

## 제5장 가계금융조사를 위한 가구패널 설계 연구

정미옥 · 백지선

### 제1절 서론

#### 1. 연구배경 및 목적

가계금융조사는 통계청의 가계자산조사, 금융감독원의 가계신용조사, 한국은행의 가구패널조사가 하나로 통합되면서 횡단 및 종단분석이 가능한 패널조사로 2010년부터 시작되었다. 가계금융조사의 목적은 가계부문의 미시적 재무 건전성을 주기적으로 파악하여 정부의 부채 및 금리 정책, 가계신용관리, 금융시스템 발전을 위한 기본정보를 제공하는 것이다. 또한 가구특성별 자산과 부채의 규모·구성 및 분포를 조사하여 각종 경제, 사회, 금융정책이나 학문연구를 위한 기초자료로 제공되고 있다.

가계금융조사는 통계청에서 실시하는 최초의 패널조사로 기초연구를 위해 우선적으로 종단조사(Longitudinal Survey) 방법론에 대한 선행 연구결과와 국내외 가계의 금융과 관련된 패널조사 사례의 검토가 필요하다. 가계금융조사와 같은 종단조사를 위한 설계 방법에는 과거의 고정패널방법 외에 고정패널의 단점을 극복하기 위해 변형된 여러 가지의 설계방법이 있다. 향후 패널자료를 지속적으로 구축해 나가기 위해서 가계금융조사의 실시 초기인 현 시점에 적절한 패널 설계 방법을 결정해야 할 필요가 있다.

이에 본 연구에서는 선행연구들을 통하여 패널조사 설계방법들의 특징을 파악하고 장·단점을 비교하고자 한다. 또한 가계금융조사와 유사한 목적으로 가계의 소득, 자산, 지출 등을 패널조사 형태로 진행하고 있는 해외 각국의 사례를 검토하여 시사점을 찾는다. 사례 검토 시에는 각 조사에 대한 조사개요와 패널형태, 표본설계, 패널유지 현황, 횡단 및 종단 가중치 작성과정 등을 위주로 살펴보았다. 이러한 사례연구를 바탕으로 가계금융조사가 국가통계로서의 정확도와 신뢰성을 확보하면서도 패널을 효율적으

로 운영·관리할 수 있도록 패널설계의 형태를 제안하고자 한다.

## 2. 패널조사 설계 개요

시간의 경과에 따른 측정을 위한 조사 설계방법으로 Duncan(1987)은 반복조사(Repeated Survey), 고정패널조사(Fixed Panel Survey), 연동패널조사(Rotating Panel Survey), 분할패널조사(Split Panel Survey)와 같이 크게 네 가지로 구분하였다. 반복조사는 동일한 모집단으로부터 선택된 표본에서 유사한 내용을 서로 다른 시점에 조사하는 방법이다. 반복조사에서는 어떠한 개체(표본)가 연속적인 자료수집에 한번 이상 포함되어 있음을 보장하지는 않는다. 반면 고정패널조사는 동일한 표본에서 유사한 내용을 서로 다른 시점에 조사하는 방법이다. 고정패널조사의 경우 패널이 지속됨에 따라 패널손실과 노후화의 문제가 발생한다. 이와 같이 원칙적으로 조사기간 동안 표본 개체들을 패널로 유지하는 고정패널조사와는 달리 연동패널조사는 표본 개체들의 패널 수명을 제한하는 방법이다. 연동패널조사에서는 패널 지속기간이 지나면 해당 패널을 탈락시키고, 새로운 개체를 추가하는 방식으로 운영한다. 이 방법은 일정기간마다 패널을 교체함으로써 모집단 대표성을 확보할 수 있기 때문에 횡단면 모수 추정에 주로 사용된다. 분할패널조사는 고정패널조사에 반복조사 또는 연동패널조사를 결합한 조사방법으로 각 설계형태에 따른 장·단점을 절충하여 설계할 수 있다.

시간경과에 따른 무엇을 측정하고자 하는지에 따라 설계형태는 달라져야 한다. 조사 목적에 맞는 설계형태를 결정하기에 앞서 시간경과에 따른 측정 시 일반적으로 관심을 갖는 조사목적들 7가지 정도로 다음과 같이 정리할 수 있다. ㉠ 월별 실업률과 같은 특정 시점에서의 모수의 추정을 제공한다. ㉡ 일 년 동안의 격주 질병율과 같이 일정기간 동안의 평균적인 모수의 추정을 제공한다. ㉢ 두 달간 실업률 변동과 같이 종합적인 수준에서의 순 변동을 측정한다. ㉣ 한 개인의 평균 변동이나 전이테이블과 같이 개인 수준에서 변동의 다양한 요인들을 측정한다. ㉤ 분기별 근로소득을 누적하여 연간 근로소득을 작성하는 것처럼 시간경과에 따른 개인들에 대한 자료 구축한다. ㉥ 질병 평균 지속시간과 같이 빈도, 시기, 주어진 기간 동안 사건 발생을 측정한다. ㉦ 희귀질병 환자 자료와 같은 희귀 모집단에 대해서 시간경과에 따른 표본을 축적하고자 한다.

<표 5-1>에 요약된 바와 같이 반복조사의 경우 ㉠, ㉡, ㉤의 목적에 적합한 설계방법이지만, ㉢, ㉣의 목적에 대해서는 조건부로 적합하며, ㉣, ㉥를 위해서는 부적합한 방법이다. 고정패널조사는 ㉣, ㉥의 목적에 대해서는 적합하고, ㉢, ㉣는 반복조사에 비해 적합한 편이며, ㉤는 조건부 적합, ㉠, ㉡는 부적합한 측면이 있지만 다른 방법을 통하면 부분적으로 극복 가능하다. 연동패널조사는 ㉠, ㉡, ㉢의 목적에는 적합하고, ㉣, ㉥에 대해서도 제한적으로 적합하며, ㉣는 반복조사에 비해 적합한 편이고, ㉤는 조건부로 적합하다. 분할패널조사는 고정패널조사에 반복이나 연동을 적절히 조합하게 된다면



〈표 5-1〉 조사 목적에 대한 조사 설계의 특성

조사 목적	반복조사 (Repeated Survey)	고정패널조사 (Fixed Panel Survey)	연동패널조사 (Rotating Panel Survey)	분할패널조사 (Split Panel Survey)
㉓특정 시점의 모수 추정	-자동적으로 모집단 변화가 고려됨	-모집단 변화를 고려하는 매커니즘 필요	-연동주기동안 모집단 변화를 고려하는 매커니즘 필요	-패널표본에 대해서는 모집단 변화를 고려하는 매커니즘 필요
	㉔모수의 평균 추정	-값의 변화와 모집단의 변화가 결합된 효과를 추정	-모집단 변화를 고려하는 매커니즘 필요 -웨이브들 간의 양의 상관관계에 의해 변동의 분산이 감소됨	-패널표본에 대해서는 모집단 변화를 고려하는 매커니즘 필요 -패널표본에 대해서는 웨이브들 간의 양의 상관관계에 의해 변동의 분산이 감소됨
㉕순 변동 추정			-추정방법이 사용될 수 있음	
㉖개인 변동의 총량 파악 그 외 요인 추정	-불가능	-적합	-연동주기보다 짧게는 가능 -몇몇 연동그룹들에 대해서만 사용이 가능	-패널표본에 대해서는 적합 -반복 조사 표본에 대해서는 불가능
	㉗시간경과에 따른 개인 자료 구축			
㉘지정된 기간 동안 발생한 사건에 대한 자료 수집	-과거 사건의 요약(telescoping)을 위한 수단은 아님 -장기간 과거 사건정보에 대해서는 회상오류(recall error)가 발생가능	-첫 번째 웨이브 이후의 사건들로 telescoping의 제한을 둘 수 있음 -장기간 과거 사건정보는 여러 웨이브들의 정보를 결합하여 구축할 수 있음	-각 연동그룹별로 첫 번째 웨이브 이후의 사건들로 telescoping의 제한을 둘 수 있음	-패널표본에 대해서는 회상에 제한을 둘 수 있지만 -반복조사 표본에 대해서는 그렇지 못함
	㉙시간경과에 따른 표본 축적	-사건들의 telescoping에 수단이 될 수 없음에도 불구하고 -정적인 특성들과 새로운 사건들에 대해서는 매우 훌륭한	-정적인 특성들에 대해 사용될 수 있음 -경계를 둔 새로운 사건들에 대해서 유용함	-반복조사에는 매우 훌륭한 -패널표본의 경우 새로운 사건들에 대해서는 유용하지만 정적인 특성들에 대해서는 유용하지 않음

출처 : Duncan and Kalton(1987)

7가지 목적에 대해 모두 적용될 수 있다.

각각의 설계형태는 자료수집과 비용, 응답자 부담, 분석과 추정, 모집단 대표성, 자료 품질 측면에서 다른 특성을 보인다. 먼저 자료수집 범위와 비용 측면에서 살펴보면, 반복조사는 동일한 표본이 한번 이상 포함되면서 발생하는 문제가 적고 덜 복잡하다. 반면 패널조사는 동일한 개체에 대해 웨이브마다 서로 다른 내용의 자료를 수집할 수 있기 때문에 각 표본 개체에 대하여 더 광범위하고 많은 내용을 조사할 수 있다. 또한 최초조사를 면접조사로 수행한 다음 재조사는 전화조사로 수행하는 경우처럼 만일 재조사를 위한 비용이 최초조사보다 저렴하다면 패널조사가 반복조사에 비해 비용이 저렴하다. 그리고 연동표본조사의 경우 중간에 패널의 교체가 있기 때문에 패널을 조사기간 내내 추적해야하는 고정패널조사에 비해 추적조사로 인한 큰 비용을 줄일 수 있다. 그러나 연동표본조사와 분할패널조사는 연동주기와 중복형태에 따라 표본설계 작업 수행이 빈번해지면서 이로 인해 발생하는 비용이 추가될 수 있다.

응답자 부담측면에서는 반복조사가 패널조사에 비해 응답자의 부담이 적다. 고정패널조사의 경우 패널의 교체나 보충이 없기 때문에 응답자 부담으로 인한 응답거부나 탈락이 조사지속 여부를 결정하는데 치명적으로 작용할 수 있다. 연동패널조사는 연동 지속기간이 정해져 있기 때문에 고정패널조사에 비해서는 응답자 부담이 적다고 할 수 있다.

분석 및 추정에 대해서 패널조사는 개인 변동, 순 변동, 행동모델 등의 분석이 가능하기 때문에 반복조사에 비해 분석에 대한 매우 높은 잠재력이 있다. 연동패널조사는 연동패널의 수명 기간에 대해서만 종단분석이 가능한 반면 분할패널조사는 고정패널 부분을 이용하면 장기간 종단분석도 가능하다. 또한 연동패널조사는 추정방법과 연동체계를 적절히 적용하면 변동의 추정과 횡단 모수 추정의 정확도를 향상시킬 수 있기 때문에 반복조사보다는 특정 시점과 모수의 변화에 대해 일반적으로 좀 더 정확한 추정값을 제공한다.

반복조사는 매조사마다 시간경과에 의한 모집단의 새로운 개체를 반영함으로써 모집단의 대표성이 확보 가능하다. 그러나 패널조사는 무응답으로 인한 패널손실이 존재하며 시간이 지나면서 모집단이 노후화되어 대표성 확보를 위한 노력이 지속적으로 필요하다. 연동패널조사는 패널교체를 통해 모집단의 새로운 개체 반영이 가능하고, 고정패널조사에 비해 패널손실의 문제도 줄일 수 있다. 분할패널조사는 반복 또는 연동조사가 추가된 형태이므로 모집단 변동에 대한 최신의 표본을 유지 가능하다.

자료품질에 대한 설계형태별 잠재적 효과는 <표 5-2>와 같다. 반복조사는 패널의 조건화 효과(conditioning effect)가 나타나지 않지만, 패널조사에서는 발생가능하다. 여기서 조건화 효과란 한 웨이브에서의 응답이 이전 웨이브에서의 참여에 의해 영향을 받게



〈표 5-2〉 자료 품질에 대한 조사 설계의 잠재적 효과

잠재적 문제점 혹은 이점	반복조사 (Repeated Survey)	고정패널조사 (Fixed Panel Survey)	연동패널조사 (Rotating Panel Survey)	분할패널조사 (Split Panel Survey)
최초 웨이브 무응답 편향 발생	-가능	-가능	-가능	-가능
이후 웨이브 무응답 편향 발생	-재조사가 없기 때문에 불가능	-가능, 그러나 무응답 조정을 위해 이전 웨이브 응답 이용 가능	-패널표본에 대해서는 가능, 그러나 무응답 조정을 위해 이전 웨이브 응답 이용 가능	-패널표본에 대해서는 가능, 그러나 무응답 조정을 위해 이전 웨이브 응답 이용 가능
무응답 손실로 인한 후속 웨이브 표본크기의 부적당함	-재조사가 없기 때문에 불가능	-가능	-각 연동그룹에 대해 최초 웨이브 이후에 가능	-패널표본에 대해 최초 웨이브 이후에 가능
이후 웨이브의 조건화 편향 발생	-재조사가 없기 때문에 불가능	-가능	-각 연동그룹에 대해 최초 웨이브 이후에 가능	-패널표본에 대해서만 최초 웨이브 이후에 가능
이전 조사 결과의 이용	-재조사가 없기 때문에 불가능	-최초 웨이브 이후에 가능	-각 연동 그룹에 대해 최초 웨이브 이후에 가능	-패널표본에 대해서만 최초 웨이브 이후에 가능
이후 웨이브에서의 응답 동기부여의 변화	-재조사가 없기 때문에 불가능	-최초 웨이브 이후에 가능	-각 연동 그룹에 대해 최초 웨이브 이후에 가능	-패널표본에 대해서만 최초 웨이브 이후에 가능

출처 : Duncan and Kalton(1987)

되면서 나타나게 된다. 즉, 이전 웨이브에서의 경험이 동일한 주제에 대한 다음 번 인터뷰에서의 응답에 영향을 미치게 되어 그들의 응답이 최초로 조사에 참여한 사람과의 응답과 체계적으로 다르게 나타남을 의미한다. 연동패널조사가 고정패널조사에 비해 패널 조건화 효과를 줄일 수 있으며, 분할패널조사에서는 고정패널조사의 패널 조건화 효과와 패널손실로부터 발생하는 편향을 확인할 수 있다.

## 제2절 고정패널조사 사례

### 1. 미국 PSID(Panel Study of Income Dynamics)

#### 가. 조사개요

소득과 빈곤의 역학 연구를 위해 미시간대학교(University of Michigan) 조사연구센터(SRC)에서는 PSID를 조사하고 있다. PSID는 최초의 패널조사로, 1968년 처음 조사를 시작한 이후로 1997년까지 매년 조사를 하였고 1999년부터는 2년마다 조사를 수행하고 있다. 패널조사가 시작될 당시 표본의 크기는 4,802가족(families)이었으며, 시간 경과에 따른 가족의 분가, 표본의 보충 등으로 인해 표본 수가 증가하여 2005년에는 8,041가족이 조사에 참여했다. 조사방법은 CATI(Computer Assisted Telephone Interviewing)이며, 조사내용과 응답에 소요되는 시간은 2005년 기준으로 <표 5-3>과 같다.

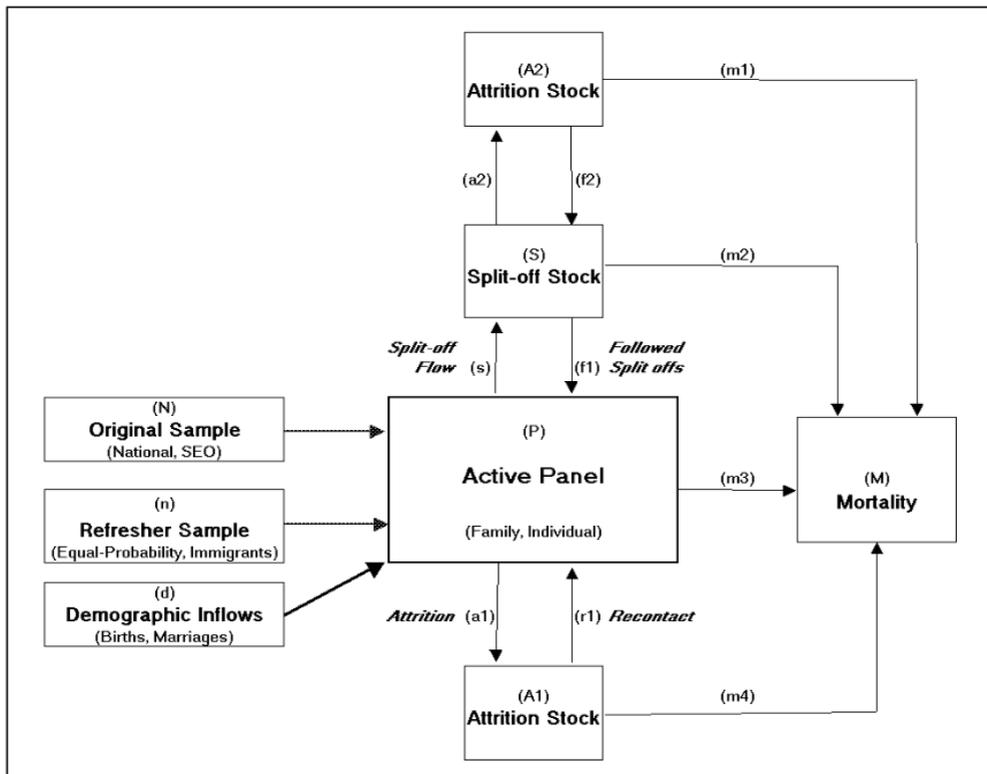
<표 5-3> PSID 조사항목 및 응답 소요시간(2005)

섹션	내용	응답시간(분)
A	주택, 유틸리티, 컴퓨터 사용	4.9
B, C, D, E	고용	20.2
F	집안일, 식품비용, 식량 지원	2.6
	교통, 교육비용, 소비	6.9
G	소득	6.6
H	일반건강	1.1
J	건강상태, 태도, 비용	10.7
K, L	결혼/출산	1.1
M	새 가장/배우자 배경	2.3
P	자선 및 자원봉사, 신앙, 원조	6.8
R	연금	3.5
W	복지	2.8
	자산, 저축	5.2
총 질문지 응답 소요시간		74.7
그 외 가족구성 정보 및 연락처 관련 질문		8.2
<b>총 응답 부담</b>		<b>82.9</b>

출처 : Katherine McGonagle and Robert Schoeni(2006)



PSID에서는 한 번 조사에 참여했던 응답자가 다음번 조사에 다시 응하는 비율이 96~98%로 매우 높은 재인터뷰율을 유지하고 있다. 이는 장기패널의 성공에 매우 결정적인 요소이기 때문에 PSID팀은 높은 응답률 유지를 위해 다각적인 노력을 하고 있다. 한 조사원이 매 웨이브마다 동일한 가족을 조사할 수 있도록 조사원을 확보하고, 조사 참여를 유도하기 위한 설득 편지를 보낸다. 다음번 자료 수집 시기와 내용을 응답자에게 지속적으로 고지하며, 감사의 인사와 패널조사의 연구적 가치에 대한 내용을 담은 연간 뉴스레터를 보내고 있다. 또한 2005년 기준으로 조사에 참여 시 응답자에게 \$60의 사례비를 지급하고, 응답자가 변경된 연락처 정보를 알려주면 \$10를 주는 등의 다양한 인센티브 전략을 세워놓고 있다. 이와 같은 PSID 고정패널조사의 강점은 동일 가구에서 반복적으로 측정함으로써 측정의 정도를 향상시킬 수 있다는 것과 긴 시계열 자료를 이용한 경제행위, 코호트 분석, 장기예측 등의 연구가 가능하다는 것이다. 반면 표본의 마모가 누적되면서 모집단 대표성을 상실하게 된다는 것이 매우 큰 한계점이라 할 수 있다.



출처 : Katherine McGonagle and Robert Schoeni(2006)

[그림 5-1] PSID 패널구조

## 나. 표본설계

1968년 PSID의 표본설계 당시 목표 표본 수는 약 5,000가족이었다. 이 중 약 2,000가족은 SEO(Survey of Economic Opportunity)의 표본으로부터 부표집(sub-sampling)하였는데, 저소득층을 과대표집(over-sampling)하였다. 나머지 약 3,000가족은 SRC에서 전국을 대표할 수 있도록 2단계 집락추출로 표본을 선정하였다. 이후 선정된 패널을 유지해오다가 1997년에 설계를 변경하였는데, 이때의 설계변경 특징은 SEO 표본의 축소와 1968년 이후 이민가구에 대한 보충 표본을 도입하는 것으로 설명할 수 있다. PSID는 고정패널조사로 패널의 구조는 [그림 5-1]과 같다.

## 다. 종단가중치 작성

종단가중치는 2단계의 과정을 통해 작성된다. 1단계에서는 조사에 한번이라도 참여했던 모든 개인에 대해 개인가중치가 작성된다. 2단계에서는 작성된 해당 해의 조사에 참여한 모든 가족 구성원의 개인가중치를 평균하여 가족가중치가 작성된다. 즉, 2단계의 가족가중치는 1단계에서 작성된 개인가중치를 기초로 구해진다.

개인가중치를 구하기 위해서는 총 3단계의 과정을 거친다. 첫 번째 단계에서 초기가중치 배정을 위해 이전 웨이브와 해당 웨이브에서의 상태에 따라 개인을 10개의 층으로 구분하고 층별로 초기가중치를 부여한다. 이때 종단가중치는 표본가구원에게만 부여하며, 비표본가구원에게는 종단가중치가 부여되지 않는다. 또한 이전 웨이브에서 무응답이었다가 다시 응답한 재접촉가구원에게는 이용 가능한 가장 최근의 가중치(reference weight)를 종단가중치로 부여한다. 표본가구원과 재접촉가구원에 대한 자세한 초기가중치 부여 규칙은 <표 5-4>와 같다.

두 번째 단계는 시간의 경과에 따라 발생한 패널 마모(attrition)에 대한 조정이다. 즉, 이전 웨이브에서 응답하고, 현 웨이브에서 ‘응답’ 또는 ‘현재 퇴장’인 그룹에 대해 가중치 조정을 한다. 이때 응답상태(응답/사망/무응답)와 생존여부(사망/생존)에 대한 조건부확률을 통계적 모형을 이용하여 추정하는데, 다항 로지스틱 모형을 사용하여 응답상태 확률을 추정한다.

세 번째 단계에서는 가중치를 표준화(scaling)한다. 이 단계는 개인과 가족가중치를 1968년 1차 웨이브의 초기가중치 총계에 맞춰 조정하는 단계이다. 이와 같은 1차 웨이브에 대한 벤치마킹(benchmarking)이외의 보정에 대해서는 자료를 이용하는 분석자의 몫으로 남기면서 PSID팀에서는 보정에 대한 방법론만을 제시해 주고 있다. 그 중 하나의 방법은 센서스(Census)나 CPS(Current Population Survey) 추정값을 벤치마킹하는 방법으로 수식으로 표현하면 다음과 같다.



$$W_{i,pop} = W_i \cdot \frac{N_{pop}}{\sum_{i=1}^n W_i}$$

여기서,  $W_i$ 는 PSID가중치이고,  $N_{pop}$ 은 목표모집단인 센서스나 CPS의 총계이다. 다른 하나의 방법은 가중치를 표본크기에 맞게 조정하는 방법으로 아래의 수식과 같다.

$$W_{i,norm} = W_i \cdot \frac{n}{\sum_{i=1}^n W_i}$$

여기서,  $n$ 은 분석에 포함된 표본가구나 개인의 수이며,  $\sum_{i=1}^n W_{i,norm} = n$ 이다.

<표 5-4> PSID 초기 가중치 부여 규칙

층		자료수집 여부		종단(초기)가중치
		$t_0$	$t$	
과거 퇴장(Past Exit)		no	no	Zero
미래 진입(Future Entry)		no	no	Zero
현재 진입 (Recent Entry)	표본(출생)	no	yes	가구주와 부인 가중치의 평균 (가구주 부인 없는 경우엔 가구주 가중치의 1/2)
	표본(진입)	no	yes	가구주와 부인 가중치의 평균 (가구주 부인 없는 경우엔 가구주 가중치의 1/2)
	비표본	no	yes	Zero
	새표본/ 표본등장	no	yes	a) 새로 추가된 표본 : 선택확률의 역수 b) 표본 등장 : 가족 구성원들 가중치의 평균
재진입/재접촉(Reentry/recontact)		no	yes	이용 가능한 가장 최근 가중치(Reference weight)
현재 퇴장 (Recent Exit)	$t_0$ 와 $t$ 사이 사망	yes	no	Zero
	$t_0$ 응답 & $t$ 무응답	yes	no	Zero
응답 (Response)	$t_0$ 와 $t$ 에 응답	yes	yes	두 종류의 가중치가 있음 i) 조정가중치 : $t_0$ 가중치*패널마모 조정 ii) 이월가중치 : $t_0$ 가중치

주)  $t_0$ : 이전 웨이브,  $t$ : 현 웨이브  
출처 : Elena Gouskova et al.(2008)

## 2. 영국 BHPS(British Household Panel Study)

### 가. 조사개요

영국의 에섹스대학교(University of Essex) 내 연구소(ISER)에서는 영국의 개인과 가구 수준에서의 사회 경제적 변화를 이해하고자 BHPS를 조사하고 있다. 1991년 조사를 시작한 이래로 약 5,000가구에 대해 매년 조사를 실시한다. 조사방법은 PAPI(Paper And Pencil Interviewing)로 시작하였으며 9차 웨이브부터는 CAPI(Computer Assisted Personal Interviewing)로 변경하였다. 전국을 대표하는 16세 이상 개인에 대해 가구구성, 노동력, 소득, 재산, 주택, 건강, 사회-경제적 변수들을 조사한다.

### 나. 표본설계

BHPS에서는 표집틀(sampling frame)인 우편번호 주소 파일(PAF)로부터 2단계 층화 집락계통추출법에 의해 표본을 선정하였다. 1단계에서는 1차추출단위(PSU: Primary Sampling Unit)인 우편번호 섹터를 250개 선정하였다. 우편번호 섹터란 최소 500가구가 포함되도록 구성된 일종의 우편번호 집락이다. 이와 같이 구성된 PSU에 대해 1981년 인구센서스 결과로부터 지역과 가구주의 사회경제적 그룹(직업이나 경영위치) 비율과 같은 사회인구학 변수를 사용하여 1차(major) 층화를 수행하였다. 다음으로 1차 층 내에서 연금수혜연령(여자는 60세, 남자는 65세) 비율에 의해 2차(minor) 층화를 하였다. 각 층이 구성된 후 비도시지역은 농업종사자 비율, 도시지역은 연금수혜연령 미만인 1인가구의 비율을 분류기준으로 설정하고, 이 분류기준에 의해 정렬 후 계통추출하였다.

2단계에서는 2차추출단위(SSU: Second Sampling Unit)인 주소를 8,167개 선정하였는데, PSU내에서 평균적으로 33개의 주소가 선정될 수 있도록 계통추출하였다. 선택된 주소 내에 3가구 이하의 가구가 거주하는 경우에는 모든 가구를 표본으로 선정하고, 주소 내 3가구를 초과하는 경우에는 3가구만을 랜덤하게 추출하였다.

패널형태는 고정패널로 원표본(original sample) 가구원으로부터 출생하거나 입양된 가구원, 원표본가구로 전입해온 가구원, 원표본가구가 전출하여 속하게 된 가구의 가구원을 모두 원표본으로 간주하였다. 또한 임시적으로 원표본 가구에 거주하게 된 사람의 일부도 원표본으로 간주하고 추적조사를 함으로써 장기 고정패널에 의한 표본마모를 일부를 보충할 수 있도록 하였다. 1차 웨이브의 응답자에 대해 표본마모 현황은 <표 5-5>와 같다. 1차 웨이브 응답자 중 2차 웨이브 부적격자는 1.4%였으며, 적격자 중 응답을 한 경우가 2차에서 87.7%, 10차에서는 70.04%였다. 또한 항목무응답의 경우에는 핫덱(Hot-Deck)과 회귀대체방법 이용하여 무응답 대체를 수행하였다.



〈표 5-5〉 BHPS 1차 웨이브 응답자에 대한 마모 현황

	% ineligible	% of eligible responding	% of eligible responding at all waves	Wave on wave response rate (interviewed all waves up to previous)
Wave 2	1.4	87.7	87.7	87.7
Wave 3	2.9	81.5	79.1	90.3
Wave 4	4.3	79.9	74.8	94.9
Wave 5	5.6	76.8	70.6	94.8
Wave 6	6.9	77.3	68.7	97.6
Wave 7	8.4	76.0	66.7	97.6
Wave 8	9.5	74.1	64.7	97.4
Wave 9	10.5	72.1	62.4	97.0
Wave 10	12.0	70.4	60.0	96.7
Wave 11	12.8	68.4	59.3	96.1
Wave 12	13.7	66.6	57.1	96.5
Wave 13	14.8	64.9	55.1	96.8

출처 : Nicholas Buck et al.(2006)

#### 다. 1차 웨이브 가중치 작성

1차 웨이브 가중치 작성을 위한 초기가중치는 설계가중치이다. 등확률추출법(EPSEM : Equal Probability SElection Mechanism)이 표본설계의 기본원칙이지만, BHPS 표본선정 과정 중 EPSEM에서 다소 벗어나 비등(unequal)확률 추출을 한 일부 지역에 대해서는 비등확률을 조정된 가중치를 초기가중치로 사용하였다.

두 번째 단계에서는 가구 무응답을 조정하였다. 무응답의 종류를 거처종류를 아는 경우와 모르는 경우와 같이 두 개로 구분하고, 각각의 경우에 대해 무응답층을 구성하여 층별로 가중치 조정을 하였다. 거처의 종류를 모르는 경우는 지역과 PSU의 사회경제적 층으로 무응답층을 구성하였으며, 거처의 종류를 아는 경우는 지역과, PSU의 사회경제적 층, 거처 종류(4개)로 무응답층을 구성하였다. 또한 응답한 가구 내에서 무응답한 개인에 대한 조정을 수행하였다. 이때는 로짓모형(logit model)을 이용하여 성향 가중치로 조정하였는데, 모형설정 시 지역, 임차유형, 부, 가구원수, 혼인상태, 취업상태, 성, 연령, 성과 연령의 교호작용을 변수로 이용하였다. 이와 같은 무응답 조정으로 인한 분산증가를 억제하기 위해 가중치 조정 시 최대 가중치를 1.75로 제한하였다.

세 번째 단계에서는 가중치 절단과 사후층화, 스케일링(scaling) 조정을 수행하였다. 가구 및 개인가중치 모두 2.5보다 큰 경우 가중치를 제한하는 가중치 절단을 수행하였다. 다음으로 센서스 자료를 이용하여 사후층화를 실시하였는데, 가구 가중치를 위해 임차유형, 가구원수, 자동차수를 층화변수로 이용하였다. 그리고 개인과 개인응답자(16세

이상) 가중치를 위해서는 임차유형, 가구원수, 자동차수, 성, 연령을 사용하였다. 마지막으로 가중치의 총합이 총 표본수가 되도록 아래 식과 같은 계수( $I$ )를 구하여 스케일링 조정을 하였다.

$$I = \frac{n \sum w_i^2}{(\sum W_i)^2}$$

이와 같은 과정을 통해 가구(AHHWGHT), 개인(AXEWGHT), 개인응답자(AXRWGHT)에 대한 세 종류의 가중치가 최종적으로 작성된다.

## 라. 2차 웨이브 이후 종단가중치 작성

2차 웨이브 이후 종단가중치의 경우 가구는 종단개체로 정의되지 않아서 종단가중치는 작성되지 않으며, 개인과 개인응답자에 대해서만 종단가중치가 산출된다.

먼저, 개인응답자 종단가중치(wLRWGHT) 작성과정을 살펴보면, 초기가중치는 이전 웨이브에서 사후층화 이전의 개인응답자 종단가중치가 된다. 초기가중치가 준비가 되면 1단계로 부적격(조사종료) 표본을 포함한 두 웨이브의 모든 응답자들에 대하여 패널 마모 가중치 조정을 한다. 그리고 2단계에서는 응답거절만을 고려한 무응답 가중치 조정을 한다. 이때 CHAID(Chi-squared Automatic Interaction Detection) 알고리즘에 의해 무응답층을 구성을 하게 되는데, 지난 웨이브부터의 이사여부, 성, 연령, 취업상태, 소득, 인종, 교육상태, 지역, 임차유형, 자동차수 등을 층 변수로 사용한다. 마지막으로 1차 웨이브의 특성과 1991년 모집단의 주변분포(성, 연령, 임차유형, 자동차수, 가구원수)가 같아질 수 있도록 사후층화 조정을 한다.

다음으로 개인 종단가중치(wLEWGHT)는 개인응답자 종단가중치(wLRWGHT)와 거의 동일한 방법으로 가중치 작성한다. 단, 가중치 작성을 위한 층을 가구와 가구주 특성 변수를 이용하여 구성한다.

## 마. 2차 웨이브 이후 횡단가중치 작성

2차 웨이브 이후의 횡단가중치는 가구(wXHWGHT), 개인(wXEWGHT), 개인응답자(wXRWGHT)에 대해 가중치가 작성된다. 먼저 사후층화 이후의 1차 웨이브 가중치를 초기가중치로 원표본에 부여한다. 다음으로 패널마모에 대한 조정을 하는데, 종단가중치 작성 시 이용한 방법과 유사하게 층을 구성하며, 가중치 작성을 위한 층은 가구와 가구주 특성변수를 이용한다. 개인응답자 가중치의 경우 모든 응답자에 대해 거부, 부적격에 의한 조정이 수행되며, 가구가중치의 경우 개인 횡단가중치를 총 가구 수에 의해 조정하



여 작성된다.

### 3. 재정패널조사

#### 가. 조사개요

재정패널조사의 조사목적은 국민생활의 기초가 되는 조세와 재정의 분배 및 재분배 효과 검증을 위한 기초 원시 자료의 구축이다. 2008년도에 시작한 패널조사로서 한국조세연구원이 주관하고, TNS Korea에서 조사를 수행한다. 조사주기는 1년이며, 전국 일반 가구에 거주하는 15세 이상 구성원(제주도 제외)을 조사대상으로 한다. 1차년도 표본크기는 5,010가구이며, 가구현황, 각종 과세현황, 환급금, 보조금, 가계지출, 소득, 자산 및 부채, 복지, 경제활동, 소득공제 등의 내용을 PAPI 방식으로 수집한다.

#### 나. 표본설계

표본설계는 2단계 집락추출방법이며, 2005 인구주택총조사 90% 조사구 자료에 2005년 이후의 신축 아파트 단지를 추가하여 표집틀로 사용하였다. PSU인 1,002개의 조사구를 선택하기 위해 시도, 동·읍면부 여부와 주택유형(아파트/그외)으로 조사구를 층화한 후 계통추출하였다. 계통추출을 위한 정렬변수로는 시군구 행정코드, 주택 전용면적을 이용하였다. SSU는 가구로 총 5,010가구가 선택되었는데, 조사구 당 5가구를 계통적으로 선정하였다. 이때 부적격과 조사불능(거절, 부재 등) 표본가구에 대해 다른 가구로 대체(substitution)해줄 수 있도록 하였다. 계통추출에 의해 표본으로 지정된 가구에 최소한 4회 이상의 재방문을 하도록 하였으며, 그래도 조사가 불가능한 경우에는 가구리스트를 기준으로 체계적 대체( $n \rightarrow n+1 \rightarrow n-1$ )를 하였다.

또한 고소득층과 저소득층을 각각 300가구(60개 조사구)씩 과대표집하였다. 고소득층 지역 과대표집을 위해 부동산 가격 등과 같은 외부 정보를 기초로 서울의 서초구, 송파구, 강남구와 성남시의 분당구를 고소득층 지역으로 선정하였다. 그런 다음 각 지역의 조사구를 주택 평균면적을 기준으로 정렬 후 상위 50%에 해당하는 조사구를 과대 추출하였다. 저소득층 지역은 보건복지부 기초수급자 통계의 정보를 기초로 읍면동별 기초수급 가구 비율이 15%이상인 14개 읍면동을 저소득층 지역으로 선정하였다. 그리고 나서 각 지역의 조사구를 주택 평균면적을 기준으로 정렬 후 하위 50%에 해당하는 조사구를 과대 추출하였다.

패널형태는 고정패널이지만, 2차 웨이브에서 흔히 발생하는 표본손실에 따른 패널축소 문제를 줄이고, 패널규모를 5,000가구 수준으로 유지하기 위해 2차 웨이브에서 무응

답이 발생한 가구를 대체해주었다. 즉, 2차 웨이브에서 탈락한 618가구를 620개의 대체 가구로 보충하였다. 재정패널조사의 패널 유지율은 <표 5-6>과 같다.

<표 5-6> 재정패널조사 패널 유지율

지역	2차(2009) 웨이브			3차(2010) 웨이브 <sup>1)</sup>		
	조사 성공	조사 실패 (탈락)	유지율(%)	조사 성공	조사 실패	유지율(%)
서울	871	204	81.02	968	98	90.81
부산	341	44	88.57	362	22	94.27
대구	256	29	89.82	276	10	96.50
인천	206	40	83.74	232	14	94.31
광주	171	14	92.43	175	11	94.09
대전	161	20	88.95	176	6	96.70
울산	127	13	90.71	134	10	93.06
경기	752	124	58.84	820	73	91.83
강원	165	10	94.29	167	8	95.43
충북	142	29	83.04	167	6	96.53
충남	189	16	92.20	193	10	95.07
전북	237	18	92.94	244	13	94.94
전남	248	7	97.25	248	10	96.12
경북	266	24	91.72	273	18	93.81
경남	264	26	91.03	285	10	96.61
계	4,396	618	87.67	4,720	319	93.67

1) 2차 웨이브 조사 성공 가구 기준  
출처 : 한국조세연구원(2010)

#### 다. 횡단가중치 작성

횡단가중치 작성을 위한 초기가중치는 설계가중치로,  $h$ 번째 층의  $i$ 번째 조사구의  $j$ 번째 가구에 대한 설계가중치는  $w_{hij} = \frac{N_h}{n_h} \times \frac{N_{hi}}{n_{hi}}$ 이다. 여기서,  $N_h$ 는  $h$ 층의 모집단 조사구 수,  $n_h$ 는  $h$ 층의 표본 응답 조사구 수,  $N_{hi}$ 는  $h$ 층  $i$ 조사구의 모집단 가구 수,  $n_{hi}$ 는  $h$ 층  $i$ 조사구의 표본 응답 가구 수이다.

이와 같이 초기가중치가 설정되면 다음으로 가구형태(원패널, 대체, 분가가구)에 따른 가중치를 산출하게 된다. 원패널가구에는 이전 웨이브의 최종가중치를 부여하고, 대체가구에게는 대응되는 탈락가구의 이전 웨이브 최종가중치를 부여한다. 분가가구에 대



해서는 분가의 사유에 따라 가중치가 다르게 부여되는데, 분가사유가 결혼인 경우에는 분가 이전 원 가구의 가중치의 1/2, 기타인 경우는 원 가구의 가중치, 분가사유가 결혼인 가구가 무응답한 경우에는 응답가구의 가중치를 배분해준다.

1차 웨이브의 경우엔 가구대체를 허용했기 때문에 무응답 조정이 없으며, 2차 웨이브에서도 마찬가지로 무응답 조정을 하지 않았다. 3차 웨이브에서는 무응답 조정을 수행하였는데, 원 패널가구, 대체가구, 1차년도 분가가구에 대해서는 무응답을 조정하였다. 그러나 2차년도 분가가구는 분가 사유에 따라 위의 방식대로 다르게 부여하였다.

마지막으로 종단가중치 산출시 사용한 방법과 동일한 레이킹(raking) 보정을 통해 최종적으로 횡단가중치가 작성된다.

## 라. 종단가중치 작성

종단가중치를 위한 초기가중치는 종단가중치 작성을 시작하고자 하는 연도의 최종 가중치( $w_{start}$ )가 된다. 만일, 2010년도에 과거 3개년도(2008년~2010년)에 대한 종단가중치를 작성하고자 한다면  $w_{start} = w_{2008}$  이 되며, 2010년도에 2개년도(2009년~2010년)의 종단가중치를 작성한다면  $w_{start} = w_{2009}$  이 된다.

다음 단계에서 로지스틱 회귀모형을 기반으로 한 무응답 조정을 하는데, 아래의 수식과 같이 표현할 수 있다.

$$w_t^{(1)} = w_{start} \times \hat{p}_t^{-1}$$

여기서,  $\hat{p}_t^{-1}$ 는 로지스틱 회귀모형을 통해 예측된 응답확률로, 로지스틱 회귀모형은 가구주성별, 가구원수, 가구주연령, 지역, 가구연간소득 변수를 이용하여 아래와 같이 구성하였다.

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1(\text{가구주성별}) + \beta_2(\text{가구원수}) \\ + \beta_3(\text{가구주연령}) + \beta_4(\text{지역}) + \beta_5(\text{가구연간소득})$$

마지막으로 벤치마킹 보정 단계에서는 해당년도의 가구추계 자료를 기준으로 지역(16개 시도), 가구주 성별(남/여), 연령(0~39세/40~49/50~59/60~), 가구원수(4인 미만/4인 이상)의 주변분포를 이용한 레이킹 비(raking ratio)기법을 적용하였다. 또한 이와 같은 레이킹 조정 이후 종단가중치의 극단값 발생을 제어하기 위해 가중치의 크기 제한하였다. 가중치 제한의 최대값을 2008년에는 9,000, 2009년에는 12,500으로 설정하였다.

## 제3절 연동패널조사

### 1. 미국 SIPP(Survey of Income and Program Participation)

#### 가. 조사개요

미국 센서스국(Census Bureau)에서는 개인과 가구의 소득, 프로그램 참여, 그리고 이에 대한 주요 결정요인에 대한 정확하고 포괄적인 정보를 제공하기 위해 SIPP조사를 수행하고 있다. 1984년 처음으로 조사가 시작한 이래로 1996년에는 중단면 추정값의 품질을 향상시키기 위해 설계를 변경하였다. 1996년 설계변경의 주요 내용으로는 초기표본의 수를 약 37,000가구로 증가시키고, 32개월(8차 웨이브)의 중복(overlapping) 패널 대신 4년(12차 웨이브)의 단일 패널로 변경하였으며, CAI(Computer-Assisted Interviewing)를 도입하고, 빈곤층 밀집지역에서 가구를 과대표집하였다. 표본의 크기는 약 4만 가구이며, 조사주기는 4개월이다. 이때 조사원의 업무 부담을 매월 균등하게 배분하기 위해서 전체 표본을 1/4씩 랜덤하게 선정해 4개의 연동그룹을 구성하여 4개월 동안 매월 한 개의 연동그룹을 조사할 수 있도록 하였다. 설문지는 매조사마다 질문하는 핵심(core)파트와 조사의 빈도와 주기는 일정하지 않지만 특정주제에 대한 깊이 있는 질문이 가능한 주제별 모듈(topical module)로 구성되어 있다. 핵심파트는 인구통계학적 특성, 노동력 참여, 프로그램 참여, 이전소득과 다양한 프로그램에서 받은 비화폐성 혜택을 포함한 근로소득과 불로소득, 자산 소유권, 사적의료보험 등의 항목으로 구성되어 있다. 주제별 모듈에서는 자산과 부채, 학교 입학, 결혼, 출산, 이주, 장애, 및 직업 기록 등을 조사한다.

#### 나. 표본설계

1996년 표본설계의 내용을 위주로 살펴보면, SIPP의 표본설계는 2단계 집락추출방법으로 1단계에서는 PSU인 지역을 선정한다. 총 322개의 PSU 중에 112개는 자기대표성이 있는(SR : Self-Representing) PSU이고, 자기대표성이 없는(NSR : non-SR) PSU가 210개이다. SR PSU는 모두 선정되며, NSR PSU는 4개의 센서스 지역(S/NE/MW/W)과 15개의 층화 변수에 의해 105개 층으로 구분되고, 각 층에서 1990년 센서스의 주택수에 비례하는 확률로 2개씩 추출된다. 이때 선형 알고리즘을 적용함으로써 1980년 센서스와 1990년 센서스로부터 선정된 PSU가 최대한 중복될 수 있도록 하였다. 1996년 표본설계에서는 이전의 설계에 비해서 PSU 수가 증가(1985-1991 : 230개, 1992-1993 : 284개)하였는데, 이로 인하여 PSU간 분산은 줄어들고, 추정값의 분산추정에 대한 자유도는 늘어나는 효과가 나타났다.



2단계에서는 주소로 집락을 구성하고 SSU인 주소 집락을 선정하였다. 그러나 집락 내 예상된 거처보다 많은 거처가 있을 때에는 집락 내에서 부표집(sub-sampling)을 실시하였다. 먼저 주소로 집락을 구성하기 위해 5개의 개별적인 추출틀을 이용하였다. 첫 번째는 Unit 추출틀(the address enumeration districts frame)로 건축허가가 발급되고 최소한 96%의 주소가 완전한 지역에 있는 센서스조사구 내에서 거처 주소가 있는 추출틀이며, 하나의 거처가 집락이 된다. 두 번째는 Area 추출틀(area EDs frame)인데, Unit 추출틀의 나머지 센서스조사구들로부터 건축허가가 발급되지 않았거나 4%이상의 주소가 누락된 지역의 주소 추출틀이다. 대부분 시골지역이 해당되며, 4개 거처가 예상되도록 집락을 구성하였다. 세 번째로 Group Quarters 추출틀(special places frame)은 기숙사나 호텔, 보호시설과 같이 센서스에서 주택으로 카운트되지 않은 거처의 주소 추출틀로서 하나의 거처에 상당하는 집락으로 구성되었다. 네 번째는 New Construction 추출틀로 지난 센서스조사 이래로 건축허가가 발급된 새로운 건축물에 대한 추출틀이며, 4개 거처 또는 8개 거처가 한 집락이 되도록 구성되어 있다. 다섯 번째는 Coverage Improvement 추출틀로 센서스에서는 누락되었지만 사후조사에서 발견된 거처의 주소(0.1%) 추출틀로서 집락은 구성되지 않는 추출틀이다.

SIPP는 저소득 가구의 표본크기를 증가시켜 빈곤 추정값의 분산을 줄이기 위해서 저소득 가구에 대한 과대표집을 실시하였다. 센서스 정보를 이용하여 각 PSU내의 거처를 빈곤층과 아닌 경우로 구분하고, 각 PSU에서 고소득층에 비해 저소득층을 1.6배 높게 선정될 수 있도록 선정하였다.

패널의 형태는 연동패널로 하나의 패널을 4년 동안 유지하다가 4년이 지나면 일시에 표본을 교체하는 단일 패널로 설계되어 있다. 이 단일 패널에서 표본을 1/4씩 랜덤하게 추출하여 동일한 크기가 되도록 4개의 연동그룹을 구성하였다. 각 연동그룹은 4개월마다 인터뷰에 참여하게 되고, 매월 하나의 연동그룹을 조사하는 방식으로 운영하고 있다.

가구 무응답은 A타입과 D타입으로 구분하여 관리하는데, 가구 무응답과 표본마모율은 <표 5-7>과 같다. 여기서 A타입은 가구의 주소는 찾았으나 응답은 얻지 못한 경우이며, D타입은 이주한 곳을 알 수 없거나 인터뷰 불가능한 주소로 간 경우에 해당한다. 또한 개인에 대한 무응답을 두 가지로 구분하고, 각 무응답에 대하여 다른 대처 방법을 적용하고 있다. 만약 대상기간의 일부동안 가구원이었던 사람이 인터뷰날짜에도 가구원이지만 응답하기를 거부하는 경우이거나 인터뷰가 가능하지 않아서 임시적으로 응답을 얻을 수 없는 경우에는 가구원 무응답으로 인정하고, 가중치조정 통해 무응답 조정을 한다. 또한 대상기간동안 가구원이었던 사람이 이주했거나 인터뷰 날짜에 더 이상 가구원이 아닌 경우에는 Z타입으로 규정하고, 무응답 대체(imputation)를 적용한다. 항목 무응답은 여러 단계의 무응답 대체 과정을 거치게 되는데, 자세한 내용은 Westat(2001)을 참

고하길 바란다.

<표 5-7> SIPP 가구 무응답과 표본마모율

Wave	1990 Panel			1991 Panel			1992 Panel			1993 Panel			1996 Panel		
	Type A	Type D	Loss												
1	7.3	—	7.3	8.4	—	8.4	9.3	—	9.3	8.9	—	8.9	8.4	—	8.4
2	10.9	1.5	12.6	12.3	1.5	13.9	12.8	1.7	14.6	12.4	1.7	14.2	13.1	1.3	14.5
3	11.5	2.6	14.4	13.1	2.7	16.1	13.1	2.8	16.4	12.9	2.9	16.2	15.6	1.9	17.8
4	12.5	3.4	16.5	13.6	3.6	17.7	13.8	3.6	18.0	13.9	3.8	18.2	17.6	3.1	20.9
5	13.6	4.6	18.8	14.5	4.2	19.3	14.9	4.7	20.3	14.9	4.7	20.2	20.4	3.8	24.6
6	14.1	5.3	20.2	14.4	5.1	20.3	15.3	5.4	21.6	15.9	5.5	22.2	22.2	4.4	27.4
7	14.3	5.9	21.1	14.7	5.6	21.0	16.0	5.9	23.0	17.2	6.2	24.3	23.8	4.8	29.9
8	14.4	5.9	21.3	14.5	5.9	21.4	16.9	6.7	24.7	17.5	6.9	25.5	24.2	5.4	31.3
9	—	—	—	—	—	—	17.7	7.3	26.2	18.2	7.5	26.9	25.0	5.6	32.8
10	—	—	—	—	—	—	17.5	7.6	26.6	—	—	—	26.1	6.0	34.0
11	—	—	—	—	—	—	—	—	—	—	—	—	25.5	6.2	35.1
12	—	—	—	—	—	—	—	—	—	—	—	—	—	6.2	35.5

출처 : U.S. Bureau of the Census(1998)

#### 다. 1차 웨이브 횡단가중치 작성

SIPP에서는 개인, 가구(household), 가족(family), 세대별 가족(subfamily)에 대한 횡단가중치를 <표 5-8>과 같이 작성하고 있다. 1차 웨이브 횡단가중치의 초기가중치(BW)는 선택확률의 역수이다. 표본으로 선정된 가구 내 가구원은 자동으로 선정되므로 가구가중치와 가구원가중치는 동일하다. 초기가중치가 준비되면 집락 내 부표집에 대한 조정을 한다. 예상보다 집락 내 더 많은 가구가 발견되어 부표집한 경우에 대한 중복조정인자(DCF: Duplication-Control Factor)를 계산하며, DCF는 1~4의 값을 갖는 정수이다. 다음으로 총 512개 조정그룹에 대하여 무응답 조정(NAF)을 수행한다. 조정변수는 PSU내 과대표집(빈곤) 층 여부, 센서스 지역, 인종(흑인/NO), 주택소유여부, 거주지 상태(MSA<sup>1)</sup>도시/MSA시골/NonMSA 센서스지역/그외), 가구원수(1/2/3/4명이상)를 이용하였다. 이때 조정그룹 내에 가구 수가 30보다 적거나 조정인자가 2를 초과할 경우엔 인접한 셀로 병합한다. 인접 셀은 1979년 빈곤율 추정값을 기초로 정의되며, 센서스국에서 병합순서를 지정해 준다. 마지막으로 모집단 총계에 대한 조정인 칼리브레이션(calibration) 조정(SSCA)을 수행한다. CPS의 성별, 연령, 인종, 라틴아메리카 계통, 가족관계, 가구형태에 대해 벤치마킹하고 있다. 이때 각각의 연동그룹은 모집단 총계의 1/4에 해당하므로, 연동그룹별

1) Metropolitan Statistical Area



로 개별 조정한다. 이를 위해 총 5단계 주요한 알고리즘을 적용하는데, 각 단계별로 비조정, 레이킹, 셀병합, 총계계산 등의 작업이 반복적으로 수행된다. 이러한 일련의 과정을 통해 최종적으로 횡단가중치(FNLWGT = SSCA\*BW\*DCF\*NAF)가 작성된다.

〈표 5-8〉 SIPP 가중치 종류

Variable Name	Description
<b>Core Wave Files</b>	
WPFINWGT (FNLWGT)	Reference month, final weight of person
WHFNWGT (HWGT)	Reference month, final weight of household
WFFINWGT (FWGT)	Reference month, final weight of family
WSFINWGT (SWGt)	Reference month, final weight of related subfamily
WPFINWGT (P5WGT) <sup>d</sup>	Interview (5th) month, final weight of person
WHFNWGT (H5WGT) <sup>d</sup>	Interview (5th) month, final weight of household
<b>Topical Module Files</b>	
WPFINWGT (FINALWGT)	Prior to 1996: interview month, final weight of person. 1996+: 4th reference month, final weight of person
<b>Full Panel Files<sup>e</sup></b>	
WPFINWGT (FNLWGT)_x	Calendar year x, final weight of people in the calendar year cohort
PNLWGT (Not kept for 1996 panel)	Final weight for people in full panel cohort

출처 : U.S. Bureau of the Census(1998)

## 라. 2차 웨이브 이후 횡단가중치 작성

2차 웨이브 횡단가중치의 초기가중치(IW)는 1차 웨이브에서 칼리브레이션 조정단계 이전의 가중치(BW\*DCF\*NAF)이다. 초기가중치가 준비되면 이주자에 대한 조정(MW)을 한다. 예를 들어, 가구원수가 총 2명인 원표본가구에 3번째 가구원이 이주해온 경우에 3명의 가구원이 기존 가구 초기가중치의 2/3만큼씩 가중치를 나눠 갖는 식의 조정을 하게 된다. 그리고 다음 단계에서는 무응답 조정(LWNIA)을 하는데 조정 셀은 총 109개로, 주택소유여부, 가구주특성(여성가구주/배우자없음/65세이상가구주/기타), 인종과 라틴아메리카계 여부, 교육정도, 가구원수, 빈곤상태, 소득형태, 금융자산형태 등의 조정 변수가 고려된다. 마지막으로 칼리브레이션 조정(SSCA)을 하는데, 1차 웨이브에서 횡단가중치를 작성할 때와 동일한 방식으로 수행되며, 참조 월만 달라진다. 이러한 과정을 통해 이주자 조정이 필요 없는 경우에는 초기가중치에 무응답조정과 칼리브레이션 조정만을 수행한 가중치(FNLWGT = IW\*LWNIA\*SSCA)가, 이주자 조정이 필요한 경우에는 이주자에 대한 조정까지 수행한 가중치(FNLWGT = MW\*LWNIA\*SSCA)가 최종적으로 작성된다.

### 마. 종단가중치 작성

SIPP의 종단가중치는 크게 두 가지 종류로 구분된다. 하나는 전체패널가중치(full panel weight)로 1차 웨이브에 표본이었으며 모든 월에 대해 자료가 수집된 패널들에 대한 가중치이다. 1차 웨이브 이후에 표본으로 들어온 사람들은 이 가중치가 작성되지 않는다. 또 다른 하나는 해당년가중치(calendar year weight)로 해당 년의 1월에 인터뷰를 하고 해당 년의 모든 월에 대해 자료가 수집된 사람들에게 대한 가중치이다.

종단가중치 작성을 위한 초기가중치는 full panel 가중치는 패널시작점, calendar year 가중치는 해당 해의 칼리브레이션 조정 전 횡단가중치( $BW*DCF*NAF$ )가 된다. 무응답 조정은 2차 웨이브 이후의 횡단가중치 작성 시 무응답 조정과 유사한 방법으로 149개 조정 셀에 대해 진행된다. 마지막 단계인 칼리브레이션 조정에서는 full panel 가중치는 해당 패널 1차 웨이브가 설계된 시점의 모집단을 기준으로, calendar year 가중치는 해당 년의 모집단을 기준으로 조정된다.

## 2. 캐나다 SLID(Survey of Labour and Income Dynamics)

### 가. 조사개요

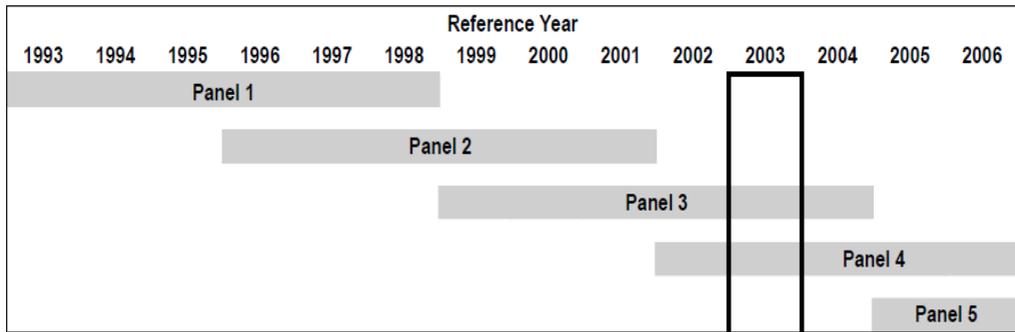
캐나다 통계청에서는 개인의 경제적 웰빙의 변화와 그 변화에 영향을 미치는 요인을 측정하기 위해 SLID를 조사하고 있다. 1993 시작한 이래로 매년 35,000가구에 대해 CATI방법으로 조사를 한다. 조사대상은 원가구에 속한 아이가 16세 이상이 되면 조사에 참여를 시키고, 70세 이상이 되면 노동 관련 질문에서는 제외시킨다. 또한 연 2회의 조사를 1월과 5월에 수행하고 있는데, 1월에는 노동, 교육, 가족관계 등에 대한 항목을 조사하며, 5월에는 소득에 대한 내용을 수집한다. 두 번의 조사에 참여해야 하는 응답자 부담을 고려하여 소득항목에 대해서는 직접 응답을 하거나 통계청에 소득세 자료 접근 권한을 주는 방식을 응답자가 선택할 수 있도록 하였다. 이에 따라 실제로 응답자가 직접 소득항목을 응답하는 경우는 약 20%에 불과하며, 80% 정도는 소득세 자료를 통해 자료가 수집되고 있다. 그러나 소득세 자료를 통한 자료수집의 비중이 커지면서 비용감소를 위해 2005년부터 1월과 5월 조사가 통합되었다.

### 나. 표본설계

SLID를 위한 표본은 독립적으로 설계되지 않으며 통계청의 노동력조사(LFS: Labour Force Survey)로부터 부표집된다. LFS는 1/6 연동표본제를 채택하고 있는 조사로 매월



표본교체가 이뤄진다. SLID는 이러한 LFS 표본가구 중에서 1월과 2월 교체되어 나가는 가구들 중 마지막 월에 응답한 가구를 표본으로 선정하고 있다. 이를 위해 LFS 마지막 조사에서 SLID에 대한 소개를 하고, 연락처를 수집하는 등의 사전 인터뷰를 수행하면서 표본가구에 대한 기본정보를 수집한다. 패널형태는 연동패널조사로 패널의 지속기간은 6년이다. [그림 5-2]에서 볼 수 있듯이, 전체 표본은 크기가 유사한 2개의 패널로 나누어지고, 각각의 패널은 약 17,000가구(34,000명)로 구성되어 있다. 이러한 2개의 패널은 항상 중복되게 운영되며, 3년마다 기존의 한 패널을 탈락시키고, 새로운 패널을 선정하여 교체해준다. 이와 같은 SLID의 연동모형은 Sylvie Michaud et.al.(1992)이 수행한 연구를 기초로 선택되었다. 이 연구에서는 총 4만 가구에 대한 3가지 연동모형을 검토하였는데, 1안은 매년 1/5(8천가구)을 연동, 2안은 2년마다 2만 가구 연동, 3안은 응답률에 따라 2년 또는 3년마다 2만 가구를 연동하는 방법을 고려했다. 이와 같은 세 가지 제안들의 장·단점을 비교한 후 3안(3년)을 가장 적합한 연동모형으로 최종 선택하였다.

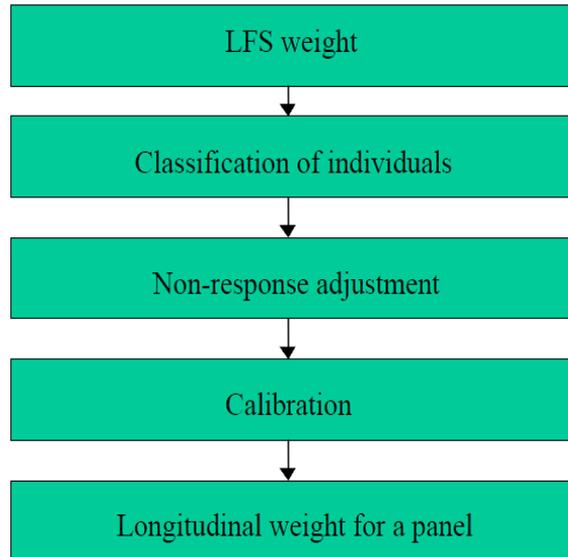


출처 : Sylvie LaRoche(2007)

[그림 5-2] SLID 패널형태

#### 다. 종단가중치 작성

SLID의 종단가중치 작성과정은 하나의 패널에 대해서 [그림 5-3]에서 보는 바와 같이 크게 5단계로 진행된다. 1단계에서는 종단가중치 작성을 위한 초기가중치를 선택확률의 역수로 설정한다. 패널  $i$ 에 대한 초기 종단가중치는  $w_{initial, p_i} = w_{LFS} * 3$ 인데,  $w_{LFS}$ 는 LFS의 가중치이고 LFS의 6개의 연동그룹 중 2개(1/3)를 선택했기 때문에 3을 곱해준다. 또한 동일 가구 내에서는 모두 동일한 가구원 초기가중치를 갖는다.



출처 : Jean-Francois Naud(2004)

[그림 5-3] SLID 종단가중치 작성과정

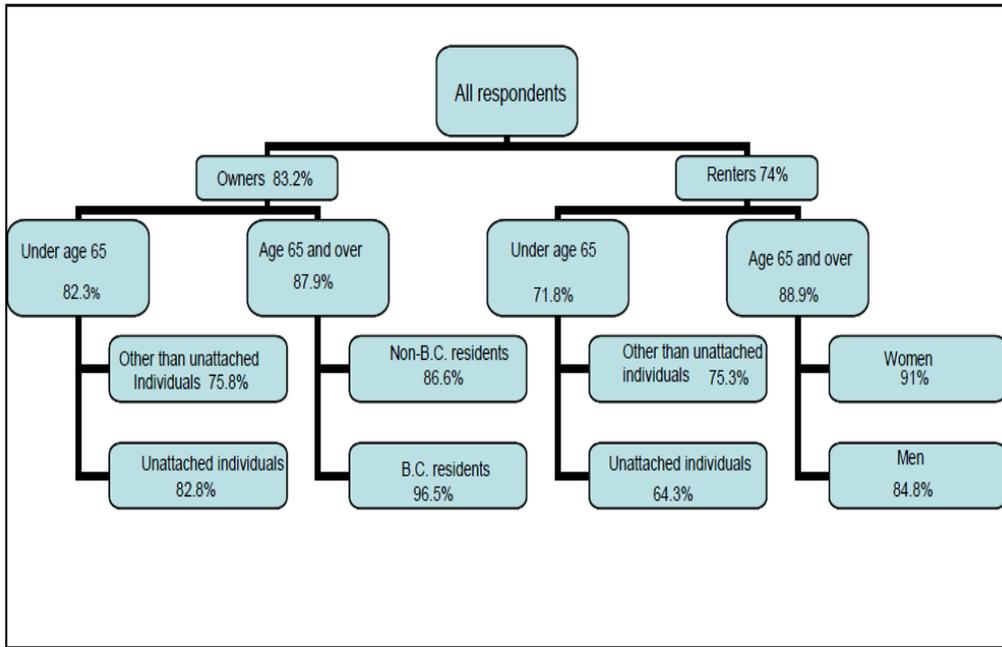


2단계에서는 종단가중치 부여를 위한 개인 분류를 수행하고 이 분류에 따라 가중치를 부여한다. 모든 가구원이 두 번(1월, 5월)의 조사에 무응답한 가구를 무응답 가구로 정의하고 있으며, 종단가중치는 0의 값을 준다. 반면 응답가구의 무응답 가구원은 응답으로 정의하며 무응답 대체를 수행한다. 또한 사망, 6개월 이상 기관거주, 캐나다에 거주하지 않은 경우에는 횡단면 부적격자로 분류하고, 종단면표본으로 선택된 표본의 경우에 한하여 종단가중치를 부여한다.

3단계에서는 무응답 조정을 실시하는데, 두개의 패널에 대해 개별적으로 조정을 한다. 무응답 조정을 위해 먼저 로지스틱 회귀모형으로 무응답층 구성을 위한 변수를 선택한다. 선택된 변수를 이용하여 CHAID 알고리즘을 적용하여 [그림 5-4]와 같이 무응답층을 구성한다. 그렇게 구성된 무응답층별로 아래의 수식에 의해 가중치를 조정한다.

$$w_{adjust} = \begin{cases} 0, & \text{무응답 가구의 개인} \\ \frac{w_{initial}}{R_{GHR}}, & \text{응답가구의 어른과 부적격자} \\ \frac{w_{initial}}{\bar{R}_{GHR,adults}}, & \text{응답가구의 아이} \end{cases}$$

여기서,  $R_{GHR}$ 는 어른이 속해 있는 무응답층에서의 가중 응답률이고,  $\bar{R}_{GHR,adults}$ 는 아이가 살고 있는 가구의 어른 구성원들의 가중 응답률 평균이다.



출처 : Sylvie LaRoche(2007)

[그림 5-4] SLID 무응답층

4단계에서는 영향력이 큰 관측값을 조정한다. 다시 말해, 횡단면 소득추정값과 분산 추정값에 영향력이 큰 관측값에 대해서 조정을 해준다. 영향력이 큰 관측값을 선정하는 방법은 뒤에서 나오게 될 횡단면가중치 작성방법에서 자세하게 설명하도록 하겠다. 이 단계를 통해 작성되는 가중치는  $w_{infl} = w_{adjust} * b_{infl, min}$ 으로,  $b_{infl, min}$ 는 현 웨이브와 이전 웨이브의 조정인자들 중 최소값으로 정의하고 있다.

마지막으로 5단계는 칼리브레이션 조정으로 역시 두 패널을 독립적으로 조정한다. GREG(Generalized REGression)방법을 이용하여 성-연령, 가구원수, 경제개체 수, 임금그룹(7개) 등의 변수에 대한 조정을 수행하는데, 2단계의 작업을 거친다. [그림 5-2]의 연동 모형의 패널 4에 대한 칼리브레이션 조정 단계에 대한 예가 <표 5-9>와 같다. 먼저 1차 웨이브 모집단에 대해 위의 4개의 변수를 모두 포함하여 조정을 한다. 그런 다음 목표 모집단인 0차 웨이브 모집단에 대해 임금그룹을 제외한 나머지 변수로 조정을 한다.

<표 5-9> SLID 패널 4의 종단가중치를 위한 칼리브레이션 조정

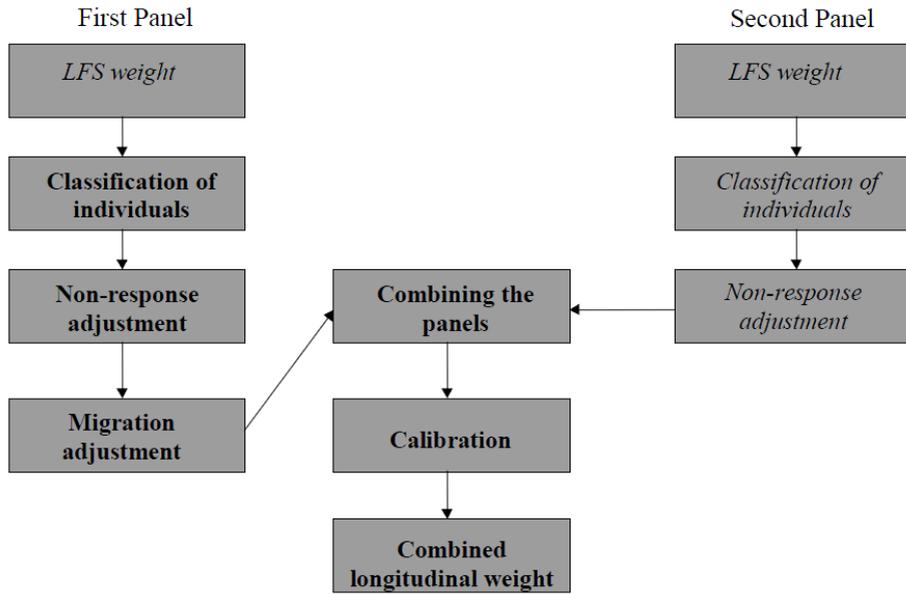
Control Totals			
<b>First calibration</b> Wave 1 Reference period: December 31, 2002	Age-sex group	1 to 7 (both sexes) 8 to 18 19 to 25 26 to 35	36 to 45 46 to 55 56 to 65 66 and over
	Number of households by size	Size=1 and size=2	
	Number of economic entities by size	Size=1 and size=2	
	Wages and salaries classes	7 classes defined by 10 <sup>th</sup> , 25 <sup>th</sup> , 50 <sup>th</sup> , 65 <sup>th</sup> , 75 <sup>th</sup> and 95 <sup>th</sup> , 98 <sup>th</sup> or 99 <sup>th</sup> percentile	
<b>Second calibration</b> Wave 0 Reference period: December 31, 2001	Age-sex group	0 to 6 (both sexes) 7 to 17 18 to 24 25 to 34	35 to 44 45 to 54 55 to 64 65 and over
	Number of households by size	Size=1 and size=2	
	Number of economic entities by size	Size=1 and size=2	

출처 : Sylvie LaRoche(2007)



## 라. 종단가중치 통합

SLID는 2개의 패널이 중복되는 연동패널형태이므로 앞에서 살펴본 바와 같이 각 패널에 대한 종단가중치를 개별적으로 작성하고 종단분석을 하게 된다. 그러나 이 두 패널 자료를 모두 이용하여 종단면 분석을 하게 되면 분석을 위한 표본의 크기가 커지게 되어 추정의 정도를 향상시킬 수 있다. 따라서 두 패널이 겹치는 기간인 3년 동안의 종단 분석을 위하여 종단가중치 통합을 수행하고 있다. [그림 5-5]의 종단가중치 통합과정을 보면, 먼저 시작한 첫 번째 패널에 대해서는 이주자 조정단계를 마치고, 두 번째 패널은 무응답 조정단계까지 수행한 다음 두 패널을 결합하여 칼리브레이션을 해줌으로써 종단가중치를 통합하고 있다.

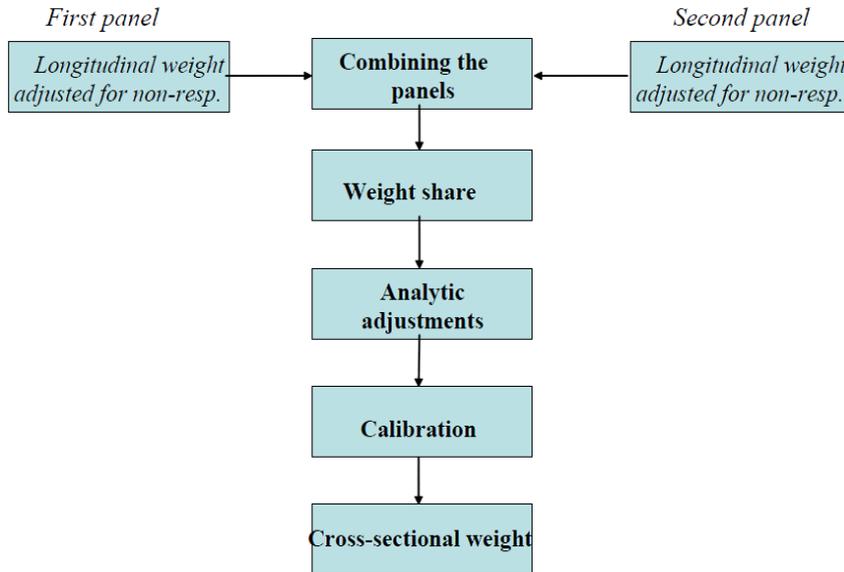


출처 : Jean-Francois Naud(2004)

[그림 5-5] SLID 종단가중치 통합과정

#### 마. 횡단가중치 작성

SLID에서는 두 종류의 횡단가중치를 제공하고 있다. 가구, 가족단위의 분석을 위한 가구의 가구원별 가중치가 동일한 통합가중치와 개인단위의 분석을 위하여 한 가구의 가구원별 가중치를 다르게 부여한 개별(노동)가중치이다. 개별가중치는 1월의 노동조사에 응답한 경우만을 대상으로 가중치를 작성하며, 1월 노동조사와 5월 소득조사에 모두 응답한 경우는 가중치를 0으로 주고 대상에서 제외한다. 이러한 개별가중치는 통합가중치에 비해 노동 특성에 대해 좀 더 좋은 추정값을 제공한다고 알려져 있다. 이와 같은 횡단가중치는 종단면 개인들 중에서 해당 년의 12월 31일 목표 모집단에 속한 경우에만 작성되고, 그렇지 않은 경우에는 가중치가 0이 된다. 횡단가중치 작성과정은 [그림 5-6]과 같다.



출처 : Jean-Francois Naud(2004)

[그림 5-6] SLID 횡단가중치 작성과정

횡단가중치(통합, 개별) 작성을 위한 초기가중치는 각 패널로부터 무응답 조정까지 마친 종단가중치인  $w_{year, adjust}$  이다. 카운티(county)의 경계를 넘어 이주한 표본의 경우 지역별 추출확률이 다르기 때문에 이주에 대한 조정이 필요하게 된다. 이를 위해 전입자의 가중치가 카운티 가중치 분포의 95백분위수보다 큰 경우에는 95백분위수에 해당하는 값으로 조정하는 방식을 사용한다. 카운티 간 이주 조정된 가중치는 횡단 초기가중치에 조정인자  $a_{mig}$  를 곱해준  $w_{year, mig} = w_{year, adjust} * a_{mig}$  이 된다.

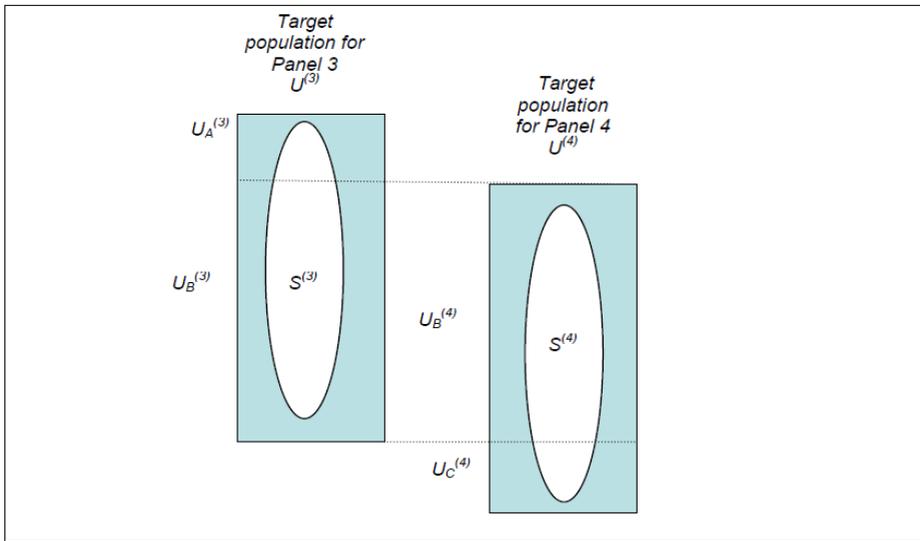
다음 단계는 아래 수식과 같이 패널할당인자( $p$ )를 구하고, 두 패널을 통합하여 추정하는 단계이다.

$$p = \left( \frac{n_1}{d_1} \right) \left( \frac{n_1}{d_1} + \frac{n_2}{d_2} \right)^{-1} = \frac{n_1}{n_1 + n_2 \left( \frac{d_1}{d_2} \right)}$$

여기서  $n_i$ 는 패널  $i$ 의 종단면 개인 수(16세 이상의 응답자)이고,  $d_i$ 는 패널  $i$ 의 설계효과(design effect)이다. 이와 같이 패널할당인자가 구해지면, [그림 5-7]의 패널중복여부를 고려하여 아래와 같이 가중치를 각각 조정하여 준다.

$$w_{year,p} = \begin{cases} pw_{year,mig}, & \text{중복된 앞선 패널의 개인 } (S_B^{(3)}) \\ (1-p)w_{year,mig}, & \text{중복된 후속 패널의 개인 } (S_B^{(4)}) \\ w_{year,mig}, & \text{중복되지 않은 두 패널의 개인 } (S_A^{(3)}, S_C^{(4)}) \end{cases}$$

또한 두 패널로부터 구해진 추정값과 패널할당인자를 이용하여  $\hat{Y} = p\hat{Y}_3 + (1-p)\hat{Y}_4$ 의 복합추정 형태로 횡단면 추정값을 구할 수 있다.



출처 : Sylvie LaRoche(2007)

[그림 5-7] SLID 목표 모집단과 패널 중복 형태

다음은 가중치 공유단계로 처음에는 표본으로 선택되지 않았지만 전입이나 출생 등으로 인해 나중에 표본으로 추가된 동거자에게 가중치를 부여한다. 이를 위해 동거자 상태를 IP(initially present)와 IA(initially absent)로 구분하였는데, IP는 후속 패널이 설계될 당시 목표 모집단에 포함되었던 동거자로 정의하였고, IA는 그와 반대되는 개념으로 정의하였다. 먼저 통합가중치의 경우는 가구 내 가구원의 가중치가 모두 같기 때문에 아래 식과 같이 동거자에게 동일한 가중치를 공유해준다.  $n_{L,h}$ 는  $h$ 가구의 종단 가구원수이고,  $n_{IP,h}$ 는  $h$ 가구의 IP 동거자 수이다.

$$w_{year,shared} = \frac{\sum_h w_{year,p}}{n_{L,h} + n_{IP,h}}$$

또한 개별(노동)가중치는 가구의 구성 상태나 가구원의 상태에 따라 아래와 같이 다른 가중치를 부여하게 된다.

$$w_{year,shared} = \begin{cases} w_{year,p} & : \text{종단표본으로만 구성} \\ & \text{\&모든 동거자가 IA인 가구의 가구원} \\ w_{year,p} \left( \frac{n_{L,h}}{n_{L,h} + n_{IP,h}} \right) & : \text{IP동거자가 최소 한명 이상인 가구의} \\ & \text{종단표본 가구원} \\ \frac{\sum_h w_{year,p}}{n_{L,h} + n_{IP,h}} & : \text{모든 동거자} \end{cases}$$

다음으로 영향력이 큰 관측값을 조정하는 단계를 거친다. 소득과 같이 꼬리가 긴 분포를 갖는 변수들에 대해 영향력이 큰 관측값이 발생이 가능한데 이러한 자료를 그대로 이용하면 추정 시 문제가 생길 수 있다. 특히 성별이나 연령별 등과 같이 어떤 작은 영역별로 추정했을 시 문제가 커질 수 있으므로 이러한 관측값의 영향을 줄일 수 있는 조정방법이 필요하다. 1단계에서는 소득에 대한 영향력이 큰 관측값을 결정하는데, 소득에 대한 한계치를 중위수+k(75백분위수-중위수)와 같이 지정하고, 관측값이 한계치를 넘으면 영향력이 큰 관측값으로 판단한다. 여기서 k는 상수로 퀘백과 온타리오는 75, 나머지 주는 25로 설정되어 있다. 2단계에서는 0~1의 값을 갖는 가중치 조정인자( $b_{infl}$ )를 작성하는데, 표본과 소득세의 분포를 비교하여 소득이 \$99,999 이상 차이가 나는 경우에만 가중치로 조정한다. 2003년의 경우 66명에게 적용되었으며, 영향력이 큰 관측값은 전체에서 매우 적은 비중을 차지한다. 이러한 조정과정을 통해  $w_{year,infl} = w_{year,shared} * b_{infl}$ 가 작성된다.

마지막 단계인 칼리브레이션은 2단계의 조정과정을 거쳤던 종단가중치와는 다르게 조정과정이 1단계로 마무리된다. 횡단가중치 작성 시에는 <표 5-10>과 같이 인구통계학적 변수와 임금그룹에 대한 변수를 모두 고려한 칼리브레이션 조정을 수행한다.

<표 5-10> SLID 2003년 횡단가중치를 위한 칼리브레이션 조정

		Control totals	
Reference period: December 31, 2003	Age-sex group	0 to 6 (both sexes)	35 to 44
		7 to 17	45 to 54
		18 to 24	55 to 64
		25 to 34	65 and over
	Number of households by size	Size=1 and size=2	
Number of economic entities by size	Size=1 and size=2		
Wages and salaries classes	7 classes defined by the 10 <sup>th</sup> , 25 <sup>th</sup> , 50 <sup>th</sup> , 65 <sup>th</sup> , 75 <sup>th</sup> and 95 <sup>th</sup> , 98 <sup>th</sup> or 99 <sup>th</sup> percentiles		

출처 : Sylvie LaRoche(2007)



## 제4절 분할패널조사

### 1. 이탈리아 SHIW(Survey of Household Income and Wealth)

#### 가. 조사개요

이탈리아 은행에서는 가구의 경제적 행위에 대한 정보를 수집하여 소득과 재산이 어떻게 분배되고 있는지를 추정하기 위하여 SHIW를 실시하고 있다. 1965년 처음 시작된 이래로 1987년까지는 횡단조사로 진행하였으나 현상의 변화를 분석하기 위해서 1989년부터 종단을 고려한 표본설계로 변경하였다. 약 8천 가구의 표본에 대해 조사전문기관에 위탁하여 181명의 조사원이 조사를 수행한다. 조사방법은 CAPI로 약 79.5%가 조사되고 있으며, 약 20.5%는 PAPI로 조사된다. 조사주기는 2년으로 소득, 재산, 노동, 교육 등의 항목을 수집하는데, 인터뷰 소요시간은 평균 55분이다. 조사표는 주제가 바뀌지 않는 핵심파트와 주제별 섹션으로 구성되어 있다. 설문지 일부 섹션에 대해서는 응답자 부담을 줄이기 위해 부표본 집단을 랜덤하게 선정하여 조사한다. 예를 들어, 질문지를 두 개의 세트로 구분하여 가구주의 태어난 해의 짝·홀수 여부로 응답해야 할 질문지 세트가 결정되는 방식을 취하고 있다.

#### 나. 표본설계

표본설계는 2단계 추출로 먼저 PSU인 도시(municipality)를 선정하였다. 도시를 지역(북부/중부/남부)과 인구수(4만 이상/미만)로 층화를 하고, 각 층 내에서 인구 4만 이상인 도시는 모두 포함하고, 작은 도시는 확률비례추출(PPS: probability proportional to size)로 선택하였다. 다음으로 선택된 PSU내에서 SSU인 가구를 랜덤하게 선정하였다.

패널형태는 분할패널로서 <표5-11>에서 보는 바와 같이 전년도 표본의 일부(30~45%)를 패널로 선택하고 나머지는 매 웨이브마다 새롭게 표본을 설계하여 목표 표본크기를 보충해주는 형태로 운영하고 있다. 전년도 표본의 일부를 선정할 때는 지난조사의 표본 도시들 중에서 패널도시를 먼저 선정하고 해당 도시 내에서 이전 조사에 참여했던 가구를 모두 패널로 선정하고, 그 이외에는 직전조사에 인터뷰를 마친 가구들 중에서 랜덤하게 선택한다. 새로 보충되는 비패널 가구들은 패널과 비패널 도시들의 등록부에서 랜덤하게 선택된다.

〈표 5-11〉 SHIW의 표본가구 수(1987~2008)

Year of first interview	Year of survey										
	1987	1989	1991	1993	1995	1998	2000	2002	2004	2006	2008
1987.....	8,027	1,206	350	173	126	85	61	44	33	30	28
1989.....		7,068	1,837	877	701	459	343	263	197	159	146
1991.....			6,001	2,420	1,752	1,169	832	613	464	393	347
1993.....				4,619	1,066	583	399	270	199	157	141
1995.....					4,490	373	245	177	117	101	84
1998.....						4,478	1,993	1,224	845	636	538
2000.....							4,128	1,014	667	475	398
2002.....								4,406	1,082	672	525
2004.....									4,408	1,334	995
2006.....										3,811	1,143
2008.....											3,632
<b>Sample size</b>	<b>8,027</b>	<b>8,274</b>	<b>8,188</b>	<b>8,089</b>	<b>8,135</b>	<b>7,147</b>	<b>8,001</b>	<b>8,011</b>	<b>8,012</b>	<b>7,768</b>	<b>7,977</b>
<i>Panel households as a % of total households</i> .....		14.6	26.7	42.9	44.8	37.3	48.4	45.0	45.0	50.9	54.5

출처 : Bank of Italy(2010)

SHIW에서는 항목 무응답이 발생했을 때 무응답 대체를 한다. 무응답으로 인한 결측치의 값을 추정하기 위해 회귀모형을 사용하며, 대체해주는 값이 평균에 집중되는 것을 방지하기 위해 랜덤항목 추가한다. 여기서 랜덤항목은 평균이 0이고 회귀모형의 잔차와 같은 값의 분산을 갖는 정규분포로부터 랜덤하게 발생시켜 준다. 이와 같은 방법을 통해 실제 측정되었던 자료의 분산과 평균을 유지시킬 수 있다. SHIW는 무응답 대체를 하고 있지만 이와 같이 대체되는 데이터의 양은 매우 적은 편이다.

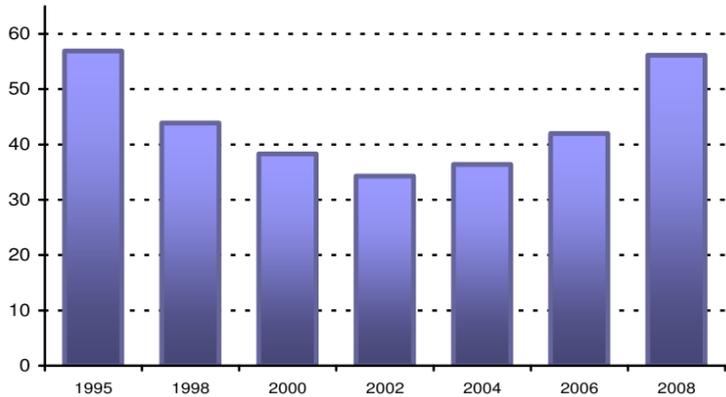
단위 무응답의 경우를 살펴보면, 2008년도에 7,977가구(패널 4,345, 비패널 3,632)의 최종 응답을 얻기 위해 접촉한 총 가구 수는 14,209가구(패널 5,477, 비패널 8,732)였다. 전체 응답률은 56.1%로 79.3%의 응답률을 보이고 있는 패널가구가 41.6%보다 비패널보다 응답률이 높다. 그 외 조사실패의 이유로는 거절이 32.4%, 부재<sup>2)</sup>가 11.5%를 차지하고 있었으며, 부적격<sup>3)</sup>도 5.2% 수준이었다.

[그림 5-8]에서 보는 바와 같이 SHIW는 35~55% 수준의 낮은 응답률을 보이고 있다.

- 2) 서로 다른 날, 다른 시간대에 건 3번의 전화에 모두 연결되지 않은 경우
- 3) 주소지에서 가구를 확인할 수 없는 경우(잘못된 주소, 사망, 주소 변경)



이처럼 조사 수행 시 접촉할 수 없는 가구들은 동일한 도시 내에서 랜덤하게 선택된 다른 가구들로 대체할 수 있도록 하고 있다. 이와 같은 가구대체는 조사자의 영향을 최대한 배제할 수 있도록 설정된 엄격한 프로토콜에 의해 수행되고 있다. 또한 대체로 인한 문제발생을 줄이기 위해 조사의 후반부에 응답자의 특성들을 기반으로 하는 사후층화를 통해 보정해준다.



출처 : Bank of Italy(2010)

[그림 5-8] SHIW 응답률(1995~2008)

#### 다. 횡단가중치 작성

횡단가중치 작성을 위한 첫 번째 단계인 초기가중치는 설계가중치(design weights)이며 아래의 식과 같다.

$$w_h^{(0)} = \frac{N'_h}{n'_h} \frac{N_h}{N'_h} = \frac{N_h}{n'_h}$$

여기서,  $N_h$ 는  $h$ 층의 총 거주 인구,  $N'_h$ 는  $h$ 층의 표본 도시 인구,  $n'_h$ 는  $h$ 층에서 선택된 가구 수이다.

둘째로 무응답에 대한 조정을 하게 되면 가중치는 아래의 식으로 표현할 수 있다.

$$w_h^{(1)} = w_h^{(0)} \frac{n'_h}{n_h} \frac{n_h}{m_h} = \frac{N_h}{m_h}$$

여기서  $n_h$ 는  $h$ 층의 적격 가구 수,  $m_h$ 는  $h$ 층의 응답 가구 수이며,  $n'_h \geq n_h \geq m_h$ 이다.

세 번째 단계에서는 패널마모<sup>4)</sup> 조정 가중치(attrition-adjusted weights)를 아래와 같이 작성하는데, 이 가중치는 패널에게만 적용된다.

$$w_c^{(2')} = w_c^{(1)} \alpha_c$$

여기서,  $\alpha_c$ 는 패널마모 사후 층  $c$ 에 대한 조정인자로  $c = 1, \dots, 12$ 이다. 패널마모에 대한 조정을 위해 패널가구들을 사후층화를 하는데, 사후 층은 소득(4그룹)과 가구의 고용 상태(3그룹)의 총 12개로 구성된다. 이와 같이 사후 층이 구성된 다음 각 층별로 현 웨이브의 분포가 이전 웨이브에서 관측된 전체 표본의 분포와 같아지도록 조정인자를 계산한다.

네 번째 단계에서는 자기상관 교정 가중치(autocorrelation corrected weights)를 작성한다. 패널의 경우 같은 가구에서 연속적으로 조사를 하므로 소득이나 자산과 같은 변수에 대해 과거에 조사된 정보를 이용하면 효율적인 추정값을 얻을 수 있다. 주요 변수 간 양의 자기상관관계로 자기상관 교정 계수를 구하여 아래와 같이 조정하게 된다.

$$w_p^{(2'')} = w_p^{(2')} \beta_p$$

여기서,  $\beta_p$ 는 패널과 비패널의 자기상관 교정 계수이고,  $p$ 는 패널여부에 대한 지시변수이다.

※ 참고 :  $\bar{y}_t^* = \frac{Q(1-Q\rho^2)}{1-Q^2\rho^2} \bar{y}_t^q + \frac{P}{1-Q^2\rho^2} [\bar{y}_t^p + \rho(\bar{y}_{t-1} - \bar{y}_{t-1}^p)]$ ,

여기서  $w_p^{(2'')} = \frac{Q(1-Q\rho^2)}{1-Q^2\rho^2}$ 로 놓으면,

$\Rightarrow \bar{y}_t^* = w_p^{(2'')} \bar{y}_t^q + (1-w_p^{(2'')}) \bar{y}_t^p + (1-w_p^{(2'')}) \rho (\bar{y}_{t-1} - \bar{y}_{t-1}^p)$  : 복합추정량  
만일 사후층화가 만족스럽게 되었다면,  $\bar{y}_{t-1} \cong \bar{y}_{t-1}^p$ 이 되고,  
경험적으로  $0.4 \leq \rho \leq 0.6$ 이고,  $\tilde{\rho} = 0.5$ 로 놓으면,

$\Rightarrow \bar{y}_t^* = w_p^{(2'')} \bar{y}_t^q + (1-w_p^{(2'')}) \bar{y}_t^p \quad \tilde{w} = \frac{Q(1-\tilde{\rho}^2 Q)}{1-\tilde{\rho}^2 Q^2} \Big|_{\tilde{\rho}=0.5}$

$\Rightarrow w_p^{(2'')} = \begin{cases} w_p^{(2')} \frac{1-\tilde{w}}{1-Q} & \text{for panel households} \\ w_p^{(2')} \frac{\tilde{w}}{Q} & \text{for non panel households} \end{cases}$

4) 무응답과 부적격(사망, 이주 등)에 의한 표본 감소를 일컬음

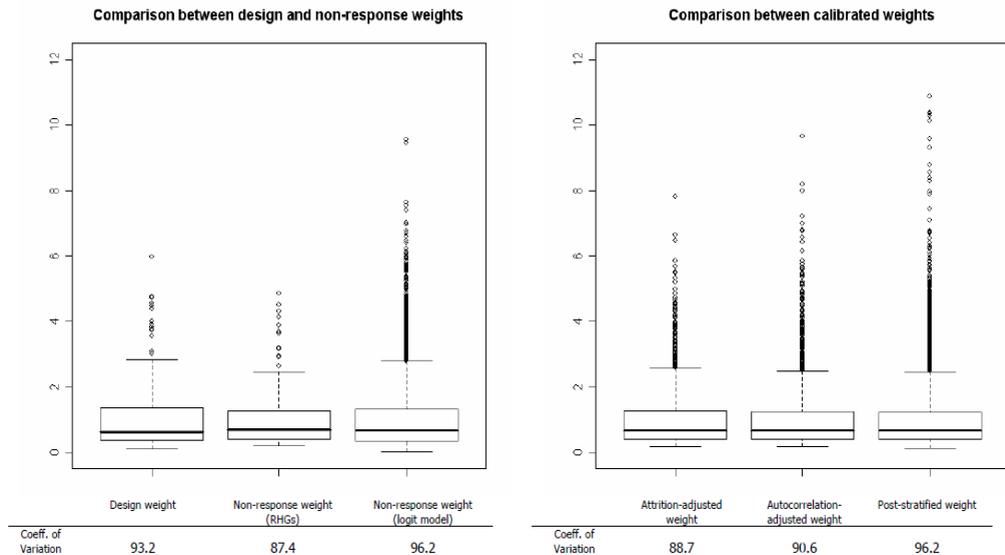
마지막으로 사후층화 가중치(post-stratification weights)를 작성하는데, 레이킹 비 방법을 사용한다. 조정변수는 성, 연령 4개(26세 미만/26~45/46~65/65세 이상), 지역 3개(북/중/남부), 도시규모 4개(인구 2만 미만/2~4만/4~50만/50만 이상)그룹이다.

$$w_j^{(2''')} = w_j^{(2'')} \gamma_j$$

여기서,  $\gamma_j$ 는  $j$ 번째 사후 층의 인자이다. 이와 같이 총 5단계의 과정을 통해 횡단가중치가 최종적으로 작성되며, 각 단계를 거치면서 변화하는 횡단가중치 분포의 형태를 [그림 5-9]에서 확인할 수 있다.

다음으로 분산추정은 잭나이프(Jack-knife)방법을 이용하고 있다. 잭나이프 분산추정시 반복의 횟수(c)는 "pseudo"층의 수와 같다. 각 반복에 대해서 첫 번째 "pseudo" PSU의 가중치는 0으로 설정해 놓고, 나머지의 표집가중치는 cancelled unit의 가중치를 보상해 주기 위한 인자만큼 키워준다. 이러한 가중치를 사용하여 각 반복에 대해  $\hat{\theta}_{(i)}$  계산한다. 각 층마다 분산추정에 대한 두 개의 unit을 포함하고 있기 때문에 표준오차(standard error)의 추정값은 전체 표본에 대한 추정값( $\hat{\theta}$ )와 반복의 추정값의 차이의 제곱합의 제곱근으로 아래와 같이 계산된다.

$$stderr_j = \sqrt{\sum_{i=1}^c (\hat{\theta}_{(i)} - \hat{\theta})^2}$$



출처 : Ivan Faiella and Romina Gambacorta(2007)

[그림 5-9] SHIW 가중치 작성과정 단계별 횡단가중치 분포

## 라. 종단가중치 작성

1차 웨이브의 종단가중치는 횡단가중치와 동일하며, 현 웨이브( $t$ ) 종단가중치는 이전 웨이브( $t-1$ ) 종단가중치에 조정인자를 곱한 형태인  $w_{i,t}^{(L)} = w_{i,t-1}^{(L)} \times a_{i,t-1 \Rightarrow t}$ 이다. 여기서, 조정인자는  $a_{i,t-1 \Rightarrow t} = a_{i,t-1 \Rightarrow t}^{(pr)} \times a_{i,t-1 \Rightarrow t}^{(nr)} \times a_{i,t-1 \Rightarrow t}^{(ps)}$ 와 같이 3부분으로 분해된다. 또한 가구원 가중치는 가구 가중치와 동일하다.

먼저 포함확률 조정인자  $a_{i,t-1 \Rightarrow t}^{(pr)} = \frac{p}{p+fq}$ 인데,  $p$ 는 패널 가구 수,  $q$ 는 비패널 가구 수 그리고  $f$ 는 비패널 비율이다.

※ 참고 :  $P_t^* = (P_{t-1} \cup f_t Q_{t-1})$  : target panel sample

패널마모를 보상하기 위해 표본을 추가 : sample refreshing

$$\pi_{i,t-1 \Rightarrow t}^{(ref)} = pr_{S_{t-1}}(i \in P_t^*) = pr_{S_{t-1}}(i \in P_{t-1}) + pr_{S_{t-1}}(i \in f_t Q_{t-1}) = \frac{p_{t-1} + f_t q_{t-1}}{n_{t-1}}$$

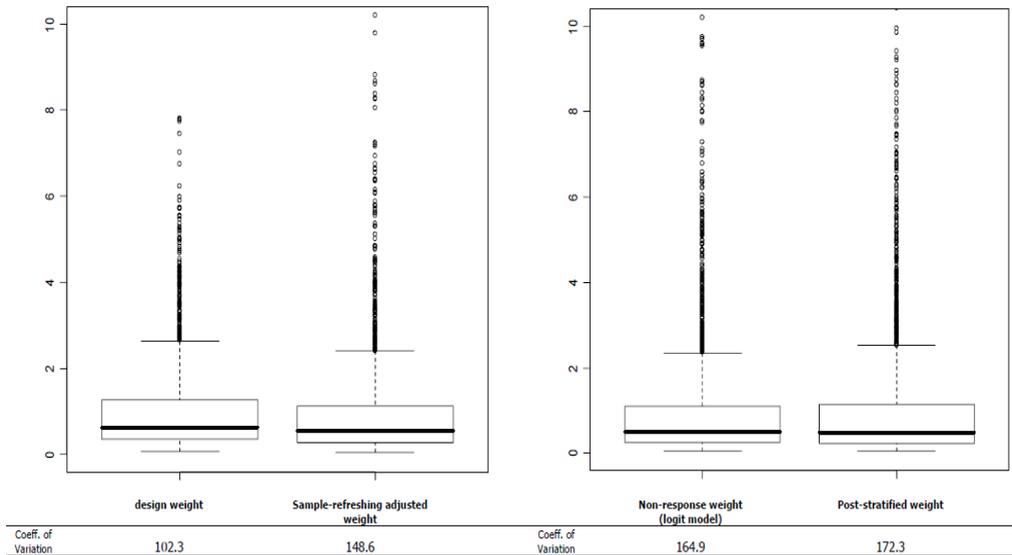
$$\pi_{i,t-1 \Rightarrow t}^{(noref)} = pr_{S_{t-1}}(i \in P_t^*) = pr_{S_{t-1}}(i \in P_{t-1}) = \frac{p_{t-1}}{n_{t-1}}$$

즉,  $\pi_{i,t-1 \Rightarrow t}^{(ref)} = \frac{p+fq}{p} \pi_{i,t-1 \Rightarrow t}^{(noref)}$  이므로  $\frac{p+fq}{p}$ 의 역수만큼 줄여줘야 함

다음으로 무응답 조정인자는  $a_{i,t-1 \Rightarrow t}^{(nr)} = 1/\hat{\pi}_{i,t-1 \Rightarrow t}$ 이고, 여기서  $\hat{\pi}_{i,t-1 \Rightarrow t}$ 는 전체 응답 성향 추정값이다. 응답 성향 추정값은 로지스틱 모형(2002년도, 2004년도 응답 패턴)을 이용하여 추정하게 되는데, 사용변수는 조사자에 의해 제공받은 면접조사 시 분위기와 같은 협조적인 습성과 가구 구성원들 최고령 나이, 도시의 크기 및 지리적 위치와 같은 비접촉 패턴이다.

마지막으로 사후층화 인자  $a_{i,t-1 \Rightarrow t}^{(ps)}$ 는 한 웨이브에서 다른 웨이브로 모집단 특성이 전이되는 것을 보정하기 위한 목적으로 도입되었다. 이 단계에서는 횡단가중치에서 사후층화 가중치 단계에 사용된 동일 모집단 총계에 대해 맞추는 조정을 한다. 이와 같은 종단가중치 작성과정 단계별 가중치 분포의 변화를 [그림 5-10]에서 확인할 수 있다.





출처 : Ivan Faiella and Romina Gambacorta(2007)

[그림 5-10] SHIW 가중치 작성과정 단계별 종단가중치 분포

## 2. 스페인 SHF(Survey of Household Finances)

### 가. 조사개요

스페인 은행에서는 스페인 경제의 금융 계정의 집계 데이터를 보완하고, 스페인 가구의 투자 및 자금 조달 결정과 이들의 금융 포지션을 만드는 연구를 위해 SHF를 실시하고 있다. SHF로 인하여 수행되는 연구들은 스페인 경제의 더 나은 이해와 공공 정책을 적절히 디자인하는 역할을 하고 있다. 이 조사는 이탈리아 은행의 SHIW를 기초로 하여 2002년에 처음으로 시작하였다. 조사주기는 3년으로 2005년에 2차, 2008년에 3차 웨이브가 진행되었다. 표본크기는 약 7천 가구로 CAPI방법에 의해 조사를 수행하고 있다. 인터뷰에 소요되는 시간은 중앙값이 55분이고 전체 응답자의 약 95%가 2시간 미만에 조사를 완료한다. 조사되는 내용으로는 인구통계, 실물자산 및 그와 관련된 채무, 기타 부채, 금융자산, 연금 및 보험, 고용시장 상황과 근로소득, 직전연도의 비근로소득, 결제 수단, 소비와 저축 등이 있다.

## 나. 표본설계

SHF의 표집틀 작성을 위해 통계청에서는 먼저 주소에 의해 정의된 가구단위의 도시등록부(Municipal Register)를 국세청으로 이송한다. 그런 다음 국세청에서 재산과 소득세금환급액에 기초한 정보인 다음의 세 변수를 각 주소마다 추가하여 표집틀을 구축한다. 첫 번째 변수는 가구 총 과세재산을 기준으로 구성된 재산층(wealth strata)<sup>5)</sup>이다. 두 번째 변수는 소득세는 있으나 재산세는 없는 가구에 대하여 가구가 속해있는 전국 소득분포의 사분위수이며, 세 번째 변수는 1인당 가구소득이다. 이와 같은 소득관련 변수로 정렬을 한 다음 계통추출(systematic sampling)을 하는 방식으로 표본을 선정한다. 각 웨이브마다 실제로 설계된 내용은 뒤에서 좀 더 설명하겠다.

SHF 표본설계의 특징은 부유층에 대한 과대표집이라 할 수 있다. 일반적으로 부유층은 인구에 있어서는 낮은 비율을 차지하고 있지만, 전체 가구 재산에 대해서는 큰 비중을 보인다. 인구에 대한 대표성 확보뿐만 아니라 재산의 총계를 내기 위하여 과대표집을 실시하였다. 이를 위해 스페인 은행은 국세청과 통계청의 협조를 받아 개인의 재산세 정보를 기초로 과대표집을 실시하였다. <표 5-12>에는 2002년과 2005년 과대표집 비율이 나와 있는데, 2005년 상위 1%의 부유층의 경우 랜덤추출에 비해 약 9배정도를 과대표집한 것을 알 수 있다.

<표 5-12> SHF 과대표집 비율

Net worth decile group	EFF 2002		EFF 2005	
	Number of observations	Oversampling rate <sup>1</sup>	Number of observations	Oversampling rate
Bottom 50%	1,878	0.73	2,234	0.75
50% to 90%	1,944	0.94	2,036	0.85
90% to 95%	429	1.67	481	1.61
95% to 99%	524	2.55	675	2.83
Top 1%	368	7.16	536	8.99

출처 : Olympia Bover(2008)

5) 층1(재산세 정보 없음)/ 층2(재산세 2천만 이하)/ 층3(재산세 2~5천만)/ 층4(재산세 5천만~1억)/ 층5(재산세 1억~2억)/ 층6(재산세 2억~5억)/ 층7(재산세 5억~20억)/ 층8(재산세 20억 이상 페세타)



1차 웨이브인 2002년 표본설계 내용을 좀 더 살펴보면, 도시의 인구 규모별로 표본설계를 다르게 수행하였다. 먼저 인구 10만 이상의 대도시에 대해서는 8개의 재산층 내에서 랜덤하게 표본을 선택하였다. 반면, 인구 10만 이하의 소도시는 다음과 같이 2단계 집락설계를 하였다. 먼저 PSU인 도시는 인구수를 기초로 확률비례추출을 하고, SSU인 가구는 PSU내 세금과일 정보에 의해 선정하였다. 특히 나바르와 바스크지방은 국세청에서 개인 과세정보를 관리하지 않기 때문에 도시 크기에 따른 6개 층을 구성하여 2단 층화 집락설계를 하였다. <표 5-13>의 1차 웨이브에서 조사완료를 위해 접촉을 시도한 가구 수와 실패이유를 살펴보면, 5,143가구의 응답을 얻기 위해 총 19,901가구를 접촉하였으며, 그 중 5,722가구는 거절, 6,670가구는 부재였다. 또한 앞서 정의된 재산층별 참여율을 <표 5-14>에서 살펴보면, 재산세 수준이 높은 부유층일수록 협조율이 낮음을 확인할 수 있다. 이와 같은 낮은 조사 성공률로 인하여 SHF에서는 원래의 표본가구에 대해 4개까지 표본대체를 허용하고 있다. 그러나 이와 같은 표본대체를 엄격하게 관리하고 있는데, 대도시는 도시 내에서, 소도시는 PSU 내에서 원래의 표본가구와 지리적으로 멀리 떨어지지 않으면서도 소득 사분위수, 재산층, 일인당 가구소득으로 순위를 매긴 후 순위 바로 위의 2가구와 바로 아래의 2가구로 대체하는 규칙에 의해 수행된다.

<표 5-13> SHF 1차 웨이브(2002) 가구 접촉시도

Completed	5,143
Refused	5,722
Never at home	6,670
Out of scope (wrong address, not a housing unit, empty dwelling, deceased, others out of scope)	1,797
Discarded after supervision	569
Total	19,901

출처 : Olympia Bover(2008)

〈표 5-14〉 SHF 1차 웨이브(2002) 재산층별 협조율

	Never at home <sup>1</sup>	Cooperation rate <sup>2</sup>
Total	33.5%	47.3%
Stratum 1	31%	53.6%
Stratum 2	38.9%	45.3%
Stratum 3	32.9%	44.7%
Stratum 4	35.5%	46.5%
Stratum 5	37%	38.5%
Stratum 6	38%	36.1%
Stratum 7	40.1%	37.8%
Stratum 8	39.8%	29.4%
Navarre and Basque Country	26%	46%

1. Defined as (Never at home/Total attempted contacts)

2. Defined as (Completed/Completed+Refused)

출처 : Olympia Bover(2008)

2차 웨이브인 2005년에는 1차 웨이브 표본 5,143가구를 다시 인터뷰하는 것을 원칙으로 삼았다. <표 5-15>에서 보는 바와 같이, 1차 웨이브의 5,143개 표본가구의 약 50%에 해당하는 2,580가구가 2차 웨이브에 패널로 참여함으로써 2차 웨이브 표본(5,962)의 약 43%가 패널가구로 선정되었다. 그리고 나머지 표본을 보충하기 위해 1차 웨이브와 전체적인 특성이 동일한 표본을 설계하는 것을 목표로 설정하였다.

우선 지방의 수도나 인구 10만 이상의 대도시에서는 패널 표본을 보완하기 위한 과대표집을 수행하였는데, 8개 부유 층 내에서 가구를 랜덤하게 선정하였다. 반면 인구 10만 이하의 소도시에 대해서는 과대표집을 하지 않고, 1차 웨이브와 동일한 방식으로 2단계 집락 설계를 하였다. 즉, 인구수에 의한 확률비례추출로 PSU를 선정하고, 패널 표본을 보완하기 위해 PSU당 7가구가 조사될 때까지 랜덤하게 SSU인 가구를 선정하였다. 이와 같은 표본설계의 결과로 3,382가구가 최종적으로 비패널가구로 보충되었다. 이러한 SHF의 패널형태는 고정패널과 웨이브마다 표본을 새롭게 설계하여 보충해주는 분할패널형태를 보이고 있다.

2005년 2차 웨이브의 재산층별 협조율은 <표 5-16>과 같다. 1차 웨이브에 참여했던 패널의 협조율은 67.1%로 2차 웨이브에 처음으로 참여하게 된 비패널의 38.6%에 비해 협조율이 매우 높게 나타나고 있다. 또한 패널의 경우 2차 웨이브의 재산층별 협조율은 층2를 제외한 나머지 재산층별로 비슷한 수준을 보이고 있다. 그러나 비패널의 경우



2002년과 마찬가지로 재산세를 적게 내는 층에서 부유층으로 갈수록 협조율이 떨어짐을 확인할 수 있다.

SHF는 항목무응답에 대해 확률적인 방법을 이용한 다중 무응답 대체를 실시하고 있다. 다중대체에 의해 항목무응답을 처리하게 되면 대체되는 값의 경우의 수(m)에 따라 여러 개(m)의 데이터 셋이 작성된다. 이와 같은 다중대체에 의한 자료의 분석은 m개의 데이터 셋에 대해 각각 분석을 한 다음 분석결과를 종합하는 과정을 거쳐야 하므로 이용자 입장에서는 번거로울 수 있다. 그러나 최근에는 STATA와 같이 다중대체를 분석할 수 있는 소프트웨어가 개발되면서 분석을 용이하게 할 수 있다. 이러한 무응답 대체를 위해 SHF에서는 미국의 SCF(Survey of Consumer Finances)에서 개발한 소프트웨어를 이용하고 있다. SCF의 다중 무응답 대체 프로그램인 Fritz(Federal Reserve Imputation Technique Zeta)는 순차적이고 반복적인 구조인 깃스 샘플링(Gibbs sampling)방법을 사용하여, 대체하고자 하는 변수의 형태에 따라 대체방법을 다르게 적용하고 있다. 연속형변수에 대해서는 회귀모형 예측에 의한 랜덤방법(randomizations from regression predictions), 이항변수는 다양한 선형확률모형(the linear probability model), 다항변수는 핫덱(hot deck)방법의 일종을 통해 무응답 대체를 하고 있다.

〈표 5-15〉 SHF 2차 웨이브(2005) 가구 접촉시도

	TOTAL	PANEL	NON-PANEL
Completed	5,962	2,580	3,382
Refused	6,634	1,263	5,371
Never at home	1,602	479	1,123
Out of scope (wrong address, not a housing unit, empty dwelling, deceased, others out of scope)	1,275	363	912
Discarded after supervision	189	24	165
Total	15,662	4,709	10,953

출처 : Olympia Bover(2008)

〈표 5-16〉 SHF 2차 웨이브(2005) 재산층별 협조율

	TOTAL		PANEL		NON-PANEL	
	Never at home <sup>1</sup>	Co-operation rate <sup>2</sup>	Never at home	Co-operation rate	Never at home	Co-operation rate
Total	10.2	47.3	10.2	67.1	10.2	38.6
Stratum 1	8.4	54.9	9.3	68.8	7.9	47.2
Stratum 2	14.3	47.3	12.5	90.9	14.5	42.4
Stratum 3	12.5	46.1	12.6	69.4	12.5	35.3
Stratum 4	11.6	46.9	12.1	67.7	11.3	36.1
Stratum 5	12.5	36.3	13.0	62.1	12.4	29.5
Stratum 6	14.1	35.3	11.2	63.5	14.9	26.5
Stratum 7	14.3	31.4	10.4	58.8	15.2	25.0
Stratum 8	13.0	31.1	10.0	66.7	13.7	21.8
Navarre and Basque Country	5.1	44.3	7.4	61.0	4.3	38.5

1. Defined as (Never at home/Total attempted contacts).

2. Defined as (Completed/Completed+Refused).

출처 : Olympia Bover(2008)

#### 다. 종단가중치 작성

종단가중치는 패널가구에 대해서만 작성이 되며, 여기서는 2차 웨이브를 기준으로 설명하고자 한다. 종단가중치 작성을 위한 초기가중치는 1차 웨이브에서 무응답 조정을 한 설계가중치가 된다. 초기가중치가 준비되면, 1차 웨이브의 모집단을 참고로 하여 2차 웨이브 무응답에 대한 조정을 수행한다. 무응답 조정변수는 도시크기, 재산층, 소득 사분위수를 이용하였다. 다음으로 2005 센서스에 대한 조정 가중치를 작성하게 되는데, 이를 위해 칼마방법(Calmar procedure)을 도입하였다. 칼마방법이란 프랑스 INSEE에 의해 개발된 방법으로 선형거리함수(linear distance function)에 의해 가구와 개인과 같이 서로 다른 수준의 조정을 동시에 가능하게 하는 것이 특징이다. 조정변수는 성별, 도시 크기별 연령, 도시 크기별 가구원수를 사용하였다.

이와 같은 종단가중치는 국세청과 통계청에서 작성되고 있다. 과세자료 변수를 이용하여 조정하게 되는 설계가중치와 무응답가중치는 비밀보호의 문제로 인하여 통계청에



서 조정방법을 제시해주면 국세청에서 가중치를 작성한다. 또한 센서스에 대한 조정가중치는 모집단 정보가 있는 통계청에서 작성된다.

## 라. 횡단가중치 작성

횡단가중치는 패널과 비패널 표본을 두 개의 독립표본으로 간주하고 각각 가중치를 작성한다. 횡단가중치의 초기가중치는 비패널의 경우에는 설계가중치(추출률의 역수)에 무응답을 조정한 가중치가 되며, 패널의 경우에는 종단가중치가 된다. 다음으로 패널과 비패널의 초기가중치를 통합하는 과정을 거치는데, 부표본의 상대적 크기에 의해 조정한다. 이 방법은 동일한 모집단을 대표하는 두 독립 표본에 대한 최소분산추정량으로 알려져 있다. 마지막으로 센서스에 대한 조정 가중치를 칼마조정에 의해 수행하는데, 성별, 도시 크기별 연령, 도시 크기별 가구원수가 조정변수가 된다.

# 제5절 가계금융조사의 패널설계

## 1. 가계금융조사 개요

### 가. 조사개요

가계금융조사는 통계청의 가계자산조사, 금융감독원의 가계신용조사, 한국은행의 가구패널조사를 하나로 통합하여 횡단 및 종단분석이 가능하도록 기획된 패널조사이다. 2010년 제1회 가계금융조사를 시작으로 매년 약 1만 가구에 대하여 가구구성, 자산 및 금융자산운용 계획, 부채 및 부채상환능력, 소득 및 주요지출 등 130여개 항목을 조사한다. 가계금융조사는 가계부문의 미시적 재무 건전성을 주기적으로 파악함으로써 정부의 부채 및 금리 정책, 가계신용관리, 금융시스템 발전을 위한 기본정보 제공을 목적으로 한다. 또한 가구특성별 자산과 부채의 규모·구성 및 분포를 조사하여 각종 경제, 사회, 금융정책이나 학문연구를 위한 기초자료로 제공된다. 조사방법은 PAPI이며, 원하는 가구에 한하여 인터넷조사 방식으로 자료를 수집하고 있다.

### 나. 표본설계

가계금융조사의 표본설계는 2단계 층화집락추출방법에 의해 수행된다. 2010년 1차 웨이브에서의 표본설계 내용을 살펴보면, 우선 PSU인 조사구를 선정하기 위해 전국을

시도(16개), 동·읍면부(2개), 주택유형(2개)에 의해 50개의 층으로 층화를 하였다. 그리고 각 층별 가구 수에 비례하여 표본크기를 배정하였다. 층별 조사구를 주택유형, 주택 평균 연건평으로 정렬한 후 가구 수를 기준으로 확률비례 계통추출하여 1,040개의 조사구를 선택하였다. 다음으로 SSU인 가구를 선정하고자 먼저 표본 조사구에 대한 현지 확인을 한 후 조사구마다 10가구를 계통추출하였다. 특히 1차 웨이브에서는 고소득층에 대한 과대표집을 위하여 대형아파트(전용면적 132㎡ 이상) 조사구를 80개 과대표집하였고, 조사구별로는 5가구를 계통추출함으로써 총 400가구를 과대표집하였다.

2차 웨이브에서는 패널유지율 90%를 목표로 하고, 2011년 조사 시 예상되는 불응가구를 감안하여 100개 조사구의 2,000가구에 해당하는 표본을 추가로 설계하였다. 이때 버블 세븐지역에 대하여 18개 조사구가 배정되도록 할당하고, 이를 특별조사구로 지정하였다. 7개의 특별조사구 지역은 서울특별시의 강남구, 서초구, 송파구, 양천구 목동과 경기도의 성남시 분당구, 안양시 평촌동, 용인시이다. 또한 조사구 관리를 고려하여 조사구당 가구는 동일하게 20가구를 추출하였다.

#### 다. 횡단가중치 작성

1차 웨이브의 횡단가중치는 다음과 같이 작성된다. 우선 설계가중치는 아래의 식과 같다.

$$w_{hij}^1 = \frac{S_h}{n_h m_{hi}}$$

여기서,  $h$ 는 설계층을 나타내는 첨자이고,  $i$ 는  $i$ 번째 표본조사구,  $j$ 는  $j$ 번째 표본가구를 나타낸다.  $S_h$ 는 설계당시  $h$ 층 내 전체 가구 수이고,  $n_h$ 는  $h$ 층 내 표본조사구 수이며,  $m_{hi}$ 는  $h$ 층,  $i$ 번째 조사구내 표본 가구 수이다. 즉, 일반조사구는  $m_{hi}=10$ 이고, 대형아파트조사구는  $m_{hi}=5$ 이다.

1차 웨이브에서는 불응이나 부적격 가구에 대해 가구 대체를 허용하였기 때문에 무응답 조정을 수행하지 않았다. 다음으로 시도, 동읍면부 여부, 주택유형(일반/중소형아파트/대형아파트)의 조정변수로 아래와 같이 사후층화 조정을 하였다.

$$w_{hij, g_1}^1 = w_{hij}^1 \times F(g_1), \quad F(g_1) = \frac{X_{g_1}}{\hat{X}_{g_1}}$$

여기서,  $X$ 는 모집단 추정값,  $\hat{X}$ 는 표본에서 조사된 가구의 총 가중값,  $F$ 는 사후층화 조정인자,  $g_1$ 는 대형아파트 과대표집 조정을 위한 사후층화 조정층이다.



## 2. 가계금융조사 패널설계안

가계금융조사에 대한 패널형태를 제안하기 위해 다음과 같은 7가지 사항을 주요하게 고려하여 검토하였다. 첫째, 가계금융조사는 횡단면 추정이 주된 목적이다. 둘째, 횡단면 추정과 동시에 종단면 분석도 가능해야 한다. 셋째, <표 5-17>의 해외 패널 유지율을 감안하여 각 웨이브에서의 패널 유지율을 5차 80%, 10차 70%로 예상하였다. 넷째, 종단분석 시 필요한 최소 시계열을 10년으로 가정하였다. 다섯째, 가구 무응답은 허용하고 가구 대체는 원칙적으로 허용하지 않는다. 여섯째, 국가통계로서 통계의 품질과 정확성을 확보해야 한다. 일곱째, 패널의 유지나 비용과 같은 조사의 효율성과 응답자 부담 등을 함께 고려한다.

일반적으로 패널유지 기간과 표본 수 사이에는 트레이드오프(trade-off) 관계가 있기 때문에 패널유지 기간은 ‘합리적으로 긴(reasonably long)’ 수준으로 정의해야 한다. 여기서 ‘합리적으로 긴’ 수준이라는 것은 예상 패널 유지율, 자료 이용자의 요구, 종단분석의 목적과 자료의 특성 등을 종합적으로 고려하여 판단할 사항이다. 그러므로 본 연구에서 가정한 10년의 기간은 목적이나 정책에 의해 변경될 수 있을 것이다.

<표 5-17> 해외 패널 유지율

웨이브	1	2	3	4	5	6	7	8	9	10
PSID1)	100	88.1	85.0	82.9	80.8	78.5	76.3	74.4	71.9	69.7
BHPS1)	100	87.7	81.5	79.9	76.8	77.3	76.0	74.1	72.1	70.4
SIPP2)	91.6	85.5	82.2	79.1	75.4	72.6	70.1	68.7	67.2	66.0

1) 1차 웨이브 응답완료 기준 2) 1996년 패널

가계금융조사를 위한 패널설계 형태를 제안하기에 앞서 아래에서 설명할 여러 가지의 이유로 고정패널 설계는 다른 방법들에 비해 가계금융조사에 적합하지 않다고 판단하였다. 먼저 횡단추정에 대한 측면에서 고정패널은 표본의 노후화로 인하여 횡단시점의 모집단 대표성 확보가 어렵다. 즉, 고정패널로 선정된 표본은 표본설계 당시의 모집단을 대표하며, 표본설계 이후에 변화된 모집단에서는 표본 선정이 불가능하므로 시간이 갈수록 패널 노후화 현상이 심해진다. 또한 최초 표본 설계 시점과 횡단면 추정 시점의 차이로 인해 모집단 구조 및 추정값에 차이가 발생할 수 있고, 패널손실이 심할 시에는 표본 수 감소와 편향 발생 등으로 횡단면 추정이 불가능 할 수도 있다.

다음으로 종단면 분석 측면에서는 고정패널은 장기적인 분석이 가능하지만, 패널을 얼마만큼 잘 유지할 수 있는지가 관건이다. 패널의 수명이 정해져 있지 않기 때문에 패

널이 유지되는 한 장기적인 분석이 가능하다는 매우 큰 장점도 있다. 그러나 응답자의 참여의지가 결국 조사 자체의 존폐를 결정지을 수 있을 만큼 매우 치명적으로 작용할 수 있다.

패널유지 측면에서도 고정패널은 패널 마모로 인하여 유지 및 관리가 힘들고 시간이 갈수록 추적비용이 점차 증가하게 된다. 시간이 지나면서 응답거부나 추적불능, 부적격 등으로 인한 패널 손실이 누적되면서 최초의 목표보다 표본 수가 적어질 수 있다. 게다가 패널이 장기화되면 표본가구가 점차 집락의 범위를 벗어나 지리적으로 분산되고, 그에 따르는 추적조사 비용과 업무 부담이 증가하게 된다. 결과적으로 높은 패널 유지율만이 횡단 추정 및 종단 분석에 타당한 결과를 제공할 수 있기 때문에 고정패널의 경우에는 다른 패널형태에 비하여 패널 관리에 각별한 노력이 필요하다.

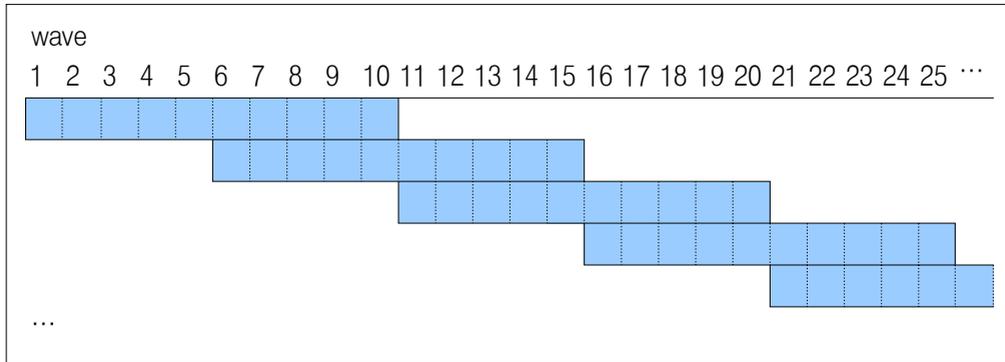
품질측면에서 고정패널은 여러 가지 편향이 발생할 수 있으나 이를 측정할 수 있는 방법이 없다. 우선 패널손실이 장기적으로 누적되면서 무응답 편향이 점차 증가할 수 있다. 또한 동일한 패널에 대한 반복조사로 응답자가 조사의 용어나 개념이 익숙해지고 조사원과의 친밀감이 형성되면서 단기적으로는 응답의 정확성이 향상될 수 있다. 그러나 장기적인 관점으로는 조사의 익숙함이 응답에 대한 태만으로 바뀌게 될 가능성이 있다. 이와 같은 조건화 효과는 경우에 따라서 긍정적일 수도 부정적일 수도 있지만, 이러한 효과의 측정이 불가능하다.

조사 효율성 측면을 살펴보면, 우선 추정식이 간단하고, 추가적인 설계 및 조사 설득이 불필요하다는 장점이 있다. 즉, 1차 웨이브에서 표본설계 및 설계가중치가 결정되기 때문에 1차 웨이브 이후에 추가되는 설계에 대한 부담이 적고, 가중치 작성 및 추정식이 복잡하지 않다. 그리고 동일한 가구를 장기간 조사하게 되므로 조사초기에 참여에 대한 가구 설득과 지속적인 참여를 유도를 통해 응답자와의 친밀감이 일단 한번 형성되면 추가되는 신규 가구에 대한 설득 작업 등이 필요 없기 때문에 상대적으로 조사 관리자의 부담을 덜 수 있을 것이다.

그러나 응답자 측면에서는 표본으로 선정된 가구의 응답부담이 매우 크다. 최초 표본으로 선택된 가구가 동일한 조사에 장기간 응답하게 되므로 표본가구에 응답부담이 크고, 해당 가구에만 부담이 과도하게 집중되는 경향이 있다. 마지막으로 이용자 측면에서는 자료에 대한 이해가 쉽고, 분석이 용이하다는 장점이 있지만, 그보다 더 중요한 자료의 신뢰성과 정확성이 먼저 보장되어야 한다.



### 가. 제안1 : 연동패널



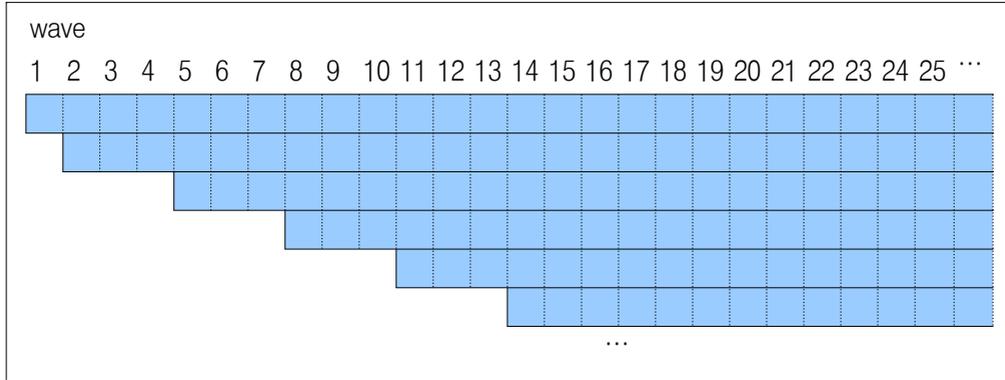
[그림 5- 11] 가계금융조사의 패널형태 제안1 : 연동패널

가계금융조사를 위한 패널형태로 제안하고자 하는 첫 번째 방법은 10년 주기 연동패널이다. 이 방법은 [그림 5-11]에서 보는 바와 같이 두개의 패널그룹을 동시에 운영하면서 두 패널은 5년씩 겹치게 조사하는 방법이다. 가계금융조사의 추출틀인 센서스 주기가 5년이므로 최신의 모집단이 확보될 때 마다 신규 표본을 설계가 가능하기 때문에 교체 주기를 5년으로 설정하였다. 또한 각 패널은 5천 가구씩 구성되어 10년 동안 유지되고 10년이 지나면 새로운 패널로 교체된다. 즉, 현재의 1만 가구 패널 중 6차 웨이브에서 5천 가구를 랜덤하게 선정하여 유지시키고 나머지는 탈락시킨 후에 5천 가구의 연동패널 그룹을 새롭게 설계하여 추가하면 된다.

이와 같은 방법의 장점을 정리해보면 다음과 같다. ① 표본가구의 응답부담이 고정패널에 비하여 경감되는 효과가 있다. ② 최신의 모집단으로부터 패널을 선정함으로써 노후화를 방지할 수 있다. ③ 횡단 추정 측면에서 패널 마모로 인한 편향을 일정 분량의 표본교체로 인하여 부분적으로 상쇄시킬 수 있다. ④ 패널의 1/2 교체로 전면교체에 비해 횡단 추정 시 시계열 단절을 줄일 수 있다. ⑤ 패널 보충이 5년마다 이뤄지므로 패널 손실에 대해 유연하게 대처가 가능하게 되어 적정 표본 수를 유지할 수 있다. ⑥ 연동계획이 표본 수에 있어 체계적이기 때문에 조사자의 업무량을 매년 유사하게 계획할 수 있다. ⑦ 패널 수명이 정해져 있기 때문에 응답자의 부담도 최대 10년이며, 패널의 추적 및 관리가 상대적으로 수월하기 때문에 비용측면에서도 효율적이다.

반대로 단점은 ① 종단분석이 가능한 기간은 최대 10년이며, 그 이상의 장기분석은 불가능하다. ② 다음번 연동그룹 교체 이전에 표본 손실이 예상보다 클 경우에는 횡단추정의 정도가 낮아질 수 있다.

## 나. 제안2 : 분할패널(고정+반복)



[그림 5- 12] 가계금융조사의 패널형태 제안2 : 분할패널(고정+반복)

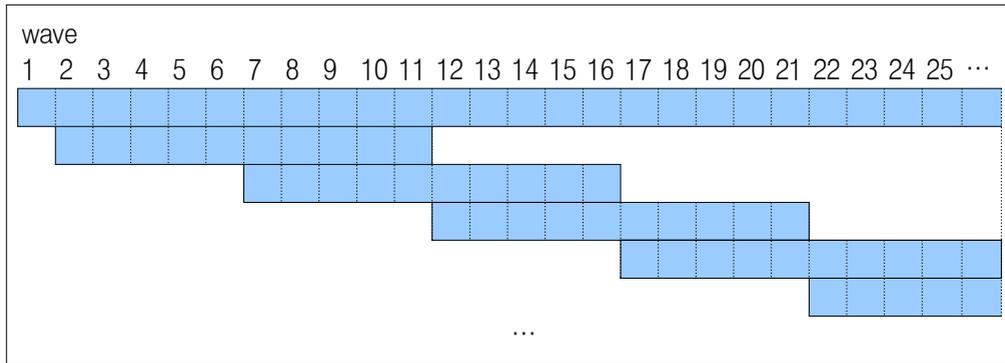
가계금융조사의 패널형태로 제안하는 두 번째 방법은 분할패널로서 [그림 5-12]와 같이 고정패널 부분을 유지해 가면서 2~3년마다 일정 비율만큼의 새로운 패널을 보충하는 형태이다. 이미 가계금융조사에서는 목표 표본 수를 달성하기 위해 2차 웨이브에 약 2,000가구 추가하였다. 이와 같은 형태에서 고려해야 할 점은 보충되는 표본의 비율과 표본을 보충을 해주는 주기이다. 먼저 보충비율은 새로운 패널분에 대한 조사구 확인, 가구명부 작성, 신규가구 접촉 등으로 인한 업무 부담이나 비용의 증가와 직결되는 사항이므로 적절한 비율 설정이 필요하다. 그리고 보충주기에 대해서는 기존의 고정패널의 유지율과 새로운 패널 보충으로 인한 표본설계 및 신규가구 조사 업무 부담, 모집단 확보 주기 등을 함께 검토하여 결정해야 할 것이다.

이와 같은 패널형태의 장점은 다음과 같다. ① 고정패널 부분에 의해 장기적인 종단 분석이 가능하다. ② 최초 표본설계 이후 변화된 모집단으로부터 설계한 최신의 보충패널에 의해 패널 노후화를 방지할 수 있으며 횡단 추정값의 정도를 제고할 수 있다. ③ 패널 보충비율을 설정하기에 따라 횡단 추정 시 시계열 단절을 최소화할 수 있다. ④ 목표 표본 수를 달성하기 위해 가장 쉬우면서도 유연한 접근 방식이다.

반면 단점으로는 ① 최초 조사를 위해 투입된 노력(가구설득, 명부작성 등)과 비용이 재조사에 비해 매우 크다면 비효율적이다. ② 만일 보충비율을 패널 유지율에 근거하여 세우게 된다면 업무량 계획을 수립하기 어려워지고, 지역 및 조사자별 업무 불균형 문제가 발생할 수 있다. ③ 장기간 패널가구로 참여하는 가구의 응답부담이 크며, 조건화 편향도 발생 가능하다. ④ 여러 개의 패널그룹을 동시에 운용해야 하므로 관리가 복잡하며 어렵고, 표본가구가 지리적으로 분산될 시 그에 따르는 추적조사 비용과 업무 부담이 증가하게 된다. ⑤ 보충되는 패널에 대한 가중치 작성방법이 복잡해지고, 기존 패널과의

통합하는 방법에 대한 연구가 필요하다.

### 다. 제안3 : 분할패널(고정+연동)



[그림 5- 13] 가계금융조사의 패널형태 제안3 : 분할패널(고정+연동)

가계금융조사를 위한 패널형태로 제안하는 세 번째 방법은 분할패널로서 [그림 5-13]과 같이 고정패널을 유지하는 동시에 10년 주기의 연동패널을 운영하는 방법이다. 이를 위해 현재 조사 중인 1만 가구의 패널 중 5천 가구를 랜덤하게 선정하여 고정패널로 유지시키고 나머지는 탈락시킨 후에 2천 5백가구의 연동패널 그룹 2개를 새롭게 설계하여 추가한다면 이와 같은 형태를 구성할 수 있을 것이다.

이러한 방법의 장점으로는 ① 고정패널 부분에 의해 부분적으로 장기적인 종단분석이 가능하고, 연동패널에 의해서도 10년간은 종단분석이 가능하다. ② 노후한 패널을 5년마다 최신 모집단으로부터 교체 가능하나 연동패널이 차지하는 비중에 대해서만 가능하므로 다소 제한적이며, 2안에 비해 패널손실에 대한 대처가 유연하지 못하다. ③ 횡단 추정 시 발생 가능한 시계열 단절을 1안에 비하여 완화시킬 수 있다. ④ 연동패널에 대한 비율이 1안에 비해 낮으므로 최초 조사에 따르는 업무 부담을 경감시킬 수 있다. ⑤ 연동패널과 고정패널의 장점을 최대한 접목시키는 방법을 고안함으로써 조사 효율과 추정의 정도를 높일 수 있다.

반면 단점으로는 ① 3개의 패널그룹을 동시에 운용해야 하므로 패널 관리가 힘들다. ② 추정식과 가중치 작성이 복잡하다.

### 3. 결론

이상으로 살펴본 바와 같이 제안된 3가지 방법들은 나름대로의 장점과 단점을 가지고 있다. 그러나 본 연구에서는 가계금융조사를 위한 패널형태로 1안을 최종적으로 제안하고자 한다. 가계금융조사의 주목적은 횡단추정이며, 응답자 부담, 패널 유지 및 관리, 정확성 및 품질 측면에서도 가장 적합한 형태로 판단된다. 그러나 만일 가계금융조사의 자료를 이용하여 10년 이상의 종단분석이 필요하다면 3안의 형태도 가능한 방법이 될 것이다.

향후 패널형태에 대한 설계가 결정이 되고 난 다음에는 횡단 및 종단 가중치 작성방법과 추정방법, 자료의 품질 등에 대한 연구가 지속적으로 필요하다.

## 참고문헌

- 강석훈(1997), “유럽의 패널조사 현황과 시사점 - 가구패널조사와 국제비교학적인 관점을 중심으로”, KLIPS 워킹페이퍼.
- 한국조세연구원(2008), 「재정패널 기초분석보고서」.
- \_\_\_\_\_ (2009), 「2차(2009)년도 재정패널 기초분석보고서(2009)」.
- \_\_\_\_\_ (2010), 「2010년 재정패널 보고서」.
- 통계청(2011), 2011년 가계금융조사 추가표본 설계보고서, 내부보고서.
- \_\_\_\_\_ (2011), 2011 가계금융조사 종합 실시 계획(안), 내부보고서.
- Bank of Italy(2010), “Household Income and Wealth in 2008”, Supplements to the Statistical Bulletin.
- Elena Gouskova, Steven Heeringa, Katherine McGonagle, Robert Schoeni and Frank Stafford(2008), “Panel Study of Income Dynamics Revised Longitudinal Weights 1993-1995”, Panel Study of Income Dynamics Technical Paper Series, University of Michigan.
- Greg J. Duncan and Graham Kalton(1987), “Issues of Design and Analysis of Surveys Across Time”, International Statistical Review, Vol. 55, No. 1, pp.97-117.
- Ivan Faiella and Romina Gambacorta(2007), “The weighting process in the SHIW”.
- Jean-Francois Naud(2004), “Combined-panel longitudinal weighting Survey of Labour and Income Dynamics 1996-2002, Income Research Paper Series.
- Katherine McGonagle and Robert Schoeni(2006), “The Panel Study of Income Dynamics: Overview and Summary of Scientific Contributions After Nearly 40 Years”, Panel Study of Income Dynamics Technical Paper Series, University of Michigan.
- Marcia Freed Taylor, John Brice, Nick Buck and Elaine Prentice-Lane(2010), 「British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices」, Colchester: University of Essex.
- Nicholas Buck, Jonathan Burton, Heather Laurie, Peter Lynn and S C Noah Uhrig(2006), 「Quality Profile: British household Panel Survey」, Colchester: University of Essex.
- Olympia Bover(2004), “The Spanish Survey of Household Finances(EFF): Description

and Methods of the 2002 Wave”.

\_\_\_\_\_ (2008), “The Spanish Survey of Household Finances(EFF): Description and Methods of the 2005 Wave”.

Statistics Canada(2010), Survey of Labour and Income Dynamics: 2008 Survey Overview, <http://www.statcan.gc.ca/pub/75f0011x/75f0011x2010001-eng.htm>.

Sylvie LaRoche(2007), “Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics”, Income Research Paper Series.

Sylvie Michaud, Lecily Hunter and Georges Lemaitre(1992), “Sample Rotation for the Survey of Labour and Income Dynamics”, Income Research Paper Series.

U.S. Bureau of the Census(1998), 「SIPP Quality Profile」 .

Westat(2001), 「Survey Of Income and Program Participation Users’ Guide」 .



## <부 록>

<부표 5-1> 가구패널 사례 비교

	PSID	BHPS	제정패널	SIPP	SLID	SHIW	SHF
주관	미시간대학교 연구소(SRC)	에섹스대학교 연구소(ISER)	한국조세연구원	미국 센서스국	캐나다 통계청	이탈리아 은행	스페인 은행
시행	"	"	TNS Korea	"	"	전문기관 위탁	전문기관 위탁
시작년도	1968	1991	2008	1984	1993	1989	2002
조사주기	1년(~1997) 2년(1999~)	1년	1년	4개월	1년	2년	3년
조사내용	주택, 고용, 소득, 건강, 결혼, 연금, 복지, 저축 등	가구구성, 노동력, 소득, 재산, 건강, 그 외 사회적 경제적 변수	과세, 환급금, 보조금, 가계지출, 자산, 부채, 경제활동, 소득공제 등	(핵심) 노동, 프로그램참여, 소득, 자산, 사회적보호(주제) 자산, 부채, 임학, 결혼, 출산, 이주, 장애 등	노동, 교육, 가족 관계, 소득 등 * 소득은 소득세 자료 접근에 대한 동의를 얻어 직접 조사 대신 행정자료 이용(80%)	소득, 재산, 노동, 교육	실물자산, 채무, 부채, 금융자산, 고용, 소득, 소비, 저축, 보험, 연금 등
조사방법	CATI	PAPI → CAP(9차웨이브)	PAPI	CAI	CATI	CAPI(79.5%) PAPI(20.5%)	CAPI
표본크기	5,000 Families	5,000여 가구	5,000여 가구	40,000여 가구	35,000여 가구	8,000여 가구	7,000여 가구
패널형태	고정패널	고정패널	고정패널	연동패널 4년 단일패널	연동패널 6년 중복패널	분할패널 전년도 표본의 일부(30~45%)+패널보충	분할패널 전년도 표본 유지+패널보충

〈부표 5-1〉 가구패널 사례 비교(계속)

	PSID	BHPS	재정패널	SIPP	SLID	SHIW	SHF
과대표집	(저소득층) SEO표본에서 2,000가구	(저소득층) 1차 웨이브 이후 표본보충	(고소득층) 부동산가격정보 300가구 (저소득층) 기초수급자통계 300가구	(저소득층) 센서스 정보 각 PSU에서 고소 득층에 비해 1.6배 높게 선정	없음	없음	(고소득층) 재산세 정보 상위1% 고소득자 들은 랜덤표본에 비해 약 9배 높게 선정
표본설계	2단계 집락추출	2단계 집락추출 (PSU)ZIP집락 (SSU)주소 PSU당 33개 주소	2단계 집락추출 (PSU)조사구 (SSU)가구 조사구당 5가구	2단계 집락추출 (PSU)도시 (SSU)주소집락	LFS표본 부표집 2개의 패널이 3년 씩 겹침 *LFS표본은 1/6연 동표본(54,000가 구)으로 총 6개의 연동그룹 중 1,2월 에 교체되는 2개 의 연동그룹을 패 널로 선정	2단계 추출 (PSU)도시 (SSU)가구	2단계 집락추출 (PSU)도시 (SSU)가구
표집틀	SEO표본 + 미시 간대학교 SRC 자 체 표집틀	우편번호 주소 파 일(PAF)	2005 인종 90% 조 사구 + 신축 아파 트 단지	5개의 표집틀 센서스 조사구(단 위, 지역, 커버리 지 보완) + 시설 조사구 + 신규 건 축물		등록자료	도시 등록부 + 국 세청의 재산과 소 득 정보
단위무응 답 조정	-응답확률을 다항 로지스틱 모형으 로 추정	(가구)지역, PSU 의 사회경제적층 으로 무응답 층 구성 (개인)로지모형을 가 이용하여 성향 가 중치에 의한 조정	로지스틱 회귀모 형의 예측응답확 률에 의한 조정	무응답 조정 증별 가중치 조정	-로지스틱 회귀모 형으로 변수선택 -CHAID로 무응답 층 구성 -무응답 조정 증별 가중치 조정	(황단)무응답 조정 증별 가중치 조정 (중단)로지스틱 모 형의 응답성향 추 정에 의한 조정	무응답 조정 증별 가중치 조정

<부표 5-1> 가구패널 사례 비교(계속)

	PSID	BHPS	재정패널	SIPP	SLID	SHIW	SHF
횡단 가중치	중단가중치와 구분없음	-초기가중치 -무응답 조정 -사후층화 조정 -가중치 표준화	-설계가중치 -가구형태 분류 -무응답 조정 -칼리브레이션 (Raking Ratio)	-초기가중치 -부표집 조정 -이주자 조정 -무응답 조정 -칼리브레이션 (Raking Ratio)	-초기가중치 -무응답 조정 -주 이주 조정 -패널조합 -가중치 공유 -영향력이 큰 관 측치 조정 -칼리브레이션	-설계가중치 -무응답 조정 -패널마모 조정 -자기상관 수정 -사후층화 조정	-초기가중치 -무응답가중치 -칼리브레이션 (Linear distance function)
중단 가중치	-초기가중치 -패널마모 조정 -가중치 표준화	-초기가중치 -패널마모 조정 -응답거부 조정 -사후층화 조정	-초기가중치 -무응답 조정 -칼리브레이션 (Raking Ratio) -극단치 제한	-초기가중치 -무응답 조정 -칼리브레이션 (Raking Ratio)	-초기가중치 -중단 개인분류 -무응답 조정 -영향력이 큰 관측 값 조정 -칼리브레이션 (GREG)	-초기가중치 -패널포함확률 조 정 -무응답 조정 -사후층화 조정	-초기가중치 -가중치 통합(패널 +비패널) -칼리브레이션 (Linear distance function)
패널손실 (유지율)	(2차) 88.1% (5차) 80.8% (10차) 69.7%	(2차) 87.7% (5차) 76.8% (10차) 70.4%	(2차) 87.67% (3차) 93.67%* *2차 성공 기준	(90패널) 78.7%* (93패널) 73.1%* (96패널) 64.5%* *최종웨이브	(99패널) 76.5%* *5차 웨이브	(08패널) 56.1%* 패널 79.3%, 비패널 41.6% *응답률	(05패널) 44.3%* 패널 61.0%, 비패널 38.5% *응답률
비고	이주민에 대한 표 본 관리	각 단계마다 분산 증가 여제를 위해 가중치의 크기를 제한	2차 웨이브에서 패 널규모 유지를 위 해 단위무응답발생 시 가구 대체 (620 가구)	-96년 패널설계 변 경 (32개월 중복패 널 → 48개월 단일 패널) -개인, 가구, 가족 가중치 별도제공	-중단가중치 통합 (패널중복기간 3년) -횡단가중치 두 중 류(통합, 노동)	-패널마모를 보완 하기 위해 표본을 추가하는 형태 -분산추정(잭나이 프)	-표본대체는 엄격 한 규칙을 적용하 여 허용(4개가구, 소득정보 이용)