http://www.kli.re.kr/klosa

KLoSA
Korean Longitudinal Study of Ageing

고령화연구패널조사(KLoSA)

2008년 제2차 기본조사 사용자안내서

한국노동연구원
KOREA LABOR INSTITUTE

머리말

한국노동연구원은 압축적 고령화를 겪고 있는 우리나라의 고령화 과정을 파악하고, 이에 대비한 정책 수립과 학술 연구에 활용하기 위하여, 45세 이상 중고령자를 대상으로 하는 패널자료를 구축하기로 하였다. 이 패널조사의 명칭은 '고령화연구패널조사(KLoSA: Korean Longitudinal Study of Ageing, 클로사)'로 명명하였다. 2006년에 만 45세이상 10,254명의 패널 구축과 함께 제1차년도 기본조사를 마쳤고, 2007년 직업력 조사를 완료하였으며, 2008년 12월 제2차 기본조사 실사를 완료한 후 2009년 7월 1일 제2차 베타 버전 데이터를 공개하게 되었다.

고령화연구패널조사(이하 KLoSA)는 조사의 내용과 방법, 양 측면에서 중요한 의의를 지닌 조사이다. 내용면에서는, 빠르게 진행되는 우리나라 고령화에 대한 정책 및 다학제적 연구를 위한 기초자료로서 의의가 있다. KLoSA의 조사 내용은 사회적, 경제적, 육체적, 정신적 삶의 여러 측면을 다루고 있어서 사회학, 경제학, 보건의료학, 사회복지학, 노년학, 심리학, 가족학, 인구학 등 다양한 학문 분야를 포괄하고 있다. 또한 미국과 유럽 여러 나라의 고령자대상 패널자료와의 비교연구를 염두에 두고 설계되었다. 조사방법 상의 의의는 우리나라에서는 처음으로 컴퓨터를 이용한 대인면접(CAPI) 방식으로 수집한 패널자료라는 점에서 찾을 수 있다. 우리나라에서도 컴퓨터를 이용한 자료수집은 전화조사(CATI)나 웹조사(CAWI) 등의 형식으로 많이 활용되고 있으나, 복잡한 설문의 대인면접에서는 시도된 적이 없었다. 또한 2007년 직업력 조사는 응답자가 쉽게 기억할 수 있도록 캘린더와 CAPI를 혼합한 방식으로 면접을 실시하여 회고조사에서의 새로운 방법을 시도하여 보았다.

본 사용자안내서는 2006년 제1차 고령화연구패널조사와 2007년 직업력 조사 그리고 2008년 제2차 기본조사 베타버전 자료에 대한 기본적인 소개와 표본추출방식, 그리고 가중치 적용방식을 설명하고 각 영역별 세부적인 내용과 주의할 점을 다루었다. 각 장별로 세부적인 내용과 구체적인 내용들은 테크니컬 리포트 형식으로 업데이트 되고 있으며, KLoSA와 관련된 여러 가지 관련 기록들은 홈페이지(http://klosa.kli.re.kr)를 통해서 공개되고 있다.

2009년 7월

한국노동연구원
고령화연구패널조사팀

목 차

Ⅰ. 고령화연구패널조사(KLoSA) 개요

Ⅱ. 표집과 가중치

Ⅲ. 다중대체(Multiple Imputation) 방법과 자료사용법

Ⅳ. 데이터 이용방법 / 47

Ⅴ. 2006년 제1차 기본조사 자료의 세부적 내용과 특징

Ⅵ. 2007년 직업력 조사

Ⅶ. 2008년 제2차 기본조사 자료의 세부적 내용과 특징 / 77

I

고령화연구패널조사 개요

1 배경 및 목적

고령화연구패널조사는 빠르게 진행되고 있는 우리나라의 인구고령화에 대비하기 위해서는 정책 및 제도연구가 필요하고, 이런 연구를 위해서 고령자의 실태와 행위양식에 관한 기초자료의 축적이 선행되어야한다는 문제의식에서 출발하였다. 따라서 '고령화연구패널조사'의 기본 목적은 우리나라 중고령인구의 경제활동에 대한 정확한 실태조사를 통해 향후 고령사회로 변화해 가는 과정에서 개인의 행동을 예측하고 이를 토대로 효과적인 사회경제 정책을 수립하고 시행하는 데 활용될 기초자료 생산에 있다. 특히, 중고령자의 일회성 횡단면 조사보다는 은퇴 전후의 소득 변화를 비롯하여 사회 제도 및 정책 변화가 개인의 의사결정에 미치는 효과를 시계열적으로 포착할 수 있는 패널자료의 구축이 필요하다고 판단하였다.

2 조사설계 상의 특징

이 조사를 설계함에 있어서 두 가지를 특히 유의하였다. 첫째, 고령화 관련 자료는 노동분야 관련 연구 뿐 아니라 다양한 학제적 연구에 널리 활용될 수 있어야한다는 점이다. 고령화연구패널조사(KLoSA)는 중고령자의 고용 현황과 소득수준 및 자산규모에 관한 정보뿐만 아니라 가족관계와 건강, 주관적인 의식세계를 파악하는 항목들을 포함시킴으로써 노동경제학, 사회학, 사회복지학, 보건학, 가족학, 노년학 등 다양한 분야의 연구자들이 활용할 수 있는 자료를 생산하고자 하였다. 둘째, '고령화연구패널조사'는 미국의 HRS: Health and Retirement Survey 자료를 벤치마킹하고, 영국(ELSA: English Longitudinal Study of Ageing), 유럽(SHARE: Survey of Helath, Ageing and Retirement in Europe)등 이미 고령자 패널조사를 실시하고 있는 선진국들과 수 차례의 자문회의를 통해 이들 국가간 비교연구가 가능하도록 설계하고자 하였다. 우리나라의 고령자의 실태는 다른 나라 고령자와의 비교연구를 통해서 그 특성이 더욱 분명하게 드러날 것이기 때문이다.

3 조사 대상

고령화연구패널조사의 대상자는 대한민국 제주도를 제외한 지역에 거주하는 45세(1962년 이전생) 이상의 중고령자 개인을 대상으로 표본수 약 10,000명을 목표로 조사하였다. 2006년 고령화연구패널 기본조사의 대상자는 일반가구 거주자를 대상으로 표집 및 조사를 실시하였다.

해외 고령자 패널조사는 50세 이상을 대상으로 하고 있으나, KLoSA에서 조사대상을 45세 이상 중고령 연령층으로 확대한 이유는 1990년대 말 외환위기 이후 40대 중반부터 주된 일자리에서 퇴직하는 경향이 심화됨에 따라 중년 연령기의 주된 일자리 변동이 매우 중요한 사회적 이슈로 떠올랐기 때문이며, 나아가 중년기의 경제활동이 고령기의 일자리나 노후생활과 갖는 관계를 살펴보는 데에도 유리하기 때문이다.

2006년 제 1차 기본조사 결과 10,254명의 패널을 구축하게 되었고, 이들 패널을 대상으로 짝수년 해는 기본조사를 홀수년 해에는 특정한 주제를 다룬 조사를 실시하게 된다. 2007년에는 이들 10,254명의 패널을 대상으로 직업력 조사를 실시하였다.

4 주요 설문 내용

◆ 2006년 제1차 기본조사의 주요조사 내용은 다음의 <표 I-1>과 같다.

<표 I-1> 2006년 제1차 기본조사 영역별 주요 내용

세부 영역	주 요 내 용
CV. 커버스크린	• 가구원에 대한 정보(성별, 태어난 해, 혼인상태, 응답자와의 관계 등)
A. 인구	• 응답자의 생년월일, 성별, 학력, 종교, 혼인상태 • 배우자 정보(사별, 실종, 별거 등의 이유로 응답자와 비동거 또는 배우자가 45세 미만인 경우): 배우자의 생년월일, 학력, 고용상태 • 종교, 사회적 관계 및 활동
Ba. 가족 (자녀와 손자녀)	• 자녀: 성별, 연령, 학력, 근로상태, 주택소유, 결혼상태, 자녀 수 • 비동거 자녀인 경우: 응답자와의 근접성, 접촉빈도, 자녀에게 받은/준 경제적 도움 • 손자녀: 손자녀수, 손자녀 보살핌 여부, 돌봄노동시간, 돌봄노동기간
B. 가족(부모와 형제자매, 기타 가족수발)	• 부모: 생존여부(사망시 당시 나이), 연령, 학력, 근로상태, 주택소유여부 • 비동거 부모: 응답자와의 근접성, 접촉빈도, 부모에게 받은/준 경제적 도움 • 형제자매: 응답자와의 관계, 연령, 결혼상태 • 기타가족: 기타가족에게 받은/준 경제적 도움 • 가족중 ADL/ IADL이 어려운 사람 여부, 수발시간, 수발기간
Ca. 건강상태	• 주관적 건강상태 • 만성질환 진단여부, 진단시기, 치료, 그로 인한 일상생활 어려움 • 교통사고/낙상/골절 경험

세부 영역	주 요 내 용
	• 노후징후(시력,청력,치아상태, 통증) • 신체계측(BMI), 건강관련습관(운동, 영양, 흡연, 음주) • 우울증 측정
Cb. ADL/IADL, 수발노동	• 일상생활수행능력(ADL/IADL)상태 • 일상생활수행에 도움이 필요한 응답자의 경우, 수발자 정보(3명까지): 응답자와의 관계, 간병일수 및 시간, 간병비용 지불여부 등 • 현재 수발도움이 필요 없지만, 미래에 가능한 간병수발자 예측
Cc. 의료보장과 시설이용	• 건강보험이용: 건강보험, 의료급여, 민간의료보험의 가입여부, 건강보험료 부담자, 건강보험료, 체납여부, 체납기간, 건강검진 여부 등 • 의료시설이용: 입원, 치과, 보건소, 한방병원, 기타 외래진료 및 왕진, 처방약, 의료기구 구입여부 및 본인 지불 비용 등
Cd. 인지력	• K-MMSE한국형 인지능력 측정 척도 사용
Ce. 신체기능 측정	• 악력 측정
D. 고용(근로형태 구분) 문항번호 D001-D010	• 10개의 문항을 통해 현재 근로형태 구분
D. 고용(임금근로자) 문항번호 100-200번대	• 개인 일자리에 대한 세부사항: 사업장 정보, 고용형태, 직종, 직위, 근로계약, 근로시간, 정기휴무, 임금결정방식 4대 사회보험가입 및 복리후생 및 노동조합 관련 등 • 일자리에 대한 인식 및 만족도, 향후 희망하는 일자리 등
D. 고용(자영업자) 문항번호 300번대	• 자영업에 대한 세부사항: 자기사업을 하는 이유, 사업장 정보, 근로시간, 일수, 근로시간에 대한 인식, 정기휴무일, 일자리에 대한 인식 및 만족도, 향후 희망하는 일자리 등
D. 고용(무급가족종사자) 문항번호 400번대	• 무급가족종사에 대한 세부사항: 무급가족종사를 하는 이유, 사업장 정보, 근로자 현황, 사업장 매출액, 근로시간, 근로일수, 정기휴무일, 일자리에 대한 인식 및 만족도, 향후 희망하는 일자리 등
D. 고용(구직자) 문항번호 500번대	• 구직에 대한 세부사항: 1주일/1개월 기준 구직 여부, 구직의사, 취업가능성, 취업을 하지 않는 이유, 일자리를 찾는 이유, 구직활동, 구직활동의 어려움, 은퇴계획, 노동경험, 가장 최근 일자리 등
D. 고용(은퇴자) 문항번호 600번대	• 은퇴자에 대한 세부사항: 은퇴시기, 은퇴이유, 은퇴 당시 배우자의 경제활동 상태, 소일거리에 대한 사항, 은퇴 및 경제활동에 대한 인식, 가장 최근 일자리 등
D. 고용(가장 최근 일자리) 문항번호 700번대	• 응답자가 구직자이거나 은퇴자인 경우 가장 최근 일자리에 대한 정보
E. 소득	• 근로소득: 임금, 자영업, 농어업, 부업 등 • 연금소득: 국민연금, 특수직역연금, 개인연금 등 • 사회보장소득: 사회보장소득여부 및 종류 • 기타 수입 및 소득 • 지난 1년간 가구총소득
F. 자산	• 현재 거주주택 정보 • 거주주택 외 부동산 자산: 부동산 소유, 임대/임차, 사업체/농장 • 금융자산(곗돈포함) • 기타 비금융자산 • 상속/증여 • 부채 • 가구 총자산

세부 영역	주 요 내 용
G. 주관적 기대감	• 경제적 상황에 대한 주관적 판단: 유산증여/상속, 예상근로시간, 앞으로의 근로활동 등 • 연령별 기대수명 • 생활수준의 변화 및 정부에 대한 기대감 • 삶의 만족도

※ 고령화연구패널조사 제1차 기본조사 1.0 버전 자료에서는 CV 영역에서 조사한 정보는 제공하지 않는다. 그 이유는 2006년 제1차 기본조사는 패널구축 단계이므로, 해당 가구에 대상자가 살고 있는지 아닌지 여부를 묻고, 대상자가 살고 있지만 부재중인 경우 대상자가 아닌 가구원이 CV영역에 응답을 했기 때문에 CV의 응답자에 따라서 가구원의 구성이 달라지기 때문이다. 즉 CV 영역 응답자와 패널 대상자가 일치하지 않아 자료로 제공하지 않는다. 그러므로 패널 구축이후 제2차 기본조사에서부터 CV 영역의 자료를 제공한다.

◆ 2007년 직업력조사의 주요내용은 다음의 <표 I-2>과 같다.

<표 I-2> 2007년 직업력 조사의 주요내용

세부 영역	주 요 내 용
1. 월급을 받는 상시 임금 근로자	• 근로기간, 직산업분류, 사업장 규모, 퇴직이유 • 45세이후에 해당한 경우: 퇴직1년전 월평균 급여 • 45세때 해당한 경우: 사업장위치, 근로시간, 퇴직금, 국민연금가입년도
2. 일당을 받는 일용 임금 근로자	• 근로기간, 직산업분류, 월평균 근로일, 년평균 근로월수 • 45세이후에 해당한 경우: 평균 일당 • 45세때 해당한 경우:사업장 위치, 근로시간
3. 점포가 있는 자영업자	• 운영기간, 직산업분류, 고용임금근로자정보, 무급가족종사자규모, 사업을 그만둔 이유 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 사업장 위치, 주평균 근로일과 시간, 자기사업을 택한 이유
4. 점포가 없는 자영업자	• 운영기간, 직산업분류, 월평균 근로일, 년평균 근로월수,사업을 그만둔 이유 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 사업장 위치, 주평균 근로시간
5. 농,축,임, 어업 종사자	• 종사기간, 직산업분류, 고용임금주로자, 무급가족종사자 규모, 사업을 그만둔 이유 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 사업장 위치, 주평균 근로시간, 자기사업을 택한 이유
6. 무급가족종사자	• 종사기간, 직산업분류, 사업장 대표, 고용임금근로자정보, 무급가족종사자 규모, 그만둔 이유 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 사업장 위치, 주평균 근로시간, 무급가족종사일을 택한 이유
7. 동시에 다양한 일을 한 경우	• 근로기가, 직산업 분류, 월평균 근로일, 년평균 근로월수 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 일평균 근로시간
8. 구직	• 비근로기간

세부 영역	주 요 내 용
9. 가사 10. 요양 11. 교육 12. 군대 13. 기타	• 생계유지방법 • 소득원(동거가족 중 누구의 소득원) • 생계지원(비동거 가족의 지원인 경우) • 월평균 수입

※ 2007년 직업력조사의 자료는 두 가지이다. 우선 일자리특성을 나타내는 자료에서는 <표 Ⅰ-2>의 내용을 중심으로 한 자료이다. 다음으로 개인특성을 나타내는 자료가 있는데 이것은 간단한 개인특성과 15세부터 105세까지의 패널들이 각 연령별 근로여부를 더미변수로 새로 생성한 자료이다.

◆ 2008년 제2차 기본조사의 주요조사 내용은 다음의 <표 I-3>와 <표Ⅰ-4>와 같다. <표Ⅰ-3>은 생존자를 대상으로 한 기본조사 설문 내용이고, <표Ⅰ-4>는 패널 대상자 중 사망자에 대한 대리 인터뷰로 이루어진 Exit Interview 설문 내용이다.

<표 I-3> 2008년 제2차 기본조사 영역별 주요 내용(생존자 대상)

세부 영역	주 요 내 용
CV. 커버스크린	• 가구원에 대한 정보(성별, 태어난 해, 혼인상태, 응답자와의 관계 등) • 2006년 제1차 기본조사 이후 이사한 경우 관련 정보(년월, 이유)
A. 인구	• 응답자의 생년월일, 성별, 학력, 종교, 혼인상태 • 배우자 정보(사별, 실종, 별거 등의 이유로 응답자와 비동거 또는 배우자가 45세 미만인 경우): 배우자의 생년월일, 학력, 고용상태 • 종교, 사회적 관계 및 활동
Ba. 가족 (자녀와 손자녀)	• 자녀: 성별, 연령, 학력, 근로상태, 주택소유, 결혼상태, 자녀 수 • 비동거 자녀인 경우: 응답자와의 근접성, 접촉빈도, 자녀에게 받은/준 경제적 도움 • 동거 자녀: 자녀에게 받은/준 경제적 도움 • 손자녀: 손자녀수, 손자녀 보살핌 여부, 돌봄노동시간, 돌봄노동기간
B. 가족 (부모와 형제자매, 기타 가족수발)	• 부모: 생존여부(사망시 당시 나이), 연령, 학력, 근로상태, 주택소유여부 • 비동거 부모: 응답자와의 근접성, 접촉빈도, 부모에게 받은/준 경제적 도움 • 형제자매: 응답자와의 관계, 연령, 결혼상태 • 기타가족: 기타가족에게 받은/준 경제적 도움 • 가족중 ADL/ IADL이 어려운 사람 여부, 수발시간, 수발기간
C. 건강	• 주관적 건강상태 • 만성질환 진단여부, 진단시기, 치료, 그로 인한 일상생활 어려움 • 교통사고/낙상/골절 경험 • 노후징후(시력,청력,치아상태, 통증) • 신체계측(BMI), 건강관련습관(운동, 영양, 흡연, 음주) • 우울증 측정 • 일상생활수행능력(ADL/IADL)상태 및 수발노동 • 의료보장과 시설이용관련 • 건강보험이용 관련:건강보험, 의료급여, 민간의료보험의 가입여부, 건강보험료 부담자, 건강보험료, 체납여부, 체납기간, 건강검진 여부 등 • 의료시설이용: 입원, 치과, 보건소, 한방병원, 기타 외래진료 및 왕진, 처방약, 의료기구 구입여부 및 본인 지불 비용 등 • K-MMSE한국형 인지능력 측정 척도 사용

세부 영역	주 요 내 용
	• 악력 측정
D. 고용(근로형태 구분) 문항번호 (D001-D085)	• 제1차 기본조사 당시 확인 • 제1차 기본조사 이후 일자리 변동사항
D. 고용(임금근로자) 문항번호 100번대	• 신규 임금근로자 일자리 특성 • 현재 일자리에 대한 세부사항: 사업장 정보, 고용형태, 직종, 직위, 근로계약, 근로시간, 정기휴무, 임금결정방식 4대 사회보험가입 및 복리후생 및 노동조합 관련 등 • 일자리에 대한 인식 및 만족도, 향후 희망하는 일자리 등
D. 고용(자영업자) 문항번호 200번대	• 신규 자영업자 일자리 특성 • 자영업에 대한 세부사항: 자기사업을 하는 이유, 사업장 정보, 근로시간, 일수, 근로시간에 대한 인식, 정기휴무일, 일자리에 대한 인식 및 만족도, 향후 희망하는 일자리 등
D. 고용(무급가족종사자) 문항번호 300번대	• 신규 무급가족종사 일자리 특성 • 무급가족종사에 대한 세부사항: 무급가족종사를 하는 이유, 사업장 정보, 근로자 현황, 사업장 매출액, 근로시간, 근로일수, 정기휴무일, 일자리에 대한 인식 및 만족도, 향후 희망하는 일자리 등
D. 고용(구직/비근로자) 문항번호 400번대	• 구직 및 비근로자에 대한 세부사항: 1주일/1개월 기준 구직 여부, 구직의사, 취업가능성, 취업을 하지 않는 이유, 일자리를 찾는 이유, 구직활동, 구직활동의 어려움, 은퇴계획, 노동경험, 가장 최근 일자리 등
D. 고용(은퇴자) 문항번호 500번대	• 신규 은퇴자 관련 • 은퇴자에 대한 세부사항: 은퇴시기, 은퇴이유, 은퇴 당시 배우자의 경제활동 상태, 소일거리에 대한 사항, 은퇴 및 경제활동에 대한 인식, 가장 최근 일자리 등
D. 고용(소일거리) 문항번호 600번대	• 소일거리 관련
E. 소득 및 소비(소득) 문항번호 100번대	• 근로소득: 임금, 자영업, 농어업, 부업 등 • 연금소득: 국민연금, 특수직역연금, 개인연금 등 • 사회보장소득: 사회보장소득여부 및 종류 • 기타 수입 및 소득 • 지난 1년간 가구총소득
E. 소득 및 소비(소비) 문항번호 200번대	• 소비: 생활비, 식비, 외식비, 공교육비, 사교육비, 주거비, 보건의료비, 피복비 • 생활비 부족분 충당방법 • 저축액
F. 자산	• 현재 거주주택 정보 • 거주주택 외 부동산 자산: 부동산 소유, 임대/임차, 사업체/농장 • 금융자산(곗돈포함) • 기타 비금융자산 • 상속/증여 • 부채 • 가구 총자산
G. 주관적 기대감 및 삶의 질	• 국민연금 및 특수직역연금: 가입 및 납부여부, 예산 수혜시점, 예상 금액수준 • 기초노령연금: 신청 및 향후 신청여부, 수급여부, 수급액수 • 경제적 상황에 대한 주관적 판단: 유산증여/상속, 예상근로시간, 앞으로의 근로활동 등 • 연령별 기대수명

세부 영역	주 요 내 용
	• 생활수준의 변화 및 정부에 대한 기대감 • 삶의 만족도

<표 I-4> 2008년 제2차 기본조사 Exit Interview 영역별 주요 내용(사망자 대상)

세부 영역	주 요 내 용
Xa. 대리응답자 및 사망자 기본정보	• 대리응답자의 기본정보: 사망자와의 관계, 교류 정도 • 사망자의 당시 상황: 사망 원인, 사망일, 당시 상황 등
Xc. 건강	• 사망자의 당시 건강상태 • 사망자의 당시 인지기능 • 사망자의 당시 증상 및 건강행위 • 사망자의 당시 기능장애와 간병 • 사망자의 당시 의료이용 • 장례 및 유족의 후유증세
Xd. 고용	• 마지막 조사 이후의 근로여부 • 사망자의 마지막 일자리 정보
Xf. 유산과 부채	• 유산(현재 살고 있는 집, 그 외 부동산, 금융, 보험 그 외 기타 자산) • 부채 • 유언과 상속

5 조사 주기 및 기간

패널조사는 하나의 현상을 일정한 시간적 간격을 두고 반복적으로 측정하여 시간의 흐름에 따라 그 변화를 살펴보고자 하는 것이다. 고령화연구패널조사는 2006년을 시작으로 매 2년 간격으로 짝수 년도에는 반복적으로 측정할 기본적인 사항을 조사하는 '기본조사'를 수행한다. 매 홀수년도에는 필요한 경우 기본조사에 포함되지 않는 내용을 중심으로 특정한 주제를 정하여 실시할 예정이다.

6 경 과

한국노동연구원은 급속하게 진행되고 있는 우리나라 고령화에 관한 정책 및 학술 연구에 활용할 기초자료를 마련하고자 2005년부터 중고령자를 대상으로 하는 패널조사를 실시하고자 기초연구를 진행해 왔다. 이 패널조사의 우리말 명칭으로는 '고령화연구패널조사'를, 영문으로는 'Korean Longitudinal Study of Ageing'을, 영어 약칭으로는 'KLoSA(클로사)'로 하였다.

고령화연구패널조사 사업을 실시하게 된 그간의 경과는 다음과 같다.

- 고령화가 초래할 사회적 변화에 대한 관심이 높아지면서 2004년 초 당시 '대통령자문 고령화 및 미래사회위원회'에서 사업의 필요성을 인정하고, 노동부가 2005년부터 한국노동연구원에 고용보험기금 출연금을 지원하여 고령화에 관한 패널조사 사업을 추진키로 확정하였다. 2005년 초부터는 패널조사를 위한 기초연구가 본격적으로 시작되었다.

- 고령화연구패널조사 제1차 기본조사를 실시하기 전 기초연구를 진행하면서 3차례에 걸친 예비조사를 실시하였다. 제1차 예비조사(2005년 10월)에서는 면접 현장에서 검증이 필요하다가 판단된 설문 문항과 척도를 중심으로 중고령자 약 500여명의 표본을 대상으로 인쇄된 설문지를 이용하여 면접조사를 실시하였다. 제2차 예비조사(2006년 2월)는 서울 지역에 거주하는 중고령자 30명을 대상으로 실시하였는데, 우리나라 패널조사에서 최초로 시도하는 컴퓨터를 이용한 대인면접방식(Computer Assisted Personal Interviewing:CAPI)이 면접원과 응답자간의 실제 환경에서 가능한지를 판단하는 것이 주된 목적이었다. 제3차 예비조사(2006년 4월~5월)는 인구주택총조사(통계청)의 조사구를 기본틀로 이용하여 표본을 추출하고, 면접원이 대상 가구를 접촉하는 과정 등 전과정을 제1차 기본조사에서 실시하는 것과 똑같은 과정으로 수행하여 최종적인 테스트를 마치게 되었다.

- 이 과정에서 해외의 고령자패널과의 긴밀한 연계가 있었다. 미국의 HRS, 영국의 ELSA, 유럽의 SHARE 책임자들과 2005년 서울, 2006년 미국 나파밸리에서 두 차례에 걸친 자문회의를 통해 CAPI 조사에 대한 노하우와 고령자패널들의 국제비교를 위한 데이터를 고민을 함께 하였다. 또한 실무진들의 미국의 HRS Super visor 면접원 교육에 참여, Blaise CAPI Conference 참여 등을 통해 패널조사에 있어서 새로운 기술을 도입할 수 있었다.

- 고령화연구패널조사 제1차 기본조사 실사(fieldwork) 기간은 2006년 8월부터 12월까지 약 5개월이 소요되었다. 제1차 기본조사 실사를 마치고, 간단한 1차 데이터 클리닝 과정을 거쳐 2007년 3월 19일 '고령화연구패널 제1차 기본조사 데이터 베타(Beta) 버전'을 출시하여 고령화 연구에 관심을 가지고 있는 많은 연구자들이 우선 사용해 보도록 하였다. '베타(Beta)버전' 이후에 '제2차 데이터 클리닝 과정'과 'Multiple Imputation(무응답에 대한 보정)'을 거쳐 2007년 11월 '2006년 고령화연구패널 제1차 기본조사 데이터 1.0버전'을 발표하게 되었다.

- 2007년에는 홀수년도로 특정한 주세로 조사하기 위한 주제로 '생애사 조사'를 설계하였다. 그러나 예비조사 결과 전체 설문시간과 조사에 있어서 영역별 신뢰도 문제가 발생하여 생애사 조사의 영역중 '직업력' 한 영역만을 집중 조사하기로 결정하였다.

- 그러므로 2007년 8월부터 2008년 1월까지 직업력 조사가 실시되었다. 직업력 조사에서 응답자의 근로기간에는 종사상 지위별 세분화하여 그 직업별 특징을 파악하였고, 비근로기간에는 어떠한 일을 했는지를 파악하여 전 생애에 걸친 일자리를 캘린더 형식으로 채워나감으로써 생애 주기별 특징이 반영하였다. 직업력 조사는 데이터 클리닝 과정을 거쳐 2008년 11월 '직업력 조사 데이터 1.0 버전'을 발표하게 되었다. 직업력 자료는 응답자들이 보다 다양한 연구를 할 수 있도록 두 가지 자료로 이루어졌다. 일자리특성을 나타내난 자료는 설문문항에 충실한 자료이고, 개인특성을 나타내는 자료는 설문 문항에 일부와 15세 이후부터 연령별 근로 여부를 더미처리해 준 변수로 구성하였다.

- 2008년 7월부터 11월말까지 제2차 기본조사를 실사를 마쳤고, 이 때 CAPI를 이용하여 제1차 기본조사의 개별 응답자 정보를 pre-loading 시켜 지난 조사와 차이가 크게 나는 응답을 double check 하였다. 이 과정에서 제1차 기본조사에서 잘못된 응답을 다시 수정하여 2008년 말 '2006년 고령화연구패널 제1차 기본조사 데이터 1.1버전'을 출시하였고, 2009년 7월 '제1차 기본조사 데이터 1.2 버전'으로 업데이트 하였다.

- 2008년말에 실사 완료된 제2차 기본조사는 2009년 데이터 클리닝 작업을 마치고, 2009년 7월 1일 '제2차 기본조사 데이터 베타 버전'을 발표하게 되었다.

7 조사 방법

고령화연구패널조사는 컴퓨터를 이용한 대인면접법(CAPI: Computer Assisted Personal Interviewing)을 기본으로 활용하였다. 그러나 홀수년도에 진행되는 주제에 따라서는 다른 방법이 병행될 수 있다.

- 2006년 제1차 기본조사 : CAPI
- 2007년 직업력 조사: 제1차 기본조사 정보를 이용한 개인별 캘린더 + CAPI
- 2008년 제2차 기본조사: 제1차 기본조사와 직업력조사의 개인별 내용을 Pre-lading 한 CAPI

◆ 컴퓨터를 이용한 대인면접법(CAPI)

고령화연구패널조사의 면접방식은 컴퓨터를 이용한 대인면접(CAPI)으로, 사회조사에서 전통적으로 활용하던 종이와 연필을 이용한 방식(Paper and Pencil Interviewing: PAPI)이 아니라 면접원이 노트북 컴퓨터를 지참하고 조사대상자에게 컴퓨터 화면에 나오는 질문을 읽어준 후 그 응답을 키보드나 마우스를 이용하여 직접 입력하는 방식이다. 서구 국가에서는 정부가 수행하는 통계조사를 중심으로 표본 규모가 크고 반복적으로 이루어지는 조사에서 CAPI를 활용하는 사례가 많다. 설문구조가 복잡하고 분량이 많을 뿐만 아니라 대규모 표본을 대상으로 장기간 반복적인 조사를 수행해야 하는 패널조사에 적합하다고 판단되어 CAPI 방식을 채택하였다. 프로그램은 네덜란드 통계청에서 개발하여 전세계 30여개국 공공부문에서 수행하는 조사에 널리 쓰이고 있는 블레이즈(Blaise)[1]를 사용하였다.

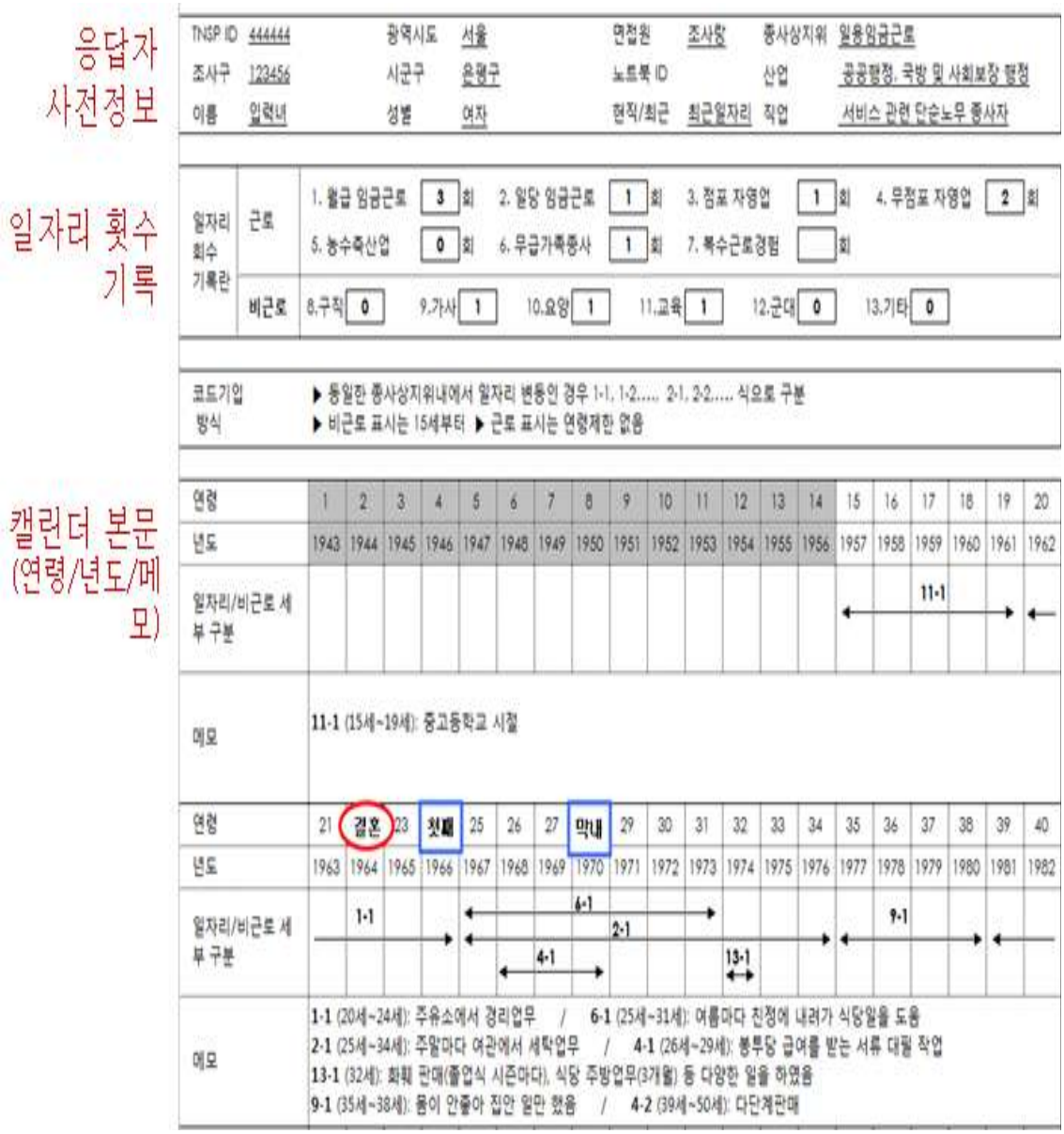
◆ 패널조사에 있어서 CAPI 조사방법의 장점

- CAPI는 설문 논리에 따라 경로를 지정하여 면접원이 해당 문항을 찾아서 기입할 필요가 없고, 응답자의 응답에 따라 컴퓨터 프로그램으로 자동으로 문항이동이 이루어지므로, 잘못된 경로로 갈 수 있는 오류를 방지할 수 있다.
- CAPI는 문항에 따라 응답범위를 사전에 지정하여 입력 오류의 가능성을 감소시키고, 문항간 일관성을 유지시켜 준다.
- CAPI에서는 필요에 따라 설문의 순서를 무작위로 배치할 수 있으며, 응답자 상황과 대답에 따른 맞춤식 설문문항으로 면접할 수 있다.
- 자동 계산 및 채점 기능을 이용하여 사용할 수 있다. 인쇄된 설문지를 이용할 경우에는 면접원이 계산기를 이용하여 계산한 후 정/오답을 판단해서 적어야 하나, 고령화연구패널조사 CAPI는 응답자가 말한 숫자를 입력하면 컴퓨터가 정/오답을 판단하여 자동으로 입력된다.

1) 1986년 네덜란드 통계청이 개발한 통계조사종합시스템으로 자세한 내용은 인터넷 홈페이지를 참조(http://www.blaise.com/onlinehelp)

◆ 2007년 직업력 조사에서 사용한 캘린더 방식

[그림Ⅰ-1] 직업력 조사에 사용된 응답자 개인별 캘린더 예시



직업력 조사는 전 생애에 걸쳐 자신의 직업을 회고하는 것이 중요하므로, 캘린더를 이용하여 응답자의 기억을 돕고 이후에 캘린더와 컴퓨터를 함께 놓고 설문을 진행하는 방식을 택했다. 이때 이 캘린더는 제1차 기본조사를 바탕으로 개인마다 기본적인 정보를 각각 담아서 응답자 개별 캘린더를 면접원이 가지고 방문하였다. 캘린더를 통해 우선 결혼, 첫째자녀, 막녀자녀 출생의 시점을 회상하고 이 시점을 기준으로 회고적으로 생애 직업력을 작성하도록 하였다. 이 캘린더 작업을 완성한 후 CAPI를 이용하여 각 직업별 설문을 진행하였다.

◆ 2008년 제2차 기본조사에서 Pre-loading을 이용한 CAPI

기본조사에서 면접방법은 CAPI 이다. 2006년 제1차 조사와 달라진 부분은 우선 GPMS(Global Panel Management System)를 이용하여 면접원들이 자신이 조사해야 할 응답자와의 연락, 약속잡기, 설문완료 등을 보다 편리하게 활용하고, 중앙에서 면접원들의 개별 사항을 관리할 수 있는 시스템을 개발하고 적용시켰다는 점이다. 또한 2006년 제1차 조사와 2007년 직업력 조사에서 수집한 개별 정보를 미리 컴퓨터에 pre-loading 시켜 기존 응답과 일정수준 이상이 차이가 나는 경우 팝업창을 이용하여 다시 확인하고 잘못 되었다면 수정할 수 있는 설문 로직을 CAPI 프로그램에 장착하였다.

8 데이터 배포

데이터, 설문지, 코드북, 유저가이드, 실사과정과 관련된 자료, 데이터를 이용한 연구논문 등은 인터넷 홈페이지(http://klosa.kli.re.kr)에 직접 접속하거나 한국노동연구원 홈페이지(http://www.kli.re.kr)에 들어가 '고령화연구패널조사' 배너를 클릭하여 무료로 제공받을 수 있다.

◆ 고령화연구패널조사 자료를 다운로드 받을 때 주의해야 할 사항

- 우선 이용자의 컴퓨터에서 **'팝업 차단 해지'**를 해 주어야한다는 점이다. 팝업이 차단되어 있는 컴퓨터에서는 로그인을 했음에도 불구하고 데이터가 있는 '노동연구원 통계자료 시스템'으로 들어갈 수 없기 때문에 데이터를 다운로드 받을 수 없다.
- 설문지, 코드북, 유저가이드 등 관련자료는 사용자등록 절차 없이 바로 다운로드 받을수 있지만, **데이터의 경우 간단한 사용자등록(회원가입) 절차를 거쳐야 다운로드를 받을 수 있다.** 회원가입한 데이터 사용자들께는 향후 학술대회나 데이터의 업데이트 상황 등 기타 필요한 정보를 이메일이나 등록한 주소지를 통해 발송할 예정이다.
- 데이터는 국문과 영문, SPSS 버전과 SAS 버전으로 제공되므로, 사용자의 편의에 따라서 선택해서 다운로드 받으면 된다.
- 2006년 제1차 기본조사 '전체자료 내려받기'를 통해 자료를 전체로 내려받을 수도 있고, '부분자료 내려받기'를 통해서 데이터의 일부 문항을 선택적하여 내려 받을 수도 있다. 그러나 많은 사용자들이 부분자료 내려받기를 이용하기 보다는 전체받기만 이용하기 때문에 '2007년 직업력 조사'의 이후의 데이터는 '전체받기'만 가능하도록 구성하였다.
- 영문 설문지와 코드북, 유저가이드는 영문 홈페이지에서 별도로 자료를 다운로드 받을 수 있다.

9 예산지원, 조사 주관 및 수행 기관

고령화연구패널조사는 노동부가 고용보험기금을 출연하여 예산을 지원하였다. 한국노동연구원의 '고령화패널연구팀'이 다양한 분야의 국내자문위원 및 해외자문위원과 교류하면서 조사를 주관하였다. 실사는 TNS Korea에서 수행하였다.

제1차 기본조사 설문내용은 장지연박사와 이진국교수가 주도하여 노동연구원 고령화패널연구팀이 개발하였으며, 이 과정에서 각 분야 전문가 50여명의 자문을 청취하였다. CAPI 개발을 위해서는 미국의 RAND 연구소, 오하이오 주립대 CHRR 연구소, 네덜란드 통계청의 기술지원을 받았다. 표집과 가중치부여는 이계오·김영원 교수팀이 수행하였으며, 무응답보정은 송주원 교수팀이 맡았다. 건강영역은 서울대학교 보건대학원의 장숙랑 박사님을 중심으로 자문을 받았다.

II

표집과 가중치

표집과 가중치에 관련된 자세한 내용은 KLoSA 홈페이지 '조사설계' 섹션을 참조하십시오.
(※ http://klosa.kli.re.kr 또는 www.kli.re.kr 배너 클릭)

1 2006년 제1차 기본조사 표집과정

고령화연구패널조사의 모집단은 원칙적으로 전국에 거주하는 45세 이상 국민이다. 그러나 조사 편의상 제주도를 제외하였으며, 제1차 기본조사에서는 시설거주자를 제외하여 일반가구에 거주하는 사람으로 대상을 제한하였다.

표집틀(Sampling frame)은 2005년 인구주택총조사의 조사구이며, 전체 조사구 가운데 섬지역 조사구와 시설단위 조사구를 제외한 261,237개 보통조사구 및 아파트조사구를 추출단위 조사구로 설정하였다. 제1차 기본조사에서는 10,000명을 최대 유효 표본크기로 정하고, 표본조사구당 패널을 6가구로 구축하고자 하였는데, 2000년 인구주택총조사에서 나타난 가구당 45세 이상 평균 인구 1.67명을 감안하여 1,000개의 표본조사구를 추출하였다.

1) 조사모집단의 층화

표본조사구를 추출하기 전에 우선 모집단을 지역과 주거형태별로 층화하였다. 지역은 15개 특광역시와 도별로 동부와 읍면부로 층화한 후, 각 지역층 내에서 일반주택조사구와 아파트조사구로 층화하였다.

한편 1,000개 표본조사구 가운데 인구수가 적은 시도에서도 신뢰성 있는 통계를 생산할 수 있도록 15개 시도별로 15개 조사구를 우선 할당한 후, 나머지 775개는 인구수를 기준으로 15개 시도에 할당하였고, 각 시도내에서도 인구수를 기준으로 동부와 읍면부로 나누어 할당하였다.
지역 및 주거형태별 표본조사구 할당 결과는 다음의 <표 1>과 같다. 1,000개 조사구 가운데 409개는 아파트조사

구에 할당하고 나머지 591개는 보통조사구에 할당하였다. 특히 동부에서는 아파트조사구에 363개, 보통조사구에는 440개가 할당하였고, 읍면부에서는 아파트조사구에 46개, 보통조사구에는 151개가 할당하였다.

<표 Ⅱ-1> 15개 시도별 표본조사구 할당 결과

	전체 조사구			동부 조사구			읍면부 조사구		
	아파트	보통	합계	아파트	보통	소계	아파트	보통	소계
서울	62	115	177	62	115	177	0	0	0
부산	29	44	73	29	42	71	0	2	2
대구	24	32	56	22	30	52	2	2	4
인천	25	32	57	25	30	55	0	2	2
광주	21	18	39	21	18	39	0	0	0
대전	19	20	39	19	20	39	0	0	0
울산	15	17	32	13	14	27	2	3	5
경기	89	99	188	78	78	156	11	21	32
강원	15	24	39	12	12	24	3	12	15
충북	16	23	39	13	11	24	3	12	15
충남	16	30	46	9	8	17	7	22	29
전북	19	26	45	17	14	31	2	12	14
전남	14	31	45	11	9	20	3	22	25
경북	19	40	59	14	17	31	5	23	28
경남	26	40	66	18	22	40	8	18	26
합계	409	591	1,000	363	440	803	46	151	197

2) 표본 추출

이렇게 할당된 1,000개 조사구는, 지역 및 주거형태별로 층화된 모집단 조사구를 행정코드 순서대로 정렬한 후 계통추출법을 적용하여 할당된 수만큼 추출하였는데, 표본조사구가 변동되었을 상황을 대비하여 20%의 예비표본조사구를 합하여 추출하였다. 그리고 이렇게 추출된 표본조사구 가운데 20%에 해당하는 예비표본조사구는 다시 계통표집법으로 본표본조사구와 예비표본조사구를 분류하였다.

이와 같은 방식으로 1,000개의 표본조사구를 확정한 후, 2005년 인구주택총조사의 가구명부를 이용하여 서울 조사구는 15개 가구, 광역시 및 경기도는 13개, 나머지 도 지역은 12개 가구를 단순무작위 방식으로 표본가구를 추출하였다.

이렇게 추출한 표본가구를 면접원이 정해진 순서대로 선정한 가구를 방문하여 가구원 가운데 만 45세 이상인 사람이 1명 이상 거주하고 있으면 조사대상 적격가구로 판정하고 그 가구에 거주하는 모든 만 45세 이상가구원에 대하여 면접조사를 실시하고, 만약 45세 이상이 거주하지 않는다면 부적격가구로 판정하고 다음 표본가구를 방문하면서 패널가구 및 패널을 구축하도록 하였다.

2 2006년 제1차 기본조사 패널구축 과정 및 응답율

실사과정과 응답율에 관련된 자세한 내용은 KLoSA 홈페이지 '실사과정' 섹션을 참조하십시오.
(※ http://klosa.kli.re.kr 또는 www.kli.re.kr 배너 클릭)

1) 패널구축 과정

면접원은 115명을 교육하여 107명을 실사에 투입하였다. 면접이 진행되면서도 9명의 면접원이 중도에 탈락하여 실사 완료 때까지 실사에 참여한 면접원은 모두 98명이었다.

2006년 7월에 실시한 면접원 교육은 전체 교육 1일과 지역별 교육 2일 등 총3일로 구성되었다. 전체 교육에서는 KLoSA 개요 및 중요성에 대한 소개, 직업/산업 분류, 인지능력 측정 등에 대한 내용을 다루었다. 지역별 교육에서는 패널 접촉 요령, CAPI를 위한 노트북사용법, 설문내용, 모의 면접 등을 다루었다. 지역별 교육은 서울 2회, 부산 1회, 대구 1회, 광주 1회, 대전 1회 등 모두 6차례 이루어졌다.

면접원이 표본으로 추출된 가구를 찾아가기에 앞서 2006년 6월 말부터 7월 말까지 조사대상가구로 선정된 가구에게 우편으로 편지를 발송하였다. 편지는 조사에 대한 안내서와 공문(노동부, 한국노동연구원, 서울대학교 보건대학원 등의 기관장 명의)을 담아 면접원이 방문할 것임을 알리고 조사에 대한 협조를 구하였다. 또한 표본조사구가 속한 읍·면·동 사무소에도 등기 우편으로 조사 안내서 및 공문을 보내어 관할 지역에서 조사가 실시될 것임을 알렸다.

면접원이 가구를 방문하여 패널을 구축하는 과정은 다음과 같다. 면접원은 방문 순서가 정해진 가구 명부와 표본조사구 지도, 그리고 표본가구별 설문내용이 입력된 노트북 컴퓨터를 지참하고, 지도를 보고 순서에 따라 정해진 주소의 가구를 방문하였다.[2] 그 가구에 살고 있는 거주자를 만나 만 45세 이상 중고령자가 살고 있는지 파악하여 조사 적격가구 여부를 판정하였다. 그 가구에 거주하는 45세 이상 중고령자를 모두 면접하는 방식으로 진행되었다. 만약 45세 이상 중고령자가 거주하지 않으면 부적격가구로 판정하고 다음 순서의 가구를 방문하도록 하였다. 각 조사구별로 지역에 따라 12~15개 가구를 방문하여 6개 가구를 패널로 구축하도록 하였는데, 만약 제공한 12~15개 가구 내에서 6개 가구를 패널로 구축하지 못하면 5~8개의 추가 리스트를 제공하여 면접을 계속 진행하도록 하였다. 그러나 실사 결과 모든 조사구에서 6개 패널가구를 구축하지 못하고 조사구 상황에 따라 최소 1개 가구에서 최대 12개 가구가 패널로 구축되었다. 6개 가구가 패널로 구축된 조사구는 422개(42.2%)였고, 조사구당 구축된 패널가구수가 4~8개 가구인 조사구는 모두 812개 조사구였다.
한편 실사를 진행하면서 적격가구가 1개 가구도 없거나 재개발로 인해 철거중인 조사구, 그리고 가구 방문을 원천

2) 조사구 지도와 가구 명부는 통계청으로부터 제공받았다. 통계청에서 가구 명부와 조사구 지도를 복사하는 데에 약 1주일(2006년 5월 22일부터 26일까지)이 소요되었으며, 가구 명부를 입력하여 파일로 정리하는 데는 약 2주일(2006년 5월 29일부터 6월 13일)이 소요되었다.

적으로 거절당하여 실사 진행이 불가능한 조사구는 예비조사구로 대체하였다. 최초 1,000개 조사구 중 적격가구가 없거나 실사 진행이 불가능할 것으로 보이는 조사구를 대체한 경우는 실사 진행기간 중 모두 32회였다. 조사구를 대체할 수밖에 없었던 이유는 다음과 같다.

- 재개발로 인해 철거 혹은 철거중인 경우와 다른 이유로 철거된 조사구 7개
- 해당 조사구의 표본 가구가 모두 부적격인 조사구 21개
- 조사구 내 표본 가구 중 적격가구가 2개 이하인 조사구 2개
- 해당 조사구의 가구 방문를 원천적으로 거절 당한(관리사무소 출입거부) 조사구 2개

표본가구가 모두 부적격가구인 이유로 예비조사구로 대체했으나, 대체된 조사구 역시 표본가구가 모두 부적격가구로 확인되어 다시 조사구를 대체한 경우가 1건이 있어, 최초 선정된 조사구 1,000개 가운데 조사구 대체 없이 조사된 조사구는 969개 조사구였다.

2) 패널구축 결과

고령화연구패널조사 패널 구축 및 제1차 기본조사 결과는 다음 <표 Ⅱ-2>과 같다.

<표 Ⅱ-2> 고령화연구패널 제1차 기본조사 실사결과

시도	진행 조사구	조사성공 조사구	성공가구	성공가구원	성공가구원 /성공가구
전체	1000	999	6,171	10,254	1.7
서울	177	177	1,076	1,767	1.6
인천	57	57	400	556	1.4
경기	188	188	1,170	1,935	1.7
강원	39	39	215	391	1.8
부산	73	73	450	743	1.7
울산	32	32	188	318	1.7
경남	66	66	390	676	1.7
대구	56	56	337	562	1.7
경북	59	59	361	602	1.7
광주	39	39	233	401	1.7
전북	45	45	292	485	1.7
전남	45	45	293	480	1.6
대전	39	39	243	390	1.6
충북	39	39	235	392	1.7
충남	46	45	288	556	1.9

고령화연구패널 제1차 기본조사 실사 결과 999개 조사구에서 6,171가구에서 45세이상에 해당하는 10,254명의 가구원에 대한 면접조사가 완료되었다. 면접에 성공한 가구내 면접성공 평균 가구원 수는 1.7명으로 조사되었다.

전체 조사구에서 45세 이상의 인구가 1명이상인 적격가구 7,574가구 중 가구원 1명 이상을 면접 성공한 가구가 6.171가구로 가구 성공률은 81.5%로 나타났다. 적격가구원 전체 13,602명 중 실사 기간내에 면접을 완료한 가구원은 10,254명으로 개인 성공률은 75.4%를 보였다.

3) 성공률 및 응답률

가구성공률: 2006년 고령화연구패널 제1차 기본조사 가구응답률은 적격가구 7,574가구 중 가구내 45세이상 가구원 1명 이상을 면접 성공한 가구는 6,171가구로 가구성공률은 81.5% 로 집계되었다.

이러한 실사 결과를 가지고 KLoSA 제1차년도 기본조사 응답률은 다음과 같은 식에 의해 계산하였고, 그 결과는 다음 <표 Ⅱ-3>과 같다.

응답률 계산을 위해서 우선 미확인가구의 적격가구율을 추정하는 것이 필요한데, 일반적으로는 확인가구의 적격률을 적용한다. 이번 조사에서 확인가구의 적격가구율은 64.2%로, 이를 미확인가구 1,789가구에 이를 적용하면 1,149가구가 45세 이상 중고령자가 거주하는 적격가구일 것으로 추정할 수 있다.[3] 이러한 방식으로 가구응답률을 계산하는 식은 다음과 같다. 다음 식에서 e는 추정 적격률이다.

$$\text{가구응답률} = \frac{\text{응답가구수}}{\text{응답가구수} + \text{면접거부가구수} + e \cdot \text{미확인가구수}}$$

<표 Ⅱ-3> 중고령자를 대상으로 하는 패널조사들의 제1차년도 조사 응답률

국가, 조사약칭(1차년도)	가구응답률	개인응답률
미국, HRS(1992)	80.2%	81.6%
영국, ELSA(2002)	69.9%	96.5%
유럽11개국, SHARE(2004)	55.4%	86.3%
우리나라, KLoSA(2006)	70.7%	89.2%

출처: 미국(HRS). Steven Heeringa and Judith Connor. *Technical Description of the Health and Retirement Survey Sample Design.* Institute for Social Research, University of Michigan. 1995.
영국(ELSA). Michael Marmot, James Banks, Richard Blundell, Carli Lessof and James Nazroo (eds.). *Health, Wealth and Lifestyles of the older Population in England: The 2002 English Longitudinal Study of Ageing*. Institute for Fiscal Studies. 2003.
유럽연합(SHARE). Börsch-Supan, A., Brugiavini, A., Jürges, H., Mackenbach, J., Siegrist, J. and Weber, G. (eds.). *Health, Ageing and Retirement in Europe: First Results from the Survey of Health, Ageing and Retirement in Europe*. Mannheim Research Institute for the Economics of Aging, University of Mannheim. 2005.

3) 응답률을 계산할 때 항목 무응답(item nonresponse)이 있는 부분 응답자를 어떻게 처리할 것이냐, 즉 응답으로 할 것인가 단위 무응답으로 할 것인가의 문제도 있다. 보수적으로 응답률을 계산할 때는 무응답으로 처리하기도 한다. 그러나 KLoSA에서는 주요 변수의 결측값에 대해서는 대체(imputation)를 실시하기 때문에 항목 무응답을 가진 응답자 역시 응답자로 간주하였다.

3 2006년 제1차 기본조사 가중치 부여와 모수 추정

1) 가중치 사용의 필요성

일반적으로 표본조사의 목적은 어떠한 특정 항목에 대하여 모집단의 특성인 총계, 평균 또는 비율 등을 정확하게 추정하는 것이지만 표본추출과정에서 시도별로 또는 주택유형별로 표본을 할당한 후에 각 층별로 계통추출법을 적용하여 최종 추출단위인 가구를 선정하였기 때문에 표본가구들이 동일한 추출확률로 선정되지 않았다. 모평균이나 모비율 추정에서 단순 표본평균이나 표본비율을 사용한다면 추정량은 불편성을 갖지 못하여 편향이 포함된 추정값을 산출하게 되므로 조사대상별로 추출률과 응답률 및 기타외부정보를 이용한 벤치마킹 조정값을 고려하여 가중치를 부여하여 불편추정량(Unbiased Estimator) 산출하도록 해야 한다.

표본조사에서 표본추출과정, 조사과정 및 정확한 외부정보의 이용 등을 반영하여 모수를 추정하기 위해서 표본조사가구에 대한 가중치를 부여한다. 특히 표본설계 과정에서 모든 조사단위들이 동일한 추출률을 갖도록 하였을지라도 조사과정에서 완전한 응답을 얻지 못할 수도 있고 표본설계에 사용한 정보와 조사시점에서 정보간의 차이가 있을 경우에는 이들을 반영하여 조사시점에서 모집단의 특성을 제대로 추정할 수 있도록 가중치를 산출하여 모수추정에서 사용해야한다.

여기에서는 우선 고령화연구패널 제1차 기본조사에서 가중치 계산에 대해서만 언급을 하겠다. 본 패널조사에서 편향이 없는 모수추정량을 계산할 수 있도록 하기 위해서 가중치를 계산하였으며 세부적인 절차와 내용은 아래와 같다.

2) 가중치 산출과정

고령화 연구 패널 조사의 경우의 가중치는 크게 세 부분으로 나누어 계산하였으며 그 내용은 다음과 같다.

가중치=추출률의 역수*응답률의 역수*벤치마킹 가중치

위 내용에 대해서 세부적으로 살펴보면 다음과 같다.

◆ 추출률의 계산

- 고령화 연구 패널 조사에서는 표본설계 시 1차 추출단위로 조사구를 추출하였으며, 2차 추출단위로는 가구를 추출하였기 때문에 추출률은 조사구 추출률과 가구 추출률로 구성된다.

$$f = f_1 \times f_{2 \cdot 1}$$

- 여기서 f는 추출률이며, f_1은 조사구 추출률, $f_{2\cdot1}$은 조사구내의 가구 추출률을 나타낸다. 또한 고령화 연구 패널 조사는 시도별, 동부/읍면부별, 아파트/일반 조사구별로 층화추출 되었기 때문에 이를 고려한 추출률은 다음과 같이 계산된다.

$$f = f_{1i} \times f_{2\cdot1i}$$

$$f_{1i} = \frac{n_i}{N_i}, \quad f_{2\cdot1i} = \frac{m_{j\cdot i}}{M_{j\cdot i}}$$

- 여기서 첨자 i는 층을 나타내며, N_i는 i번째 층내의 전체조사구수, n_i는 i번째 층내의 표본조사구수를 나타내며, $M_{j\cdot i}$는 i번째 층내의 j번째 조사구내의 적격가구수 , $m_{j\cdot i}$는 i번째 층내의 j번째 조사구내에서 조사한 가구수를 나타낸다.

- 또한 $M_{j\cdot i}$는 다음과 같이 세부적으로 계산할 수 있다.

$$M_{j\cdot i} = P_{j\cdot i} \times H_{j\cdot i}$$

- 여기서 $P_{j\cdot i}$는 i번째 층내의 j번째 조사구의 적격가구의 비율을 나타내며, $H_{j\cdot i}$는 i번째 층내의 j번째 조사구내의 전체가구수를 나타낸다.

◆ 응답률의 계산

- 고령화 연구 패널조사에서 응답률은 적격가구내에서 조사대상인 45세 이상의 인구수 중 조사완료 인원수의 비로 나타낸다. 즉 적격가구내에서 조사대상인 45세 이상의 인구수를 R_{ijk}로 나타내고, 적격가구내에서 조사완료 인원수를 r_{ijk}로 나타내면 응답률은 r_{ijk}/R_{ijk}로 나타낼 수 있다.

◆ 설계가중치

- 본 표본설계의 경우 층화 및 층별 표본배정이 이루어지는 동시에 각 조사구의 적격 가구수를 사전에 파악할 수 없기 때문에 결과적으로 등확률추출법(epsem)에 의한 표본추출이 불가능하므로 자체가중(self-weighted) 표본설계를 적용할 수 없다. 이런 경우에 표본추출에 따른 설계 가중치를 계산해야 한다.
- 설계가중치는 추출률과 응답률을 고려한 가중치로 추출률의 역수와 응답률의 역수의 곱으로 표현된다.
- 본 표본설계의 경우는 층화이단집락추출에 해당하기 때문에 2단계에 걸친 추출확률을 기초로 설계가중치를 계산해야한다. 여기서 조사구내에서 적격가구의 추출확률을 산출하기 위해서 각 표본 조사구내의 전체 가구수 및 적격 가구수 현황을 정확히 파악해야한다.

$$\cdot\ w_{ijk} = \frac{1}{f_1} \times \frac{1}{f_{2\cdot1}} \times \frac{R_{ijk}}{r_{ijk}}$$

여기서 첨자 i는 층을 나타내고, j는 조사구를 의미하며, k는 가구를 나타낸다.

◆ 최종 가중치

최종 가중치는 설계가중값과 벤치마킹 보정계수를 이용하여 구해진다.

$$w^*_{ijkl} = BF \times w_{ijkl}$$

여기서 첨자 l은 개인을 의미하며, BF는 벤치마킹 보정계수로 다음과 같이 구해진다.

$$BF = \frac{i\text{시도}, j\text{성별}, k\text{연령 그룹의 상주추계인구}}{i\text{시도}, j\text{성별}, k\text{연령 그룹의 설계가중값의 합}}$$

3) 가중치 이용한 추정식의 계산

◆ 총 계

$$\hat{Y} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij} \cdot y_{hij} = \sum_{i=1}^{H} \hat{Y}_h$$

$$V(\hat{Y}) = \sum_{h=1}^{H} V(\hat{Y}_h)$$

위 식에서 $\hat{Y}_h$는 h번째 층의 총계를 나타내는데 그것의 분산추정량은 다음과 같다.

$$V(\hat{Y}_h) = \frac{n_h(1-f_h)}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi\cdot} - \overline{y_h})^2$$

여기서, $y_{hi\cdot} = \sum_{j=1}^{m} w_{hij} \cdot y_{hij}$ 이며, $\overline{y_h} = \frac{\sum_{i=1}^{n_h} y_{hi\cdot}}{n_h}$ 이다.

◆ 평균(비율)

평균에 대한 전국 추정량과 그 분산추정량의 추정식들은 다음과 같다. 비율 추정도 평균 추정의 일종인데 다만 응답 데이터인 y가 0이나 1의 값을 갖는 이항변수이다. 따라서 비율추정에 대해서도 아래의 식을 그대로 사용할 수 있다.

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij} \cdot y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij}}$$

$$V(\hat{\bar{Y}}) = \sum_{h=1}^{H}\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}(e_{hi\cdot} - \overline{e_{h\cdot}})^2$$

여기서 $e_{hi\cdot} = \left[\sum_{j=1}^{m} w_{hij}(y_{hij} - \hat{\bar{Y}})^2\right] / w_{\cdot\cdot}$ 이고, $\overline{e_{h\cdot}} = \left(\sum_{i=1}^{n_h} e_{hi\cdot}\right) / n_h$ 이다.

표준오차는 다음과 같은 식으로 나타낼 수 있다.

$$StdErr(\hat{\bar{Y}}) = \sqrt{V(\hat{\bar{Y}})}$$

4 2007년 직업력조사 가중치 부여와 모수추정법

1) 개요

직업력에 관한 조사에서 조사대상자를 고령화연구패널의 구성원으로 제한하여 조사모집단을 고령화연구패널가구에서 2006년의 패널조사의 참여한 사람으로 정의함으로써 조사된 데이터로부터 모집단의 특성을 추정하기 위해서 표본과 모집단간의 관계를 설명할 수 있는 가중치를 사용해야할 것이다. 가중치는 표본조사에서 응답한 조사단위가 모집단의 어느 정도를 대변할 수 있는지를 나타낸다고 할 수 있으므로 2006년의 패널조사에서 표본과 2007년 직업력 조사에서 표본의 관계를 응답률을 통해서 추론하고 또한 2007년의 모집단에 대한 직업력의 추정은 2007년 기준의 모집단 특성을 반영하여야만 모수 추정에서 편향을 줄일 수 있을 것이다. 모수추정에서 편향을 최소화하고 표본조사결과를 모집단의 특성에 근접하도록 조정하여 해석할 수 있도록 하기 위해서는 2006년의 패널조사 결과와 연계한 종단면 가중치와 2007년 모집단의 특성과 관계를 반영한 종단면가중치를 사용해야하므로 이 두 가지 가중치 산출방법을 설명하고 가중치산출결과를 산출하여 정확한 모수추정과 종단면적 분석을 통해서 직업력에 관한 사항을 정확하게 추정하고 정책수립에 사용할 수 있는 정확한 기초 정보를 제공하고자한다.

2) 가중치 산출 방법

가중치는 과거의 고령화연구패널조사의 결과와 비교분석을 통해서 변동상황을 파악하는데 이용될 횡단면 가중치와 조사시점에서 모집단의 특성을 표본조사결과로부터 정확하게 추정하는데 사용하는 종단면 가중치로 구분할 수 있는데 본 연구에서는 두 가지 가중치 계산절차를 설명하고 조사된 데이터에 대한 가중치의 산출 하겠다.

◆ 종단면 가중치

먼저 종단면 가중치를 살펴보면 2007년 직업력 조사시점에서 조사된 데이터와 기준년도의 모집단 특성을 분석하여 구조적 특성을 동일하게 하는 것으로 설계가중치와 벤치마킹 가중치로 구분할 수 있다

설계가중치는 표본 추출률과 응답률을 반영하여 표본구조와 모집단의 구조를 비교하여 동일하게 조정하는 가중치이고 벤치마킹가중치는 표본조사를 통해서 얻어진 데이터내의 개인별 특성인 성별 연령대별 특성을 조사시점의 모집단 특성인 성별 연령대별로 일치시킬 수 있는 가중치이다.

2006년도의 고령화연구패널의 가중치 산출과정 중 벤치마킹가중치에서 성별 연령대별 범주구분을 45세-59세와 60세 이상으로 2개로 나누었는데 데이터 분석에서 세분화된 연령대의 분포가 모집단에서 상이한 차이를 나타낸다는 연구결과를 참고로 좀더 연령대를 세분화하여 벤치마킹 가중치를 조정하는 방안을 연구하고 다음에는 이 결과를 기준으로 2006년의 조사데이터와 2007년 조사데이터간의 연계분석에 사용할 수 있는 2007년 기준의 횡단면가중치를 계산하는 절차를 설명하겠다.

- **2006년도 벤치마킹가중치 수정**

기존의 2006년도 가중치는 15개 시도별, 동부/읍면부, 주택유형별로 층화한 후에 조사구를 추출하고 조사구 내에서 22가구정도를 선정한 후에 45이상의 가구원들을 조사대상으로 하였는데 이들에 대한 추출률과 응답률을 반영한 설계가중치를 계산하였으며 설계가중치를 15개 시도별로 성별, 연령대(45세-59세,60세 이상)로 합계를 산출하고 이것과 추계상주인구수를 비교하여 벤치마킹가중치를 보정하였는데 연령대를 10세 간격인 45세-54세, 55세-64세, 65세-74세와 75세 이상으로 세분화하여 벤치마킹가중치를 수정하였다.

계산절차는 기존의 설계가중치에 벤치마킹보정가중치를 곱하여 최종가중치를 산출하였다.

$$W_{hijk} = W_{06hijk} \cdot BF_{hijk}$$

여기서 W_{06hijk}는 2006년의 설계가중치 이고, BF_{hijk}는 기존 설계가중치의 15개 시도별 성별*연령대(4개 범주별)합계와 2006년 기준 추계상주인구수(15개시도별*성별*4개 연령범주별)를 비교한 보정치이다.

- **2007년 종단면 가중치**

2007년 직업력 조사에서 응답률을 반영한 가중치로서 15개 시도별 동부/읍면부와 주택유형별(아파트와 일반주택으로만 구분)층 내에서 조사구, 가구수 및 인구수를 기준으로 응답률를 산출한 후에 이의 역수를 2006년의 최종가중치에 곱하여 계산한 가중치가 2007년 종단면 가중치이다.

15개 시도별 동부/읍면부와 주택유형별(아파트와 일반주택으로만 구분)층별로 응답률 계산에 이용한 표본조사구의 수가 아래 표에 주어졌으며 조사구별로 2006년 조사된 인구수와 2007년 조사 성공 인구수를 대비하여 응답률을 산출한 후에 응답률의 역수를 2006년 개인별 가중치의 곱으로 표현된다.

< 표 9 > 시도별 주거형태별 표본할당 결과

	전체 조사구			동부 조사구			읍면부 조사구		
	아파트	보통	합계	아파트	보통	소계	아파트	보통	소계
서울	62	115	177	62	115	177	0	0	0
부산	29	44	73	29	42	71	0	2	2
대구	24	32	56	22	30	52	2	2	4
인천	25	32	57	25	30	55	0	2	2
광주	21	18	39	21	18	39	0	0	0
대전	19	20	39	19	20	39	0	0	0
울산	15	17	32	13	14	27	2	3	5
경기	89	99	188	78	78	156	11	21	32
강원	15	24	39	12	12	24	3	12	15
충북	16	23	39	13	11	24	3	12	15
충남	16	30	46	9	8	17	7	22	29
전북	19	26	45	17	14	31	2	12	14
전남	14	31	45	11	9	20	3	22	25
경북	19	40	59	14	17	31	5	23	28
경남	26	40	66	18	22	40	8	18	26
합계	409	591	1,000	363	440	803	46	151	197

◆ **횡단면 가중치**

- 조사해당년도인 2007년을 기준으로 조사된 인원에 대한 가중치를 산출하는 것임.
- 2007년의 가구 수와 상주인구수가 있다면 사후층화법으로 가중치를 산출할 수 있으나 정확한 정보가 없기 때문에 통계청의 추계상주 인구수를 기준으로 2007년의 종단면 가중치를 이용하여 계산함.

2007년 종단면 가중치를 15개 시도 내에서 성별 * 연령대별(남/여 * 4개 연령대 범주)로 종단면 가중치 합계와 2007년 추계상주 인구수의 비로 보정함.

$$W_{07횡hl} = W_{07종hl} \cdot \frac{07추계상주인구_{성연령시도}}{W_{07종}합계_{성연령시도}}$$

여기서 h는 층(동부/읍면부, 일반주택/아파트)을 나타내고 l은 층 내의 개인별을 나타낸다. 단 시도별로 2007년의 종단면 가중치를 시도 내에서 성별, 연령대별로 합계를 계산하고 이것과 2007년 시도 내의 성별, 연령대별 추계상주 인구수간의 차이를 보정하는 작업이 2007년의 종단면 가중치 산출이다.

3) 가중치 산출 예제

앞서 살펴본 직업력 가중치 산출 과정을 한 개의 조사구(서울시 종로구 사직동)에 대하여 가중치 계산절차를 살펴보면 다음과 같다.

◆ 2006년 응답자 중에서 2007년 직업력 조사에 응답한 사람의 비율(응답률)의 역수를 2006년 벤치마킹 가중치에 곱해준다.

- 서울시 종로구 사직동의 조사구 11010530061의 경우 2006년 10명 중 9명이 응답하여 응답률은 9/10 이며, 이의 역수는 1.111111111 이다. 1.111111111을 2006년의 벤치마킹 가중치에 곱해준다. 이는 2007년 직업력 조사의 설계(종단면)가중치 이다.

◆ 2007년 직업력 조사의 벤치마킹 가중치를 구하기 위해 (1)에서 계산한 2007년 설계(종단)가중치의 층별(시/도, 성별, 연령그룹(4개 범주)) 합계와 2007년 추계상주인구의 층별(시/도, 성별, 연령그룹(4개 범주)) 합계의 비를 2007년 설계(종단면)가중치에 곱해준다.

- 서울시 종로구 사직동의 11010530061 조사구의 개인 아이디 33061010002002 응답자는 서울, 남자, 연령그룹 1(46세-55세)이며, 이에 해당하는 2007년 설계(종단면) 가중치는 3102.661003 이다. 또한 해당 층의 2007년 설계(종단면) 가중치 합계는 938026.2586이며, 2007년 추계상주인구는 734096.8이다. 여기서 추계상주인구에 소수점아래 수치가 있는 이유는 추계상주인구가 5세 간격으로 주어져서 연령대구분에 맞추어 산출하였기 때문이다.

따라서 해당 아이디의 2007년 직업력 조사의 벤치마킹 가중치는 2428.134067 (= 3102.661003*(734096.8/938026.2586)이 된다.

- 여기서 유의해야 할 사항은 2007년 추계인구의 연령 범주는 45-54세, 55-64세, 65-74세, 75세 이상으로 구분되어 있으며, 2007년 조사의 경우 2006년 조사할 때보다 연령이 1씩 증가하였으므로 연령 그룹을 구분할 때 이에 유의해야 한다. 또한 추계 상주인구도 45-54세, 55-64세, 65-74세, 75세 이상의 10살 간격의 형태로 나와 있기 때문에 해당 연령 그룹에 맞게 곱해주어 상주 추계인구를 수정해야 한다.

4) 가중치 이용 모수 추정법

직업력 조사항목 중에서 모집단의 특성인 총계, 평균 또는 모 비율을 추정하는데 가중치를 사용하는 절차와 해석에 대한 내용을 설명하고자한다.

가중치 두 종류 중에서 어떤 가중치를 사용해서 분석해야할 것인지를 결정하는 것은 분석할 내용에 따라 결정하게 된다. 2007년 직업력 조사 내용에 대한 모수추정에서는 2007년 횡단면 가중치를 사용해야하고 2006년의 고령화연구패널조사의 내용과 2007년 직업력 조사내용을 연계한 분석이나 인과분석에서는 2007년 종단면 가중치를 사용해야한다. 아래 주어진 모수추정에서는 2007년 횡단면가중치를 사용하면 될 것이다.
우선 모 총계에 대한 추정량과 분산추정은 아래 식으로 표현할 수 있다.

$$\hat{Y} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij} \cdot y_{hij} = \sum_{i=1}^{H} \hat{Y}_h$$

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \hat{V}(\hat{Y}_h)$$

위 식에서 $\hat{Y}_h$는 h번째 층의 총계를 나타내는데 그것의 분산추정량은 다음과 같다.

$$\hat{V}(\hat{Y}_h) = \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(y_{hi\cdot} - \overline{y_h})^2$$

여기서, $y_{hi\cdot} = \sum_{j=1}^{m} w_{hij} \cdot y_{hij}$ 이며, $\overline{y_h} = \frac{\sum_{i=1}^{n_h} y_{hi\cdot}}{n_h}$ 이다.

다음에는 평균의 추정량과 그 분산추정량의 추정식들은 다음과 같다. 비율 추정도 평균 추정의 일종인데 다만 응답 데이터인 y가 0이나 1의 값을 갖는 이항변수이다. 따라서 비율추정에 대해서도 아래의 식을 그대로 사용할 수 있다.

$$\hat{Y} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij} \cdot y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij}}$$

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi.} - \overline{e_{h.}})^2$$

여기서 $e_{hi.} = \left[\sum_{j=1}^{m} w_{hij}(y_{hij} - \hat{Y})^2 \right] / w_{..}$ 이고, $\overline{e_{h.}} = \left(\sum_{i=1}^{n_h} e_{hi.} \right) / n_h$ 이다.

표준오차는 다음과 같은 식으로 나타낼 수 있다.

$$StdErr(\hat{Y}) = \sqrt{\hat{V}(\hat{Y})}$$

수식적으로 위와 같이 나타낼 수 있으나 실제 데이터를 이용한 계산은 SAS의 PROC surveymeans를 사용할 것을 권장하며 실제 자료분석에서 적용할 수 있도록 사용방법을 아래와 같이 제시한다.

5) SAS의 PROC SURVEYMEANS 프로시저

◆ SAS의 PROC SURVEYMEANS 프로시저

SAS의 SURVEY 프로시저는 실제 표본조사에서 표본추출과 모수추정 등을 용이하게 할 수 있는 효율적인 프로시저이다. 버전 9.1에는 SURVEYMEANS, SURVEYREG, SURVEYSELECT, SURVEYFREQ, SURVEYLOGISTIC 등의 총 5가지의 프로시저를 제공한다.

이 중에서 SURVEYMEANS 프로시저는 표본조사 데이터로부터 모집단의 평균과 총합의 추정에 이용되며 변수의 추정값, 추정오차와 신뢰구간 등을 쉽게 구할 수 있는 기능을 제공한다.

기본적인 SAS의 문법은 다음과 같다.

```
      PROC SURVEYMEANS <option> <statistic keyword>;
    BY variables;
    CLASS variables;
    CLUSTER variables;
    DOMAIN variables <variable* variable variable *variable*
variable >;
    STRATA variables </option>;
    VAR variables;
    WEIGHT variables;
    ODS OUTPUT name=;
RUN;
```

① option

ALPHA = value, DATA = SAS data set, MISSING

ORDER = DATA or FORMATTED or INTERNAL

RATE = value SAS data set, TOTAL = value SAS data set

② statistic keyword

ALL, CLM, CLSUM, CV, CVSUM, DF, LCLM, LCLMSUM, MAX, MEAN, MIN, NCLUSTER, NMISS, NOBS, RANGE, RATIO, STD, STDERR, SUM, SUMWGT, T, UCLM, UCLMSUM, VAR, VARSUM

③ ODS OUTPUT name= 사용자 정의

이처럼 PROC SURVEYMEANS 프로시저는 많은 내용을 포함하고 있으며, 필요한 모수추정량들 중에서 총계, 모평균과 모비율에 대하여 어떻게 사용하는지 알아보도록 하겠다.

고령화 연구 패널의 표본설계는 층화 집락 추출법을 적용하여 조사대상가구를 추출하였고 여기서 사용한 층화변수는 지역(15개), 동부/읍면부, 주택유형(아파트/일반) 등 3개 변수이고, 집락 변수로는 조사구가 사용되었다. 일반적으로 표본조사 데이터에서 모수추정을 하는 경우 모집단의 총합, 평균, 비율을 추정하는 것이 일반적이므로 본 절에서는 위의 3가지 모수를 추정하는 경우에 대해서 알아보기로 한다.

◆ 총계추정

직업력 조사 데이터의 변수 중 "퇴직당시 일주일 평균근로시간"이 있다. 이 변수에 대하여 표본설계를 고려하여 모집단의 총합을 추정해 보자. 고령화 연구 패널조사의 데이터는 표본설계에서 15개 지역과 동부/읍면부, 아파트/일반주택 등이 층화변수로 사용되었으며 각 층별로 조사구를 추출한 후에 조사구내에서 적절가구를 선정하여 조사하였으므로 분석은 층화집락추출법에 맞도록 이루어져야 할 것이다.

먼저 데이터의 변수명을 정의하면 다음과 같다.

J119: 퇴직당시 일주일 평균 근로시간

region1: 지역(15개 시도)

region2: 동부/읍면부

enu_type: 주거형태(아파트/일반)

josagu: 조사구

wgt: 가중치

총합추정을 하는 SAS 문법은 다음과 같다.

```
proc surveymeans data=grand.w1070329 sum;
       strata region1 region2 enu_type;
       cluster josagu;
       var J119;
       weight wgt;
       ods output statistics=mystat1;
run
```

위 문장에서 총합을 구하기 위해 sum을 지정해주었으며, strata는 층화변수를 지정해주는 것이며, cluster는 집락변수를 지정, weight는 가중치를 지정해주는 것이다. ods output에서는 출력된 통계량을 mystat1에 데이터로 저장하는 것이다.

필요에 따라서 class 또는 by를 이용하여 범주형 변수에 대해서도 결과를 나타낼 수 있다. 단 by문장을 이용하려면 그 변수에 대하여 미리 데이터가 정렬되어 있어야 한다. 이는 SAS의 PROC SORT 프로시저를 이용하면 쉽게 해결할 수 있다.

ods output 문장은 때때로 아주 유용하게 사용된다. 특히, 보고서에 삽입할 결과를 원하는 양식이나 필요한 결과만을 출력할 때 keep, drop, set, merge 등의 문장을 이용하여 효과적인 작업을 할 수 있다.

◆ 평균추정

본 절에서는 평균추정에 대하여 알아보도록 한다. 분석 변수는 앞에서 사용한 "퇴직당시 일주일 평균근로시간"으로 예를 들겠다.

이를 위한 SAS 문장은 다음과 같다.

```
proc surveymeans data=grand.w1070329 mean;
       strata region1 region2 enu_type;
       cluster josagu;
       var J119;
       weight wgt;
       ods output statistics=mystat1;
run
```

앞의 총계를 추정하는 SAS 문장과 상당히 유사하며, sum 대신 mean을 사용하고 있다

◆ 비율추정

비율추정은 기본적으로 변수가 기본적으로 이진변수일 경우 가능하다. 즉 성공(1) 또는 실패(0)와 같은 형식으로 나타나는 변수이어야 한다.

중·고령화 패널 데이터의 경우 예를 들어보면 설문 응답 보기 항목이 "예", "아니오"로 구성되어 있는 경우

비율추정을 할 수 있다.

설문 문항 중 "J122문항 : 그 일자리에서는 국민연금에 가입되어 있었습니까? "에 대해 비율 추정을 해보자.

J122문항 : 그 일자리에서는 국민연금에 가입되어 있었습니까?

region1: 지역(15개 시도)

region2: 동부/읍면부

enu_type: 주거형태(아파트/일반)

josagu: 조사구

wgt: 가중치

비율추정을 위한 SAS 문법은 다음과 같다.

```
proc surveymeans data=grand.w1070329 mean;
       strata region1 region2 enu_type;
       cluster josagu;
       var J122;
       weight wgt;
       ods output statistics=mystat1;
run
```

비율추정 시 통계 키워드는 mean을 사용하며, 나머지 문장은 앞서 살펴본 총계추정이나 평균추정과 동일한 형식이다.

5 2008년 제2차 기본조사 가중치 부여와 모수추정법

※ 자세한 내용은 2009년 12월에 출시될 2008년 제2차 기본조사 버전 1.0에서 수록될 예정입니다.

III

다중대체(Multiple Imputation) 방법과 자료사용법

※ 항목무응답에 대한 다중대체 보정방법(Multiple Imputation) 관련 자세한 내용은 홈페이지 사용자안내서란 안에 "다중대체 보정방법 리포트"를 참고.
- 2006년 제1차 기본조사의 데이터는 항목 무응답을 다중대체 보정방법(Multiple Imputation)으로 대체한 별도의 5개의 imputation data set을 제공한다. 그러므로 연구자의 연구에 따라서는 5가지 data set을 결합하여 사용할 수 있다.
- 연구자들에 따라서는 Multiple Imputation을 사용하지 않고, Single Imputation만을 사용할 경우, 5개의 imputation data set 중 어느 한 가지 자료만 사용하면 된다. 그럼에도 불구하고 종종 어떤 자료를 사용해야 하는 문의가 있는데, 편의상 가장 첫 번째 imputation 자료를 사용하라고 안내하고 있다.
- 사용 방법은 원자료와 imputation 자료를 pid 기준으로 정렬한 후, merge 하여 imputation된 변수를 사용하면 된다.
- 주로 imputation 된 변수들은 금액에서 무응답인 경우가 대부분이며, 소득과 자산영역의 변수들이 주를 이루고 있다.
- 2007년 직업력 조사는 imputation 자료가 별도로 필요없으므로 제공하지 않는다.
- 2008년 제2차 기본조사의 imputation data set은 베타버전에서는 제공하지 않고, 2009년 12월에 제공될 제2차 기본조사 버전 1.0 데이터에서 제공할 예정이다.

현실에서 만나는 거의 모든 자료는 결측값을 포함하고 있다. 연구자가 변수들을 제어할 수 있는 실험과 달리 설문조사 자료의 경우 연구자가 응답 여부를 통제할 수 없고 조사 참여자의 의사에 따라 값을 관찰하거나 결측이 발생되게 된다. 특히 개인패널조사의 경우 항목에 따라 응답이 거절되는 비율이 높을 수 있으며 결측값을 무시한 통계 분석은 부적절한 결과를 도출할 수 있다. 고령화연구패널조사에서는 결측값의 적절한 분석을 위하여 결측이 발생한 주요 항목에 대하여 다중대체(multiple imputation)가 실시하였다.

1 2006년 제1차 기본조사 결측값 대체 방법

고령화연구패널조사의 경우 대부분의 변수에서 결측값(missing data)의 비율이 5%미만으로 작게 나타났으나 소득 및 자산 일부 항목에서 결측값의 비율은 10 - 20% 내외까지 증가하였으며 일부 응답자가 많지 않은 항목의

경우 약 30% 정도까지 나타났다. 따라서 결측값을 포함한 변수에 대한 적절한 분석을 위하여 결측이 발생한 주요 항목에 대하여 multiple imputation이 실시되었는데 특히 결측 비율이 높은 소득 및 자산 항목의 대체에 중점을 두고 진행되었다. 각 변수별 결측 비율이 거의 대부분의 변수에서 20% 미만으로 나타났으므로 imputation의 수는 5개로 결정하였다.

고령화연구패널조사는 전체 8개의 영역(session)으로 구성되어 있는데 결측값의 대체가 결측 비율이 높은 소득 및 자산 항목의 대체에 중점을 두고 진행되었지만 관련된 주요 변수들의 대체도 함께 실시되었다. 우선 인구영역의 주요 변수들이 5번 대체되었고 대체된 5개의 자료 각각에 대하여 주요 인구영역 변수 및 디자인 변수(design variables)들을 설명 변수로 사용하여 건강 영역 주요 변수의 대체를 실시하였다.

이렇게 대체된 5개 자료 각각에 근거하여 관련 변수를 설명 변수로 사용한 고용 영역의 현재 고용 및 퇴직 소득에 관한 대체를 실시하였다. 다음으로 각 대체된 자료에서 소득과 관련된 변수들을 설명 변수로 사용하여 소득 영역의 주요 소득 항목들에 대한 대체가 실시되었다. 이 때 자산 영역의 집 소유 여부 및 금융자산 총액도 설명변수로 포함되어 소득과 자산의 연관성을 설명하고자 하였다. 소득영역이 대체된 후 대체된 각 개인당 총소득을 계산한 후 다른 관련 변수들과 함께 설명 변수로 설정하여 주요 자산 영역 변수들에 대한 대체를 실시하였다. 마지막으로 가족대표자만이 응답한 가족 영역 중 자녀의 수 및 자녀들에게서 지원받은 액수 및 지원한 액수에 관한 항목들을 대체하였다. Imputation 모형에 사용된 설명 변수에 관한 자세한 정보는 KLoSA multiple imputation 결과보고서에 기술되었다.

결측값의 대체를 위하여 사용 가능한 여러 가지 대체 방법(imputation method) 중 고령화연구패널 자료의 결측값 대체에 적절할 것으로 사려되는 세 가지 대체법을 고려하였고 모의실험을 실시하여 가장 좋은 결과를 나타낸 대체 방법인 수정된 예측 평균에 근거한 핫덱 방법 (hotdeck based on a modified predictive mean matching)이 대체 방법으로 선택되었다. 이 방법은 Little(1988)이 제안한 일종의 핫덱 대체법(hotdeck imputation)으로서 미국 RAND의 Bell (1999)이 SAS Macro로 프로그램화하여 여러 가지 조사 연구에 적용해 왔고 우수한 결과를 보여온 대체 방법이다.

이 방법은 결측이 발생한 자료값을 자료 내 관찰된 값 들 중 하나 또는 여러 개의 값을 가지고 대체시키는 일종의 핫덱 대체법이지만 관찰값 중 하나 또는 여러 개의 값을 임의로 선택하는 랜덤 핫덱(random hotdeck) 대신 자료를 비슷한 여러 개의 하위 그룹(subclass)으로 나누어 같은 하위 그룹 내에서 핫덱 대체를 실시한다.

이 때 하위그룹은 결측이 발생한 변수에 대하여 관찰된 자료만을 대상으로 회귀모형(regression model)을 적합하여 결측이 포함된 모든 자료에 대한 예측값을 구한 후 예측값에 근거하여 층화(stratification)를 하여 구성한다. 각 층 내에서 결측값은 같은 층의 관찰자 중에서 기증자(donor)를 선택하여 기증자의 값으로 대체를 실시한다. 이 방법은 기증자를 선택하는 데 있어서 임의로 한 명 또는 여러 명의 기증자를 선택하는 랜덤 핫덱 방법보다 회귀모형의 예측력이 클수록 좋은 결과를 기대할 수 있다.

◆ 고령화연구패널조사의 대체를 실시할 때 몇 가지 특징

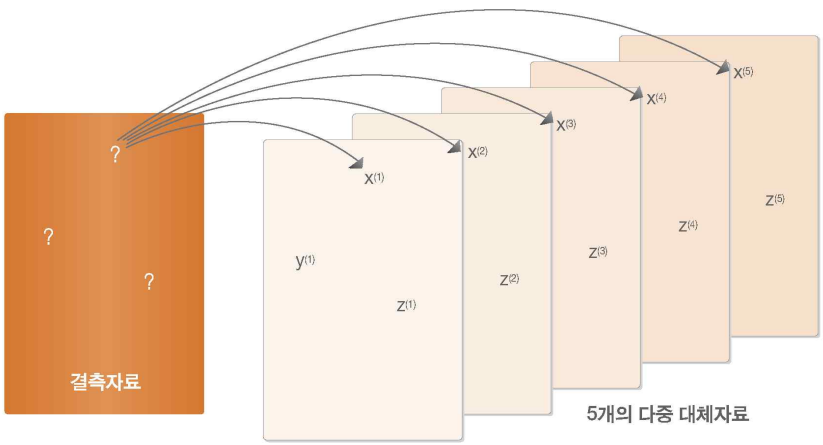
- 첫 번째로 일부 소득 및 자산 항목은 응답이 거절되거나 응답 문항 간 불일치가 나타나는 경우 대괄호 질문들(unfolding brackets)을 이용하여 얻어진 부분 정보를 포함하고 있다.
- 두 번째로 응답자가 많지 않은 일부 문항의 경우 대괄호 질문으로부터 얻어진 부차 정보에 근거한 하위그룹(subclass)에서 기증자를 발견하지 못 한 경우가 발생하였다.
- 세 번째로 일부 항목의 경우 한 사람이 여러 개의 답을 제시하는 것이 가능하였고 이 경우 동일인에 의한 여러 개의 응답은 서로 연관되어 나타날 수 있으므로 연관성을 고려하여 예측이 실시되어야 한다.
- 네 번째로 같은 영역의 연관된 문항들 사이에 일치성(consistency)을 만족시키도록 대체가 실시될 필요가 있었다. 각각의 경우 사용된 대체 방법에 대한 자세한 설명은 KLoSA multiple imputation 보고서에서 자세히 설명되었다.

2 2006년 제1차 기본조사의 Imputation 자료의 형태 및 구분

Multiple imputation은 한 개의 결측값에 대하여 한 개의 값으로 대체하는 대신 타당한 여러 개의 값을 가지고 대체하는 방법을 의미하는데 대체된 값들은 각각 다르므로 대체된 값이 참값과 다른 점을 모형에 반영시켜 한 개의 값을 대체하는 방법인 단일대체(single imputation)에서 발생하는 추정량의 편의를 보정하는 것을 가능하게 한다. 본 연구에서는 각 결측값에 대하여 5개의 값을 대체하는 multiple imputation을 사용하였다. 5개의 대체된 자료를 제공하므로 single imputation에 근거한 분석을 시행하길 원하는 연구자는 대체된 자료 중 한 개의 자료(예를 들면, 첫 번째 자료)를 선택하여 분석을 실시할 수 있다.

1) 대체된 자료의 형태

[그림 Ⅳ-1] 결측자료에 대하여 5개의 다중대체를 실시한 경우의 예



한 개의 자료에 multiple imputation이 적용되면 결과로서 여러 개의 결측값이 없는 대체된 자료가 만들어지게 된다. 고령화연구패널조사의 경우 5번의 대체가 이루어 졌으므로 대체된 항목들에 결측값이 없는 5개의 자료가 제공된다.

이 5개의 자료는 관찰된 값들은 모두 동일하지만 대체된 결측값은 같기도 하고 다르기도 한 형태를 가지고 있다. 결측된 자료에 대한 대체를 시행한 후 생성된 대체된 자료는 다음의 <그림 1>에 나타난 것과 같이 5개가 존재하게 되는 것이다.

제공된 자료는 대체된 변수만을 포함하므로, 대체가 실시되지 않은 변수들을 분석에 포함시키려면, 원자료와 결합하여 분석을 시행하면 된다. 이때 원 자료는 대체된 변수와 동일한 이른을 지닌 대체가 시행되기 이전의 결측을 포함한 원 변수들로 포함하므로, 대체된 자료를 가지고 원 변수들을 덮어 씌운 후 분석을 시행해야 한다.

예를 들면, SAS Program을 사용하여는 경우 제1차 기본조사의 원자료는 w01_v1.0k 이고, 대체된 자료는 w01_i1_v1.0k부터 w01_i5_v10k 까지 5개의 자료이므로, 각 대체된 자료를 원자료와 결합하여 대체된 변수를 포함한 전체 자료를 만들어 분석을 시행한다. SAS programe의 예제는 아래와 같다.

```
* 첫 번째 대체된 전체 자료 생성;
DATA w01_i1_v10k;
   MERGE w01_v10k w01_i1_v10k;
RUN;

        ⋮

* 다섯 번째 대체된 전체 자료 생성;
DATA w01_i5_v10k;
   MERGE w01_v10k w01_i5_v10k;
RUN;
```

위의 프로그램에서 전체 변수를 포함하는 5개의 자료를 만들기 위하여 SAS의 data step을 다섯 번 써야하는 번거로움이 있으나 아래의 SAS macro를 이용하면 5개의 자료를 간단히 생성할 수 있다.

```
* Macro를 이용하여 5개의 대체된 전체 자료 생성;
%MACRO fulldata;
%DO j = 1 %TO 5;
   DATA w01_i&j_v10k;
      MERGE w01_v10k w01_i&j_v10k;
      BY pid;
   RUN;
%END;
%MEND;
%fulldata;
```

연구자는 이 자료 각각에 대하여 원하는 분석을 반복적으로 시행할 수 있다. 각 자료에 대하여 독립적으로 분석이 시행된 후 분석 결과는 일반적으로 5개의 통계량 및 관련 분산(또는 표준 오차)으로 나타나는데 연구자는 5개의 각각 다른 통계량이 아닌 하나의 통합된 통계량을 구하는 데 목적이 있다. 각각 분석된 통계량을 통합하여 하나의 통계량을 구하는 방법은 3장에서 소개된다.

2) 대체된 결측값의 구분

결측값이 대체된 자료에서 어느 관찰값이 원래 관찰된 값이며 어느 관찰값이 대체된 값인지 구분을 할 수 있다면 유용할 것이다. 이 구분이 가능하다면 대체된 자료만을 가지고도 결측값의 대체없이 원 자료에 대한 분석을 실시하길 원하는 연구자는 원 관찰값 만에 근거한 분석을 시행하는 것이 가능할 것이고 대체된 자료값들이 관찰된 자료값들과 비슷한 지 여부 등의 추가 분석도 가능하다.

이를 위하여 고령화연구패널 자료의 경우 대체된 각 변수에 대하여 대체 여부를 나타내는 부속 변수인 깃발 변수(flag variable)가 추가되었다. 이 부속 변수는 원래 변수명에 _(underbar)를 추가시킨 변수명을 취한다. 예를 들어, 소득 부분의 작년 한 해 월평균 임금 소득액을 나타내는 변수 E003의 경우 E003_라는 변수가 추가되는데 이 변수는 아래와 같은 값들을 가진다.

0 : 응답한 관찰값이 존재함
1 : 관찰값이 모형을 통해 대체됨
2 : *Bracket* 질문에 구간 대신 값으로 응답
3 : 가족대표자의 응답을 가지고 대체
• : 이 문항에 대한 응답 대상자가 아님

즉, 대체 여부를 나타내는 깃발 변수(flag variable)가 값 “0”을 갖는 경우 해당 관찰값이 응답에 의하여 관찰된 값이라는 의미이며, “1”을 갖는 경우 관찰값이 결측되었으나 수정된 예측평균에 근거한 핫덱 방법에 근거하여 대체되었음을 의미한다. 한편 값 “2”는 대괄호 질문을 포함한 소득 및 자산 변수에서 응답이 구간으로 응답되지 않고 대략적인 값으로 응답된 경우 그 값으로 대체되었음을 의미하고 “3”은 가족대표자의 응답을 가지고 대체되었음을 의미한다. 일부 깃발 변수(flag variable)에서 보이는 결측값은 이 문항이 앞의 문항에 부속되어 있고 앞 문항의 응답 때문에 이 문항이 응답되지 않았음을 의미한다. 예를 들어 월평균 임금 소득액은 E001에서 임금 소득이 있다고 응답한 연구 대상자에게만 질문되었으므로 E001에서 임금소득이 없다고 응답한 경우 이 값은 결측으로 나타난다.

3 2006년 제1차 기본조사 Multiple Imputation 자료의 분석 방법

Multiple imputation을 통해 다중대체된 자료의 경우 결측값이 없이 대체된 한 개 이상의 자료가 제공되며 이에 따른 분석은 다중대체된 각 자료의 분석 및 분석된 자료를 통합한 결과 도출의 두 단계로 나누어지게 된다.

1) 다중대체(multiple imputation)된 자료의 분석

다중대체된 자료 각각은 결측값이 대체되어 결측값이 없는 완전한 자료 형태를 가지고 있으므로 자료 각각에 대하여 연구목적에 알맞은 분석을 시행하면 된다. 예를 들어, 회귀분석(regression analysis)을 시행하고자 한다면 동일 관심변수에 대하여 동일 설명변수를 가지고 5개 자료 각각에 대하여 회귀분석을 실시하면 된다. 이렇게 분석을 실시하는 경우 추정된 회귀계수(regression coefficients), 표준오차(standard errors), 그리고 검정통계량(test statistics)은 5개 자료 각각으로부터 약간씩 다르게 나타나는데 이는 관심 변수가 결측되어 참값을 알지 못하는 불확실성에 근거한 차이를 나타내는 것이다. 하지만 연구자의 분석 목적은 관심 자료에 대한 5개의 서로 다른 결론이 아니라 한 개의 통합된 결론을 내리는 것이므로 5개 분석의 결과를 통합하여 한 개의 결론을 도출하기 위하여 아래의 통합 과정을 거쳐야 한다.

2) 분석된 자료를 통합한 결과 도출

Multiple imputation을 m번 시행한 자료 각각에 대하여 분석을 시행한 후 얻어진 모수의 추정값들을 $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_m$이라 하자. 또한, 이 모수의 추정된 분산을 각각 W1 , W2,···, Wm이라 가정하자. 예를 들어, 회귀분석을 실시하면 i번째 자료에 근거한 회귀 분석에서 관심 설명 변수의 회귀계수의 추정값이 $\hat{\theta}_i$이 되고 그 회귀계수의 표준오차의 추정값의 제곱이 Wi가 된다. 이 경우 통합된 모수의 추정값은

$$\bar{\theta}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

으로 표현될 수 있다. 즉, 추정된 모수들의 평균값이 통합된 모수의 추정값이 된다.

통합된 모수의 분산의 추정값은 다음의 두 개의 분산 성분의 합으로 표현된다. 첫 번째 분산 성분은

$$\bar{W}_m = \frac{1}{m}\sum_{i=1}^{m}W_i$$

로서 각 모수의 추정된 분산들의 평균이다. 이 분산 성분은 대체내 분산 (within-imputation variance)으로 부른다. 두 번째 분산 성분은

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \bar{\theta}_m\right)^2$$

으로 표현되는데 이는 각 대체된 자료의 모수의 추정값들 사이의 분산을 나타내므로 대체간 분산(between-imputation variance)이라 부른다. 통합된 모수의 분산의 추정값은

$$T_m = \overline{W}_m + \frac{m+1}{m}B_m$$

으로 구할 수 있다.

자료가 충분히 큰 경우, 이 모수에 대한 분포는 다음의 t-분포를 따른다.

$$\left(\theta - \bar{\theta}_m\right)T_m^{-1/2} \sim t_\nu ,$$

여기서, t-분포의 자유도 ν 는 $\nu = \left(\nu_0^{-1} + \hat{\nu}_{\text{obs}}^{-1}\right)^{-1}$ 로 계산되는데 ν_0 는 $\nu_0 = (m-1)\left(1 + \frac{1}{m+1}\frac{\overline{W}_m}{B_m}\right)^2$ 을 나타내고 $\hat{\nu}_{\text{obs}}$ 에서 ν_{com} 이 결측값이 없을 때 모수❑에 대한 추정의 자유도를 나타낼 때 $\hat{\nu}_{\text{obs}} = \left(1-\hat{\gamma}_D\right)\left(\frac{\nu_{\text{com}}+1}{\nu_{\text{com}}+3}\right)\nu_{\text{com}}$ 을 나타낸다. 또한, 여기서 $\hat{\gamma}_m$ 은 $\hat{\gamma}_m = (1+1/m)B_m/T_m$ 으로서 결측에 의하여 손실된 모수 ❑에 대한 정보의 부분(fraction of information about ❑ missing due to nonresponse)이라 불린다. 모수의 분포가 t-분포를 따르므로 t-분포에 근거한 검정을 시행하거나 모수의 신뢰구간을 구할 수 있다. 또한 이 통합 방법은 관심 모수들에 대한 다변량 검정 및 신뢰구간의 계산 등으로의 확장도 가능하다 (Rubin, 1987).

3) 예 제

◆ Multiple imputation을 시행하여 만들어진 m개의 자료들에 근거한 m개의 분석 결과를 통합하는 과정은 연구자들이 직접 프로그램화하여 시행하기에 어려울 수 있으므로 현재 여러 가지 통계 프로그램에서는 이 결과를 통합하는 프로시저를 제공하고 있다.

- 예를 들어, SAS의 PROC MIANALYZE 프로시저는 위와 같이 분석된 자료의 모수들을 통합한 결과를 제공해 준다. 그 외에 무료 통계 프로그램인 R도 다중대체된 자료를 분석한 후 통합하는 함수를 제공하고 있다.
- 또한 Schafer(1997)가 개발한 NORM은 통계 프로그램이 필요 없이 독립적으로 시행되는 작은 크기의 프로그램으로서 위의 단계를 수행하고 통합된 결과를 제공하고 있는데 이 프로그램은 http://www.stat.psu.edu/~jls/misoftwa.html에서 무료로 다운받을 수 있다.
- 다음은 SAS에서 multiple imputation으로 대체된 여러 개의 자료를 이용한 단순 평균 계산, 층화 평균 계산, 및 회귀 분석 계수의 계산 예를 보여준다.

우선 SAS에서는 여러 개의 자료에 대하여 동일한 모형을 가지고 분석을 실시하고자 하는 경우에 여러 개의 자료를 한 개의 자료로 통합한 후 통합된 자료에 대하여 한 개의 Procedure를 이용하여 자료 별 분석을 시행하는 것이 가능하다. 이를 위하여 5개의 대체된 자료를 한 개의 자료로 통합하고 각 대체된 자료를 나타내는 변수를 가지고 각 자료를 구분하면 된다. (각 대체된 자료를 원 자료와 결합하는 프로그램은 2.1절에서 설명하였다). 제공된 대체된 자료는 몇 번째로 대체된 자료인지 나타내는 구별 변수인 w01imputation_을 가지고 있으므로 이 변수별로 분석을 시행하면 된다. 이를 위한 SAS 프로그램은 다음과 같다.

```
* 5개의 impute된 자료를 한 개의 자료로 통합;
DATA total;
   SET w01_i1_v10k  w01_i2_v10k  w01_i3_v10k  w01_i4_v10k  w01_i5_v10k;
   _imputation_=w01imputation_;
RUN;
```

여기서 새롭게 생성된 변수 _imputation_은 w01imputation_과 동일한 변수로서 각각의 자료를 분석한 후 SAS PROC MIANALYZE를 이용하여 자료를 통합하는 과정에 사용하기 위하여 생성되었다.

[예제 1] 금융자산의 단순 평균 계산

```
* 각 대체 자료별 단순 평균 계산;
  PROC SURVEYMEANS DATA = total;
     VAR w01f085;
     BY _imputation_;
     ODS OUTPUT STATISTICS = stat1;
  RUN;

* 각 대체된 자료별로 계산된 단순 평균을 통합하여 원 자료의 단순 평균 추정;
  PROC MIANALYZE DATA = stat1;
     MODELEFFECTS mean;
     STDERR stderr;
  RUN;
```

SAS Procedure SURVEYMEANS에서 BY변수를 사용하여 각 대된 자료별로 금융자산의 단순평균 및 표준 오차를 계산한 후 이들 통계량들을 자료(data set)명 stat1에 저장하였다. 이 저장된 통계량들은 Procedure MIANALYZE를 사용하여 통합되었다. 이 때 MODELEFFECTS 문에는 통합할 통계량 $\hat{\theta}_i$(여기서는, 평균)을 나타내는 변수 mean을 써 주고 STDERR문에는 Wm의 제곱근인 평균의 표준 오차를 나타내는 stderr 변수를 써 주면 된다. Procedure MIANLYZE는 다음과 같은 결과를 제공한다.

[SAS 결과 예시 1]

The MIANALYZE Procedure

Model Information

Data Set	WORK.STAT
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF
mean	75.883499	1711.837134	1802.897333	1568

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mean	0.053194	0.051716	0.989763

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
mean	1023.775585	42.460539	940.4902	1107.061	1568

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
mean	1012.992767	1033.328748	0	24.11	<.0001

표에서 보는 바와 같이 금융자산의 단순 평균은 1023.78 MW으로 타나나고 표준편차는 42.46이다. 금융자산의 평균에 대한 95% 신뢰구간은 (940.49, 1107.06)으로 계산되어지며 금융자산이 0이라는 귀무가설은 t-통계량이 24.11, p-value가 <.0001로 5% 유의수준 하에서 유의하게 나타난다.

[예제 2] 금융자산의 표본 설계 가중치를 이용한 평균 계산

이계오 등 (2006)에 설명된 표본 설계 가중치를 이용한 금융자산의 평균 추정치를 계산해 보았다.

```
* 각 대체 자료별 표본 설계 가중치를 이용한 평균 계산;
  PROC SURVEYMEANS DATA = total;
        STRATA w01region1 w01region2 w01enu_type;
        CLUSTER w01enu;
        VAR w01f085;
        WEIGHT w01wgt;
        BY _imputation_;
        ODS OUTPUT STATISTICS = stat2;
  RUN;

* 각 대체된 자료별로 계산된 표본 설계 가중치를 이용한 평균을 통합하여 원 자료의
  층화 평균 추정;
  PROC MIANALYZE DATA = stat2;
     MODELEFFECTS mean;
     STDERR stderr;
  RUN;
```

SAS Procedure SURVEYMEANS에서 BY변수를 사용하여 각 대체된 자료별로 금융자산의 표본 설계 가중치를 이용한 평균 및 표준 오차를 계산한 후 이들 통계량들을 자료명 stat2에 저장하였다. 이 저장된 통계량들은 Procedure MIANALYZE를 사용하여 통합되었다. 이 때 MODELEFFECTS 문에는 통합할 통계량 $\hat{\theta}_i$(여기서는, 표본 설계 가중치를 이용한 평균)을 나타내는 변수 mean을 써 주고 STDERR문에는 Wm의 제곱근인 평균의 표준 오차를 나타내는 stderr 변수를 써 주면 된다. Procedure MIANLYZE는 다음과 같은 결과를 제공한다.

[SAS 결과 예시 2]

The MIANALYZE Procedure

Model Information

Data Set	WORK.STAT2
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF
mean	38.539587	3521.403652	3567.651157	23804

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mean	0.013133	0.013046	0.997398

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
mean	1044.266383	59.729818	927.1921	1161.341	23804

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
mean	1035.807531	1050.103901	0	17.48	<.0001

표에서 보는 바와 같이 금융자산의 표본 설계 가중치를 이용한 평균은 1044.27 MW으로 타나나고 표준편차는 59.73이다. 금융자산의 표본 설계 가중치를 이용한 평균에 대한 95% 신뢰구간은 (927.19, 1161.34)로 계산되어지며 금융자산이 0이라는 귀무가설은 t-통계량이 17.48, p-value가 <.0001로 5% 유의수준 하에서 유의하게 나타난다.

[예제 3] 금융자산에 대한 회귀분석

금융자산과 성별, 연령의 관계를 나타내는 회귀모형을 적합한 분석을 시행하였다.

```
* 각 자료별 회귀 분석 실시;
PROC REG DATA=total OUTEST=outreg COVOUT;
   MODEL f085 = w01gender1 w01a001_age;
   BY _imputation_;
RUN;

* 각 자료별 회귀 분석 결과의 통합;
PROC MIANALYZE DATA=outreg;
   MODELEFFECTS Intercept w01gender1 w01a001_age;
RUN;
```

SAS Procedure REG에서 BY변수를 사용하여 각 imputed된 자료별로 회귀분석을 실시한 후 OUTEST 문을 사용하여 자료명 outreg에 회귀계수 및 회귀계수의 표준 오차 등을 저장한다. 여기에 저장된 통계량들을 Procedure MIANALYZE에서 통합하여 준다. 이 때 통합하고자 하는 통계량은 절편(intercept) 및 나이, 성별을 나타내는 두 변수의 계수, 즉 세 개의 회귀모형 모수가 되며 이를 MODELEFFECTS 문에 나타내준다. 여기서, Intercept는 변수명이 아니고 회귀모형의 절편을 의미한다.

'SAS결과 예시3' 에서 보는 바와 같이 회귀 모형의 절편(intercept)은 2127.83, 성별(w01gender1)과 나이(w01a001_age)의 회귀 계수는 각각 -103.22와 -13.24로 나타나며, 절편을 포함한 세 회귀모수의 표준오차는 각각 260.81, 21.51, 4.18로 추정된다. 절편 및 각 변수의 회귀 계수가 0인가를 검정하는 t-통계량은 각각 8.16, -4.80, -3.17로서 모두 5% 유의수준 하에서 통계적으로 유의하게 나타난다.

[SAS 결과 예시3]

The MIANALYZE Procedure

Model Information

Data Set	WORK.OUTREG
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF
Intercept	890.090253	66955	68024	16224
w01gender1	29.573442	427.272480	462.760610	680.15
w01a001_age	0.538980	16.850509	17.497284	2927.5

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
Intercept	0.015953	0.015823	0.996845
w01gender1	0.083057	0.079391	0.984370
w01a001_age	0.038383	0.037622	0.992532

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
Intercept	2127.827592	260.813283	1616.605	2639.050	16224
w01gender1	-103.215014	21.511871	-145.453	-60.977	680.15
w01a001_age	-13.240722	4.182976	-21.443	-5.039	2927.5

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	2081.400933	2164.672953	0	8.16	<.0001
w01gender1	-111.891438	-97.829171	0	-4.80	<.0001
w01a001_age	-14.074498	-12.248854	0	-3.17	0.0016

Ⅳ 데이터 이용방법

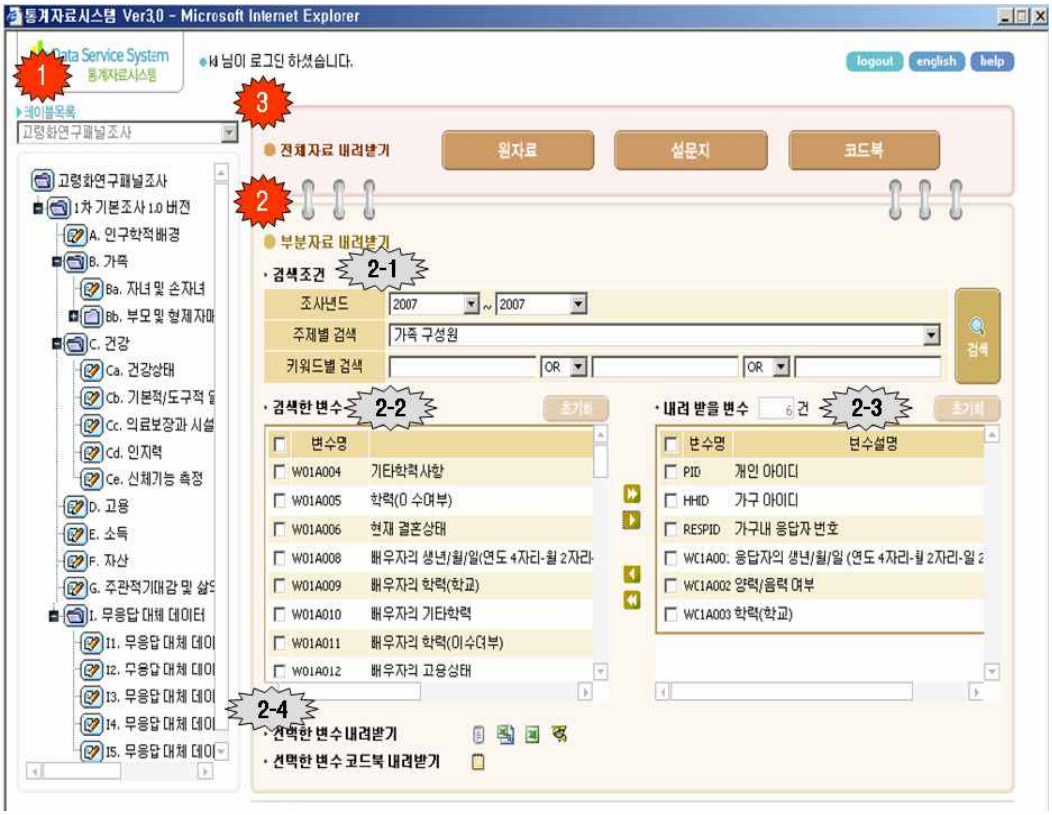
1 자료 다운로드 안내

고령화연구패널조사와 관련된 모든 자료는 인터넷 홈페이지를 통해 공개된다. 고령화연구패널조사의 인터넷 홈페이지 http://klosa.kli.re.kr 또는 한국노동연구원 홈페이지 http://www.kli.re.kr에 들어와서 "고령화연구패널조사" 배너를 통해 들어오는 방법이 있다. 보다 편리한 사용을 위해 다음의 안내를 숙지 한 후 자료를 사용하면 보다 편리하다.

◆ 원자료, 설문지, 코드북, 유저가이드 다운로드시 주의사항

- 원자료는 홈페이지에서 회원가입을 하고 로그인을 통해 무료로 다운받을 수 있다.
- 사용자안내서와 코드북, 설문지, 실사보고서, 무응답대체 보고서 등 관련 자료는 로그인 없이 홈페이지에서 바로 다운받을 수 있다.
- 관련 자료들은 별도로 인쇄를 하거나 CD 등의 형태로 배포하거나 판매하지 않으므로 연구자 각자가 인터넷을 이용하여 다운로드 받고 본인의 편의에 맞게(예: 두 쪽, 양면) 출력해서 사용하면 된다.
- 원시자료 및 설문지, 코드북, 유저가이드 등은 자료의 수정과 생성변수의 추가 등을 반영하여 업데이트 될 것이다. 이 때마다 수시로 버전(Version) 번호를 부가하여 홈페이지에 올려놓을 예정이므로 사용자들은 홈페이지에 있는 최신 버전을 사용하면 된다.
- 설문지, 코드북, 데이터 그리고 일부 자료들은 영문버전으로도 공급하여 해외에서 데이터를 이용하고자하는 연구자들에게 편의를 제공한다.
- 홈페이지 메뉴바에서 다운로드 받는 설문지, 코드북, 유저가이드와 데이터를 다운로드 받기 위해 팝업으로 들어온 통계자료시스템에서의 설문지, 코드북, 유저가이드는 동일한 파일이다.
- 제공하는 SPSS 데이터는 12.0 버전, SAS는 9.0 버전에서 작업하였으며, 그 이하의 버전에서 사용할 경우 변수명 등에서 오류가 발생할 수도 있다.

[그림 Ⅳ-1] 통계자료시스템에서 제1차 기본조사 전체 데이터와 부분 데이터 다운로드 받기



◆ 데이터 다운로드 실행 순서 및 주의사항

- 홈페이지에서 회원가입을 하고 아이디(등록한 이메일)와 비밀번호로 로그인을 후 '데이터 다운로드' 버튼을 누르면 '통계자료시스템'으로 자동연결되어 데이터를 다운 받을 수 있다.
- **주의!!** 사용자들은 자신의 컴퓨터에서 **'팝업 차단 해지'**를 해 주어야 '통계자료시스템'으로 들어갈 수 있다.
- '데이터 다운로드'를 클릭하고 들어간 첫 화면은 "통계자료시스템"으로, 주화면은 2006년 제1차 기본조사의 전체 데이터 및 부분데이터를 다운로드 받을수 있게 되었다.
- **주의!! 2007년 직업력조사를 포함하여 전체 데이터를 다운로드 받기를 원하면, '통계자료시스템'화면 상단에 [그림Ⅳ-1]의 3번부분에 있는 세 가지 갈색버튼 중에서 '원자료' 버튼을 눌러 새 팝업창이 뜨면, 사용하고자 하는 데이터를 선택하여 다운로드 받는다.**
- [그림 Ⅳ-1]은 2006년 제1차 기본조사의 부분 데이터를 다운로드 받는 방법을 알려준다. 화면은 크게 세 가지 구성되어 있는데 우선 왼쪽 상하로 연결된 박스를 1번으로 규정하고, 오른쪽 중앙에 가장 많은 면적을 차지하는 부분을 2번 그리고 화면 상단 가로로 갈색 3가지 버튼이 있는 영역을 3번으로 규정하고 각각을 설명하면 다음과 같다.
- 화면의 1번 부분은 2006년 제1차 기본조사의 영역별 데이터가 정리되어 있는 형태이다. 세부적으로 클릭해서 원하는 영역을 찾아볼 수 있다.
- 화면의 2번 부분은 1번에서 선택한 영역에서 사용자가 원하는 조건으로 데이터를 부분적 선택해서 다운로드

받는 창이다. 제1차 기본조사에 한하여, 영역별 혹은 관심분야별 키워드를 이용하여 부분적으로 데이터를 받길 원하면 "2번 부분자료 다운받기"에서 원하는 조건에 따라서 자료를 추출할 수 있다.

- 화면의 3번 부분은 데이터 전체를 다운받을 때 사용하는 메뉴들이다. 크게 원자료, 설문지, 코드북으로 구성되어 있다. 제1차 기본조사 이후의 원자료는 전체 다운로드를 받을 수 있도록 구성되기 때문에 원자료 버튼을 클릭하여 원하는 데이터를 다운로드 받을 수 있다.

2006년 **제1차 기본조사 부분자료 내려받기**

- 전체 데이터가 아닌 관심있는 변수만을 선택해서 내려받는 방식으로 [그림 Ⅵ-1] 2-1번에서 보여주듯이 크게 검색조건에 따라 주제별 혹은 키워드별로 검색하고 변수들은 모아서 내려받는 방법이 있고, [그림 Ⅵ-1] 1번에서 영역별 변수를 펼친후 영역을 선택하고, 실제 데이터의 변수설명을 보고 내려받는 방식으로 나뉜다.
- 검색조건으로 내려받기: [그림 Ⅵ-1] 2-1 참조
  주제별 검색은 미리 연구진들이 구분해 놓은 주제들이 있으므로 원하는 주제를 선택하면 '2-2검색한 변수'창에 변수가 나타나게 된다. 키워드별 검색은 사용자가 원하는 키워드를 입력하고 'and' 나 'or'을 조합하여 원하는 변수를 선택할 수 있다. 2-1을 통해 주제별 변수를 선택하면 2-2에 해당하는 변수가 생성되는데 이때 연구자가 원하는 변수를 왼쪽 콤보박스를 이용하여 선택하고, 가운데 화살표시를 이용하여 선택한 변수를 내려받기 위해 오른쪽으로 옮긴 후, 2-4에서 원하는 형식으로 선택한 변수를 내려받을 수 있다. 선택한 변수에 해당하는 코드북을 원한다면, 2-4의 코드북 내려받기를 통해 선택한 변수의 코드북을 내려받을 수 있다.
- 테이블 목록을 통해 직접 변수를 선택해서 내려받기
  테이블 목록에서 콤보박스를 통해 고령화연구패널조사가 선택되어 있는지 확인하고 폴더모양을 클릭하면 '제1차 기본조사 1.0버전'이라는 폴더가 나온다. 이것을 다시 클릭하면 데이터 A.인구학적배경 영역부터 G. 주관적 기대감 및 삶의 만족도 영역까지 각 영역이 펼쳐지는데, 여기서 폴더모양과 +표시가 남아 있는 영역은 그 안에 세부영역이 남아 있다는 표시이고 종이와 연필표시가 있으면 더 이상 들어갈 세부영역은 없다는 것을 나타낸다. 이 목록을 통해 전체 데이터의 영역을 파악한 후 내려받기 원하는 한 영역을 클릭하면, 해당하는 모든 변수들이 '2-2번의 검색한 변수'에 자동으로 생성된다. 이 후의 작업들은 앞서 설명한 검색조건으로 내려받기와 동일하다. 즉 2-2에서 선택하고 2-3에서 내려받을 변수를 모으고 2-4를 통해 내려받는 방식이다.
- **2-2번 검색한 변수 테이블 기능**
  테이블 목록을 클릭하거나 2-1의 조건을 통해 관심있는 주제의 변수들이 모이는 테이블이다. 2-2에 해당변수들이 보이면 왼쪽 네모 콤보박스를 이용하여 원하는 변수를 선택한다. 가장 윗줄의 네모를 클릭하면 전체가 선택되고 다시 클릭을 하면 전체가 해제된다. 2-2번과 2-3번 사이의 방향아이콘을 클릭하여 내려받고자 하는 변수를 오른쪽으로 옮겨 담는다. 이때 세모가 두 개 있는 아이콘은 검색한 변수에 있는 모든 변수가 옮겨지며, 세모가 한 개 있는 아이콘은 직접 선택한 변수만이 옮겨지는 기능을 한다. "초기화"버튼을 누르면 검색한 모든 변수가 없어지게 된다.
- **2-3번 내려 받을 변수 테이블 기능**
  이곳은 2-2번에서 검색한 변수 중 원하는 변수를 모아 놓는 곳이다. 이때 원하는 변수가 다른 영역에 있으면

다시 1번 테이블 목록에서부터 2-2번 검색한 변수를 거쳐 2-3번 내려 받을 변수로 여러 번 반복해서 변수를 옮겨 담을수 있고 모아 놓은 변수는 계속 축적된다. “초기화”버튼을 누르면 지금까지 모아 놓은 변수가 모두 비워지는 기능을 한다.

- **2-4번 선택한 변수 내려받기, 선택한 변수코드북 내려받기 기능**

  이 곳에서는 2-3에서 내려받고자 선택한 변수들을 각각 텍스트, CSV, 엑셀, SAS 형식중 자신이 원하는 형식으로 다운로드 받을 수 있는 있다. 그리고 선택한 변수들의 코드북만을 별도로 다운로드 받길 원하는 경우 선택한 변수 코드북 내려받기를 클릭하면 된다.

[그림 Ⅳ-1]의 3번 원자료 버튼을 클릭해서 전체 자료 내려받기 안내

- 원자료 버튼을 클릭하면 다시 팝업창이 나타나고, 자료의 목록이 다음 [표 Ⅳ-1]처럼 나타난다.
- ‘전체 자료 다운받기’에서 제공되는 데이터는 SPSS 와 SAS 형식, 각각 국문, 영문 데이터로 모두 4가지 형식(국문 SPSS, 영문 SPSS, 국문 SAS, 영문 SAS)으로 제공된다.
- 파일명은 예로, 'W01_v10k(SAS).zip' 은 제1차년도, 기본조사 1.0 버전, 한국어, SAS 파일이라는 뜻이다. 'W01_i_v11k(SAS).zip' 은, 제1차 년도, imputation 버전을 구분하기 위해 파일명 중간에 ‘i'를 넣었다. 직업력 조사의 파일을 JH 로 시작하는 파일들이다.
- 2007년 직업력조사이후의 원자료는 전체 데이터 다운로드만 가능하다. 즉, 첫 화면처럼 사용자가 원하는 조건에 따라 부분 데이터 다운로드 기능이 없다. 또한 직업력조사의 데이터의 압축을 풀면, 첫 번째 파일명은 “JH_Job~" 으로 시작되는데 이것은 **‘일자리의 특성’을 나타내는 변수들로 구성**되어 있으며, 두 번째 파일인 "JH_Ind~"는 **‘개인의 특성’을 나태내는 변수들로** 응답자의 출생년도, 성별, 결혼 당시 연도, 첫째와 마지막 자녀 출생연도 그리고 15세부터 현재 연령까지 년단위 근로여부 더미 변수가 들어있다. 두 개의 파일은 ‘PID' 변수를 이용하여 연결하여 사용할 수 있다.

[표 Ⅳ-1] 원자료 팝업창 안내

조사년도	데이터명	데이터 설명
2006	w01_v10k(SAS).zip	1차 기본조사 1.0버전(A영역~G영역)_국문 SAS 데이터
2006	w01_v10k(SPSS).zip	1차 기본조사 1.0버전(A영역~G영역)_국문 SPSS 데이터
2006	w01_v10e(SAS).zip	1차 기본조사 1.0버전(A영역~G영역)_영문 SAS 데이터
2006	w01_v10e(SPSS).zip	1차 기본조사 1.0버전(A영역~G영역)_영문 SPSS 데이터
2006	w01_i_v11k(SAS).zip	1차 기본조사 1.1버전(무응답 대체 및 변수생성)_국문 SAS 데이터
2006	w01_i_v11k(SPSS).zip	1차 기본조사 1.1버전(무응답 대체 및 변수생성)_국문 SPSS 데이터
2006	w01_i_v11e(SAS).zip	1차 기본조사 1.1버전(무응답 대체 및 변수생성)_영문 SAS 데이터
2006	w01_i_v11e(SPSS).zip	1차 기본조사 1.1버전(무응답 대체 및 변수생성)_영문 SPSS 데이터
2007	JH_v10k(SAS).zip	직업력조사 1.0버전_국문 SAS 데이터
2007	JH_v10k(SPSS).zip	직업력조사 1.0버전_국문 SPSS 데이터
2007	JH_v10e(SAS).zip	직업력조사 1.0버전_영문 SAS 데이터
2007	JH_v10e(SPSS).zip	직업력조사 1.0버전_영문 SPSS 데이터

2 기본 응답단위

◆ 기본 응답단위: 2006년 기준 '45세 이상의 개인'

- 응답 대상자는 2006년 기준으로 만45세 이상의 중고령자로, 한 가구내에서 1962년 이전 출생자는 모두 고령화연구패널조사의 응답자가 된다.
- 고령화연구패널조사는 기본적으로 개인을 패널로 하는 조사이다. 가구의 상황은 개인을 둘러싼 환경으로서 파악한다.
- 소득영역과 자산영역에서 다루는 내용은 응답자 자신의 명의를 기준으로 응답자 자신의 소득, 응답자 명의의 자산만을 자신의 자산으로 응답하도록 하였다. 가구단위의 소득과 자산은 개인의 응답내용을 근거로 산출하여 생성변수(generated variables)로 제공한다.

◆ 주의해야 할 응답 단위

- 부부가 모두 45세 이상이고, 현재 같은 가구에 동거하는 응답자인 경우, 자녀에 대한 정보는 동일하다. 부부 모두가 응답대상자가 되기 때문에 부부중 먼저 응답한 사람의 자녀정보를 나중에 응답한 응답자에게 동일하게 적용하였다. 이때 자녀와 관련된 문항에 부모중 누가 대답하였는지를 알려주는 변수는 'W01Ba_resp'이다.
- "**지난 1년**"은 조사 시점으로부터 1년 전을 의미하고, "**작년 한해(2005년)**"는 2005년 1월 1일부터 12월 31일까지를 의미한다.
- 기본조사에서 **금액**을 묻는 모든 문항의 단위는 "_____**만원**"이다.
- 직업력조사에서 **금액**을 묻는 모든 문항의 단위는 "_____**원**"이다. 현재의 시점이 아니기 때문에 '만원 단위'보다는 '원 단위'가 보다 적합하다.

3 주요 용어의 개념

◆ **루프(Loop):** 설문문항이 일정 번호부터 일정 번호까지가 한 사람 또는 한 가지 사건을 다루고 있는 경우, 응답해야 할 사람의 수나 사건의 수만큼 해당 설문 구간이 반복될 수 있는데, 이렇게 설문 문항이 그룹을 이루어 반복되는 경우 해당 문항그룹을 루프(Loop)라고 표현하였다. 예를 들면 가족영역에 '첫째 자녀'에 대한 기본적인 배경들(예; 생년월일, 학력, 혼인 여부 등)을 묻는 설문이 'Ba01부터 Ba11번까지 11문항'이라면 이 11개의 문항들은 둘째 자녀에 대한 정보를 얻고자 할 때도 동일한 11개의 문항을 묻게 된다. 그러므로 이러한 문항그룹은 응답자의 자녀수만큼 반복이 되는데 이렇게 사건이나 사람수에 의해 반복되는 문항그룹을 루프(Loop)라 한다.

◆ **범주형 전환문항(Unfolding Bracket):** 설문지에 박스로 처리되어 들어가 있는 설문문항이다. 이 설문은 응답자가 금액에 대한 응답을 '모르겠음'이나 '응답거부'로 표명했을 때, 일정 금액을 랜덤으로 되물어서 그 금액 '이상'인지 '이하'인지를 응답하도록 유도하는 방식이다. 설문지에 박스 안에 5개 문항이 모두 펼쳐져 있지만 실제 컴퓨터는 이 5개 문항 중 랜덤으로 한 문항이 떠서 해당 문항의 금액 구간을 묻는 방식으로 진행되었다. 이 범주형전환문항(Unfolding Bracket)의 기능은 응답자가 응답거절을 하는 경우 바로 다음 문항으로 넘어가지 않는다는 학습을 통해 응답거절의 빈도를 줄일수 있는 기법이기도 하고, 생각이 잘 나지 않아 넘어가는 응답자들에게는 다시 한 번 그 금액에 대한 생각을 할 수 있는 기회를 제공한다는 점에서 유용하며, 의도적으로 응답거절을 하는 응답자들에게는 missing으로 처리되는 값보다는 대략적인 구간값을 알아내어 추후 무응답의 보정처리(Imputation)을 할 때 유용한 정보로 사용되기 위해 만든 기법이다.

4 데이터 변수명 규칙

연구자가 데이터를 사용할 때에는 반드시 코드북을 기본으로 사용하는 변수와 빈도값을 확인하면서 사용해야 한다. 이 장에서는 코드북을 좀더 편리하게 이용할 수 있도록, 변수명을 부여한 기준을 설명하고자 한다.

◆ **변수명은 기본적으로 'W01+설문문항번호'의 형식을 가지고 있다. 즉 변수명 처음은 wave를 나타내는 단위이고, 바로 뒤에 설문문항번호를 취하는 형식을 기본으로 한다. 그러므로 제2차 기본조사의 경우 'W02+설문문항번호'가 기본 변수명이 된다.**

◆ **루프(Loop) 설문의 변수명 규칙: '기본변수명+_순번' (아래 박스 예시 참조)**

⇒ **상황:** 응답자가 모두 3명의 생존 자녀가 있고 그에 따른 응답을 했다면,
⇒ **설문문항 보기 예시**

Ba01. 지금부터는____님의 자녀에 관한 질문을 드리도록 하겠습니다. _____님께서는 현재 살아있는 자녀는 몇 명이십니까? (단위: 명)
_____3___명 [최소0 ~ 최대20] (0명인 경우 Ba47)

[로직: Ba01 에서 자녀수가 1 이상이면 자녀 수 만큼 Ba02 ~ Ba46 까지 반복 루프(loop), 단 Ba12~Ba46까지는 비동거자녀에게만 해당함.]
Ba02. 첫째/둘째/셋째/… 자녀의 이름은 무엇입니까?
[면접원: 커버스크린에서 쓰인 이름(관계)과 동일할 경우 커버스크린과 철자가 틀리지 않도록 주의하고, 자녀의 이름을 밝히지 않을 때는 다음과 같이 (첫째자녀, 둘째자녀..)로 적어 넣습니다.] _____

Ba03. [Ba02에서 밝힌 자녀 이름] 님은 아드님입니까, 따님입니까?
① 아들
⑤ 딸

Ba04. [Ba02에서 밝힌 자녀 이름] 님은 몇 살입니까?(연세가 어떻게 되십니까?) (단위: 세)
[면접원: 띠로 말씀하시는 분들은 띠별 생년이 나와 있는 보기카드를 참고하세요]
_____세 [최소1 ~ 최대100]

⇒ **루프(Loop) 변수명 규칙:** 루프에 해당하는 사건이나 사람수를 응답자가 이야기한 순서대로 **문항번호 뒤에 '_순서번호'를 표시한 형태로 변수명이 부여된다.** 위의 예시에서 응답자는 3명의 자녀가 있으므로 Ba02번에서 Ba04번 문항에 대하여 3번의 루프가 돌고, 첫째 자녀, 둘째 자녀, 셋째 자녀에 대한 변수명은 다음과 같이 생성된다.

⇒ **생성결과**

- 응답자가 **첫 번째 자녀**에 대한 대답을 하면 **첫 번째 루프가 돌게 되고**
  ≫ 변수명은 W01Ba02_01, W01Ba03_01, W01Ba04_01이 된다.
- 응답자가 **두 번째 자녀**에 대한 대답을 하면 **두 번째 루프가 돌게 되고**
  ≫ 변수는 W01Ba02_02, W01Ba03_02, W01Ba04_02가 된다.
- 응답자가 **세 번째 자녀**에 대한 대답을 하면 **세 번째 루프가 돌게 되고**
  ≫ 변수는 w01Ba02_03, W01Ba03_03, W01Ba04_03 가 된다.

◆ 복수응답선택 문항의 변수명 규칙: '기본변수명 + m + 순번'

- 복수응답의 문항은 응답수만큼 1과 0을 가진 변수들로 처리했다. 더미로 처리된 변수명은 문항번호 뒤에 "multiple responses"의 첫 글자 m을 따서 순서대로 m1, m2, m3...을 붙이는 형식을 취한다.

⇒ **상황:** 응답자가 ① 종교모임, ② 친목모임(계모임, 노인정 등)에 참여한다면,

⇒ **설문문항에 따른 복수응답의 변수명**

A017. _____님께서는 아래 단체 가운데 참여하고 계신 것이 있으십니까? 있으시다면 모두 말씀해 주십시오.

① 종교모임 → **변수명은** W01A017m1
② 친목모임(계모임, 노인정 등) → **변수명은** W01A017m2
③ 여가/문화/스포츠 관련단체(노인대학 등) → **변수명은** W01A017m3
④ 동창회/향우회/종친회 → **변수명은** W01A017m4
⑤ 자원봉사 → **변수명은** W01A017m5
⑥ 정당/시민단체/이익단체 → **변수명은** W01A017m6
⑦ 기타 → **변수명은** W01A017m7

⇒ **응답자 상황에 따른 결과**

- 변수명 W01A017m1 과 W01A017m2에서 각각 변수값 1을 갖고, A017m3부터 A017m7 까지는 모두 변수값 0을 갖게 된다.

◆ 루프와 복수응답 선택이 조합된 형태: '기본변수명 +_순번 +m순번'

- 가장 복잡한 변수명 조합은 위에 설명한 루프(Loop)와 복수응답선택 문항이 조합된 형태이다. 이때 앞에서 제시한 원칙대로 그대로 적용을 하되 기본변수명 뒤에 루프 변수명을 우선 적용하고 복수응답선택 변수명 조합을 나중에 붙이는 순서를 적용시키면 된다.

⇒ **상황:** 응답자가 두 명의 자녀로부터 금전적인 지원을 받았는데 첫째 자녀로부터는 정기적으로 받았고 둘째 자녀로부터는 비정기적인 금전적지원만을 받았다고 응답한 경우

⇒ **설문문항에 따른 복수응답 변수명과 변수값**

Ba15. 작년 한해(2005년) _____님께서는 (Ba02이름) 님으로부터 용돈이나 생활비 또는 병원비 등과 같은 금전적인 지원이나 선물을 받으신 적이 있으십니까? 받으셨다면 다음과 같은 경우 중 어떤 형태입니까? 모두 선택해 주세요.

① 예, 정기적으로 용돈이나 생활비 등 금전적인 지원을 받았음
→ 첫 번째 자녀인 경우 **변수명은 W01Ba16_01m1, 변수값은 1**
→ 두 번째 자녀인 경우 **변수명은 W01Ba16_02m1, 변수값은 0**

② 예, 비정기적으로 용돈이나 생활비 등 금전적인 지원을 받았음
→ 첫 번째 자녀인 경우 **변수명은 W01Ba16_01m2, 변수값은 0**
→ 두 번째 자녀인 경우 **변수명은 W01Ba16_02m2, 변수값은 1**

③ 예, 현금이 아닌 현물이나 선물 등 비금전적인 지원을 받았음
→ 첫 번째 자녀인 경우 **변수명은 W01Ba16_01m3, 변수값은 0**
→ 두 번째 자녀인 경우 **변수명은 W01Ba16_02m3, 변수값은 0**

⑤ 아니오, 금전적/비금전적 지원 또는 선물을 받은 적 없음
→ 첫 번째 자녀인 경우 **변수명은 W01Ba16_01m5, 변수값은 0**
→ 두 번째 자녀인 경우 **변수명은 W01Ba16_02m5, 변수값은 0**

◆ 코드북 및 데이터 이용시 주의사항

- 위와 같은 변수명을 기준으로 영역별 설문문항의 순서에 따라 데이터와 코드북을 구성하였다.
- **사용자들의 편리한 사용을 위해 조금 조합이 어려운 변수들이나 자주 쓰는 변수들은 각 영역의 마지막 부분에 생성변수를 만들어 제공하였다.** 그러므로 다음 장의 생성변수를 잘 살펴보고 활용하면 보다 편리하게 자료를 사용할 수 있다는 점을 밝힌다. **그러나 생성변수는 사용자의 편의를 위해서 만든 것이므로, 연구분야나 주제에 따라서 중요한 변수인 경우를 원자료를 이용해서 스스로 변수를 만들어 사용하기를 권장한다.**

5 사용자들의 편의를 위한 생성변수 구성

생성변수란 설문문항을 통해 나온 기본변수를 사용자들이 쉽게 사용할 수 있도록 흔히 사용되는 주요 변수들을 새롭게 생성하여 제공한다. 이러한 변수들은 설문에 대하여 응답자가 직접 대답한 내용을 담은 변수와 구별하기 위하여 생성변수(generated variables)라고 칭하기로 하였다. **데이터 및 코드북을 살펴보면, 각 영역별 설문문항 변수가 끝나면, 마지막부분에 해당영역에서 만들어준 생성변수를 넣었고,** 그 리스트는 다음 표와 같고 생성변수가 만들어진 과정은 '생성변수 설명' 파일에 기록하고 있으므로 자세한 내용은 이 파일을 참조하여 사용하여야한다.

A. 인구학적 배경영역	
변수명	변수설명
w01edu	응답자 학력
w01respid1	가구원 내 응답자 순번 (전체 가구원 중 응답자가 45세 이상 중에서 몇 번째 인지 알려주는 번호)
w01year	응답자의 태어난 해
w01a001_age	응답자 연령(=2006-w01A001y)
w01gender1	응답자 성별
w01respid2	응답자의 배우자임을 알려주는 번호 (값이 있으면 배우자 있고, 결측치이면 배우자 없거나 45세 이하)
w01year2	응답자 배우자의 태어난 해
w01age2	응답자 배우자의 나이(=2006-year2)
w01gender2	응답자 배우자의 성별
w01hhsize	가구원 수
w01gen_num	세대수
w01e_num	본 설문에 응답해야 할 가구원 수
w01CID	가구내에서 서로 부부임을 알려주는 변수로 가구ID에 부부임을 알려주는 식별 번호를 부여한 아이디 (예: 한 가구에 2커플이 모두 응답했고 가구번호가 1000번이라면, 10001, 10001, 10002, 10002와 같은 형식으로 커플에게 부여되는 아이디)
w01c_num	가구내에서 서로가 부부이거나 다른 세대임을 알려주는 번호로 w01CID와 다른 점은 가구ID가 부여되지 않음 (예를 들면 별거중인 응답자가 조모와 부모님과 동거한다면, 응답자는 1, 조모는 2, 아버지는 3, 어머니도 3으로 구분되는 변수값을 가짐)
w01region1	지역1: 특별시, 광역시, 도
w01region2	지역2: 동부/읍면부
w01region3	지역3 : 대도시/중소도시/읍면부
w01enu_type	거주형태
w01enu	조사구
w01mniw_y	본 설문 인터뷰 날짜 (연도)
w01mniw_m	본 설문 인터뷰 날짜 (월)
w01mniw_d	본 설문 인터뷰 날짜 (일)
w01wgt	가중치

Ca. 건강상태 영역	
변수명	변수설명
w01Ca_list	무작위 부여
w01bmi	카우프 지수(BMI)
w01body	BMI에 따른 비만 정도
w01smoke	흡연자 구분
w01smkterm	흡연기간(단위: 개월)
w01alc	음주자 구분
w01alcterm	음주기간(단위: 개월)
w01soju	소주 음주 여부
w01beer	맥주 음주 여부
w01makgeolli	막걸리 음주 여부
w01wisk	양주 음주 여부
w01wine	포도주 음주 여부
w01addic	음주태도
w01dep1	우울증 여부
w01dep2	CES-D10을 기준으로 한 우울증 여부

Cb. 일상생활 수행능력과 간병수발자 영역	
변수명	변수설명
w01adl	ADL 지수화
w01iadl	IADL 지수화

Cd. 인지력 영역	
변수명	변수설명
w01mmse	인지기능 점수
w01mmseg	인지기능 구분

Ce. 신체기능 측정 영역	
변수명	변수설명
w01mgrip	악력 지수화

D. 고용 영역	
변수명	변수설명
w01ecoact	경제활동 상태
w01empdur	조사당시 취업 중인 응답자 (임금근로자/자영업자/무급가족종사자)의 면접일까지의 취업기간
w01lastempdur	조사당시 취업 상태가 아닌 응답자의 가장 최근 일자리의 취업기간
w01D103ind	임금근로자: 산업대분류
w01D308ind	자영업자: 산업대분류
w01D405ind	무급가족종사자: 산업대분류
w01D518ind	구직자: 산업대분류
w01D707ind	가장 최근 일자리: 산업대분류
w01D103indm	임금근로자: 산업중분류
w01D308indm	자영업자: 산업중분류
w01D405indm	무급가족종사자: 산업중분류
w01D518indm	구직자: 산업중분류
w01D707indm	가장 최근 일자리: 산업중분류
w01D109occ	임금근로자: 직업대분류
w01D197occ_h	임금근로자: 희망하는 일자리에 대한 직업대분류
w01D314occ	자영업자: 직업대분류
w01D353occ_h	자영업자: 희망하는 일자리에 대한 직업대분류
w01D407occ	무급가족종사자: 직업대분류
w01D445occ_h	무급가족종사자: 희망하는 일자리에 대한 직업대분류
w01D519occ	구직자: 직업대분류
w01D610occ_p	은퇴자: 소일거리에 대한 직업대분류
w01D710occ	가장 최근 일자리: 직업대분류
w01D109occm	임금근로자: 직업중분류
w01D197occ_hm	임금근로자: 희망하는 일자리에 대한 직업중분류
w01D314occm	자영업자: 직업중분류
w01D353occ_hm	자영업자: 희망하는 일자리에 대한 직업중분류
w01D407occm	무급가족종사자: 직업중분류
w01D445occ_hm	무급가족종사자: 희망하는 일자리에 대한 직업중분류
w01D519occm	구직자: 직업중분류
w01D610occ_pm	은퇴자: 소일거리에 대한 직업중분류
w01D710occm	가장 최근 일자리: 직업중분류

E. 소득 영역	
변수명	변수설명
w01CV050_r	가구응답 대상자 중 소득을 가장 잘 아는 사람
w01CV050_rn	가구 내 소득 대표응답자(결측치 대체)
w01incfirst	가구내 조사를 참여한 45세 이상 응답자 중 개인소득 1순위자

F. 자산 영역	
변수명	변수설명
w01F235	가구 내 45세 이상이나 인터뷰하지 않은 사람의 개인총자산
w01CV051_r	가구응답 대상자 중 자산을 가장 잘 아는 사람
w01CV051_rn	가구 내 자산 대표응답자(결측치 대체)

공통으로 생성된 변수	
변수명_ct (예:w01Ba16_01ct)	변수명의 Unfolding Brackect 설문에 대한 구간 값 (예: w01Ba16_01의 Unfolding Brackect 설문에 대한 구간 값)

V

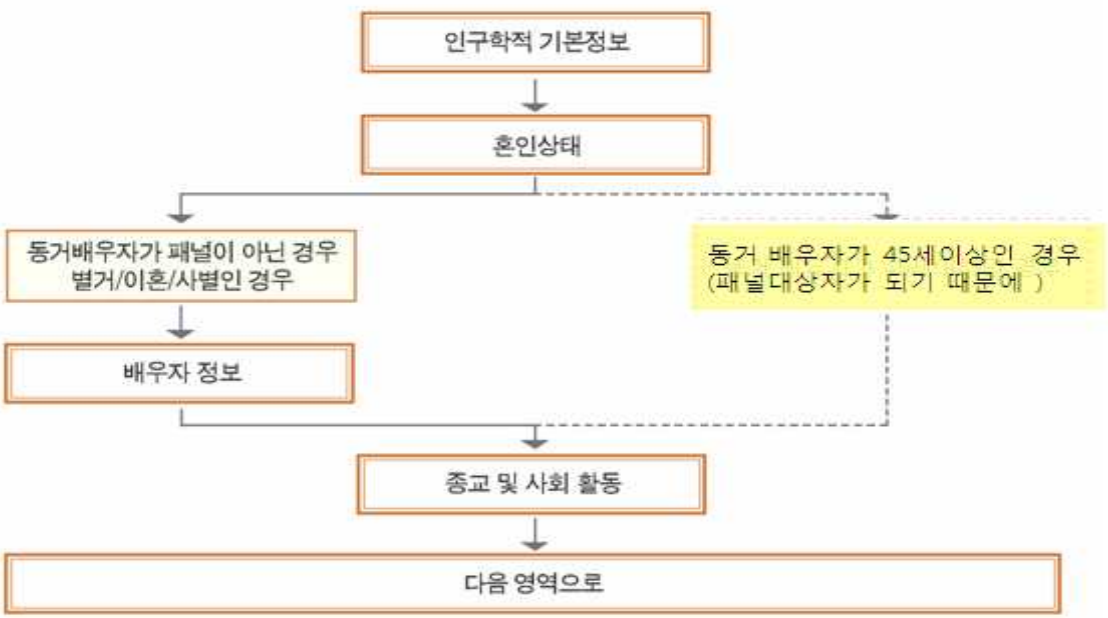
2006년 제1차 기본조사 자료의 영역별 주요내용

※ 제1차 기본조사는 패널구축과 함께 이루어졌으므로, 조사의 개요, 실사과정과 응답률, 가중치는 제1장 고령화연구패널조사 개요를 참고하면 된다. 그러므로 이 장에서는 제1차 기본조사의 자료의 영역별 설문흐름과 주요내용을 중심으로 기술한다.

1 영역별 설문흐름도 및 주의사항

1) 인구학적 배경 영역

[그림Ⅴ-1] 인구학적 배경영역의 설문구조

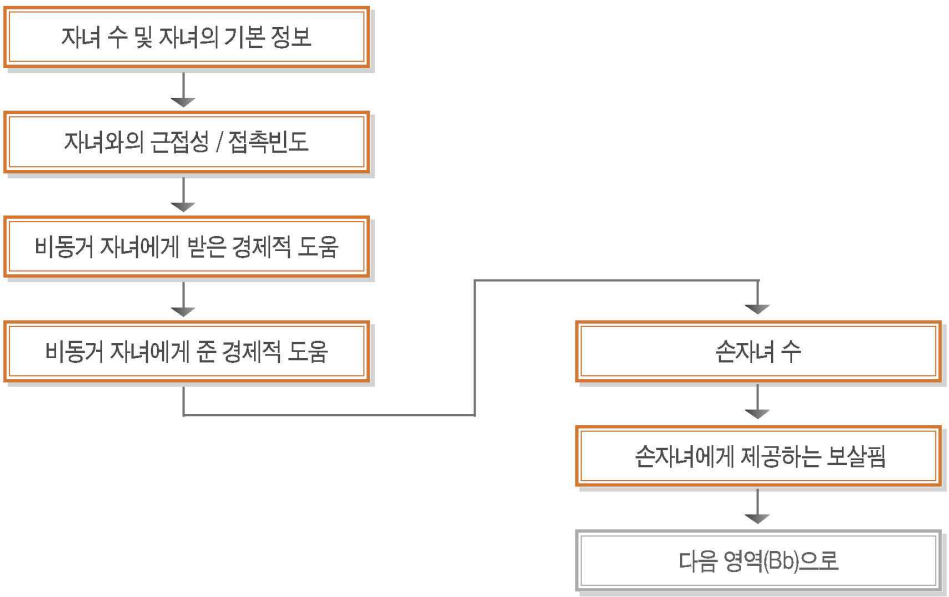


◆ 배우자 기본정보 사용시 주의해야 할 점

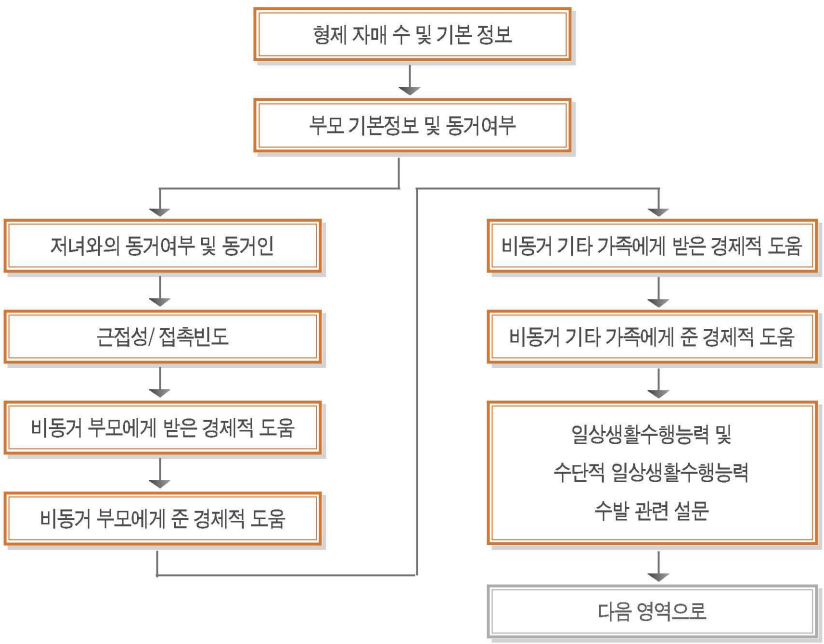
- 현재 응답자와 동거하는 배우자가 45세 이상이면 조사대상자가 되기 때문에 별도로 배우자 기본정보를 묻지 않았다.
- 그러나 응답자의 배우자가 사망, 실종, 별거중인 경우와 배우자 연령이 45세 미만인 경우 조사대상자에서 제외되기 때문에 이러한 경우에 해당하는 응답자에게만 배우자 정보(생년월일, 학력, 고용상태)를 별도로 물었다.

2) 가족영역

[그림 Ⅴ-2] 가족영역(자녀 및 손자녀) 설문구조



[그림 Ⅴ-3] 가족영역(형제자매, 부모, 기타 가족) 설문구조



◆ 가족영역에서 주의해야 할 사항 및 주요 개념

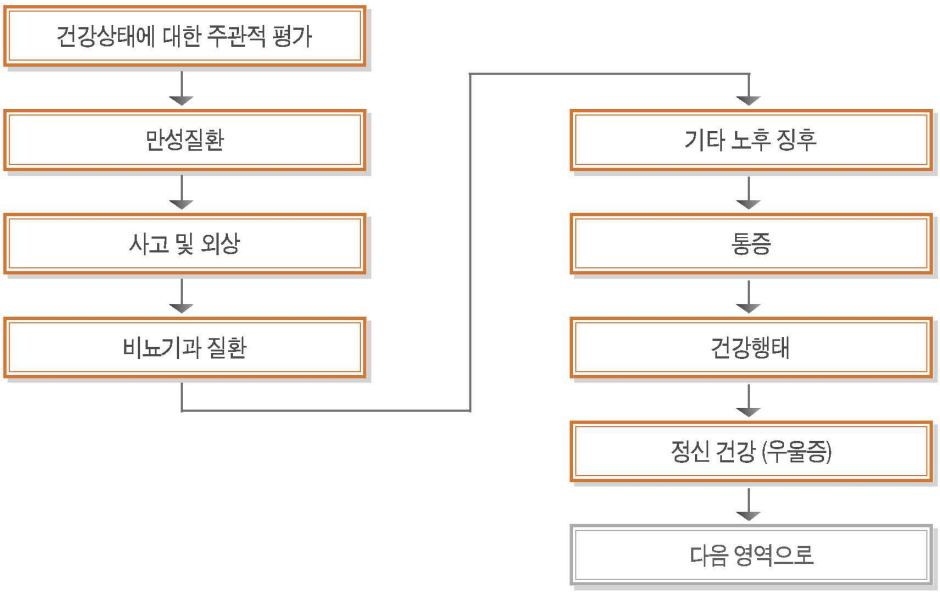
- **가족관계에서 응답자의 주관적 판단 존중 :** 가족 관계는 친자, 혈연, 호적 여부와 상관없이 응답자가 생각하는 관계를 그대로 존중하였다. 예를 들어, 혼인 여부 질문에서 "배우자 있음"이란 사실혼도 포함하며, "별거"는 불화로 인한 별거뿐만 아니라 장기수용시설, 장기해외체류 등 다른 이유의 별거를 포함시키되 응답자의 주관적인 판단을 기준으로 한다. 응답자의 아버지가 생부가 아니더라도 응답자 본인이 "아버지"라고 여기면 그대로 아버지로 인정하며, 자녀도 입양자녀, 직접 낳은 자녀, 재혼을 통해 얻은 자녀 상관없이 자신의 주관적인 응답을 존중하였다.
- **금전적 지원**(정기적 지원과 비정기적 지원 구분)**:** 금전적 지원은 현금을 포함하여, 청구된 각종 비용을 대신 지불해 준 경우, 예를 들어 의료비, 보험, 학비, 주택구입 할부금이나 전(월)세 등을 모두 금전적인 지원의 형태로 보았다.
- **정기적/비정기적인 지원:** 정기적인 지원은 일반적으로 일정한 기간을 간격(예: 한 달에 한번, 두 달에 한번 등)으로 반복적으로 이루어지는 경우를 의미하며, 비정기적인 지원은 명절이나 생신을 제외하고 불규칙적 혹은 예측하지 못하고 갑작스럽게 발생한 비용(예: 병원비, 학비, 불규칙적으로 주는 생활비 등)을 지원받거나 지원한 경우를 의미하다.
- **비금전적 지원**(현물지원)**:** 비금전적 지원은 현물이나 선물 형태로 지원을 의미하며, 직접 장을 봐 드리기, 김치나 반찬해 드리기와 같은 서비스내용도 비금전적인 지원에 포함시키다. 다만, 손자녀 돌보기, 수발들기 등의 돌봄노동은 비금전적 지원에서 제외되며 별도로 다루었다.
- **'생존한 모든 자녀'의 기본 정보:** 설문지 문항번호 **'Ba02부터 Ba11'**까지이며, 설문문항에 따른 변수명 규칙인 문항번호 앞에 'W01' 붙이기와 문항번호 뒤에 순서를 나타내는 '_01'과 같은 번호 붙이기 규칙이 적용된다(※자세한 내용은 제Ⅲ장 4절 참조). 예를 들면 문항번호 'Ba02'에 대한 변수명은 첫 번째 자녀인 경우 'W01Ba02_01', 두 번째 자녀인 경우'W01Ba02_02', 세 번째인 경우 'W01Ba02_03'이 된다.
- **'비동거 자녀'와의 접촉정도(Contact)와 소득이전(Transfer)정보 : 'Ba12부터 Ba46'**까지 설문에서는 응답자와 비동거 자녀간의 접촉정도와 소득이전 정보를 자세히 다루고 있다.
- **부모와 금전적/비금전적 지원 문항의 흐름:** 응답자와 동거하지 않는 부모에 대한 금전적/비금전적인 지원여부를 묻는 설문은 다음과 같은 조건에 따라 질문을 달리 하였다.

문항번호	질문이 나뉘는 조건
B033~B064	부모님이 모두 생존하시고, 부부가 함께 동거하는 경우
B065~B102	아버님만 생존하시거나, 부모님이 모두 생존하셔도 부부가 별거중인 경우
B103~B140	어머님만 생존하시거나, 부모님이 모두 생존하셔도 부부가 별거중인 경우

3) 건강영역

건강영역은 크게 건강상태, 일상생활수행능력, 의료보장과 시설이용, 인지력, 신체기능 측정과 같이 5개 설문영역으로 나뉘어져 있다.

[그림 Ⅴ-4] 건강상태 설문구조



◆ 주관적인 건강상태 : 'C001와 C142' 설문 순서 무작위 배치

- 본인의 건강상태를 주관적으로 평가하는 질문은 두 가지 다른 버전으로 조사하였다. C001은 "보통"을 가운데 3번에 놓고 5점 척도로 선택지를 구성하였고, C142는 "최상"을 1번에 "보통"을 4번으로 하는 5점 척도 선택지를 구성하였다. 이 두 질문 중에서 어떤 질문이 먼저 나오고 나중에 나오는지는 컴퓨터가 무작위로 결정하도록 하여 질문이 나오는 순서(위치)에 따른 응답편의를 제거하였다.

- 즉 어떤 응답자들은 첫 질문이 C142가 뜨고 마지막 질문은 C001로 끝나게 되고, C001을 먼저 응답한 사람들은 마지막에 C142에 응답하는 형식이다. 선택지에 따라서 응답자들의 주관적인 건강상태 평가가 어떻게 달라지는지를 비교하고, 다른 외국의 고령자패널조사에도 같은 설문형식이 있으므로 건강에 대한 주관적인 판단의 국제비교연구에 쓰일 수 있도록 설문을 구성하였다.

[그림Ⅴ-5] 일상생활수행능력 설문구조

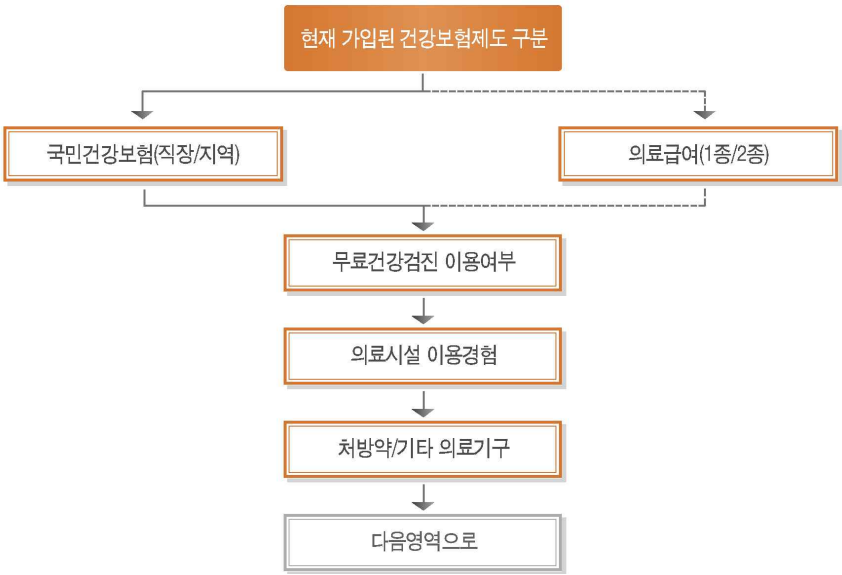
◆ 일상생활수행능력(ADL)과 수단적 일상생활수행능력(IADL) 문항에 대한 응답기준

- ‘최근 일주일 동안’의 활동을 기준으로 한다. 지금은 잠시 아프거나 다쳐 도움을 받지만, 앞으로 3개월 이내에 고쳐질 것으로 예상되는 것은 일상생활수행이 가능한 것으로 보기 때문에 ‘도움이 필요 없는 것’으로 간주하였다.
- 수단적 일상생활수행능력에서 주변에서 응답자를 돕기 때문에 혹은 자기 의향이 없어서 안 하는 경우 ‘도움이 필요하다’는 항목들에 해당되지 않는다. 즉 할 수는 있지만 안 하는 것은 도움이 필요하다로 보지 않는다. 신체적, 정신적, 인지적인 문제로 그와 같은 일이 불가능한 경우만 ‘도움이 필요하다’고 보았다.

◆ 일상생활수행능력(ADL/IADL) 수발 노동자

- B.가족영역에서의 수발노동은 응답자가 일상생활수행능력이 부족한 사람을 직접 돕는 경우를 묻는 설문이고, 본 영역에서는 응답자 본인이 일상생활수행능력이 부족하여 **‘응답자를 도와주는 사람에 대한 기본적인 사항’**을 묻는 설문이므로 두 영역을 혼동하지 않도록 주의해야 한다. 또한 본 영역에서는 가장 많이 도움을 주는 사람부터 세 번째로 도움을 많이 주는 사람까지 최대 세 명의 간병수발자를 구분할 수 있도록 설문을 구성하였다.

[그림Ⅴ-6] 의료보장과 시설이용 설문구조



◆ 의료비 지출문항 주의사항 : 응답자 본인이 지불한 금액만을 측정함

- 의료비 지출에 있어서 가장 비용이 많이 드는 상해, 수술 또는 질병으로 인한 입원비, 치과진료, 한방진료를 따로 묻고 나머지는 기타 외래진료로 구분하였다. 이에 대한 비용은 다른 가족구성원이나 각종 의료보험에서 지불한 금액을 제외하고 실제적으로 자신직접 지출한 금액을 중점으로 물었다.

[그림Ⅴ-7] 인지력 설문구조와 도구

인지력 설문 안내

- 인지능력은 응답자의 치매정도를 판단하는 측정방법 중 하나이다. 고령화과정에서 치매는 중요한 이슈이고, 45세이상 응답자의 인지능력을 정기적으로 측정하다보면 우리나라 인구의 치매 진행 과정을 연구하는 데 도움이 될 수 있다고 판단하여, 인지력 측정을 기본조사 영역에 포함시켰다.
- 정확한 인지능력 측정을 위하여 관련전문가의 도움을 받아 예비조사와 본조사에서 면접원 교육을 따로 실시하였다.
- 기억력 측정과 계산능력 측정에 면접원이 주관적으로 개입하지 않고 컴퓨터가 시간과 계산의 정답 오답을 측정할 수 있도록 CAPI 설문 프로그램을 프로그래밍하여 보다 정확하게 인지능력을 측정하고자 노력했다.

[그림Ⅴ-8] 신체기능측정 설문 구조와 도구

신체계측 설문 안내

- 고령화연구패널에서 사용한 악력계는 타니타 6103 모델을 사용하였다.
- 외국의 고령자패널조사에서는 공간이 확보가 된다면 걷기 속도를 측정하기도 하고, 한 발로 서서 균형잡기에서부터 혈액이나 타액을 채취하여 검사하기도 한다. 우리나라의 주택구조와 면접원들이 계측해 올 수 있는 현실적인 방법을 고려하여 제1차 기본조사에서는 악력을 측정하였다.
- 악력 측정은 측정을 할 수 있는 상태의 응답자인지 아닌지를 파악하고, 본인인 원하지 않는거나 현재 한쪽 손이 다치거나 아픈 상황이면 측정하지 않는 것을 원칙으로 진행하였다.

4) 고용영역

◆ 고용영역에서 주의해야 할 사항 및 주요 개념

- **돈벌이가 되는 일:** 고용영역에서의 정확한 임금근로자 기준이 아닌, 아르바이트나 비정기적인 부업, 소일거리 등 아주 작은 일이라도 소득이 발생하면 인정하는 개념.

- **은퇴:** 본격적인 소득활동을 그만두고 지금은 일을 하지 않거나 소일거리 정도의 일을 하고 있는 경우로, 앞으로도 특별한 변화가 없는 한 소일거리의 일 이상을 할 의사가 없는 경우를 말함.

- **전일제/시간제 근로:** 파트타임, 아르바이트로 일하거나, 같은 업무에 종사하는 사람들보다 적은 시간 동안 일하거나, 임금이 시간 단위로 지급되는 경우를 시간제 근로라 하며, 시간제 근로가 아닌 하루 종일 근무하는 경우를 전일제 근로로 구분함.

- **상용직 근로자 :** 근로계약이 1년 이상인 근로자거나, 정해진 계약기간 없이 본인이 원하면 계속 일할 수 있는 경우.

- **임시직 근로자** : 근로계약기간이 1개월 이상 1년 미만인 근로자거나, 근로계약기간이 없더라도 1년 이내에 이 일이 끝날 것이라고 예정된 경우. (단, 한 직장에서 오래 일하였거나 앞으로도 계속 일할 것으로 예상된다 하더라도 근로계약기간이 1년 미만이면 임시직.)

- **일용직 근로자** : 근로계약기간이 1개월 미만인 근로자거나, 매일매일 고용되어 일당제 급여를 받고 일하는 경우, 또는 일정한 장소 없이 돌아다니면서 일한 대가를 받는 경우에 해당함.

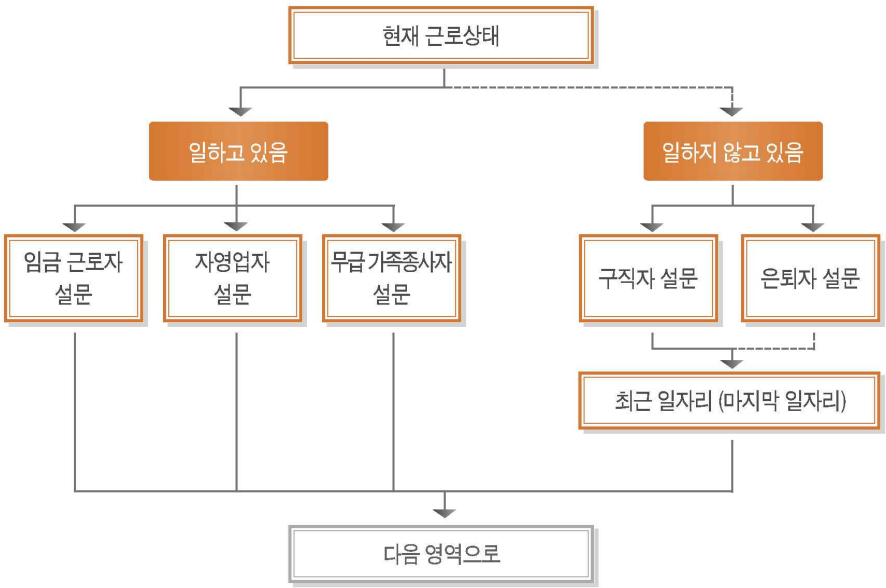
- **파견업체 근로자** : 일하는 곳에서 임금을 받는 것이 아니라 자신을 관리하는 업체에서 임금을 받고 파견법을 적용받는 근로자.

- **용역업체 근로자** : 일하는 곳에서 임금을 받는 것이 아니라 자신을 관리하는 업체에서 임금을 받고 파견법을 적용받지 않는 근로자.

- **도급제 임금방식:** 수급인이 어떤 일을 완성할 것을 약정하고, 도급인이 그 일의 결과에 대하여 보수를 지급할 것을 약정함으로써 성립하는 계약을 말함. 고용과 위임과 구별되는 점은 '일의 완성'을 목적으로 한다는 점에 있으며, 임금도 이에 따라 지급.

- **현재 근로상태에 따른 설문 구조**: 고용영역은 ***'D001부터 D010의 설문'***을 통하여 응답자의 현재 근로상태가 구분된다.

[그림Ⅴ-9] 고용영역의 설문 구조



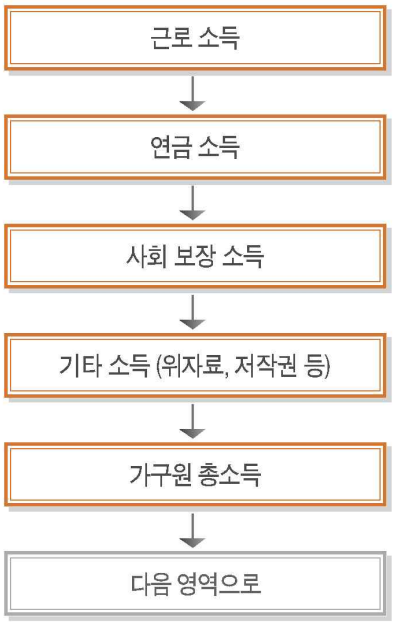
◆ **근로상태 분리지점(문항번호 D001부터 D010까지) 설문에서 주의해야 할 사항**

- 고용주와 계약을 맺고 임금을 받는 모든 경우는 임금노동자에 속한다.
- 개인 사업을 포함하여 개인의 유·무형 자산을 가지고 일을 하는 경우는 "자영업자"에 속한다. 예를 들어, 작가나 예술가 등도 자영업자에 포함된다.
- 임금을 받지 않고 가족이 경영하는 사업장에 나가서 일을 도와주는 경우는 무급가족종사자이며, 통계청 경제활동인구조사의 규정에 다라 18시간이상만 포함하였다. 가족이나 친척의 일을 돕는다 하더라도 용돈이나 수고비 등 어떠한 상태로든 정규적인 비용을 받는다면, 18시간 노동시간과 상관없이 임금노동자로 간주된다.
- 응답자가 현재 여러 가지 근로를 하고 있는 경우: '가장 주된 일자리'를 응답하도록 함.

◆ 도입부분에서 근로상태가 결정되면, 응답자가 임금근로자인 경우 D101번부터 D206번까지의 설문을, 자영업자인 경우 D300번대, 무급가족종사자는 D400번대, 구직자는 D500번대, 은퇴자는 D600번대 설문을 수행한다. ***단, D500번대 구직자 중 근로경력이 있는 응답자와 D600번대 은퇴자 응답자에게는 가장 최근일자리 D700번대의 설문이 추가되었다.***

5) 소득 영역

[그림Ⅴ-10] 소득영역의 설문구조



◆ 소득영역에서의 주의사항

- 응답자의 모든 소득은 ***'응답자 본인만의 소득'***을 기준으로 응답하도록 하였다. 즉, 남편이나 자녀 등 다른 가구원의 소득을 자신의 소득으로 응답하지 않도록 하였다. 또한 응답자의 소득은 ***근로소득을 포함하여 모든 형태의 소득에 응답***하도록 하였다. 즉, 소득영역에서는 고용영역에서 응답한 '가장 주된 일자리'에서의 소득을 포함하여 모든 일자리에서의 근로소득과, 연금소득, 기타소득 등이 포함된다.

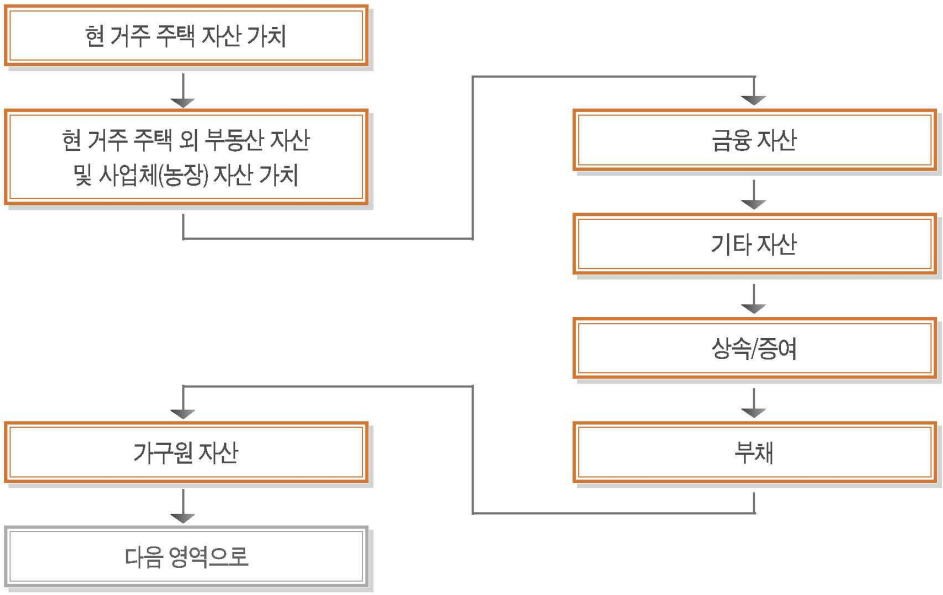
- 소득단위는 **"세후 소득"**이며, **금액은 "_____만원" 단위,** 그리고 **"작년 한해(2005년)"**는 2005년 1월 1일 부터 12월 31일까지를 의미한다.

- 소득영역 마지막 부분에 가구원 총소득을 묻는 설문이 있다(E126번). 그런데 가구내 45세이상인 응답자가 두 명이상인 경우 가구원의 총소득 설문 응답이 응답자에 따라 제 각기 다르게 나올 수 있다. 가구 총소득을 통일시키고 싶으면, 가구원 총소득의 대표응답자 변수인 'w01CV050_r'를 이용하면 된다.

◆ 소득영역은 다른 영역과 달리 자신의 소득을 밝히고 싶지 않은 경우 '모르겠음'이나 '응답거절'과 같은 무응답이 가장 발생하기 쉬운 영역이다. 그러므로 이러한 무응답을 다중대체 보정방법(Multiple Imputation)을 통해 보정한 Imputation 데이터 셋(set)을 이용하면 유용할 것이다.(※ 제4장 항목무응답과 보정방법 참고)

6) 자산영역

[그림Ⅴ-11] 자산영역의 설문구조



자산영역에서의 주의사항

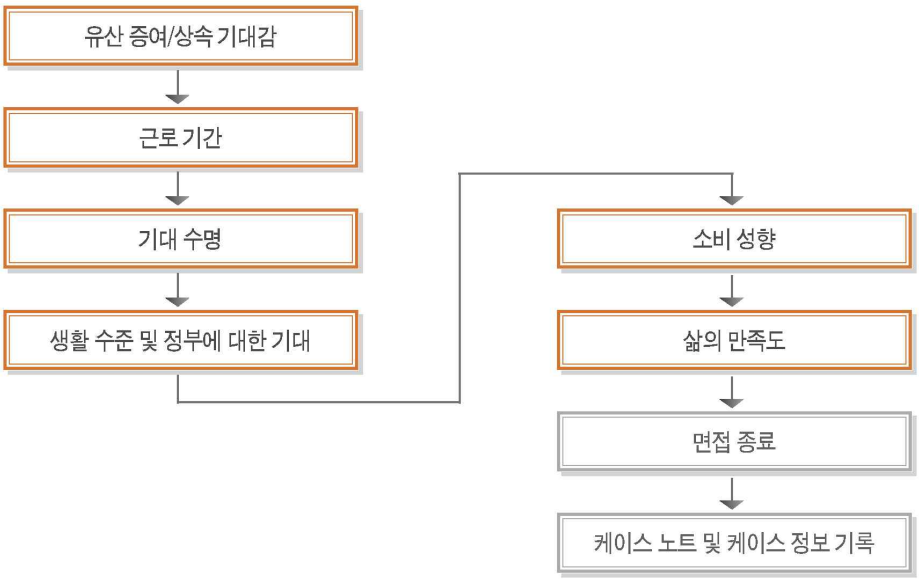
- 자산영역에서 조사하는 자산은 부동산, 사업체, 금융, 상속/증여, 부채이며, ***'명의'를 기준으로 자신의 명의로 된 자산만을 응답***하도록 하였다. 명의가 공동명의로 되어 있는 경우, 명의자수를 기입하게 함으로써 정확한 자산규모를 파악하고자 하였다.

- 문항번호 'F001부터 F048'까지의 현 거주 주택의 자산가치의 설문을 우선 묻고, 그 외의 부동산 자산에 대한 설문이 뒤따라 나오는 구조이므로 부동산 자산은 현재 거주지 자산과 그 외의 부동산 자산을 합해주어야 한다.

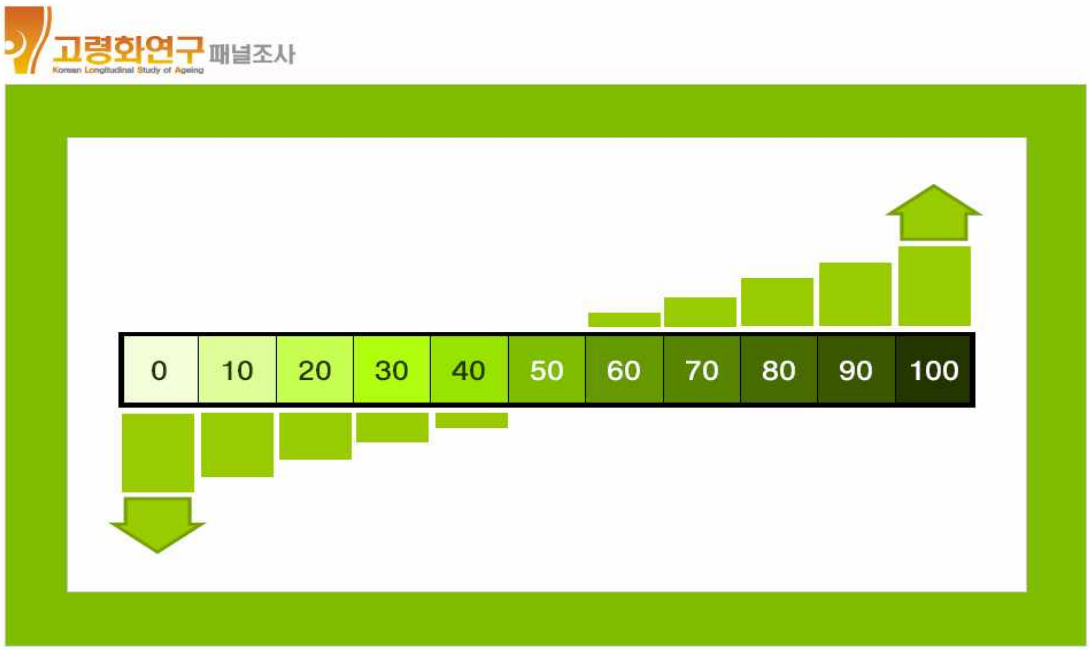
◆ 소득영역과 마찬가지로, 자산에 관한 질문은 '모르겠음'이나 '응답거절'과 같은 무응답이 발생하기 쉬운 영역이다. 이러한 무응답은 다중대체 보정방법(Multiple Imputation)을 통해 보정한 Imputation 데이터 셋(set)을 이용하면 유용할 것이다.
(※ 제Ⅳ장을 참고, 더 자세한 내용은 홈페이지 사용자관련자료에서 다운로드 받을 것)

7) 주관적 기대감과 삶의 만족도

[그림Ⅴ-12] 주관적 기대감과 삶의 만족도 설문구조 및 사용도구



[그림Ⅴ-13] 주관적 기대감과 삶의 만족도 영역에서 주로 사용한 척도 보기 그림



◆ 보기카드를 활용한 100점 척도 사용

- 선진국에서 조사하는 고령자들의 주관적 기대감과 삶의 만족도는 퍼센트(%) 개념을 도입하여 가능성을 물어보도기도 하지만 우리나라 고령자들은 퍼센트(%) 개념이 부족하여 응답이 어렵다고 판단하여 100점 보기카드를 만들어 적용하였다.

- 주관적 기대감에 대한 응답은 '**가능성**'에 대한 수치이다. '그러한 일이 전혀 일어날 수가 없다'고 생각하면 0점, '그러한 일이 반드시 일어난다'고 생각하고 0점에서 100점사이의 점수를 부과하도록 하였다.

- 삶의 만족도 문항에 대한 점수는 '**만족도**'에 대한 수치이다. 만족하지 않으면 0점, 만족하면 100점을 부과하도록 하였다.

VI

2007년 직업력 조사

1 직업력 조사 개요

1) 직업력 조사의 배경 및 목적

2006년 제1차 기본조사를 시작으로 매 짝수년에는 동일한 설문을 묻는 기본조사를 실행하여 패널자료를 구축하고 있다. 짝수해에 실행되는 기본조사를 통해서 시계열적 패널자료를 구축하고 분석하려면 적어도 3차년도 이상의 조사기간(6년)이 요구되기 때문에 연구자들은 일정시기가 되기 전까지는 패널조사를 횡단면조사로 이용할 수 밖에 없는 한계가 있다. 이를 극복하고자 홀수년도에는 특정한 주제에 대한 심도있는 설문조사를 통해 우리나라 고령화의 특징을 나타내는 기초자료의 축적을 시도하고자 한다.

그러므로 2006년 제1차 KLoSA 기본조사를 마치고 2007년 자료를 인터넷에 공개했으며, 2007년도 홀수해에는 '직업력 조사'를 실시하여 응답자의 전 생애에 걸친 근로형태를 파악하였다. 직업력 조사를 통해 응답자의 근로기간에는 종사상 지위별 세분화하여 특징을 파악하였고 비근로기간에는 어떠한 일을 했는지 캘린더 형식으로 채워나감으로써 근로기간과 비근로기간간의 구별과 생애 주기별 특징을 반영하여 응답자의 근로형태 및 여부를 패널분석을 활용할 수 있도록 하였다.

2) 조사설계 특징

이 조사를 설계함에 있어서 세 가지를 특히 유의하였다.

- 첫째, 예비조사 결과 응답자들이 '그럴 듯한 직업'만을 응답하고 그렇지 못한 직업은 본인 스스로가 직업으로 여기지 않는 경향이 있어 응답자가 응답하기 쉬운 직업구분으로 종사장 지위를 세분화 하여 설문문항을 구성하였다.

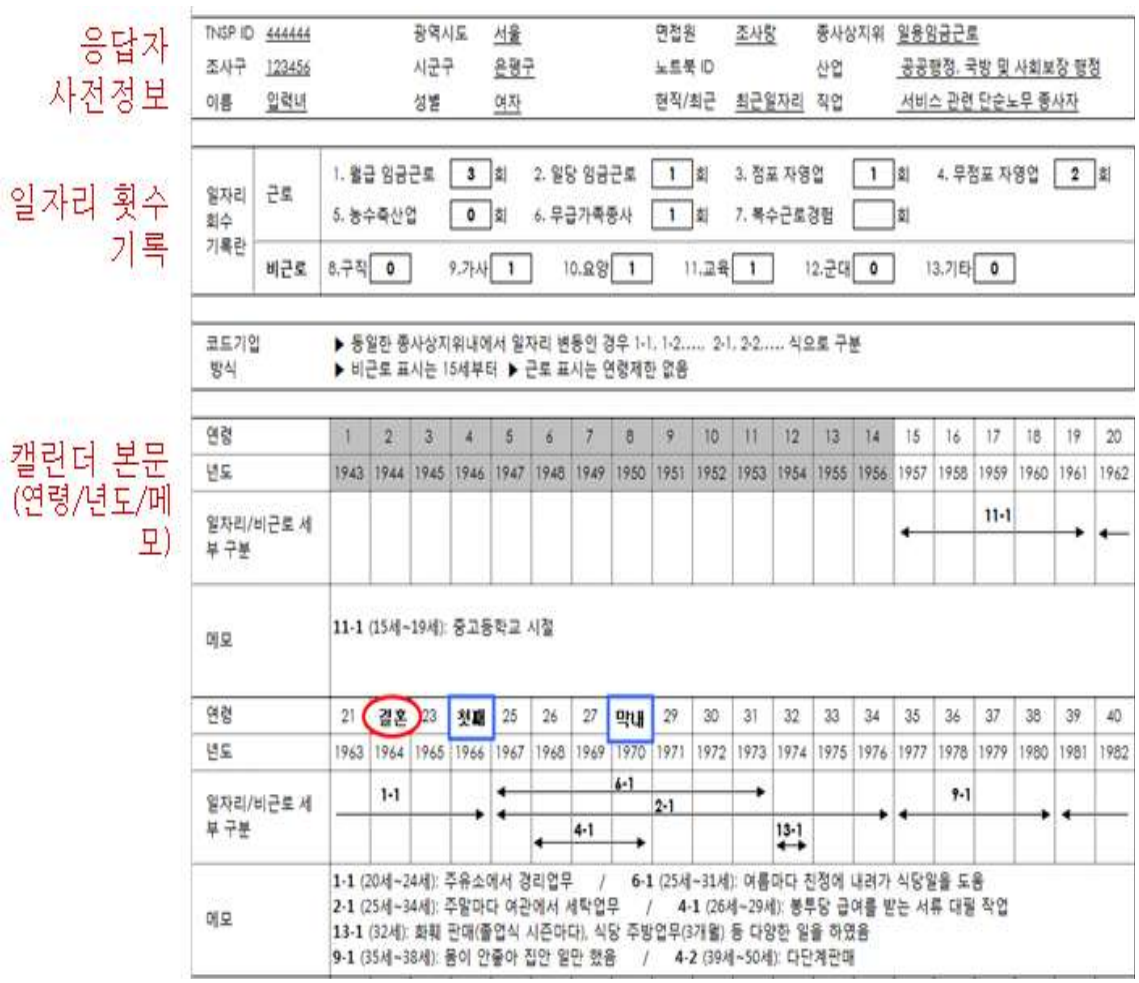
- 둘째, 전 생애에 걸친 기억인 만큼 응답자의 기억을 떠 올리는데 중심을 두어 캘린더 형식의 종이와 CAPI를 병행하는 설문 설계를 하였다. 즉, 종이캘린더를 통해 자신의 생애 주요 이벤트(예: 결혼, 출산)를 표기하고 이를 중심으로 직업을 떠 올리며 직업력을 캘린더에 그려 전체적인 상황을 먼저 파악한 후 CAPI를 이용하여 세부적인 상황을 물어 조사를 실시하였다.

- 셋째, 캘린더에 회상을 하면서 직업이 있던 시기를 우선적으로 근로기간으로 파악하고 아무런 직업이 없었던 시기는 비근로기간 항목을 두어 그 시기에 응답자는 무엇을 했는지 파악할 수 있도록 설계하여 비근로기간의 특징을 파악할 수 있도록 설계하였다.

3) 조사 대상 및 조사 방법

2006년 제1차 고령화연구패널 기본조사 결과 10,254명의 패널이 구축되었고, 이 패널을 모집단으로 전수조사 하였으며, 종이 캘린더와 노트북을 이용한 대인면접법(Computer Assisted Personal Interviewing (CAPI))을 병행하였다.

[그림2] 직업력 조사에 사용된 종이 캘린더 예시



직업력 조사는 전 생애에 걸쳐 자신의 직업을 회고하는 것이 중요하므로, 캘린더를 이용하여 응답자의 기억을 돕고 이후에 캘린더와 컴퓨터를 함께 놓고 설문을 진행하는 방식을 택했다. 이때 이 캘린더는 제1차 기본조사를 바탕으로 개인마다 기본적인 정보를 각각 담아서 응답자 개별 캘린더를 면접원이 가지고 방문하였다. 캘린더를 통해 기억을 회상한 후 다시 CAPI 설문을 통해 응답자의 응답 순서에 따라서 각각의 근로특성에 따른 설문을 진행하였다.

4) 주요 설문 구성

[그림 1] 직업력 조사의 구성도

2 실사과정 및 응답률

1) 실사과정

직업력 조사는 2007년 구축된 고령화연구패널조사의 패널 10,254명을 대상으로 전수조사하였다.

실사진행 경과를 살펴보기 앞서 면접원 교육과 관련한 내용을 간단히 소개한다. 2007년 8월말부터 9월 초까지 실시한 직업력조사 면접원 교육은 서울(경기 강원포함) 2회, 부산 1회, 대구 1회, 광주 1회, 대전 1회 등 모두 6차례 이루어졌다. 면접원교육에서는 KLoSA 개요 및 직업력 조사의 중요성에 대한 소개, 직업/산업 분류 방법, 종이캘린더를 이용하여 응답자의 기억을 돕는 방법, 설문내용 숙지, 모의 면접 등의 구성으로 다루어졌다. 지역별 교육에서는 패널 접촉 요령, CAPI를 위한 노트북사용법, 설문내용, 모의 면접 등을 다루었다.

또한 종이캘린더와 실제 데이터간의 데이터를 비교하기 위해 실사진행중 에디터를 선별하여 에디터 교육을 실시하였다. 에디터 교육은 지역사무소별 에디터를 선발하여 직산업분류, 일자리 구분 교육을 포함하여 면접원이 받은 교육 내용과 더불어 데이터 구조를 파악하고 해당 변수를 확인하는 교육내용을 포함시켜 에디터당 2일의 교육기간을 거쳤다. 에디터의 주요 업무는 면접원이 조사해온 캘린더와 데이터가 일치하지 않는 경우 면접원과 응답자에게 확인하는 작업하는 것이다.

2) 응답률

고령화연구패널 제1차 기본조사에서 구축된 10,254명의 패널에 대한 직업력조사 결과 9,026명의 면접에 성공하여 개인성공률 88.0%를 보였다. 여기서 면접조사가 불가능한 사망자(120명)과 해외이민자(6명)을 제외하면 개인성공률은 89.1%로 나타난다(표1 참조).

가구단위로 볼 경우 패널가구 6.171 가구중 6.304 가구에서 가구내 모든 패널에 조사가 이루어졌으며, 106가구에서는 패널중 일부만 조사에 응했다(표 2 참조).

<표 Ⅵ-1> 직업력조사 실사결과: 개인별

		사례수								%								
		패널	성공	사망	해외 이민	입원	조사 거절	부재 중	추적 실패	계	성공1	성공2	사망	해외 이민	입원	조사 거절	부재중	추적 실패
	전체	**10254**	9026	120	6	51	609	107	335	100.0%	88.0%	89.1%	1.2%	0.1%	0.5%	5.9%	1.0%	3.3%
광역시 도별	서울	1760	1465	18	4	4	146	18	105	100.0%	83.2%	84.3%	1.0%	0.2%	0.2%	8.3%	1.0%	6.0%
	부산	745	663	8	0	6	47	8	13	100.0%	89.0%	90.0%	1.1%	0.0%	0.8%	6.3%	1.1%	1.7%
	대구	561	499	6	0	5	38	8	5	100.0%	88.9%	89.9%	1.1%	0.0%	0.9%	6.8%	1.4%	0.9%
	인천	557	453	6	0	6	49	21	22	100.0%	81.3%	82.2%	1.1%	0.0%	1.1%	8.8%	3.8%	3.9%
	광주	401	372	3	0	1	18	3	4	100.0%	92.8%	93.5%	0.7%	0.0%	0.2%	4.5%	0.7%	1.0%
	대전	388	345	7	2	0	21	0	13	100.0%	88.9%	91.0%	1.8%	0.5%	0.0%	5.4%	0.0%	3.4%
	울산	318	284	1	0	3	18	1	11	100.0%	89.3%	89.6%	0.3%	0.0%	0.9%	5.7%	0.3%	3.5%
	경기	1934	1631	21	0	7	128	29	118	100.0%	84.3%	85.3%	1.1%	0.0%	0.4%	6.6%	1.5%	6.1%
	강원	395	355	3	0	0	28	5	4	100.0%	89.9%	90.6%	0.8%	0.0%	0.0%	7.1%	1.3%	1.0%
	충북	392	369	4	0	0	11	1	7	100.0%	94.1%	95.1%	1.0%	0.0%	0.0%	2.8%	0.3%	1.8%
	충남	560	515	12	0	2	19	1	11	100.0%	92.0%	94.0%	2.1%	0.0%	0.4%	3.4%	0.2%	2.0%
	전북	488	465	7	0	4	9	1	2	100.0%	95.3%	96.7%	1.4%	0.0%	0.8%	1.8%	0.2%	0.4%
	전남	479	451	7	0	6	14	0	1	100.0%	94.2%	95.6%	1.5%	0.0%	1.3%	2.9%	0.0%	0.2%
	경북	602	553	10	0	4	21	4	10	100.0%	91.9%	93.4%	1.7%	0.0%	0.7%	3.5%	0.7%	1.7%
	경남	674	606	7	0	3	42	7	9	100.0%	89.9%	90.9%	1.0%	0.0%	0.4%	6.2%	1.0%	1.3%

<표 VI-2> 직업력조사 실사결과: 가구별

	가구수										%								
	패널	성공가구			사망	해외 이민	입원	조사 거절	부재 중	추적 실패	성공가구			사망	해외 이민	입원	조사 거절	부재 중	추적 실패
	가구	성공계	전원성공	부분성공							성공계	전원성공	부분성공						
전체	6171	5410	5304	106	**119**	4	42	422	83	219	87.7%	86.0%	1.7%	**1.9%**	0.1%	0.7%	6.8%	1.3%	3.5%
서울	1076	887	868	19	18	3	4	102	15	67	82.4%	80.7%	1.8%	1.7%	0.3%	0.4%	9.5%	1.4%	6.2%
부산	450	401	392	9	8	0	5	33	7	8	89.1%	87.1%	2.0%	1.8%	0.0%	1.1%	7.3%	1.6%	1.8%
대구	337	297	289	8	6	0	4	27	7	4	88.1%	85.8%	2.4%	1.8%	0.0%	1.2%	8.0%	2.1%	1.2%
인천	400	324	319	5	6	0	5	36	17	19	81.0%	79.8%	1.3%	1.5%	0.0%	1.3%	9.0%	4.3%	4.8%
광주	233	216	214	2	3	0	1	11	1	3	92.7%	91.8%	0.9%	1.3%	0.0%	0.4%	4.7%	0.4%	1.3%
대전	243	212	211	1	7	1	0	14	0	9	87.2%	86.8%	0.4%	2.9%	0.4%	0.0%	5.8%	0.0%	3.7%
울산	188	167	166	1	1	0	2	11	1	8	88.8%	88.3%	0.5%	0.5%	0.0%	1.1%	5.9%	0.5%	4.3%
경기	1170	990	970	20	21	0	5	86	20	71	84.6%	82.9%	1.7%	1.8%	0.0%	0.4%	7.4%	1.7%	6.1%
강원	215	193	190	3	3	0	0	18	4	2	89.8%	88.4%	1.4%	1.4%	0.0%	0.0%	8.4%	1.9%	0.9%
충북	235	223	219	4	4	0	0	7	1	4	94.9%	93.2%	1.7%	1.7%	0.0%	0.0%	3.0%	0.4%	1.7%
충남	288	271	260	11	11	0	1	12	1	6	94.1%	90.3%	3.8%	3.8%	0.0%	0.3%	4.2%	0.3%	2.1%
전북	292	281	276	5	7	0	3	6	1	2	96.2%	94.5%	1.7%	2.4%	0.0%	1.0%	2.1%	0.3%	0.7%
전남	293	274	270	4	7	0	6	10	0	1	93.5%	92.2%	1.4%	2.4%	0.0%	2.0%	3.4%	0.0%	0.3%
경북	361	324	319	5	10	0	3	19	3	9	89.8%	88.4%	1.4%	2.8%	0.0%	0.8%	5.3%	0.8%	2.5%
경남	390	350	341	9	7	0	3	30	5	6	89.7%	87.4%	2.3%	1.8%	0.0%	0.8%	7.7%	1.3%	1.5%

3 주요내용 및 사용시 주의사항

1) 주요내용

<표 3> 직업력 조사_일자리 특성 자료의 영역별 주요내용

세부 영역	주 요 내 용
1. 월급을 받는 상시 임금 근로자	• 근로기간, 직산업분류, 사업장 규모, 퇴직이유 • 45세이후에 해당한 경우: 퇴직1년전 월평균 급여 • 45세때 해당한 경우: 사업장위치, 근로시간, 퇴직금, 국민연금가입년도
2. 일당을 받는 일용 임금 근로자	• 근로기간, 직산업분류, 월평균 근로일, 년평균 근로월수 • 45세이후에 해당한 경우: 평균 일당 • 45세때 해당한 경우:사업장 위치, 근로시간
3. 점포가 있는 자영업자	• 운영기간, 직산업분류, 고용임금근로자정보, 무급가족종사자규모, 사업을 그만둔 이유 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 사업장 위치, 주평균 근로일과 시간, 자기사업을 택한 이유
4. 점포가 없는 자영업자	• 운영기간, 직산업분류, 월평균 근로일, 년평균 근로월수,사업을 그만둔 이유 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 사업장 위치, 주평균 근로시간
5. 농,축,임, 어업 종사자	• 종사기간, 직산업분류, 고용임금주로자, 무급가족종사자 규모, 사업을 그만둔 이유 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 사업장 위치, 주평균 근로시간, 자기사업을 택한 이유
6. 무급가족종사자	• 종사기간, 직산업분류, 사업장 대표, 고용임금근로자정보, 무급가족종사자 규모, 그만둔 이유 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 사업장 위치, 주평균 근로시간, 무급가족종사일을 택한 이유
7. 동시에 다양한 일을 한 경우	• 근로기가, 직산업 분류, 월평균 근로일, 년평균 근로월수 • 45세이후에 해당한 경우: 월평균 순수입 • 45세때 해당한 경우: 일평균 근로시간
8. 구직	• 비근로기간 • 생계유지방법 • 소득원(동거가족 중 누구의 소득원) • 생계지원(비동거 가족의 지원인 경우) • 월평균 수입
9. 가사	
10. 요양	
11. 교육	
12. 군대	
13. 기타	

※ 일반적인 종사상 구분과 달리 점포 유무에 따라 구분한 이유는 직업다운 직업이 아니라고 응답자가 판단할 경우 응답하지 않는 경우를 피하기 위해서이며, 계절이나 상황에 따라 여러 가지 일을 한번에 동시한 경우도 포착하기 위해서이다.

2) 주의사항

◆ 응답단위

- **기본 응답단위: '45세 이상의 개인' 으로 제1차 기본조사의 응답단위와 동일하다.**

- **금액을 묻는 문항의 모든 단위는 "_____원"이다.** 2006 제1차 기본조사의 금액을 묻는 단위는 "만원"이었다. 그러나 직업력조사에서 응답자의 연령을 고려하면 1940년대 임금을 답하는데 있어 만원의 단위보다는 "원"의 단위가 더 잘 맞기 때문이다.

◆ 데이터 이용시 주의사항

- 2006년 제1차 기본조사의 고용영역에서의 '임금근로자', '자영업자', '무급가족종사자', '가장최근일자리'의 정보를 2007년 직업력 조사 데이터에 포함시켜 주었다. 그러나 직업력 조사는 과정에서 2006년도의 일자리가 잘못 조사된 경우 2007년도 직업력 조사시 확인하여 수정된 정보를 반영하였다. 그러므로 개인 아이디로 추적할 경우 2006년 제1차 기본조사에서의 일자리 정보와 2007년 직업력 조사에서의 정보에 차이가 있을 수 있다.
- **직업력 조사의 자료는 앞에서 설명한 바와 같이 '일자리 특성' 데이터와 '개인특성' 데이터로 분리하여 SAS와 SPSS파일 형태 배포하였다.** 첫 번째 파일명은 "JH_Job~" 으로 시작되는데 이것은 '일자리의 특성'을 나타내는 변수들로 구성되어 있으며, 두 번째 파일인 "JH_Ind~"는 '개인의 특성'을 나태내는 변수들로 응답자의 출생년도, 성별, 결혼 당시 연도, 첫째와 마지막 자녀 출생연도 그리고 15세부터 현재 연령까지 년단위 근로여부 변수가 들어있다. 두 개의 파일은 'PID' 변수를 이용하여 연결하여 사용할 수 있다.

- 그러나 직업력조사의 자료는 응답자의 개인적 특성을 나타내는 정보를 최소화하고 있으므로 보다 풍부한 분석을 위해서 변수명 'PID'를 이용하여 2006년 제1차 기본조사의 자료와 연계하여 이용할 수 있다.

- 직업력조사의 자료는 제1차 기본조사의 자료와 달리 무응답에 대한 보정(Imputation) 자료를 별도로 제공하지 않는다.

◆ 코드북 이용시 주의사항

- 제1차 기본조사에서 변수명을 설문문항에 기초한 조합방식으로 제공하였지만 직업력조사에서는 설문문항과 코드북의 변수명은 일치하지 않는 점을 유의해야 한다.

- 코드북 하늘색 바탕이 칠해진 변수명은 설문을 바탕으로 이용자들이 편리하게 사용할 수있도록 새로 만들어 준 변수를 의미한다.

- GJ001 변수를 통해 응답자의 일자리순서를 알 수 있다. 즉 GJ001의 변수값이 1 이면 응답자의 생애 첫 번째

일자리, 변수값이 2이면 생애 두 번째 일자리, 변수값이 5이면 생애 다섯 번째 일자리를 의미한다.

- GJ002 변수에서는 일자리를 종사상 지위별로 구분하는 변수이다. 응답자의 응답을 쉽게 이끌어 내기 위해 기존의 종사상 지위와는 차별을 둔 형태이다. 여기서 변수값 7번에 해다하는 "동시에 다양한 일"은 해당 시기에 변수값 1번에서 6번사이의 일을 구분하지 않고 다양한 일을 한 경우의 응답을 이끌어 낸 변수이다. 예를 들면 농사를 지으면서 주말에 장에 나가 물건을 팔고(점포없는 자영업), 때로는 일당 받는 임금근로자 일을 섞어가면서 하되 어떤 일이 주된 일자리로 잡기 어려운 경우 7번의 응답을 받았다.

- GJ004부터 GJ005y는 해당하는 기간의 시작과 끝을 알려주는 변수이다.

- GJ006부터 GJ008은 해당 일자리의 산업분류이다. 대분류로 분류했으며 제조업과 도매 및 소매업에 한해서만 한 단계 세부분류로 나누었다.

- GJ011에서 GJ015_01 해당 일자리에서의 직업대분류이다. 사무직, 서비스근로자, 농임어업 순련근로자, 단순노무자에 한해서 세부분류로 나누었다.

- GJ016c01에서 GJ016c10까지는 GJ002에서 변수값 7번에 해당하는 동시에 다양한 일을 한 경우를 직업분류상 해당 일자리에서 일 한 횟수를 계산해준 변수들이다.

- GJ021, GJ024 변수사용시 알아두어야 할 점은 직업력 조사에서는 일당을 받는 일용 임금근로자에게 일일 평균 근무시간과 일일 평균임금 문항이 있다. 그러나 제1차 기본조사에서 일용직 근로자에게 묻는 문항은 주당 평균 근무시간과 월평균 임금을 물었기 때문에 두 자료를 일치시키기 위해서 변수를 조정하였다. 따라서 직업력 조사에서의 일일 평균 근무시간(GJ021)= 주당 평균 근무시간/5 로 계산된 값이고, 일일 평균임금(GJ023)= 월평균 임금/5로 계산해 준 값이다.

VII

2008년 제2차 기본조사의 세부적 내용과 특징

1 2008년 제2차 기본조사 개요

1) 조사의 배경 및 목적

고령화연구패널조사의 기본목적인 우리나라 중고령 인구의 경제활동에 대한 정확한 실태조사와 향후 고령사회로 변화해 가는 개인의 행동을 예측하고 이를 통해 효과적인 사회경제정책 수립과 보다 건강하고 행복한 노후 생활을 위한 연구와 정책을 마련하는데 기초적인 자료를 제공하는데 있다. 이에 2006년 제1차 기본조사이후 매 2년에 한 번씩 동일한 설문을 토대로 동일한 대상자에게 반복적인 조사를 실시하여 시간이 흐름에 따른 변화를 포착할 수 있는 패널자료를 구축하고 있다. 이러한 필요성에 의해 2008년 제2차 기본조사가 이루어졌으며, 1차 기본조사를 통해 미흡했던 설문을 다듬고, 고령자를 대상으로 한 패널의 특징을 반영시키기 위해 사망자에 대한 설문을 포함시키게 되었다.

2) 조사설계 특징

제2차 기본조사를 설계함에 있어 세 가지 점에서 특징이 나타난다. 첫째, 2008년 제2차 기본조사는 제1차 기본조사의 설문을 기본으로 반복적으로 측정한다. 이때 시간이 지나도 변하지 않는 변수(예: 출생일, 성별)를 다시 확인하는 설문을 포함시켜 자료의 정확성을 한번 더 확인할 수 있도록 하였고, 시간이 지나면서 변하는 변수(예: 셋째 자녀의 사망)과 같은 경우에 대비해서 유연하게 설문을 진행시킬 수 있도록 설계에 주의를 기울였다.

둘째, 새로운 영역과 설문이 추가되었다. 우선 제1차 기본조사를 이후에 물을 수 있는 소비영역을 추가하였다. 소비영역은 지난해에 벌어들인 소득에 대한 정보가 있어야 소비를 보다 정확하게 파악할 수 있다는 자문으로 제1차 기본조사에서는 설문영역에 포함시킬 수 없었으나, 제2차 기본조사에는 소득영역 이후에 소비영역을 넣어 이에 대한 분석이 가능하도록 하였다. 또한 마지막 영역 뒤에 비넷(Vignettes) 영역을 넣어 응답자가 가상인물에 대한 상태를 설명 듣고 주관적인 판단을 하도록 하여 건강상태나 직장생활에 제한을 주는 장애 등을 판단하도록

하는 설문이다. 이 영역은 국제비교를 할 때 주관적인 판단에 대한 기준점을 마련하기 위해 사용되는데, 아직 우리나라의 고령자들에게는 설문을 이해하는데 어려움이 있었던 것으로 판단된다. 그러므로 이번 베타버전에 자료로는 제공하지 않는다. 그 밖에 기초노령연금과 같은 고령자를 대상으로 한 새로운 제도의 도입에 따른 설문을 추가하였다.

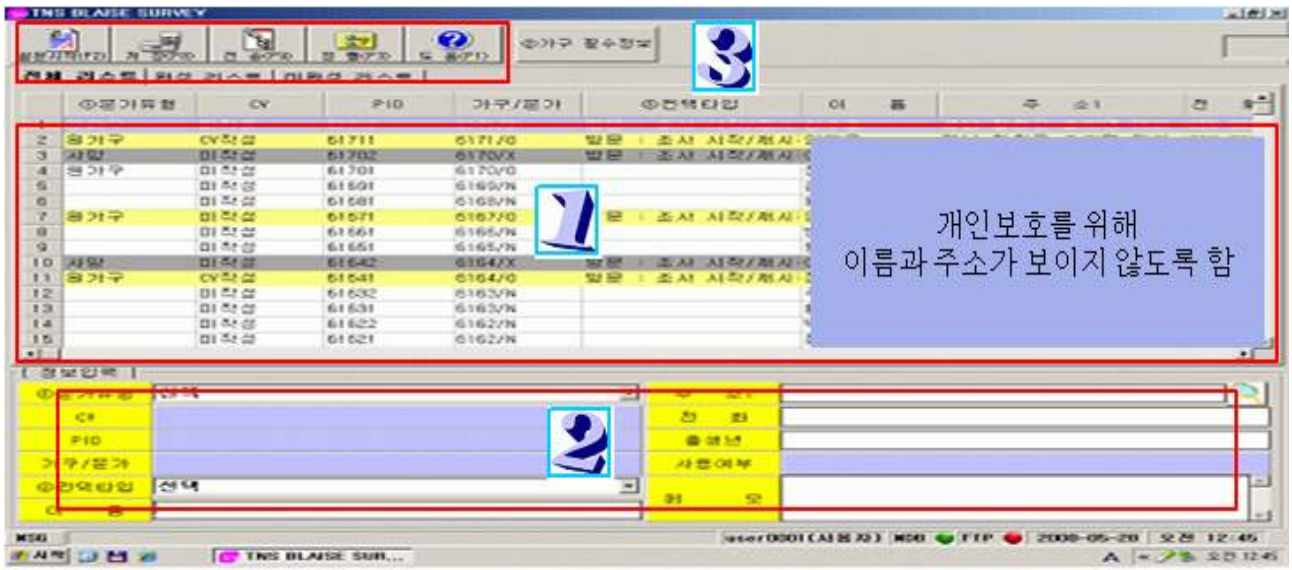
셋째, 사망자 조사(Exit Interview)를 실시하였다. 고령자를 대상으로 한 패널조사라는 특징으로 사망한 패널대상자에 대한 당시 상황이 중요하다고 판단되어 사망자 조사를 설계하였다. 사망자 조사(Exit Interview)에서는 사망자와 평소에 가장 친분이 친한 사람을 대상으로 별도의 영역과 설문을 만들어 대리응답을 하도록 하였다.

3) 조사 대상 및 조사 방법

2006년 제1차 고령화연구패널 기본조사 결과 10,254명의 패널이 구축되었고, 2007년 직업력조사에서는 이 패널을 모집단으로 전수조사 하여 9,026명의 조사를 완료하였다. 2008년 제2차 기본조사에서는 초기 패널대상자인 10,254명을 대상으로 추적을 통해 전수조사 하였다.

2008년 제2차 기본조사에서 면접방법도 CAPI이다. 2006년 제1차 기본조사에서 보다 개선된 부분은 우선 GPMS(Global Panel Management System)를 이용하여 면접원들이 자신이 조사해야 할 응답자와의 연락, 약속잡기, 설문완료 등을 보다 편리하게 활용하고, 중앙에서 면접원들의 개별 사항을 관리할 수 있는 시스템을 개발하고 적용시켰다는 점이다. 또한 2006년 제1차 조사와 2007년 직업력 조사에서 수집한 개별 정보를 미리 컴퓨터에 pre-loading 시켜 기존 응답과 일정수준 이상이 차이가 나는 경우 팝업창을 이용하여 다시 확인하고 잘못 되었다면 수정할 수 있는 설문 로직을 CAPI 프로그램에 장착하였다.

◆ GPMS(Global Panel Management System): 패널대상자 관리 프로그램



- GPMS 프로그램은 위의 그림에서 볼 수 있듯이 1번에 패널대상자들의 리스트와 주소가 있고, 조사과 완료되면 노란색으로 표시가 되도록하였다. 2번의 박스에 조사대상자와의 연락여부와 새로 바뀐 연락처등의 정보를 적고, 3번의 메뉴를 통해 설문시작, 종료, 전송을 한다.

4) 영역별 특징

2008년 제2차 고령화연구패널조사의 자료는 크게 두 가지로 구분된다. 생존자를 중심으로 한 기본조사는 데이터 이름 'W2_Beta_main' 이라는 이름으로 나가며, 사망자에 대한 조사를 한 Exit interview 는 'W02_Beta_exit' 이라는 이름으로 제공된다. **두 자료의 주요내용은 본 유저가이드 pp.7-9의 표에 정리되어 있다.**

◆ 기본조사의 영역별 특징

- 제1차 기본조사의 틀을 그대로 유지한 상태에서 설문을 구성하였으므로, 여기에서는 제1차년도 조사와 달라진 부분을 중심으로 기술하겠다.
- 세부 영역의 이름을 나타내는 영문 부호가 조금 간소화되었다. (예: 건강영역의 경우 예전에는 Ca~Ce 까지 되어 있었으나, C 영역으로 통합하였다.)
- **Ba. 가족영역에서 제2차 기본조사에서는 동거자녀와의 금전적/비금전적 이전소득 문항을 추가**하였다.
- D. 고용영역의 초반에 문항번호 D001~D085까지는 제1차 기본조사 이후 일자리 변동사항을 반복적으로 측정하였다. 그리고 조사 당시를 시점으로 주요 일자리를 기준으로 D100번대에는 임금근로자, D200번대는 자영업자와 같이 번호대를 이용하여 종사상의 지위를 구분하였다.
- **E. 소득 영역에는 소비와 저축 문항을 신설**하였으며, E100번대가 소득, E200번대가 소비 문항이다.
- **G. 영역은 주관적 기대감 및 삶의 질인데, 이 영역 도입부분에 2008년에 새로 도입된 '기초노령연금'과 관련된 설문문항이 있다.** 이것은 소득 영역으로 들어갈 수도 있었으나. 소득 영역의 응답 기준이 2007년 1월에서 2007년 12월에 해당하기 때문에 기초노령연금 소득이 있는 경우 응답기준으로 볼 때 잡힐 수 없기 때문이며, 또한 조사시점이 2008년 8월부터이기 때문에 응답자가 언제 조사에 응하는지에 따라 달라 질 수 있기 때문에 일단을 G.영역에서 다루게 되었고, 2010년 제3차년도 조사에서는 소득영역으로 문항이 들어갈 것이다.

◆ 사망자 조사(Exit interview)의 영역별 특징

- 사망자에 대한 조사는 사망자를 중심으로 가장 친근하게 지냈던 사람을 대상으로 대리응답을 받았다. 대부분이 배우자나 함께 살던 자녀와 같은 친인척이지만, 친구가 응답자가 되기도 하였다.
- 세부 영역은 대리 응답자의 기본정보와 사망자의 사당 당시 상황(원인, 사망일, 등)을 묻고, 건강과 고용상태 그리고 유산과 부채를 묻는 설문으로 구성되었다.

2 실사과정 및 응답률

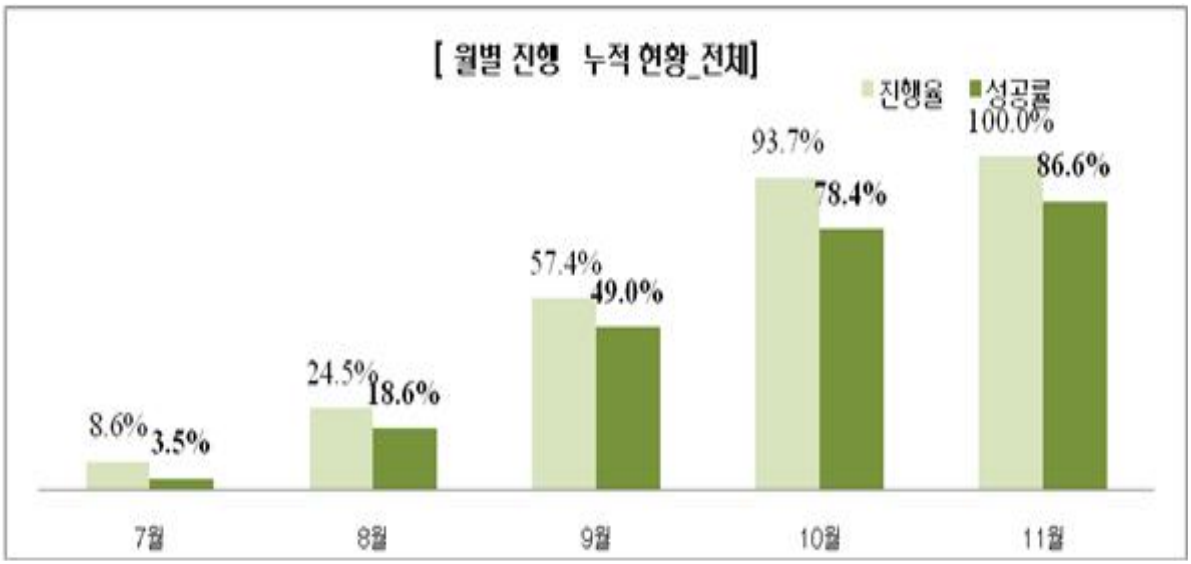
1) 실사과정

실사기간

- 2008년 7월 3일부터 면접원 교육이 시작되었으며, 7월 중순부터 실사 시작
- 실사종료 2008년 11월 30일
- 교육기간 포함 5개월

실사진행

- 면접원 교육이후 본격적인 실시진행은 7월 중순 이후이며, 7~8월은 무더위와 여름휴가 등으로 진행률이 낮다가 9~10월 집중적으로 실사가 진행되었음. 11월은 미진하거나 조사거절을 하시는 분들에 대한 설득작업을 중심으로 이루어짐.



면접원 교육

- 총 92명의 면접원이 참여하였음(수도권을 포함한 서울지역 41명, 부산, 대구, 광주, 대전 각 12~14명)
- 면접원 중 83명은 2006년 제1차 고령화연구패널조사 초기부터 함께 한 면접원이며, 9명이 신규 면접원
- 2008년 처음 투입된 신규 면접원은 별도의 CAPI 교육과 설문지 교육을 각각 하루씩 이틀간 추가적으로 받음
- 2006년부터 참여했던 면접원들은 2008 KLoSA 교육을 오전 10시부터 오후 7시까지 하루 집중교육 받음
- 면접원은 평균 111.5명의 패널을 대상으로 96.5 표본을 성공
- 권역별로는 부산지역 면접원이 133.5명으로 가장 많은 패널을 할당 받았고, 성공 패널도 118.5개로 가장 많았음

2) 응답률

2006년 제1차 고령화연구패널조사를 통해 10,254명의 패널을 구축하였고, 2007년 직업력조사를 거쳐 2008년 제1차 기본조사를 완료하였다. 그러므로 2008년 조사부터는 제1차 조사에서 구축한 원패널(10,254명)을 기준으로 했을때의 응답률을 전체현황에서 알아보고, 이 중에선 생존자를 중심으로 기본조사에 응답한 비율만을 고려하여 생존패널 응답률과 사망자에 대한 Exit Interview 성공을 여부로 사망패널 현황을 각각 정리하는 것이 필요하다. 다음으로 가구단위로 했을때의 응답 성공률을 살펴보겠다.

◆ 전체패널 응답률

- 고령화연구패널조사의 패널 10,254명중 8,875명에 대한 조사 성공으로 전체 성공률 86.6%
- 조사 거절자는 762명으로 전체의 7.4%, 거주지이전(이민, 장기 부재중 포함)에 따른 추적 실패 438명으로 전체 4.3%

[2008년 KLoSA 최종 현황_사례수]

권역	패널	성공	해외이민	입원	추적실패	조사거절	부재중	기타
전체	10,254	8,875	12	41	437	762	106	21
서울	4,646	3,763	8	16	315	456	79	9
부산	1,736	1,541	0	8	43	133	7	4
대구	1,163	1,049	0	5	30	71	7	1
광주	1,369	1,274	0	5	18	66	5	1
대전	1,340	1,248	4	7	31	36	8	6

◆ 생존패널 응답률

- 고령화연구패널조사의 패널 10,254명중 2006년 제1차 기본조사 이후 사망으로 확인된 254명을 제외한 10,000명의 패널 대상자중 8,688명에 대한 조사 성공
- 패널중 생존자 10,000명을 기준으로 성공률은 86.9%

[2008년 KLoSA 최종 현황_사례수]

권역	패널	성공	해외이민	입원	추적실패	조사거절	부재중	기타
전체	10,000	8,688	12	41	415	732	103	9
서울	4,539	3,687	8	16	306	439	77	6
부산	1,694	1,506	0	8	43	129	7	1
대구	1,132	1,028	0	5	25	67	6	1
광주	1,332	1,243	0	5	13	65	5	1
대전	1,303	1,224	4	7	28	32	8	0

사망패널 응답률

- 고령화연구패널조사의 패널 10,254명중 2006년 제1차 기본조사 이후 사망으로 확인된 254명, 이 중 187명에 대한 Exit interview 성공
- 사망자 254명을 기준으로 187명의 사망자 설문 성공으로 성공률은 73.6%

[2008년 KLoSA 최종 현황_사례수]

권역	패널	성공	추적실패	조사거절	부재중	기타
전체	254	187	22	30	3	12
서울	107	76	9	17	2	3
부산	42	35	0	4	0	3
대구	31	21	5	4	1	0
광주	37	31	5	1	0	0
대전	37	24	3	4	0	6

가구단위 응답률

[광역시도별 가구단위 조사결과]

		패널 가구	성공계	성공계 (%)	전원 성공	부분 성공	해외 이민	입원	조사 거절	부재중	추적 실패
권역별	전체	**6171**	**5248**	**85.0**	5205	43	9	36	531	85	300
	서울	**2861**	**2277**	**79.6**	2253	24	6	13	314	64	211
	부산	**1028**	**896**	**87.2**	890	6	0	7	93	5	29
	대구	**698**	**613**	**87.8**	608	5	0	4	56	5	24
	광주	**818**	**756**	**92.4**	751	5	0	5	43	4	14
	대전	**766**	**706**	**92.2**	703	3	3	7	25	7	22
광역시도별	서울	**1076**	**832**	**77.3**	821	11	3	3	116	28	95
	부산	**450**	**390**	**86.7**	387	3	0	3	43	2	12
	대구	**337**	**297**	**88.1**	294	3	0	3	32	2	6
	인천	**400**	**299**	**74.8**	296	3	2	5	51	17	28
	광주	**233**	**211**	**90.6**	211	0	0	1	16	1	5
	대전	**243**	**209**	**86.0**	209	0	2	4	11	4	10
	울산	**188**	**155**	**82.4**	155	0	0	0	21	1	8
	경기	**1170**	**960**	**82.1**	953	7	1	5	122	14	82
	강원	**215**	**186**	**86.5**	183	3	0	0	25	5	6
	충북	**235**	**220**	**93.6**	220	0	0	1	7	2	7
	충남	**288**	**277**	**96.2**	274	3	1	2	7	1	5
	전북	**292**	**272**	**93.2**	271	1	0	1	12	2	6
	전남	**293**	**273**	**93.2**	269	4	0	3	15	1	3
	경북	**361**	**316**	**87.5**	314	2	0	1	24	3	18
	경남	**390**	**351**	**90.0**	348	3	0	4	29	2	9

3 베타버전 데이터 사용안내

2008년 제2차 고령화연구패널조사 Beta 버전 출시

- 제1차 고령화연구패널조사 학술대회 참가를 위해 2009년 6월 30일까지 연구계획서를 제출한 연구자들을 대상으로 제2차 기본조사의 베타버전 자료와 제1차 기본조사 자료(버전 1.2)를 공개한다. 단, 데이터를 제외한 설문지, 코드북, 유저가이드는 홈페이지에 공개할 계획이다.
- 베타버전이 출시되는 2009년 7월 1일 이후, 고령화연구패널팀에서는 제1차 기본조사와 동일하게 무응답 문항에 대한 대체 및 보정작업(Multiple Imputation)을 진행할 계획이며, 학술대회와 여러 연구자들의 지적사항을 검토하여 데이터를 개선하여 2009년 12월에 "2008년 제2차 기본조사 버전 1.0 Data" 로 출시할 예정이다.
- 2009년 말에 출시된 제2차 기본조사 버전 1.0 자료는 홈페이지를 통해 배포할 예정이다.
- 2006년 제1차 기본조사의 자료가 업데이트 된 1.2 버전도 학술대회 참가를 신청한 연구자들에게 공개된다. 2006년 제1차 기본조사 버전 1.2 자료는 2007년 직업력 조사를 통해서 수정된 정보를 담고 있으며, 향후 제1차 기본조사의 버전이 업데이트 되는 것은 제2차 기본조사를 통해 수정된 정보를 담을 것이다. 다음의 표는 버전 1.1에서 1.2로 업데이트 되면서 변경된 정보이다.

수정된 변수	변경 전	변경 후	수정된 케이스 수
가중치 수정			10254
성별 수정			10
태어난 년도/연령 수정			340
태어난 날짜(day) 삭제			10254
고용영역 일자리 시작/끝 년도 수정			446
w01E035ct	결측	-8	2
w01E037	1	결측	1
w01D708/ w01D709	0	-8	41
w01D713	650000	65	1
	0	-8	11
	999997	-8	1
w01D411	97	-8	1

※ 베타버전의 데이터는 실사 후 두 차례에 걸친 데이터 클리닝 작업과 데이터 management 작업을 완료한 것이나, 각 분야의 연구자들이 사용하면서 발견한 새로운 문제점이 발생될 수 있습니다. 그러므로 베타버전 데이터를 사용하면서 발견한 에러나 문제점을 연구진에게 알려주시면(klosa@kli.re.kr), 버전 1.0 으로 업데이트될 때 보다 양질의 데이터로 거듭날 수 있습니다. 여러분들의 적극적인 협조를 부탁드립니다.

KLoSA USER GUIDE

고령화연구패널 2차 조사 자료에서 발생하는 결측값에 대한 다중대체

- 대체된 자료의 분석 -

고령화연구패널조사는 고령화 시대에 대비하기 위하여 한국의 고령자 모집단에 대한 실태를 파악하고 기초 자료를 수집하기 위하여 계획되었고 1차 조사가 2006년에, 그리고 제 1차 추적조사인 2차 조사가 2008년에 실시되었다. 2007년에는 1차 조사 자료에서 발생하는 결측값에 대하여 소득 및 자산 변수를 중심으로 한 다중대체(multiple imputation)가 시행되었고 2008년 시행된 2010년에는 2차 조사에서 발생하는 결측값에 대하여 소득 및 자산 변수를 중심으로 한 다중대체가 시행되었다. 본 User Guide는 2010년 시행된 다중대체을 통하여 생성된 대체(impute)된 자료에 대한 사용자들의 이해를 돕고자 쓰여 졌다. 1장에서는 우선 다중대체를 실시할 때 사용된 대체 모형에 대하여 간략히 설명하고 2장에서는 대체된 자료의 형태 및 자료에서 원래 관찰된 값과 대체된 값을 구별하는 방법에 대하여 설명한다. 3장에서는 대체된 자료를 사용하여 분석을 시행하는 방법에 관하여 설명한다. 대체 모형에 대한 자세한 설명 및 결과 비교는 KLoSA 2차 자료에 대한 다중대체 결과 보고서(report)(송주원 외, 2010)에 나타난다.

1. 결측값 대체 방법

고령화연구패널조사의 경우 대부분의 변수에서 결측값(missing data)의 비율이 5%미만으로 작게 나타났으나 자산 일부 항목에서 결측값의 비율은 10 ~ 20% 내외까지 증가하였으며 일부 응답자가 많지 않은 항목의 경우 40% 정도까지 나타났다. 따라서 결측값을 포함한 변수에 대한 적절한 분석을 위하여 결측이 발생한 주요 항목에 대하여 다중대체(multiple imputation)가 실시되었는데 특히 결측 비율이 높은 소득 및 자산 항목의 대체에 중점을 두고 진행되었다. 각 변수별 결측 비율이 거의 대부분의 변수에서 20% 미만으로 나타났으므로 대체 자료의 개수는 1차와 마찬가지로 5개로 결정하였다.

고령화연구패널조사는 전체 8개의 영역(session)으로 구성되어 있는데 결측값의 대체가 결측 비율이 높은 소득 및 자산 항목의 대체에 중점을 두고 진행되었지만 관련된 주요 변수들의 대체도 함께 실시되었다. 우선 인구영역의 주요 변수들이 5번 대체되었고 대체된 5개의 자료 각각에 대하여 주요 인구영역 변수 및 디자인 변수(design variables)들을 설명 변수로 사용하여 건강 영역 주요 변수의 대체를 실시하였다. 이렇게 대체된 5개 자료 각각에 근거하여 관련 변수를 설명 변수로 사용한 고용 영역의 현재 고용 및 퇴직 소득에 관한 대체를 실시하였다. 다음으로 각 대체된 자료에서 소득과 관련된 변수들을 설명 변수로 사용하여 소득 영역의 주요 소득 항목들에 대한 대체가 실시되었다. 이 때 자산 영역의 집 소유 여부 및 금융자산 총액도 설명변수로 포함되어 소득과 자산의 연관성을 설명하고자 하였다. 소득영역이 대체된 후 대체된 각 개인당 총소득을 계산한 후 다른 관련 변수들과 함께 설명 변수로 설정하여 주요 자산 영역 변수들에 대한 대체를 실시하였다. 마지막으로 가족대표자만이 응답한 가족 영역 중 자녀의 수 및 자녀들에게서 지원받은 액수 및 지원한 액수에 관

한 항목들을 대체하였다. 이 때, 결측이 발생한 변수의 1차년도의 응답값을 설명변수로 포함시켜 설명력을 증가시킴으로써 보다 정확한 예측값을 구할 수 있었을 것으로 기대된다. 만약 1차 조사 시 응답값이 결측인 경우 1차 년도 대체된 자료의 값을 설명 변수로 사용하였다. 즉, 1차 조사의 대체된 5개의 자료 각각을 설명 변수로 사용하여 2차 조사의 대체 자료 5개를 생성하였다. 대체 모형에 사용된 설명 변수에 관한 자세한 정보는 KLoSA 2차 자료에 대한 다중대체 결과 보고서에 기술되어 있다(송주원 외, 2010).

2006년에 실시된 고령화연구패널 제 1차년도 조사의 무응답 대체에서는 수정된 평균에 근거한 핫덱대체 방법을 사용하여 다중대체를 실시하였다. 이 방법은 Little(1988)이 제안한 일종의 핫덱대체(hotdeck imputation) 방법으로서 미국 RAND의 Bell(1999)이 SAS Macro로 프로그램화하여 여러 가지 조사 연구에 적용해 왔다. 고령화연구패널 제 1차년도 자료에 대한 모의실험에서 이 방법은 다른 여러 가지 대체법에 비하여 상대적으로 우수한 결과를 보였다. 따라서 고령화연구패널 2차 추적 조사에서 발생한 결측값도 이 모형을 사용하여 대체를 실시하였고 모의실험을 실시하여 이 모형의 적절함을 확인하였다. 이 방법은 결측이 발생한 자료값을 자료 내 관찰된 값 들 중 하나 또는 여러 개의 값을 가지고 대체시키는 일종의 핫덱대체법이지만 관찰값 중 하나 또는 여러 개의 값을 임의로 선택하는 랜덤 핫덱(random hotdeck) 대신 자료를 비슷한 여러 개의 하위 그룹(subclass)으로 나누어 같은 하위 그룹 내에서 핫덱대체를 실시한다. 이 때 하위그룹은 결측이 발생한 변수에 대하여 관찰된 자료만을 대상으로 회귀모형(regression model)을 적합하여 결측이 포함된 모든 자료에 대한 예측값을 구한 후 예측값에 근거하여 층화(stratification)하여 구성한다. 각 층 내에서 결측값은 같은 층의 관찰자 중에서 기증자(donor)를 선택하여 기증자의 값으로 대체를 실시한다. 이 방법은 기증자를 선택하는 데 있어서 임의로 한 명 또는 여러 명의 기증자를 선택하는 랜덤 핫덱 방법보다 회귀모형의 예측력이 클수록 좋은 결과를 기대할 수 있다.

고령화연구패널 1차 조사에 대한 대체(imputation)를 실시할 때 자료의 특성에 따라 대체 방법이 적절히 변형되었다. 첫 번째로 일부 소득 및 자산 항목은 응답이 거절되거나 응답 문항들 사이에 불일치가 나타나는 경우 대괄호 질문들(unfolding brackets)을 이용하여 얻어진 부분 정보를 포함하고 있으므로 이 정보를 사용하여 대체를 실시하였다. 두 번째로 응답자가 많지 않은 일부 문항의 경우 대괄호 질문으로부터 얻어진 부차 정보에 근거한 하위그룹(subclass)에서 기증자를 발견하지 못하는 경우가 발생하였고 이 경우 회귀모형을 사용한 대체가 실시되었다. 세 번째로 일부 항목의 경우 한 사람이 여러 개의 답을 제시하는 것이 가능하였고 이 경우 동일인에 의한 여러 가지 응답값은 서로 연관되어 나타날 수 있으므로 연관성을 고려하여 대체가 실시되었다. 네 번째로 같은 영역 내의 연관된 문항들 사이에 일치성(consistency)이 존재하면 이를 만족시키도록 대체가 실시되었다. 이 특징들은 고령화연구패널 2차 조사의 대체를 실시할 때도 동일하게 고려하여 1차 자료와 2차 자료 간 대체 방법의 일관성을 유지하였다. 그 외에 2차 자료의 대체 시에는 범주형 변수 일부에서 결측이

발생하였고 이를 고려하여 모형을 확장하였다. 각각의 경우 사용된 대체 방법에 대한 자세한 설명은 KLoSA 2차 자료에 대한 다중대체 결과 보고서에서 자세히 설명되어 있다(송주원 외, 2010).

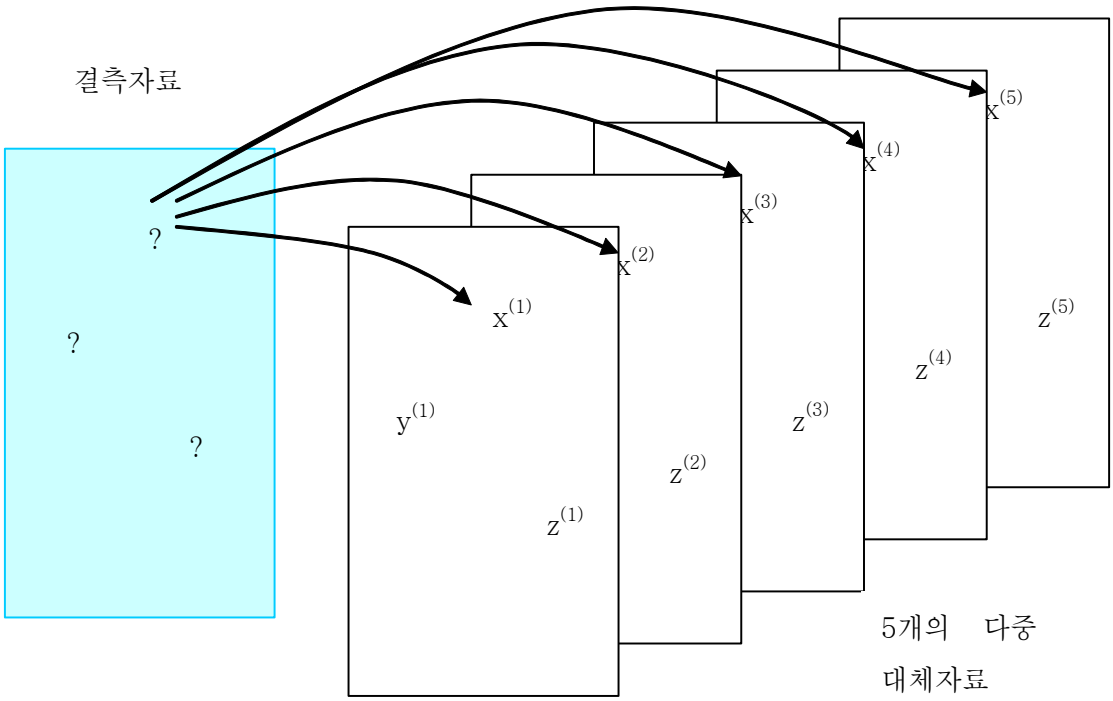
2. 결측값이 대체된 자료의 형태 및 구분

다중대체(multiple imputation)는 한 개의 결측값에 대하여 한 개의 값으로 대체하는 대신 타당한 여러 개의 값을 가지고 대체하는 방법을 의미하는데 대체된 값들은 각각 다르므로 대체된 값이 참값과 다르다는 점을 모형에 반영시켜 한 개의 값을 대체하는 방법인 단일대체(single imputation)에서 발생하는 분산추정량의 과소추정 문제를 보정하는 것을 가능하게 한다. 본 연구에서는 각 결측값에 대하여 5개의 값을 대체하는 다중대체를 실시하였다. 5개의 대체된 자료를 제공하고 각각의 자료는 단일대체된 자료이므로 단일대체에 근거한 분석을 시행하길 원하는 연구자는 대체된 자료 중 한 개의 자료(예를 들면, 첫 번째 자료)를 임의로 선택하여 분석을 실시할 수 있다.

2.1. 대체된 자료의 형태

한 개의 결측값을 포함한 자료에 다중대체를 실시하면 그 결과로서 여러 개의 결측값이 없는 대체된 자료가 만들어지게 된다. 고령화연구패널조사의 경우 5번의 대체가 이루어 졌으므로 대체된 항목들에 결측값이 없는 5개의 자료가 제공된다. 이 5개의 자료는 관찰된 값들은 모두 동일하지만 대체된 결측값은 같기도 하고 다르기도 한 형태를 가지고 있다. 결측된 자료에 대한 대체를 시행한 후 생성된 대체된 자료는 다음의 <그림 2.1>에 나타난 것과 같이 5개가 존재하게 되는 것이다. 이 대체된 5개의 자료는 연구자가 분석할 수 있도록 노동연구원 고령화연구패널 홈페이지를 통해 제공된다. 제공된 자료는 대체된 변수만을 포함하므로 대체가 실시되지 않은 변수들도 포함한 분석을 실시하기 원하는 연구자는 홈페이지에 제공되는 원 자료와 결합하여 분석을 시행하면 된다. 이 때 원 자료는 대체된 변수와 동일한 이름을 지닌 대체가 시행되기 이전의 결측을 포함한 원 변수들도 포함하므로 대체된 자료를 가지고 원 변수들을 덮어씌운 후 분석을 시행해야 한다. 예를 들어 SAS를 사용하는 경우 원 자료는 w02_main_v10k이고 대체된 자료는 w02_i1_v10k부터 w02_i5_v10k까지 5개의 자료이므로 각 대체된 자료를 원 자료와 결합하여 대체된 변수를 포함한 전체 자료를 만들어 분석을 시행하면 된다(모든 자료는 pid08의 크기순으로 배열되어 있다고 가정한다). 대체된 전체 자료를 만들기 위한 SAS program의 예제는 아래와 같다. (프로그램에서 대문자는 SAS 명령문을, 소문자는 분석자가 지정한 자료명이나 변수를 의미한다. 여기서 원 자료 및 대체된 자료명은 노동연구원에서 제공하는 자료명과 동일하게 선택하였다).

<그림 2.1> 결측자료에 대하여 5개의 다중대체를 실시한 경우의 예




```
* 첫 번째 대체된 전체 자료 생성;
DATA w02_i1_v10k;
   MERGE w02_main_v10k w02_i1_v10k;
   BY pid08;
RUN;

          ⋮

* 다섯 번째 대체된 전체 자료 생성;
DATA w02_i5_v10k;
   MERGE w02_main_v10k w02_i5_v10k;
   BY pid08;
RUN;
```

위의 프로그램에서 전체 변수를 포함하는 5개의 자료를 만들기 위하여 SAS의 data step을

다섯 번 써야하는 번거로움이 있으나 아래의 SAS macro를 이용하면 5개의 자료를 간단히 생성할 수 있다.

```
* Macro를 이용하여 5개의 대체된 전체 자료 생성;
%MACRO fulldata;
%DO j = 1 %TO 5;
   DATA w02_i&j._v10k;
      MERGE w02_main_v10k w02_i&j._v10k;
      BY pid08;
   RUN;
%END;
%MEND;
%fulldata;
```

연구자는 이 자료 각각에 대하여 원하는 분석을 반복적으로 시행할 수 있다. 각 자료에 대하여 독립적으로 분석이 시행된 후 분석 결과는 일반적으로 5개의 통계량 및 관련 분산(또는 표준 오차)으로 나타나는데 연구자는 5개의 각각 다른 통계량이 아닌 하나의 통합된 통계량을 구하는 데 목적이 있다. 각각 분석된 통계량을 통합하여 하나의 통계량을 구하는 방법은 3장에서 소개된다.

2.2. 대체된 결측값의 구분

결측값이 대체된 자료에서 어느 관찰값이 원래 관찰된 값이며 어느 관찰값이 대체된 값인지 구분을 할 수 있다면 유용할 것이다. 이 구분이 가능하다면 대체된 자료만을 가지고도 결측값의 대체없이 원 자료에 대한 분석을 실시하길 원하는 연구자는 원 관찰값 만에 근거한 분석을 시행하는 것이 가능할 것이고 대체된 자료값들이 관찰된 자료값들과 비슷한 지 여부 등의 추가 분석도 가능하다. 이를 위하여 고령화연구패널 자료의 경우 대체된 각 변수에 대하여 대체 여부를 나타내는 부속 변수인 깃발 변수(flag variable)가 추가되었다. 이 부속 변수는 원래 변수명에 _(underbar)를 추가시킨 변수명을 취한다. 예를 들어, 소득 부분의 작년 한 해 월평균 임금 소득액을 나타내는 변수 w02E003의 경우 w02E003_라는 변수가 추가되는데 이 변수는 아래와 같은 값들을 가진다.

$$\begin{cases} 0: \text{응답한 관찰값이 존재함} \\ 1: \text{관찰값이 모형을 통해 대체됨} \\ 2: \textit{Bracket}\ \text{질문에 구간 대신 값으로 응답} \\ 3: \text{가족대표자의 응답을 가지고 대체} \\ \bullet: \text{이 문항에 대한 응답 대상자가 아님} \end{cases}$$

즉, 대체 여부를 나타내는 깃발 변수(flag variable)가 값 “0”을 갖는 경우 해당 관찰값이 응답에 의하여 관찰된 값이라는 의미이며, “1”을 갖는 경우 관찰값이 결측되었으나 수정된 예측평균에 근거한 핫덱 방법에 근거하여 대체되었음을 의미한다. 한편 값 “2”는 대괄호 질문을 포함한 소득 및 자산 변수에서 응답이 구간으로 응답되지 않고 대략적인 값으로 응답된 경우 그 값으로 대체되었음을 의미하고, “3”은 가족대표자의 응답을 가지고 대체되었음을 의미한다. 일부 깃발 변수(flag variable)에서 보이는 결측값은 이 문항이 앞의 문항에 부속되어 있고 앞 문항의 응답 때문에 이 문항이 응답되지 않았음을 의미한다. 예를 들어 월평균 임금 소득액은 w02E001에서 임금 소득이 있다고 응답한 패널에게만 질문되었으므로 w02E001에서 임금소득이 없다고 응답한 경우 w02E001_의 해당 패널의 값은 결측으로 나타난다.

3. 다중대체(multiple imputation)된 자료의 분석 방법

다중대체된 자료의 경우 결측값이 없이 대체된 한 개 이상의 자료가 제공되며 이에 따른 분석은 다중대체된 각 자료의 분석 및 분석된 자료를 통합한 결과 도출의 두 단계로 나누어지게 된다.

3.1. 다중대체(multiple imputation)된 자료의 분석

다중대체된 자료 각각은 결측값이 대체되어 결측값이 없는 완전한 자료 형태를 가지고 있으므로 자료 각각에 대하여 연구목적에 알맞은 분석을 시행하면 된다. 예를 들어, 회귀분석(regression analysis)을 시행하고자 한다면 동일 관심변수에 대하여 동일 설명변수를 가지고 5개 자료 각각에 대하여 회귀분석을 실시하면 된다. 이렇게 분석을 실시하는 경우 추정된 회귀계수(regression coefficients), 표준오차(standard errors), 그리고 검정통계량(test statistics)은 5개 자료 각각으로부터 약간씩 다르게 나타나는데 이는 관심 변수가 결측되어 참값을 알지 못하는 불확실성에 근거한 차이를 나타내는 것이다. 하지만 연구자의 분석 목적은 관심 자료에 대한 5개의 서로 다른 결론이 아니라 한 개의 통합된 결론을 내리는 것이므로 5개 분석의 결과를 통합하여 한 개의 결론을 도출하기 위하여 아래의 통합 과정을 거쳐야 한다.

3.2. 분석된 자료를 통합한 결과 도출

다중대체를 m번 시행한 자료 각각에 대하여 분석을 시행한 후 얻어진 모수의 추정값들을 $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$이라 하자. 또한, 이 모수의 추정된 분산을 각각 $W_1, W_2, \ldots, W_m$이라 가정하자. 예를 들어, 회귀분석을 실시하면 i번째 자료(이 때, $i = 1, 2, \ldots, m$)에 근거한 회귀 분석에서 관심 설명 변수의 회귀계수의 추정값이 $\hat{\theta}_i$ 이 되고 그 회귀계수의 표준오차의 추정값의 제곱이 W_i가 된다. 이 경우 통합된 모수의 추정값은

$$\overline{\theta_m} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

으로 표현될 수 있다. 즉, 추정된 모수들의 평균값이 통합된 모수의 추정값이 된다.

통합된 모수의 분산의 추정값은 다음의 두 개의 분산 성분의 합으로 표현된다. 첫 번째 분산 성분은

$$\overline{W_m} = \frac{1}{m}\sum_{i=1}^{m} W_i$$

로서 각 모수의 추정된 분산들의 평균이다. 이 분산 성분은 대체내 분산(within-imputation variance)으로 부른다. 두 번째 분산 성분은

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \overline{\theta_i}\right)^2$$

으로 표현되는데 이는 각 대체된 자료의 모수의 추정값들 사이의 분산을 나타내므로 대체간 분산(between-imputation variance)이라 부른다. 통합된 모수의 분산의 추정값은

$$T_m = \overline{W_m} + \frac{m+1}{m}B_m$$

으로 구할 수 있다.

자료가 충분히 큰 경우, 이 모수에 대한 분포는 다음의 $t-$분포를 따른다.

$$\left(\theta - \overline{\theta_m}\right)T_m^{-\frac{1}{2}} \sim t_\nu,$$

여기서, $t-$분포의 자유도 ν는 $\nu = \left(\nu_0^{-1} + \widehat{\nu_{obs}}^{-1}\right)^{-1}$로 계산되는데 이 때 ν_0는 $\nu_0 = (m-1)\left(1 + \frac{1}{m+1}\frac{\overline{W_m}}{B_m}\right)^2$을 나타내고 $\widehat{\nu_{obs}} = \left(1 - \widehat{\gamma_m}\right)\left(\frac{\nu_{com}+1}{\nu_{com}+3}\right)\nu_{com}$으로 나타나는데 이 때 υ_{com}은 결측값이 없을 때 모수 θ에 대한 추정의 자유도를 나타낸다. 또한, $\widehat{\gamma_m} = \left(1 + \frac{1}{m}\right)\frac{B_m}{T_m}$으로서 무응답으로 인하여 손실된 모수 θ에 대한 정보량이라 불린다. 모수의 분포가 $t-$분포를 따르므로 $t-$분포에 근거한 검정을 시행하거나 모수의 신뢰구간을 구할 수 있다. 또한 이 통합 방법은 관심 모수들에 대한 다변량 검정 및 신뢰구간의 계산 등

으로의 확장도 가능하다 (Rubin, 1987).

3.3. 예제

다중대체를 시행하여 만들어진 m개의 자료들에 근거한 m개의 분석 결과를 통합하는 과정은 일반 연구자들이 직접 프로그램화하여 시행하기에 어려울 수 있으므로 현재 여러 가지 통계 프로그램에서는 이 결과를 통합하는 프로시저를 제공하고 있다. 예를 들어, SAS의 PROC MIANALYZE 프로시저는 위와 같이 분석된 자료의 모수들을 통합한 결과를 제공해 준다. 그 외에 무료 통계 분석 프로그램인 R도 다중대체된 자료를 분석한 후 통합하는 함수를 제공하고 있다. 또한 Schafer(1997)가 개발한 독립 프로그램 NORM은 통계분석 프로그램과의 연동없이 독립적으로 시행되는 작은 크기의 프로그램으로서 위의 단계를 수행하고 통합된 결과를 제공하는데 이 프로그램은 http://www.stat.psu.edu/~jls/misoftwa.html에서 무료로 다운받을 수 있다. 다음은 SAS에서 다중대체된 여러 개의 자료를 이용한 단순 평균 계산, 층화 평균 계산, 및 회귀 분석 계수의 계산 예를 보여준다.

우선 SAS에서는 여러 개의 자료에 대하여 동일한 모형을 가지고 분석을 실시하고자 하는 경우에 여러 개의 자료를 한 개의 자료로 통합한 후 통합된 자료에 대하여 한 개의 Procedure를 이용하여 자료 별 분석을 시행하는 것이 가능하다. 이를 위하여 5개의 대체된 자료를 한 개의 자료로 통합하고 각 대체된 자료를 나타내는 변수를 가지고 각 자료를 구분하면 된다(각 대체된 자료를 원 자료와 결합하는 프로그램은 2.1절에서 설명하였다). 제공된 대체된 자료는 몇 번째로 대체된 자료인지 나타내는 구별 변수인 w02imputation_을 가지고 있으므로 이 변수별로 분석을 시행하면 된다. 이를 위한 SAS 프로그램은 다음과 같다.

```
* 5개의 impute된 자료를 한 개의 자료로 통합;
DATA total;
   SET w02_i1_v10k  w02_i2_v10k  w02_i3_v10k  w02_i4_v10k  w02_i5_v10k;
   _imputation_=w02imputation_;
RUN;
```

여기서 새롭게 생성된 변수 _imputation_은 w02imputation_과 동일한 변수로서 각각의 자료를 분석한 후 SAS PROC MIANALYZE를 이용하여 자료를 통합하는 과정에 사용하기 위하여 생성되었다.

예제 1. 금융자산의 단순 평균 계산

```
* 각 대체 자료별 단순 평균 계산;
PROC SURVEYMEANS DATA = total;
   VAR w02f085;
   BY _imputation_;
   ODS OUTPUT STATISTICS = stat1;
RUN;

* 각 대체된 자료별로 계산된 단순 평균을 통합하여 원 자료의 단순 평균 추정;
PROC MIANALYZE DATA = stat1;
   MODELEFFECTS mean;
   STDERR stderr;
RUN;
```

<그림 3.1> 금융자산 단순 평균 계산 SAS Output

The MIANALYZE Procedure

Model Information

Data Set	WORK.STAT1
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF
mean	221.823061	4354.111277	4620.298951	1205.1

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mean	0.061135	0.059173	0.988304

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
mean	1228.257040	67.972781	1094.899	1361.615	1205.1

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
mean	1212.871470	1247.019020	0	18.07	<.0001

SAS Procedure SURVEYMEANS에서 BY변수를 사용하여 각 대된 자료별로 금융자산의 단순평균 및 표준 오차를 계산한 후 이들 통계량들을 자료(data set)명 stat1에 저장하였다. 이 저장된 통계량들은 Procedure MIANALYZE를 사용하여 통합되었다. 이 때 MODELEFFECTS 문에는 통합할 통계량 $\hat{\theta}_i$(여기서는, 평균)을 나타내는 변수 mean을 써 주고 STDERR문에는 W_i의 제곱근인 평균의 표준 오차를 나타내는 stderr 변수를 써 주면 된다. Procedure MIANLYZE는 <그림 3.1>과 같은 결과를 제공한다. 금융자산의 단순 평균은 1228.26 MW으로 타나나고 표준편차는 67.97이다. 금융자산의 평균에 대한 95% 신뢰구간은 (1094.90, 1361.62)로 계산되어 지며 금융자산이 0이라는 귀무가설은 $t-$통계량이 18.07, p-value가 <.0001로 5% 유의수준 하에서 유의하게 나타난다.

<u>예제 2</u>. 금융자산의 표본 설계 가중치를 이용한 평균 계산

이계오(2009)에 설명된 표본 설계 가중치를 이용한 금융자산의 평균 추정치를 계산해 보았다.

```
* 층화변수 생성;
 DATA total; SET total;
    strata=compress(w02region1)||compress(w02region2)||compress(w02enu_type);
 RUN;

* 각 대체 자료별 표본 설계 가중치를 이용한 평균 계산;
 PROC SURVEYMEANS DATA = total;
        STRATA strata;
        VAR w02f085;
        WEIGHT w02wgt02;
        BY _imputation_;
        ODS OUTPUT STATISTICS = stat2;
 RUN;

 * 각 대체된 자료별로 계산된 표본 설계 가중치를 이용한 평균을 통합하여 원 자료의
층화 평균 추정;
 PROC MIANALYZE DATA = stat2;
    MODELEFFECTS mean;
    STDERR stderr;
 RUN;
```

<그림 3.2> 금융자산 표본 설계 가중치를 이용한 평균 계산 SAS Output

The MIANALYZE Procedure

Model Information

Data Set	WORK.STAT2
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF
mean	144.700274	6094.662430	6268.302759	5212.7

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mean	0.028491	0.028074	0.994417

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
mean	1262.881057	79.172614	1107.670	1418.093	5212.7

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
mean	1250.056623	1278.971294	0	15.95	<.0001

표본설계에서 시도(w02region1), 동부/읍면부(w02region2), 주택유형(w02enu_type)을 층화변수로 사용하였으므로 SAS 분석을 위해 compress 함수를 사용하여 각 변수 값의 공백을 없애주고 ||을 사용하여 세 변수의 값들을 이어줌으로써 층화변수 strata를 생성하였다. 이 예제에서는 이계오(2009) 예제와 같이 층화변수 만을 고려하였지만 1차년도 자료의 조사구 변수(w01enu)를 사용하여 CLUSTER 문을 포함한 분석을 시행할 수도 있다.

SAS Procedure SURVEYMEANS에서 BY변수를 사용하여 각 대체된 자료별로 금융자산의 표본 설계 가중치를 이용한 평균 및 표준 오차를 계산한 후 이들 통계량들을 자료명 stat2에 저장하였다. 이 저장된 통계량들은 Procedure MIANALYZE를 사용하여 통합되었다. 이때 MODELEFFECTS 문에는 통합할 통계량 $\hat{\theta}_i$(여기서는, 표본 설계 가중치를 이용한 평균)을 나타내는 변수 mean을 써 주고 STDERR문에는 W_i의 제곱근인 평균의 표준 오차를 나타내는 stderr 변수를 써 주면 된다. Procedure MIANLYZE는 <그림 3.2>와 같은 결과를 제공한다.

금융자산의 표본 설계 가중치를 이용한 평균은 1262.88 MW으로 타나나고 표준편차는 79.17이다. 금융자산의 표본 설계 가중치를 이용한 평균에 대한 95% 신뢰구간은

(1107.67, 1418.09)로 계산되어 지며 금융자산이 0이라는 귀무가설은 $t-$통계량이 15.95, p-value가 <.0001로 5% 유의수준 하에서 유의하게 나타난다.

<u>예제 3</u>. 금융자산에 대한 회귀분석

금융자산과 성별, 연령의 관계를 나타내는 회귀모형을 적합한 분석을 시행하였다.

```
* 2차년도 현재 연령 계산:
DATA w01_v12k; SET w01_v12k;
   KEEP pid w01a001_Age;
RUN;
PROC SORT DATA=total;
   BY pid;
RUN;
DATA total1; MERGE total(in=in1) w01_v12k;
   BY pid;
   IF in1;
   IF w02A001 = 1 THEN w02age = w01a001_age + 2;
   IF w02A001 = 5 THEN w02age = 2008-w02A002y;
RUN;
PROC SORT DATA=total1;
   BY _imputation_ pid08;
RUN;

* 각 자료별 회귀 분석 실시;
PROC REG DATA=total OUTEST=outreg COVOUT;
   MODEL w02f085 = w02gender1 w02age;
   BY _imputation_;
RUN;

* 각 자료별 회귀 분석 결과의 통합;
PROC MIANALYZE DATA=outreg;
   MODELEFFECTS Intercept w02gender1 w02age;
RUN;
```

2차 자료에서는 1차 자료(w01_v12k.sas7bdat)에서 응답한 연령에 대한 확인을 거쳐 연령

이 잘못 응답된 경우만 수정하였으므로 1차 자료의 연령 응답이 정확한 경우 1차 자료의 응답 연령에 2를 더하고 1차 자료의 연령 응답이 부정확한 경우 2차 자료의 수정된 연령 정보를 이용하였다. 이렇게 생성된 변수는 w02age이다.

SAS Procedure REG에서 BY변수를 사용하여 각 imputed된 자료별로 회귀분석을 실시한 후 OUTEST 문을 사용하여 자료명 outreg에 회귀계수 및 회귀계수의 표준 오차 등을 저장한다. 여기에 저장된 통계량들을 Procedure MIANALYZE에서 통합하여 준다. 이 때 통합하고자 하는 통계량은 절편(intercept) 및 나이, 성별을 나타내는 두 변수의 계수, 즉 세 개의 회귀모형 모수가 되며 이를 MODELEFFECTS 문에 나타내준다. 여기서, Intercept는 변수명이 아니고 회귀모형의 절편을 의미한다. Procedure MIANLYZE는 <그림 3.3>과 같은 결과를 제공한다.

<그림 3.3> 금융자산 회귀분석 SAS Output

The MIANALYZE Procedure

Model Information

Data Set	WORK.OUTREG
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF
Intercept	122.595195	187543	187690	6.51E6
w02gender1	30.551110	1103.890668	1140.551999	3871.5
w02age	0.052628	43.370934	43.434087	1.89E6

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
Intercept	0.000784	0.000784	0.999843
w02gender1	0.033211	0.032643	0.993514
w02age	0.001456	0.001455	0.999709

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
Intercept	2393.721139	433.232233	1544.601	3242.841	6.51E6
w02gender1	-109.420725	33.772059	-175.633	-43.208	3871.5
w02age	-13.236982	6.590454	-26.154	-0.320	1.89E6

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	2384.954213	2412.581441	0	5.53	<.0001
w02gender1	-114.811754	-101.040600	0	-3.24	0.0012
w02age	-13.503984	-12.904216	0	-2.01	0.0446

회귀 모형의 절편(intercept)은 2393.72, 성별(w01gender1)과 나이(w01a001_age)의 회귀계수는 각각 -109.42와 -13.24로 나타나며, 절편을 포함한 세 회귀모수의 표준오차는 각각 433.23, 33.77, 6.59로 추정된다. 절편 및 각 변수의 회귀 계수가 0인가를 검정하는 $t-$통계량은 각각 5.53, -3.24, -2.01로서 모두 5% 유의수준 하에서 통계적으로 유의하게 나타난다.

참고문헌

송주원, 임화경, 육태미, 윤라헬, 신성원, 강승희, 윤초롱 (2010) “KLoSA Report: 고령화연구패널 2차 조사 자료에서 발생하는 결측값에 대한 다중대체”, 한국노동연구원.

이계오 (2009) 고령화연구패널조사 2차년도(WAVE I) 가중치, 고령화연구패널보고서, 노동연구원.

Bell R. (1999) Depression PORT Methods Workshop (I). RAND: Santa Monica, CA.

Little R. J. A. (1988), “Missing data adjustments in large surveys,” *Journal of Business and Economic Statistics*, 6, 287 301.

Rubin D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*, John Wiley: New York.

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data, Chapman and Hall, London, UK.