했기 때문에 본 연구에 이용되었다.

2. 패널구축 방법

이 연구는 2003년에 연동표본이 도입되기 전까지 최고 5년간을 패널구축기간으로 선정했다. 패널구축을 위해서는 먼저 가구관리명부상의 조사구번호, 구

역, 거처, 가구원번호를 연계하여 일련의 개인별 ID를 만들었다. 그리고 월별자료들을 병합한 후 같은 ID별로 자료를 정리해서 가구관리패널을 구축했다. 이 가구관리명부패널은 경활과 도시가계 모두 월별조사가 이루어졌기 때문에 원래 총 60회 차의 패널차수를 갖는다. 그러나 도시가계의 경우 조사는 월별로 이루어지더라도 결과는 분기별로 발표되고 있기 때문에, 각 분기 중에서 응답이 발생한 첫 달을 선정하여 총 20회 차의 패널을 구축하였다.

3. 표본소실

이 연구에서 표본소실은 제1회 차에 응답한 표본 중에서 이후에 응답거절, 전출, 사망 등의 사유로 무응답이 발생한 최초의 시점으로 정의하였다. 이렇게 정의할 경우 소실율을 과대평가(overestimated)하게 되는 경향이 있다. 실제 조사에서는 한번 무응답이 발생했을지라도 전출여부가 명확하지 않는 한 그 가구를 계속 조사하기 때문에 표본에서 완전히 탈락하는 것은 아니기 때문이다.

이것은 무응답의 형태 중 패널조사에서만 나타나는 차수 무응답(wave non-response)이다. 위의 패널소실 정의를 따르면 차수무응답이 발생했을 지라도 그 이후의 정보들이 패널자료상에서는 제외된다. 극단적인 예를 들자면 한 가구가 총 60회 차 중 2회 차에 부재 등의 사유로 응답을 하지 않고, 그 이후 58회 차를 모두 응답했을지라도, 본 연구에서는 2회 차에 표본소실이 발생한 것으로 간주하고, 그 이후의 정보들은 패널로 구축된 자료상에서 제외시키게 된다. 이렇게 표본소실을 엄격하게 정의한 이유는 1998년부터 5년간의 경활 및 도시가계 자료를 통한 각종 패널구축시 표본소실율의 최하한선을 제공하기 위해서이다.

본 연구의 표본소실율은 또 다른 사유로 인해서도 과대 측정될 수 있다. 경활과 도시가계에서 가구원 번호는 첫 번째 조사에서는 가구주와의 관계 순으로 부여되지만, 다음번 조사부터 변동사항이 발생했을 경우 가구원이 신규 진입한 순서대로 부여된다. 이동이 없을 경우 동일한 번호가 부여되지만, 가구원 중 일부가 일정기간 외지에 나가 있다가 다시 대상가구로 돌아오면 원래 가구원번호가 아닌 새 번호를 부여하게 된다. 현재의 가구원번호 체계가 출입(in and out) 여부를 자유롭게 반영할 수 없기 때문에 실제로는 동일인에 관한 기록일 수 있지만, 가구번호가 다를 경우 연결할 수 없기 때문에 패널로 구축된 사례에서 제외될 수 있다.

IV. 분석결과

1. 표본소실 추이 및 특성

1) 패널표본의 대표성

추출된 표본이 모집단 전체의 특성을 얼마만큼 잘 대표하는가라는 표본대표성의 문제는 횡단조사에서는 최종표본의 추출확률이 동일했는가를 의미한다. 이 대표성의 문제가 패널조사에서는 더욱 복잡해지는데, 그 이유는 모집단 자체가 시간에 따라 변화하기 때문이다. 횡단조사의 목표모집단은 시점에 따라 달라지겠지만, 패널조사의 경우는 기간에 따라 달라진다(Nathan, 1998).

동일한 대상을 일정한 간격을 두고 계속 추적 조사한다는 패널조사의 특성은 또 다른 문제를 야기 시킨다. 패널조사 첫 차수의 표본이 대표성을 확보하고 있었더라도, 무응답이 무작위로 발생한 것이 아니라면 그 이후에 조사된 표본의 대표성을 보장할 수 없다. 응답자에 관한 기록이 시간에 따라 누적되는 패널자료의 특성상 첫차수의 응답률은 상당히 중요하다. 제2회 차 이후부터 표본의 소실율은 첫차수의 응답자를 분모로 계산되기 때문이다.

이하에서는 경활패널과 도시가계패널의 연간 표본소실율 추이와 특성을 각각 살펴보고자 한다. 두 패널자료의 소실추이를 직접적으로 비교하는 것은 해석상에 주의를 요한다. 앞서 언급했듯이 통계청 가구단위 조사에서는 경활과 도시가계는 동일한 표본프레임을 사용하고, 경활표본의 일부가 도시가계 표본으로 이용된다. 그러나 경활표본에서 비농가로 분류된 표본만이 도시가계의 대상이 되기 때문에 표본 표본소실 특성에서 차이를 발생시킬 수 있다.

2) 경제활동패널

다음의 <표 1>은 경활 가구관리명부자료를 바탕으로 가구와 개인표본의 소실율 추이를 연도별로 살펴본 것이다. 최초표본의 의미는 각 연도 1월의 패널표본의 총 규모를 의미하고, 소실율은 해당연도 12월까지 탈락된 사례들의 비율이다. 패널 1차(1998년 1월)에는 총 29,271가구, 93,726명의 가구원이 응답하였다. 이들 중 마지막 60차까지 모두 응답한 가구는 46.5%(13,616가구)이고, 가구원은 36.7%(34,386명)이었다. 가구에 비해 가구원의 소실율은 높을 수밖엔 없다. 가구차원에서 한번 무응답이 발생하면 가구원 모두에게 무응답이 발생한 셈이 되

지만, 한 가구가 마지막 차수까지 잔류했을지라도 가구내 특정 가구원의 소실은 발생할 수 있기 때문이다. 연도별 패널소실율 추이를 살펴보면, 첫 1년간의 가구 소실율은 19.5%로 가장 높았고, 시간이 갈수록 점차 감소하다가 3년 이후부터는 소실율이 일정수준에서 안정화되는 것을 알 수 있다.

<표 1> 경제활동패널의 연도별 패널소실율 추이

	패널차수(년/월)					
	1차(98/1)	13차(99/1)	25차(00/1)	37차(01/1)	49차(02/1)	60차(02/12)
가구패널						
표본규모	29,271	23,570	19,448	16,902	15,081	13,616
소실율(%)		19.5	17.5	13.1	10.8	9.7
생존율(%)		80.5	66.4	57.7	51.5	46.5
개인패널						
표본규모	93,726	72,563	57,125	47,221	39,931	34,386
소실율(%)		22.6	21.3	17.3	15.4	13.9
생존율(%)		77.4	60.9	50.4	42.6	36.7

월별로 이루어지는 가구단위 패널조사가 드물기 때문에 직접적으로 비교할 수는 없지만 국내외 주요 패널조사와 비교해 볼 때 경활패널의 소실율 수준 자체는 높지 않은 것으로 보인다. 한국노동패널의 연도별 패널소실율 추이를 살펴보면, 1998년 5천 가구 패널로 시작해서 지난 6회 차까지의 가구 소실율은 61.7%였는데, 이는 신규로 대체된 표본이 감안된 수치이다(<부표 2> 참조). 패널 첫 차수에 소실율이 높고, 3년 이후 부터 안정화되는 패턴도 경활패널과 유사한 면이 있다. 이렇게 패널초기에 높은 소실율과 이후의 안정화 경향은 외국의 조사에서도 발견할 수 있다. 미국의 대표적인 패널조사인 PSID(<부표 3 참조>)는 1968년에 18,191가구를 대상으로 시작되었는데, 20년 후인 1988년에서야 표본생존율이 50%이하로 떨어질 만큼 상당히 성공적으로 패널이 정착된 조사이다. 이 조사도 첫 차수에 소실율이 11.9%로 가장 높다가 점차 떨어져서, 제3년차 이후부터는 안정화되는 것을 알 수 있다[2].

앞서 언급했듯이 패널자료에서 무응답이나 전출, 사망 등으로 절단사례들이 발생할 때 이것이 응답자의 사회인구학적 특성에 관계없이 무작위로 발생한다면

2) 초기 차수들의 패널소실율에 영향을 미치는 첫 번째 요인은 가구의 무응답이었고, 두 번째는 이주로 인한 추적불능이다. 제시된 <부표 3>을 보면 3년 차부터 가구무응답과 함께 이주로 인한 소실율이 안정화되는 것을 알 수 있다. 대부분의 패널조사들이 아주 다양한 방식으로 이주로 인한 추적불능을 줄이기 위한 대책들을 마련해 놓고 있지만 조사 초기에 이주로 인한 소실의 일정부분은 가구에서 계속 조사에 적극적으로 참여할 의사가 없었기 때문인 것으로 해석할 수 있다.

패널표본의 대표성은 심각한 문제가 되지 않는다. 패널 1차에 추출된 표본의 특성과 최종 60차까지 응답을 완료한 표본(잔류표본)의 특성이 얼마만큼의 차이를 보이는가는 패널자료의 대표성을 판단하는 기준이 될 수 있다.

<표 2> 가구원 특성별 경제활동패널의 최초표본과 최종잔류표본 구성비

계		패널차수	
		1차(%)	60차(%)
성	남	48.7	45.2
	여	51.3	54.8
연령	15-19	8.8	
	20-24	6.9	3.9
	25-29	8.9	2.9
	30-34	8.4	4.0
	34-39	9.5	6.1
	40-44	8.8	9.9
	45-49	6.0	9.6
	50-54	4.9	8.5
	55-59	4.9	7.7
	60+	11.7	8.2
	65+	----	17.2
학력	초졸 이하	41.8	50.4
	중졸	15.1	14.9
	고졸	27.3	24.2
	초대졸	6.8	4.0
	대졸이상	9.0	6.5
가구주관계	가구주	31.4	34.3
	배우자	22.7	29.0
	미혼자녀	36.1	28.7
	형제자매	1.7	0.6
	부모	4.6	5.3
	조부모	0.1	0.1
	기타 친인척	2.8	1.8
	동거인	0.7	0.1
경제활동상태	취업	42.5	44.5
	실업	2.1	1.4

경활패널에서 최초표본 가구원과 최종잔류표본 가구원의 특성을 성, 연령, 교육, 가구주와의 관계별로 살펴보자. 앞의 <표 2>를 보면 동일 가구 내에서도 가구원의 특성에 따라 표본탈락이 차별적으로 발생하고 있음을 알 수 있다. 여자보다는 남자가 표본소실이 많이 발생했다. 전체 표본에서 남자가 차지하는 비중은 최초표본에 비해 최종표본에서 7.1% 감소한데 비해, 여자의 구성비는

6.8% 증가했다. 연령별로 살펴보면 대체로 연령이 낮을수록 많은 소실이 발생했다. 1998년에 35세 이하 연령층에 속한 집단에서 가장 많은 표본소실이 발생했는데, 20-24세 연령층의 구성비는 58.2%나 감소했다. 35세 이후 연령부터는 점차 표본소실이 줄어들면서 40대 연령층에서는 별다른 변화가 없다가 그 이후 연령부터는 최종표본에 잔류하는 비중이 다른 연령층에 비해 상대적으로 높았다.

교육정도별로 두 표본간의 분포차이를 살펴보면, 초대졸자의 비중의 감소가 41.6%로 가장 많았고, 다음으로는 대졸자와 고졸자의 순으로 감소하고 있다. 가구주와의 관계별로 최초표본과 최종잔류표본을 살펴보면 가구주와 배우자가 표본에 잔류하는 비중은 상대적으로 높은 반면, 형제자매와 미혼자녀의 표본소실이 많이 발생했음을 알 수 있다.

<표 3> 가구특성별 경제활동패널의 최초표본과 최종잔류표본 구성비

계		패널차수	
		1차(%)	60차(%)
가구주 직업	입법자/고위임원/관리자	3.1	3.0
	전문가	4.7	3.8
	기술공/준전문가	7.3	5.8
	사무직원	9.3	7.4
	서비스/시장판매근로자	16.4	13.9
	농/임/어업종사자	11.8	20.2
	기능관련근로자	13.1	11.6
	장치/기계조작	11.6	11.3
	단순노무직근로자	6.8	7.2
	군인/기타	15.8	15.7
거처 종류	단독	60.5	60.9
	아파트	26.6	27.1
	연립	7.8	7.5
	다세대주택	1.8	1.7
	비주거용주택	3.2	2.7
	주택이외의 거처	0.1	0.1
조사 참여	도시가계+경활	27.9	24.4
	경활	72.2	75.6

앞의 <표 3>은 경활패널의 소실이 가구주의 직업, 거처의 종류등 가구의 특성에 따라 어떤 차이가 있는지를 살펴본 것이다. 이 표에 의하면, 농임어업종사자와 단순무직 근로자를 제외하고는 거의 대부분의 직업에서 표본소실로 인해 최종표본에서의 구성비가 감소하였다. 거처의 종류별로 살펴보면 주택거주자보다 주택이외의 거처에서 생활하는 가구에서 표본소실이 많이 발생했다. 주택에

거주하는 경우에는 단독주택과 아파트의 거주가구 보다는 연립이나 다세대 주택 거주 가구에서 표본소실이 약간 더 많이 발생했다. 통계청 가구단위 조사에서는 경활 표본의 일부가 도시가계 표본으로도 이용되는데, 조사의 부담 정도가 응답자의 최종잔류 여부에 영향을 미칠 수 있다. 응답가구의 조사유형에 따라 최초 표본과 최종표본을 비교해 본 결과 도시가계와 경활 모두를 조사하는 가구의 구성비 감소는 경활만 조사하는 가구에 비해 더 많이 감소했음을 알 수 있다.

3) 도시가계패널

분기별로 발표되는 도시가계자료의 특성을 감안하여 1998년 1월에 응답한 4,612가구 중에서 이후 각 분기별로 응답이 발생한 가장 첫 달을 선정하여 총 20회 차의 도시가계패널이 구축되었다. 다음의 <표 4>에서 도시가계패널의 5년간 최종잔류율은 가구차원에서 34.5%인 1,590가구, 가구원별로는 28.1%인 4,818명으로 경활패널보다 훨씬 낮게 나타나고 있다. 이는 도시가계와 경활간의 응답부담의 차이 때문인 것으로 보인다. 도시가계는 한 달 동안의 가계소비 및 지출에 관한 모든 현황을 조사하기 때문에 응답부담이 높은 데 반해, 경활은 일주일간의 경제활동 여부만을 조사하므로 상대적으로 응답이 용이한 편이다.

<표 4> 도시가계패널의 연도별 패널소실율 추이

	패널차수(년/분기)					
	1차(98/1)	5차(99/1)	9차(00/1)	13차(01/1)	17차(02/1)	20차(02/12)
가구패널						
표본규모	4,612	3,572	2,764	2,233	1,868	1,590
소실율(%)		22.5	22.6	19.2	16.3	14.9
생존율(%)		77.5	59.9	48.4	40.5	34.5
개인패널						
표본규모	17,158	12,769	9,506	7,360	5,906	4,818
소실율(%)		25.6	25.6	22.6	19.8	18.4
생존율(%)		74.4	55.5	42.9	34.4	28.1

도시가계패널의 가구단위 연간소실율 패턴은 일반 패널조사들과 마찬가지로 패널초기에 높고 점차 안정되어가고 있다. 가구의 패널소실이 첫 해와 그 이듬해까지 연간 22%대로 높았고, 제3년차부터 19%대로 낮아져서 마지막 해에는 16.3%까지 떨어졌다. 가구원의 연간 소실율 패턴은 가구와 유사하지만 소실율이 상대적으로 더 높아서 최종 20회차 잔류율은 28.1%이었다.

도시가계패널은 가구단위로 소득과 지출결과가 발표되기 때문에 가구특성에 따른 패널 소실 추이가 중요하다. 다음의 <표 5>에서 거처의 종류별로 최초표본과 최종잔류표본의 구성비 변화를 살펴보면 비주거용 주택과 주택이외의 거처에 거주하는 가구와 단독주택 거주가구의 패널소실이 두드러진다. 단독주택 거주가구의 경우 최초표본에서는 53.2%를 차지한 반면 최종잔류표본에서는 47.7%로 5.5% 감소했다. 응답가구의 주택소유 형태에 따라서도 최초표본과 최종표본간의 구성비 차이가 보이는데 주택을 소유하고 있는 가구는 가장 안정적으로 조사된 반면 전월세 거주 가구는 표본잔류비율이 크게 감소했다.

<표 5> 가구특성에 따른 도시가계패널의 최초표본과 최종잔류표본 (%)

계		패널차수	
		1차	60차
거처종류	단독주택	53.2	47.7
	아파트	31.7	35.1
	연립	10.0	11.4
	다세대	3.0	3.6
	기타(비주거용, 주택이외)	2.1	1.5
주택소유	자가	53.4	61.2
	전세	32.0	11.8
	월세	12.4	5.0
	기타	2.2	1.9
주된 소득	근로	61.8	60.8
	사업	26.7	27.6
	연금	0.7	1.2
	재산	1.9	2.3
	수증/보조	5.3	4.3
	기타	3.6	3.8
가계구분	봉급자	26.0	24.1
	노무자	34.8	35.5
	근로자 외	27.1	27.9
	무직	12.1	12.5

가계의 주된 소득 유형에 따라서도 최종표본잔류율은 차이가 있었다. 근로 및 사업소득가구에서는 최초표본과 최종표본간에 별다른 차이를 보이지 않았다. 연금이나 재산소득이 주된 소득원인 가구가 가장 안정적으로 조사되었다면, 수증이나 보조로 생계를 유지하는 가구는 최종표본으로 잔류하는 비율이 가장 낮았다. 가계형태에 따른 최종잔류율은 별다른 차이가 없었다. 다만 봉급자 가계는 최초표본에 비해 최종잔류비율이 다른 가계형태에 비해 유일하게 감소하였다.

2. 경제활동패널과 도시가계패널의 생명표 분석

1) 경제활동패널의 가구원 특성별 생존곡선

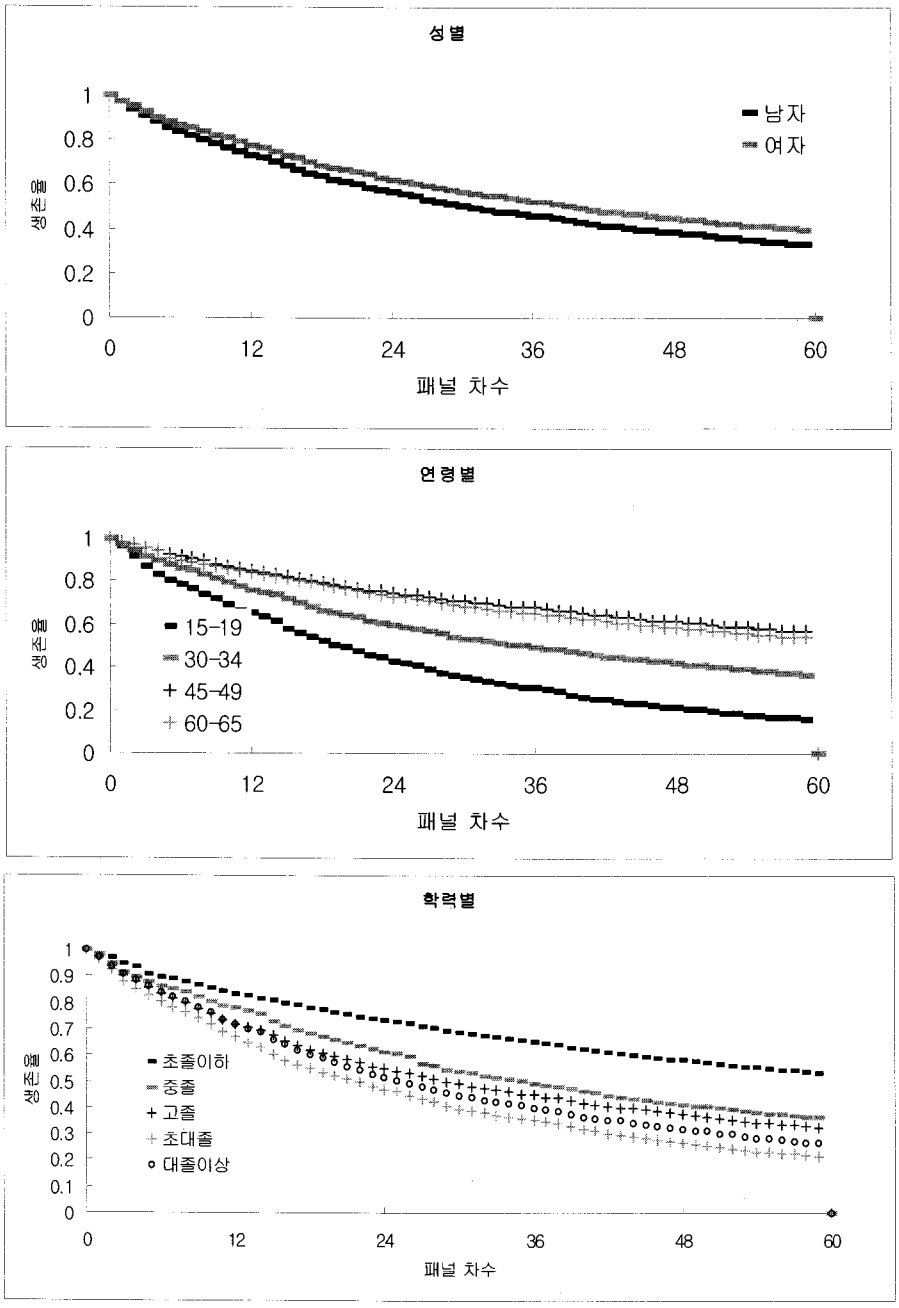
표본의 특성에 따라 패널생존기간은 얼마나 차이가 날까? 패널자료를 분석하는데 주요한 개념은 위험율(hazard rate)이다. 위험율이란 사건발생 위험에 노출되어 있는 집단(위험집합)에 속한 한 개인이 특정 시점 전까지 사건을 경험하게 될 확률이다. 이것은 특정한 시간 간격내의 사건발생율(사건 발생 사례 수/위험집단에 포함된 사례수)로, 1에서 이러한 위험율을 빼면 사건이 발생하지 않고 생존할 확률이 된다.

생명표 분석방법을 이용하면 패널표본의 생존기간별 분포를 추정할 수 있다. 표본소실이라는 사건발생 시간은 경활에서는 경우 한 달 간격으로 도시가계에서는 3개월 간격으로 집합한 후 각 간격의 중간시점에서 누적생존율을 추정할 수 있다. 생명표 방법의 핵심은 일정시점과 그 다음 시점사이의 구간별 사건발생률을 구하고, 이를 통해 일정시점까지 사건이 발생하지 않고 생존할 확률을 구하는 것이다. 다음의 <그림 1>은 경활패널의 가구원 특성에 따른 생존곡선이다.

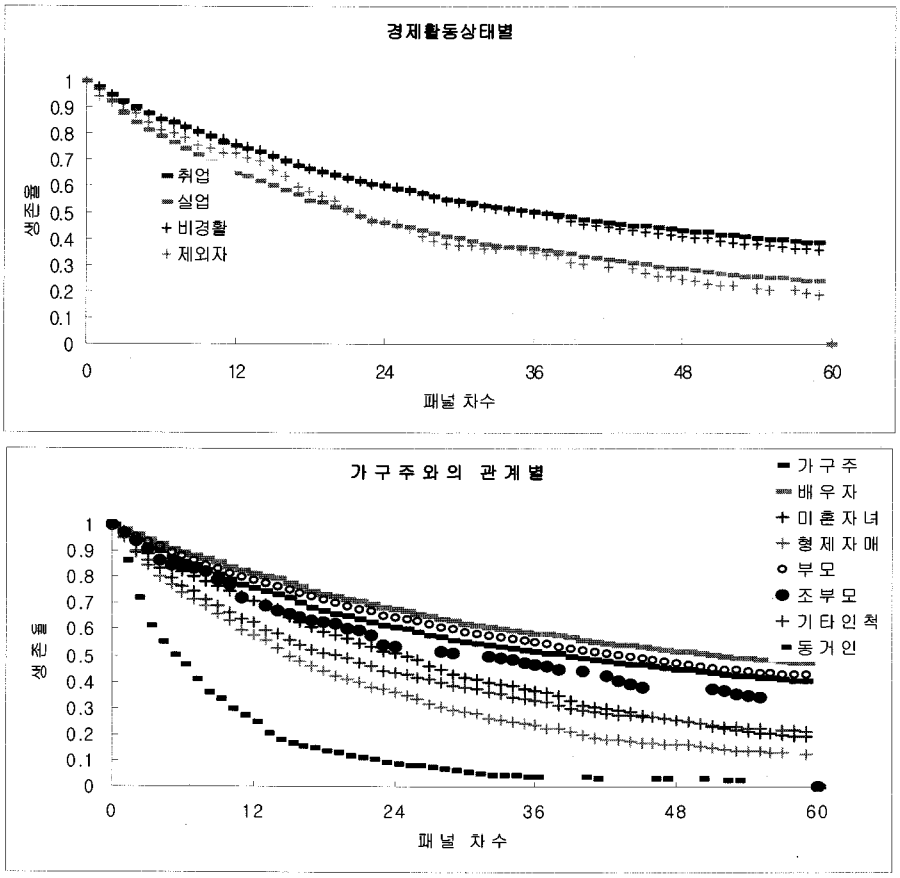
경활패널에서 응답자의 성별에 따른 생존기간 차이를 살펴보면 여자의 생존확률이 남자에 비해 항상 높았던 것을 알 수 있다. 남녀 간의 생존확률의 차이는 6회 차 이후부터 점차 벌어지기 시작해서 마지막 60차수에서는 6.4%까지 벌어진다. 남자의 경우 패널생존율은 제3회 차 조사에서 90.6%까지 떨어지고, 12회 차에 73%, 30회 차가 되면 50%밖엔 되지 않는다. 다시 말해서 경활조사에 표본으로 선정된 사람이 남자라면 조사가 시작된 달로부터 3개월 이내에 10명중 1명은 적어도 1회 이상 조사에 응답하지 않았음을 의미한다.

패널의 생존기간의 차이를 설명하는데 있어서 연령은 중요한 변수이다. 5세 계급별로 생존확률을 분석해 본 결과, 20대 전후의 연령대를 제외하고는 이웃한 연령대간에는 생존곡선이 크게 차이 나지 않았기 때문에 분석상의 편의를 위해 본문에는 연령계급을 4개 집단으로 구분한 자료가 수록되었다. 연령과 패널의 생존확률은 정의 상관관계를 나타내는데, 45세 이후부터는 생존곡선이 매우 유사한 것을 알 수 있다. 한편, 15-29세 집단의 생존곡선은 가장 급격히 떨어진다. 생존확률이 제5회 차에 80%, 제20회 차에 50%, 최종 60회 차에는 16%에 이른다. 이에 비해 30-34세 집단은 비교적 완만한 편으로 제35회 차까지는 50%의 생존확률을 보이고 있다. 그리고 60세 이상 집단의 경우는 제16회 차까지도 평균 5명중 1명 정도, 마지막 차수까지는 평균 2명 중 1명이 응답을 하지 않았다.

<그림1> 경제활동패널의 가구원 특성별 생존곡선



<그림1(계속)> 경제활동패널의 가구원 특성별 생존곡선



생존기간은 학력에 따라서도 차이가 있다. 패널의 생존확률과 학력은 연령과는 달리 전체적으로는 부의 상관관계를 보인다. 그러나 생존확률이 가장 낮은 집단은 대졸이상자가 아니라 초대졸자였다. 이것은 부분적으로는 연령효과로 인한 것으로 보이는데, 뒤에 소개될 콕스비례위험모형에서는 연령효과가 통제되자 생존확률이 가장 낮은 집단은 대졸자임이 드러났다. 초졸 이하의 학력소지자들은 제6회 차까지 90%이상의 생존확률을 보이고, 제44회 차에 60%, 제60차에 53%의 생존확률을 보인다. 이에 반해 대졸자의 생존곡선은 제4회 차에 이미 88%까지 떨어지고, 제18회 차에는 60%, 제60회 차에는 26%만이 생존했다.

응답자의 경제활동상태 또한 패널의 생존기관과 밀접한 관련이 있다. 앞의

<그림 1>에서 경제활동상태별 생존곡선을 보면 취업자와 비경활자의 생존곡선의 패턴이 유사하고, 실업자와 제외자(군인, 집단시설 수감자등)의 생존곡선이 유사하다는 점을 알 수 있다. 제12회 차 이후부터는 경제활동상태에 따른 집단간 차이가 확연히 드러나기 시작한다. 취업자의 경우 제22회 차까지의 생존확률이 90.3%인 반면, 실업자의 경우는 79%이다. 제39회 차에도 전자의 생존확률은 80.2%, 후자는 62.9%로 낮아진다.

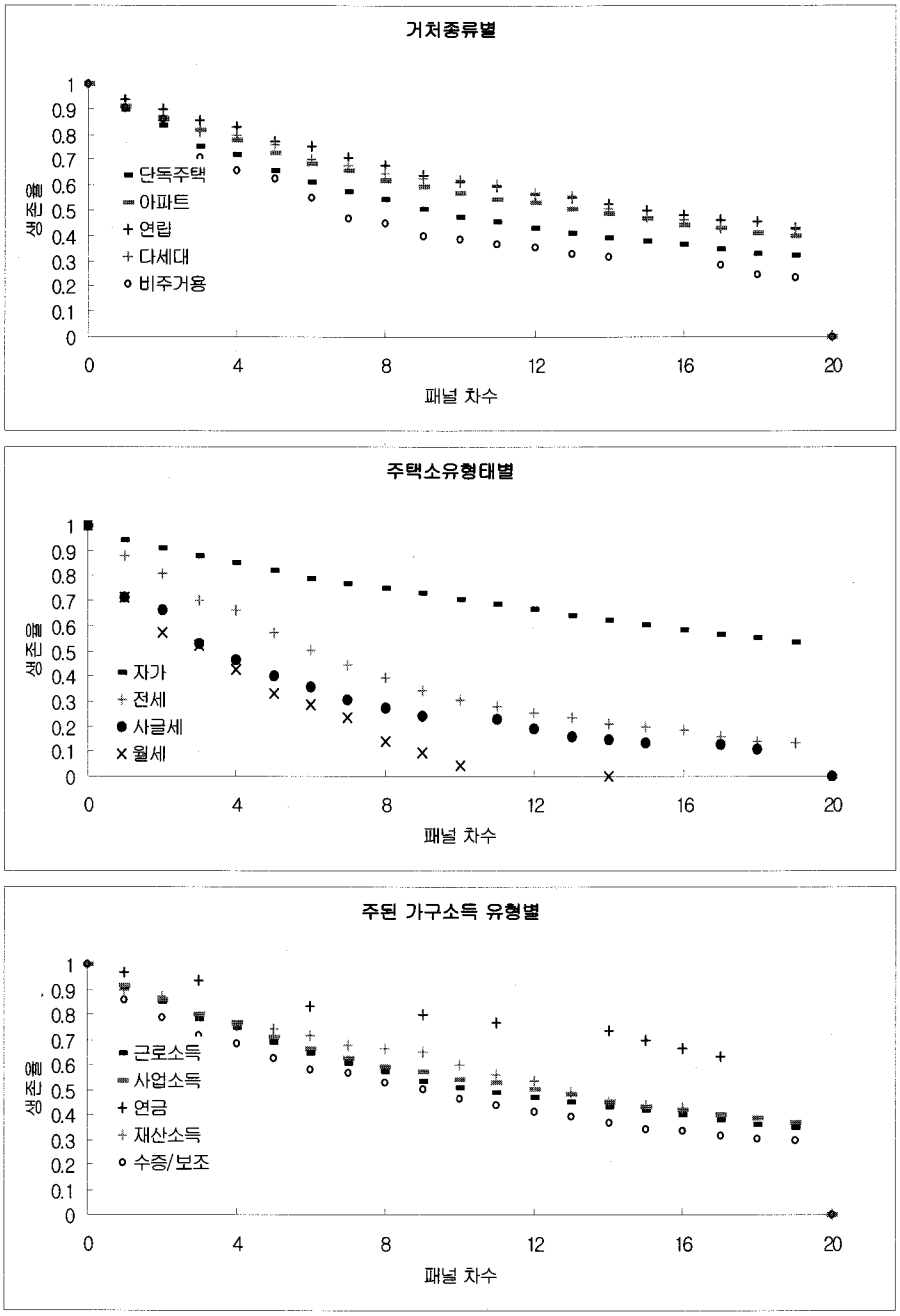
취업자의 생존확률은 제60회 차 모두 항상 높은데 반해, 실업자의 생존율은 처음에는 가장 낮지만 2년차 이후부터는 제외자 보다 높게 나타난다. 이러한 현상은 시간에 따라 변화하는 변수(time-dependent covariate)인 경제활동상태의 특성이 생명표 분석에 반영되지 않기 때문이다. 따라서 실업자와 제외자의 생존곡선상의 교차는 실제로 두 집단의 생존확률이 변화했다기 보다는 실업자의 경제활동상태가 2년 정도 후에는 변화를 경험하기 때문으로 보이는 데, 이에 대해서는 별도의 검증이 이루어져야 보다 분명한 판단이 가능할 것으로 보인다.

가구주와의 관계 또한 생존기간의 차이가 있는데, 조부모의 생존곡선이 가장 완만한 반면 기타 친인척의 생존곡선이 가장 급격하게 떨어졌다. 제43회 차까지도 조부모가 표본으로 생존할 확률은 90.2%인 반면, 기타 친인척은 59% 밖엔 되지 않았다. 생존확률의 차이는 있지만 생존곡선의 패턴이 서로 유사한 집단들이 있다. 가구주와 배우자의 곡선이 유사하고, 형제자매는 기타 친인척과 부모와 미혼자녀의 패턴이 유사한 것을 알 수 있다.

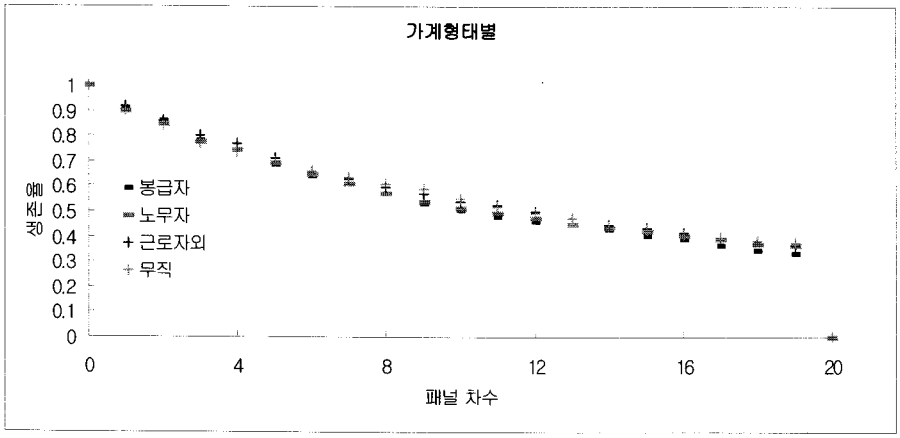
2) 도시가계패널의 가구특성별 생존곡선

도시가계 패널은 가구특성이 주요 변수이기 때문에, 거처의 종류, 주택소유형태, 주된 가구소득 유형별로 생존곡선의 차이를 살펴보았다. 다음의 <그림 2>를 보면 주택의 소유여부가 패널생존기간에 유의미한 차이를 주고 있었다. 생존기간이 긴 집단은 연립주택거주 가구였는데, 첫 회 차 조사로부터 1년 후인 제4회 차까지도 82.6%의 생존확율을 보였고, 최종 20회 차에도 43.2%가 생존했다. 이에 반해서 단독주택과 비주거용주택소유자의 생존확률은 낮은 편이다. 제2회 차 조사까지는 단독주택의 생존확률이 83%로 가장 빨리 떨어졌지만 이후부터는 비주거용주택에 거주하는 가구의 생존확률이 더 빨리 떨어져서 이 집단의 생존확률은 조사 시작 후 1년 반이 지난 제6회 차에는 55% 정도밖에 되지 않았다.

<그림 2> 도시가계패널의 가구 특성별 생존곡선



<그림 2(계속)> 도시가계패널의 가구 특성별 생존곡선



주택소유형태별로 생존곡선을 살펴보면, 자가주택 소유, 전세, 사글세, 월세 순으로 생존확률이 높았다. 자가주택 소유가구의 생존확률은 제2회 차까지 90%를 유지했고, 최종 20회 차까지 53.7%가 생존했다. 이에 반해 전세 거주가구는 제2회 차에 80%, 제20회 차에 13.1% 생존했고, 월세의 경우는 제8회 차에 이미 생존확률이 14%밖엔 되지 않았다.

가구의 주된 소득 유형에 따라서도 도시가계패널의 생존확률은 차이가 났다. 연금소득가구는 다른 가구와 달리 생존확률이 높았다. 그 외의 집단의 생존확률은 재산소득, 근로소득, 사업소득, 수증 및 보조 가구 순으로 낮아지지만, 생존곡선의 패턴이 서로 유사한 것을 알 수 있다.

3. 경제활동 및 도시가계패널의 생존기간 결정요인 분석: 콕스비례위험모형

지금까지는 가구원 특성 변수별 경활패널의 생존기간과 가구특성별 도시가계패널의 생존기간을 살펴보았다. 그러나 패널의 생존기간은 단일한 설명변수에 의해서 결정되는 현상이 아니다. 연령이나 학력과 같이 설명변수들 간에도 상관관계가 존재하기 때문에 다른 설명변수의 값을 통제한 상태에서 특정 설명변수가 표본의 생존기간이라는 종속변수에 독립적으로 행사하는 영향력을 회귀모형을 통해서 측정해 볼 필요가 있다.

본 연구에서는 가구원과 가구 차원의 설명변수들이 경활과 도시가계 패널의 생존기간에 미치는 상대적인 효과를 알아보기 위해서 콕스 비례위험 모형을 이용하였다. 한 개인이 특정시점에서 패널소실을 경험할 확률은 다음의 방정식으로 표현된다.

$$\log h_i(t) = \log h_0(t) + \beta_{1x_{i1}} + ... + \beta_k x_{ik}$$

이 때 $h_0(t)$는 설명변수들의 값에 의해서 변화하는 시간 t 에서의 기저위험(baseline hazard)이다. 콕스 비례위험모형에서 회귀계수의 해석은 설명변수가 한 단위씩 증가함에 따른 종속변수의 위험도(hazard ratio) 백분율 변화량으로 해석된다. 다음의 <표 6>은 콕스비례위험 모형을 이용하여 가구원 특성에 따라 경활패널의 표본소실 위험도를 분석한 결과이다. 분석에 포함된 성, 연령, 교육정도, 가구주와의 관계, 경제활동 상태 변수 모두가 패널소실 위험에 통계적으로 유의미한 영향력을 행사하고 있음을 알 수 있다. 특히 연령에 의한 효과가 가장 두드러지는데, 연령이 증가할수록 패널소실 확률이 감소하고 있다. 예를 들어 30세 이하의 집단에 비해 30-44세 집단은 패널 차수가 한 단위씩 높아질 때마다 패널소실확률이 42.2%(1-0.578= 0.422) 감소하고 있다.

교육과 가구주와의 관계에서 각 항목집단별 패널소실위험은 앞의 생명표 분석 결과와는 차이가 있다. 생명표 분석에서는 생존확률이 초대졸자에서 가장 낮고, 대졸자에서는 높은 U자 형태를 보였다. 반면에 콕스비례위험모형에서는 연령효과가 통제되자 교육정도와 패널소실위험이 정(+)의 상관관계를 갖고 있다는 점이 드러났다. 초졸 이하의 학력소지자에 비해 대졸이상의 학력소지자의 패널소실위험은 패널차수가 한 단위 증가할 때마다 1.28배씩 증가하고 있다.

가구주와의 관계에 따라 패널소실위험을 살펴보면, 가구주로부터 관계가 멀수록 위험도가 높아지는 것을 알 수 있다. 생명표 분석결과는 배우자, 부모, 가구주의 순으로 생존확률이 높은 반면에 미혼자녀의 생존확률은 매우 낮았다. 그러나 콕스비례위험모형의 결과를 보면 다른 변수들을 통제할 때 가구주에 비해 미혼자녀의 패널소실위험은 77.1% 수준으로 가장 높았다. 부모의 경우 또한 패널소실위험이 가구주의 124%에 해당되었다. 동거인의 패널소실 위험은 가구주에 비해 3배가 넘는 것으로 나타났다. 경제활동상태에 따른 패널소실위험을 살펴보면 취업자가 다른 집단에 비해 위험률이 가장 낮았다. 취업자에 비해 실업자의 상대적 위험도가 가장 높았고, 그 다음은 비경활자였다.

<표 6> 가구원 특성에 따른 경제활동패널 생존기간 비례위험도

변수		β
성	(남)	
	여	.933**
연령	(15-)	
	30-44	.578**
	45-59	.443**
	60+	.434**
학력	(초졸 이하)	
	중졸	1.097**
	고졸	1.117**
	초대졸	1.177**
	대졸이상	1.282**
가구주관계	(가구주)	
	배우자	.889**
	미혼자녀	.771**
	형제자매	1.071**
	부모	1.249**
	조부모	1.500**
	기타 친인척	1.025
	동거인	3.301**
경제활동상태	(취업)	
	실업	1.134**
	비경활	1.063**
	제외자	1.058
χ^2		7894.5
d.f		18
n		72,953

* p<.05, ** p<.01

패널소실의 위험도를 설명하는데 있어서 가장 직접적인 요인은 가구차원보다는 개인차원의 변수들이다. 이는 다음의 <표 7>에서 보듯이 도시가계패널의 소실위험도를 설명하는데 있어 대부분의 가구차원 변수들이 가진 효과가 통계적으로 그다지 유의미하지 않은 점을 통해서도 알 수 있다. 그러나 주택소유여부는 패널소실의 위험을 설명하는데 있어서 상당히 큰 설명력이 있는 변수임을 알 수 있다. 주택을 소유한 가구의 패널소실위험이 가장 낮았고, 세들어 사는 가구는 상대적으로 위험이 높았다. 주택을 소유한 가구에 비해 전세로 거주하는 가구는 패널 차수가 한 단위씩 높아질 때마다 패널소실확률이 2.2배 만큼 증가하고 있다. 사글세의 경우 패널소실 확률이 2.8배가 넘었고, 기타 형태의 경우는 패널소실위험이 무려 4.8배에 달하고 있다. 가계구분을 보면 봉급자가계에 비해 노무자가계의 낮은 위험도만이 통계적으로 유의미했다.

<표 7> 가구특성에 따른 도시가계패널 생존기간 비례위험도

변수		β
거처종류	(단독주택)	
	아파트	1.002
	연립	.957
	다세대	.995
	비주거용	1.099
	기타	2.893
주택소유	(자가)	
	전세	2.205**
	월세	2.155**
	사글세	2.841**
	무상	1.512*
	사택	1.146*
	기타	4.888**
주된 소득	(근로)	
	사업	.951
	연금	.880
	재산	1.209
	수증/보조	1.274
	기타	.989
가계구분	(봉급자)	
	노무자	.891*
	근로자외	.971
	무직	.840
χ^2		594.5
d.f		19
n		4,108

* p<.05, ** p<.01

V. 결론 및 제언

최근 국내에서도 사회적인 현상, 개인이나 집단의 특성이 시간에 따라 어떻게 변화하는지를 측정하고자 하는 연구들을 종종 찾아 볼 수 있다. 패널조사는 적어도 두 시점 이상의 기간 동안 동일한 대상을 상대로 같은 변수나 항목이 조사되는 것을 말한다. 이렇게 패널자료에 대한 수요가 증가하는 이유는 기존의 횡단조사가 특정 현상의 존재 여부나 변수들 간의 현재적인 관계만을 파악하는 것과는 달리, 한 요인이 특정 현상에 미친 인과적인 효과를 직접적으로 평가해 볼 수 있기 때문이다(Ruspini, 2002).

미국에서 1960년대에 처음 시작된 패널조사는 종단자료 분석기법의 발달과

함께 현재 20여 개국에서 다양한 목적과 방법으로 실시되고 있다. 패널조사는 대학이나 연구기관이 주관하는 경우들도 많지만, 국가통계기관에서 패널조사를 실시하는 경우도 있다. 가구소득조사를 위한 캐나다의 Survey of Labor and Income Dynamics, 프랑스 Lorraine 지역 거주자들의 특성을 연구하는 French Household Panel, 개인과 가구의 사회경제적인 특성을 조사하는 Dutch Socio Ecnomic Panel 등은 국가통계기관이 주관하는 대표적인 패널조사들이다.

패널자료가 가진 폭넓은 분석 가능성에도 불구하고, 동일대상을 지속적으로 추적할 때 발생하는 비용과 표본설계 및 조사과정의 복잡성은 패널조사를 개발하는데 가장 큰 장애요인으로 작용하고 있다. 그러나 반복 횡단조사로 설계된 자료를 가구 및 개인의 ID를 이용해 상이한 시점들을 연계해서 사용할 수 있다면 별도의 비용 없이도 패널자료를 구할 수 있을 것이다. 경활과 도시가계의 경우 인구주택총조사에 의해 표본프레임이 확정된 후 특정가구가 표본으로 추출되면 다음번 표본개편까지는 동일한 대상을 상대로 반복적으로 조사가 실시된다. 따라서 조사결과를 최장 5년, 2003년 연동표본이 도입된 이후에는 최장 3년간의 패널자료로 구축해 볼 수 있다.

본 연구에서는 1998년에서 2002년까지 경활자료를 이용하여 60회 차의 월별 패널, 도시가계자료는 20회 차의 분기별 패널로 구축한 후, 두 패널자료가 가진 표본의 대표성을 분석해 보았다. 경활패널의 최종 표본소실율은 53.5%, 도시가계패널은 63.3%였다. 연도별 패널소실율 추이는 두 조사 모두 비슷한데, 첫해의 소실이 가장 많이 발생하고, 점차 감소하다가 3년 이후부터는 안정되는 것을 알 수 있다.

경활과 도시가계 패널의 생존기간을 살펴보면, 패널의 소실에 특정한 유형이 있다는 것을 알 수 있다. 경활의 경우 여자보다는 남자가 패널소실위험율이 높았고, 장년층보다는 젊은층이 패널소실 위험율이 높았다. 학력이 높을수록 패널소실확률도 함께 증가했으며, 취업자보다는 실업자의 패널소실확률이 높게 나타났다. 도시가계의 경우 세들어 사는 가구의 패널소실위험율이 주택을 소유한 가구에 비해 높게 나타났고, 노무자가구보다 봉급자가구가 패널소실위험이 더 높았다.

개인차원에서 패널소실기간의 차이를 설명하는 가장 효과적인 변수는 연령이었던 반면에, 가구 차원에서는 주택소유형태였다. 이러한 결과는 Clarke and Tate(1999)의 연구와도 일치한다. 이들은 횡단조사로 설계된 영국의 Labour Force Survey 자료를 이용해서 최장 12개월간의 패널자료를 구축한 후 표본소실의 발생하고 있는 유형을 분석한 결과 연령과 주택소유형태가 가장 중요한 변

수였다고 주장했다. 정리하자면 패널소실은 지리적 이동성이 높은 젊은 연령층과 주거상태가 상대적으로 불안정한 가구에서 주로 발생한다는 것이다. 또한 응답자 부담도 하나의 변수로 작용해서 경활과 도시가계를 모두 응답하는 가구의 경우 패널생존확률이 낮았다.

이상의 연구결과와 5년간의 패널차수(경활 60회, 도시가계 20회)를 고려해 볼 때 전체적인 표본소실의 수준 자체는 표본의 대표성을 위협할 정도로 심각한 것은 아니라고 판단된다. 그 이유는 본 연구에서 차수 무응답까지도 패널소실에 포함시키는 엄격한 기준을 적용했음에도 경활에서는 최초패널의 과반수 가까이가 제60회 차까지, 도시가계패널에서는 최초패널의 1/3이상이 제20회 차 조사까지 생존했기 때문이다. 문제는 경활과 도시가계자료에서 패널의 소실이 무작위로 발생하고 있는 것은 아니라는 데 있다. 경활에서 제1회 차 가구와 총 60차까지 생존한 가구의 가구주 직업을 비교해 보면 거의 대부분의 직업에서 최종표본 구성비가 감소한 반면, 농임어업종사자와 단순무직 근로자만이 그 구성비가 크게 증가했다. 이것은 경활처럼 횡단자료로 설계된 자료를 사후에 종단적으로 활용할 때 발생하는 문제점으로, 실제 자료가 축적되는 동안 패널의 소실을 고려한 적극적인 표본교체가 불가능했기 때문에 갖게 되는 한계점이다.

경활과 도시가계자료를 패널자료로 이용하는데 있어서 다음의 몇 가지 사항을 주의해야 할 것으로 보인다. 첫 번째는 경활과 도시가계가 일종의 거처패널이기 때문에 봄과 가을의 이사철에 무응답 증가로 인해 패널소실에 계절성이 나타난다는 점이다. 하지만 사후적으로 패널로 구축된 경활과 도시가계자료는 이러한 계절성을 조절할 수가 없기 때문에, 자료 분석에 어떠한 영향을 미치게 될지는 별도의 연구가 필요한 사안이다.

두 번째는 가중치 문제이다. 횡단조사로 설계된 경활과 도시가계조사에서는 무응답이나 사후 층화로 인해서 표본의 추출확률이 불균등해지는 것을 보정하기 위해서 사후에 가중치를 조정하고 있다. 두 조사를 패널로 구축하게 되면 가중치 조정문제가 더욱 복잡해진다. 예를 들어 표본소실이 위의 분석결과와 같이 연령과 주택소유 유형에 따라서 상이하게 발생하고 있다고 하자. 연령은 기존의 가중치를 부여할 때 성과 지역과 함께 통제변수로 사용되고 있던 변수이다. 그러나 주택소유유형은 통제변수에 포함되지 않았다. 가중치를 부여하여 조정하기 위해서는 연령과 주택소유 유형분포에 관한 모수값을 알아야만 한다. 제1차 조사가 실시된 시점의 주택소유 유형분포는 알 수 있겠지만, 모수값은 시간에 따라 변동하므로 제60차 조사시점의 모수값은 알 수 없다. 이 때 제1차 조사에서 사용된 가중치를 제60차 조사까지 계속 사용할 경우 기존의 횡단설계의 가중치

체계를 통해 발표된 결과치와 패널자료를 통해 분석된 결과치가 다를 수 있다. 따라서 패널자료 구축시 어떠한 방식으로 무응답에 대한 가중치를 조정해야 공표했던 결과치와 유사한 결과를 도출할 수 있을지는 별도의 심층적인 연구를 필요로 하는 사안이다.

세 번째는 무응답을 처리하는 일관된 기준과 이에 대한 상세한 설명이 필요할 것으로 보인다. 예를 들어, 도시가계에서는 단위무응답으로 인해 발생하는 대표성의 문제를 완화하기 위해 응답가구의 값을 복제하여 일정규모의 표본규모를 유지하는 방식을 사용해 왔다. 동일한 표본주기 내에서도 복제방식이 약간씩 달라졌다. 예를 들어, 1998년과 1999년에는 응답한 가구의 항목별 응답값 뿐만 아니라 가구일련번호까지 모두 무응답가구 자료에 복제했기 때문에 가구일련번호가 중복되는 사례들이 있었다. 이러한 문제점을 보완하기 위하여 2000년부터는 응답가구의 값만 무응답 가구자료에 복제하고 무응답 가구가 가진 원래 일련번호는 유지시켰다. 따라서 도시가계자료를 제공할 때는 이러한 기준변화에 대한 설명과 함께 원래 값과 복제된 값을 데이터 상에 표기되어야 할 것이다.

이상으로 경활과 도시가계자료를 패널로 구축하고, 분석하는 과정에서 해결해야만 할 과제들을 살펴보았다. 그러나, 패널조사 자료를 통해 얻을 수 있는 풍부한 분석의 가능성과 정책이나 프로그램의 효과를 객관적으로 측정할 수 있는 수단으로서 패널자료에 대한 수요는 앞으로 점점 더 확대될 전망이다. 또한, 기존의 자료에 대한 새로운 이용수요를 창출한다는 면에서도 패널자료 구축에 대한 연구는 지속적으로 수행되어져야 할 것으로 보인다.

참고문헌

김대일 · 남재량 · 류근관 (2000) "한국노동패널 표본의 대표성과 패널조사 표본이탈자의 특성연구" 《노동경제논집》 23:1-33.

남재량 (1997) "우리나라 실업률 추세변화에 관한 연구" 서울대학교 경제학박사 학위논문.

남재량 · 류근관 (1999) "우리나라 여성 노동력 상태의 동태적 특성연구" 《한국사회과학》 21:115-159.

남재량. · 이창용 (2001) "외환위기와 실업률 변화에 대한 연구" 경제학 공동학술대회 발표논문.

대우경제연구소 (1999) 《대우패널데이타》 대우경제연구소.

류재우 · 배무기 (1984) "한국의 노동시장 플로우와 실업" 《노동경제논집》 10:55-75.

이병희 · 정재호 (2002) "경제위기 이후의 빈곤구조 분석: 반복빈곤 및 고용과의 관계를 중심으로" 《동향과 전망》 52:128-150.

한국노동연구원 (2005) 《6차년도 패널 보고서: 한국 가구와 개인의 경제활동》 한국노동연구원.

황덕순 (2002) "빈곤에 대한 동태적 분석" 《소득불평등 및 빈곤의 실태와 정책과제》 한국노동연구원.

Clarke, Paul and Pam Tate (1999) "Methodological Issues in the Production and Analysis of Longitudinal Data from the Labour Force Survey" *GSS Methodology* Series No. 17. Office for National Statistics. U.K..

Davies, Richard B. (1994) "From Cross-Sectional to Longitudinal Analysis" Angela Dale and Richard B. Davies(eds) *Analysing Social and Political Change* London: Sage.

Diggle, Peter J., Patrick J. Heagerty, Kung-Yee Liang, and Scott Zeger (2002) *Analysis of Longitudinal Data* NY: Oxford University Press.

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

Gong, Greg X (2004) "Planning and Design of Household Panel Surveys for Enhancing Poverty Statistics" Paper presented at the UNESCAP Subcommittee on Statistics, 1st Session, Bangkok. Thailand.

Kuzmicich, Geoff and Wigbout Max (2001) "A Longitudinal Look at Some Data of the Household Labour Force Survey" *Research Report* 16. Statistics New Zealand.

Laux, Richard and Emma Tonks (1998) "Longitudinal Data from the Labour Force Survey." *Labour Market Trends* 106:175-188.

Menard, Scott (2002) *Longitudinal Research* (2nd ed.) London: Sage.

Nathan, Gad (1998) "A Review of Sample Attrition and Representativeness in Three Longitudinal Surveys" GSS Methodology Series No.13. Office for National Statistics, U.K..

Neumark, David and Daiji Kawaguchi (2004) "Attrition Bias in Labor Economics Research Using Matched CPS Files" *Journal of Economic and Social Measuremen* 29:445-472.

Rowe, Geoff and Huan Nguyen (2004) "Longitudinal Analysis of Labour Force Survey Data" *Survey Methodology* 30:105-114.

Ruspini, Elisabetta (2002) *Introduction to Longitudinal Research* London: Routledge.

< 부표 1> 경제활동의 노동력 상태별 유동율

패널기간	성별	취업유동률	실업유동률	비경활유동률
1985-1989	남성	0.056	0.681	0.097
	여성	0.136	0.844	0.090
1993-1997	남성	0.034	0.586	0.075
	여성	0.080	0.727	0.062

자료: 남재량 · 류근관 (1999). "우리나라 여성 노동력 상태의 동태적 특성연구" 《한국사회과학》.

<부표 2> 한국노동패널(KLIPS)의 표본소실율 추이

	패널차수(년)					
	1차(98)	2차(99)	3차(00)	4차(01)	5차(02)	6차(03)
가구패널						
표본규모	5,000	4,508	4,266	4,248	4,298	4,592
소실율(%)		12.4	11.4	11.0	7.8	7.5
생존율(%)						61.7
개인패널						
표본규모	13,107	12,537	12,186	12,678	13,264	14,961
소실율(%)		16.5	14.3	12.5	8.9	7.6

자료: 김대일 · 남재량 · 류근관 (2000) "한국노동패널 표본의 대표성과 패널조사 표본이탈자의 특성연구" 《노동경제논집》.

<부표 3> 미국 Panel Study of Income Dynamics의 패널 소실율과 유형, 1968-72, 1987-89

		조사년도							
		1968	1969	1970	1971	1972	1987	1988	1989
가구생존율(%)		100.0	88.1	85.0	82.9	80.8	52.2	50.6	49.0
소실 사유	무응답	-	9.9	2.2	1.3	1.3	2.2	1.9	2.3
	사망	-	0.5	0.5	0.6	0.8	0.1	0.1	0.9
	이주	-	1.6	1.0	0.7	0.8	0.5	0.4	0.3

자료: Fitzgerald, John, Peter Gottschalk, and Robert Moffitt (1998) "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics" *Journal of Human Resources*.

A Longitudinal Look at Economically Active Population Survey and Household Income and Expenditure Survey: Potential and Limitation

Ji-Youn Lee · Jin Kim

This study attempts to create a longitudinal dataset by linking tdata on the identical individuals across the monthly sample household management lists of the Economically Active Population Survey(EAPS) and the Household Income and Expenditure Survey(HIES). Using the data constructed through such process, the study also tryies to analyze the duration of longitudinal responses and the characteristics of nonrespondents. Between 1998 and 2002, longitudinal response rates had declined to 46% of total EAPS and 34% of total HIES. The fact that nonresponse was not a random phenomenon leads to concerns about the representativeness of the remaining sample. Using Cox's proportional hazard model the study revealed that the duration of longitudinal responses is affected by the ownership of house and the age of the respondent.

Key Words: panel data, panel attrition, Cox's proportional hazard model, the economically active population survey, the household income and expenditure survey