

KLoSA

Korean Longitudinal Study of Ageing

2006년 고령화연구패널조사 1차 기본조사 사용자 안내서



머 리 말

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)’로 명명하였다. 2006년에 약 1만여명의 중고령자 패널 구축과 함께 제1차년도 기본조사를 완료하였고, 지금은 개인직업력 조사를 진행 중에 있다.

고령화연구패널조사(이하 KLoSA)는 조사의 내용과 방법, 양 측면에서 중요한 의미를 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, 보건의학, 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 설문문의 대인면접에서는 시도된 적이 없었다.

본 사용자안내서는 고령화연구패널조사에 대한 기본적인 소개를 시작으로 표본추출방식과 가중치 적용방식을 설명하고 각 영역별 세부적인 내용과 주의할 점을 다루었다. KLoSA 자료는 한국노동연구원의 데이터서비스시스템을 통하여 무료로 다운받을 수 있는데, 구체적인 방법은 본 안내서의 후반부에 자세히 소개하였다.

각 장별로 세부적인 내용과 구체적인 내용들은 테크니컬 리포트 형식으로 업데이트 될 예정이며, KLoSA와 관련된 여러 가지 관련 기록들은 홈페이지(<http://klosa.kli.re.kr>)를 통해서 공개될 것이다. 그동안 『노동리뷰』 지면을 통해서 KLoSA의 설문내용, 예비조사, 제1차 기본조사 실시 계획 등에 대하여 소개해 왔다. KLoSA 설계에 대한 전반적인 내용에 대해서는 『노동리뷰』에 발표한 이전 글들을 참고하면 도움이 될 것이다.

2007년 11월

한국노동연구원
고령화연구패널조사팀

목 차

I. 고령화연구패널조사 개요

1. 배경 및 목적	3
2. 조사설계 상의 특징	3
3. 조사 대상	4
4. 주요 설문 내용	4
5. 조사 주기 및 기간	6
6. 경 과	6
7. 조사방법	7
8. 데이터 배포	8
9. 예산지원, 조사 주관 및 수행 기관	8

II. 표집과 가중치

1. 표집과정	9
1) 조사모집단의 층화	9
2) 표본 추출	10
2. 실사과정 및 응답률	11
1) 실사과정	11
2) 최종결과	12
3) 성공률 및 응답률	13
3. 가중치 부여와 모수 추정	14
1) 가중치 사용의 필요성	14
2) 가중치 산출과정	14
3) 가중치 이용한 추정식의 계산	16

III. 고령화연구패널조사 자료의 세부적 내용과 특징

1. 기본 응답단위	17
2. 전 영역에서 공통으로 사용된 주요 용어의 개념	18
3. 영역별 설문흐름도 및 주의사항	19
1) 인구학적 배경 영역	19
2) 가족영역	20
3) 건강영역	22
4) 고용영역	26
5) 소득영역	28
6) 자산영역	29
7) 주관적 기대감과 삶의 만족도	30
4. 데이터 변수명 규칙	32

IV. 항목 무응답과 보정방법(Imputation)

1. 결측값 대체 방법	34
2. 결측값이 대체된 자료의 형태 및 구분	36
1) 대체된 자료의 형태	36
2) 대체된 결측값의 구분	38
3. 다중대체(Multiple imputation)된 자료의 분석 방법	39
1) 다중대체(Multiple imputation)된 자료의 분석	39
2) 분석된 자료를 통합한 결과 도출	39
3) 예 제	40

V. 생성변수 목록 / 45

VI. 데이터 이용방법 / 49

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 정보 뿐만 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 미국, 영국, 유럽 등 이미 고령자 패널조사를 실시하고 있는 선진국들과 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 일반 가구 거주자를 대상으로 표집 및 조사를 실시하였다.

외국 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.

4 주요 설문 내용

KLoSA의 주요조사 내용은 다음의 <표 I-1>과 같다.

<표 I-1> 영역별 주요 내용

세부 영역	주요 내용
CV. 커버스크린	<ul style="list-style-type: none"> • 가구에 대한 정보(성별, 태어난 해, 혼인상태, 응답자와의 관계 등)
A. 인구	<ul style="list-style-type: none"> • 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)	<ul style="list-style-type: none"> • 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)	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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
Cb. ADL/IADL, 수발노동	<ul style="list-style-type: none"> • 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

세부 영역	주요 내용
Cc. 의료보장과 시설이용	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> MMSE-K 한국형 인지능력 측정 척도 사용
Ce. 신체기능 측정	<ul style="list-style-type: none"> 약력 측정
D. 고용(근로형태 구분) 문항번호 D001-D010	<ul style="list-style-type: none"> 10개의 문항을 통해 현재 근로형태 구분
D. 고용(임금근로자) 문항번호 100-200번대	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 응답자가 구직자이거나 은퇴자인 경우 가장 최근 일자리에 대한 정보
E. 소득	<ul style="list-style-type: none"> 근로소득: 임금, 자영업, 농어업, 부업 등 연금소득: 국민연금, 특수직역연금, 개인연금 등 사회보장소득: 사회보장소득여부 및 종류 기타 수입 및 소득 지난 1년간 가구총소득
F. 자산	<ul style="list-style-type: none"> 현재 거주주택 정보 거주주택 외 부동산 자산: 부동산 소유, 임대/임차, 사업체/농장 금융자산(갯돈포함) 기타 비금융자산 상속/증여 부채 가구 총자산
G. 주관적 기대감	<ul style="list-style-type: none"> 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 버전에서는 CV 영역에서 조사한 정보를 사용자가 편리하게 사용하도록 정리하여 제공함.

5 조사 주기 및 기간

패널조사는 하나의 현상을 일정한 시간적 간격을 두고 반복적으로 측정하여 시간의 흐름에 따라 그 변화를 살펴 보고자 하는 것이다. 고령화연구패널조사는 2006년을 시작으로 매 2년 간격으로 짝수 연도에는 반복적으로 측정할 기본적인 사항을 조사하는 ‘기본조사’를 수행한다. 2006년 제1차 기본조사는 2006년 8월부터 12월까지 6개월 간 진행되었다.

매 홀수년도에는 필요한 경우 기본조사에 포함되지 않는 내용을 중심으로 특정한 주제를 정하여 실시할 예정이다. 2007년에는 개인직업력(Job History) 조사를 수행한다.

6 경과

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 필요하다고 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 대면면접방식(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.

고령화연구패널조사 제1차 기본조사 실사(fieldwork) 기간은 2006년 8월부터 12월까지 약 5개월이 소요되었다. 향후 고령화연구패널조사의 ‘기본조사’는 매 짝수년도에 시행될 예정이다. 홀수년도에는 기본조사에는 포함되지 않는 내용을 중심으로 특정한 주제를 정하여 실시할 것이다.

고령화연구패널조사는 2006년 12월 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.

데이터와 유저가이드, 설문지, 코드북 등 조사와 관련된 모든 자료들은 고령화연구패널조사 홈페이지(<http://klosa.kli.re.kr> 또는 한국노동연구원www.kli.re.kr)에서 연결된 고령화연구패널조사 배너를 통해서 무료로 다운로드 받을수 있다.

7 조사 방법

◆ 노트북을 이용한 대인면접법(Computer Assisted Personal Interviewing(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 맞춤형 설문 문항으로 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 입

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

력하면 컴퓨터가 정/오답을 판단하여 자동으로 입력된다.

8 데이터 배포

데이터, 설문지, 코딩가이드, 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.

원시 데이터의 경우 간단한 사용자등록(회원가입) 절차를 거쳐야 다운로드를 받을 수 있다. 회원가입한 데이터 사용자들끼는 향후 학술대회나 데이터의 업데이트 상황 등 기타 필요한 정보를 이메일이나 등록된 주소지를 통해 발송할 예정이다.

데이터를 다운로드 받을 때 한 가지 주의해야 할 사항은, 먼저 이용자의 컴퓨터에서 ‘팝업 차단 해지’를 해 주어야 한다는 점이다. 팝업이 차단되어 있는 컴퓨터에서는 로그인을 했음에도 불구하고 데이터가 있는 ‘노동연구원 통계자료 시스템’으로 들어갈 수 없기 때문에 데이터를 다운로드 받을 수 없다.

‘통계자료 시스템’에서 데이터를 다운로드 받을 때는 ‘전체자료 내려받기’를 통하여 2006년 기본조사 자료 전체를 내려받을 수도 있고, ‘부분자료 내려받기’를 통해서 데이터의 일부 문항을 선택적하여 내려 받을 수도 있으므로 이용자가 원하는 대로 이용하면 된다.

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.

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.

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 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 아파트조사구에 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 예비표본 조사구를 합하여 추출하였다. 그리고 이렇게 추출된 표본조사구 가운데 20%에 해당하는 예비표본조사구는 다시 계통표집법으로 본표본조사구와 예비표본조사구를 분류하였다.

이와 같은 방식으로 1,000개의 표본조사구를 확정된 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 실사과정 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) 실사과정

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.

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.

한편 실사를 진행하면서 적격가구가 1개 가구도 없거나 재개발로 인해 철거중인 조사구, 그리고 가구 방문을 원천적으로 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 방문을 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) 최종결과

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 e 는 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., Brügiavini, 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 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 모수 추정에서 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.

$$\text{가중치} = \text{추출률의 역수} * \text{응답률의 역수} * \text{벤치마킹 가중치}$$

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.

$$f = f_1 \times f_{2.1}$$

- 여기서 f 는 추출률이며, f_1 은 조사구 추출률, $f_{2 \cdot 1}$ 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.

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

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

- 여기서 첨자 i 는 층을 나타내며, N_i 는 i 번째 층내의 전체조사구수, n_i 는 i 번째 층내의 표본조사구수를 나타내며, $M_{j \cdot i}$ 는 i 번째 층내의 j 번째 조사구내의 적격가구수, $m_{j \cdot i}$ 는 i 번째 층내의 j 번째 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 i 번째 층내의 j 번째 조사구의 적격가구의 비율을 나타내며, $H_{j \cdot i}$ 는 i 번째 층내의 j 번째 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.

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

여기서 첨자 i 는 층을 나타내고, j 는 조사구를 의미하며, k 는 가구를 나타낸다.

◆ 최종 가중치

최종 가중치는 설계가중값과 벤치마킹 보정계수를 이용하여 구해진다.

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

여기서 첨자 i 는 개인을 의미하며, BF 는 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

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

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

위 식에서 \widehat{Y}_h 는 h 번째 층의 총계를 나타내는데 그것의 분산추정량은 다음과 같다.

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

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 y 가 0이나 1의 값을 갖는 이항변수이다. 따라서 비율추정에 대해서도 아래의 식을 그대로 사용할 수 있다.

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

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

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

표준오차는 다음과 같은 식으로 나타낼 수 있다.

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

고령화연구패널조사 자료의 세부적 내용과 특징

1 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.
- 금액을 묻는 모든 문항의 단위는 " 만원"이다.

2 전 영역에서 공통으로 사용된 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.
- ◆ **범주형전환문항(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.

각 영역에는 그 영역에서 조사된 정보를 바탕으로 생성한 ‘**생성변수(Generated Variables)**’가 추가되어 있다. 생성변수란 설문문항을 통해 나온 기본변수를 사용자가 쉽게 사용할 수 있도록 조정하여 만들어준 변수를 말한다. (※ 제 V장 생성변수 목록과 홈페이지 메뉴중 사용자 안내서에서 ‘생성변수 사용방법’파일을 참고.)

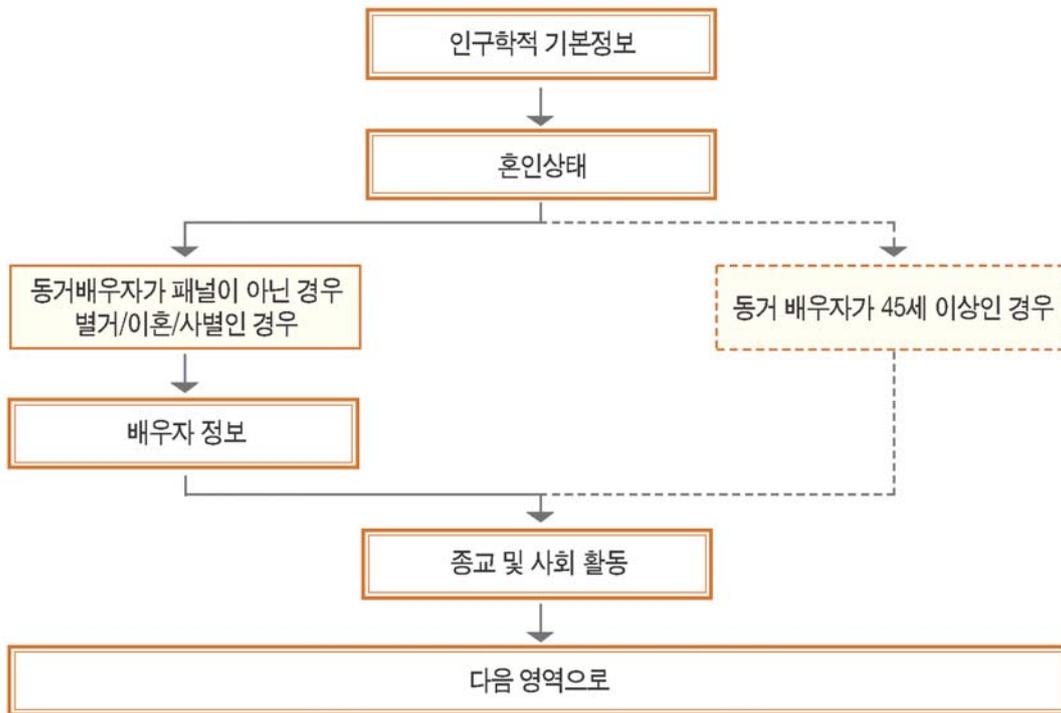
응답자가 ‘모르겠음’이나 ‘응답거절’로 대답한 문항들에 대하여는 다중대체 보정방법(Multiple Imputation)을 통해 보정한 ‘Imputed data set’을 제공한다.

(※ 제 IV장에 참조, 더 자세한 내용은 홈페이지 메뉴중 사용자 안내서이 있는 “다중대체 보정방법 리포트” 파일 참고)

3 영역별 설문흐름도 및 주의사항

1) 인구학적 배경 영역

[그림 III-1] 인구학적 배경영역의 설문구조



◆ 배우자 기본정보 사용시 주의해야 할 점

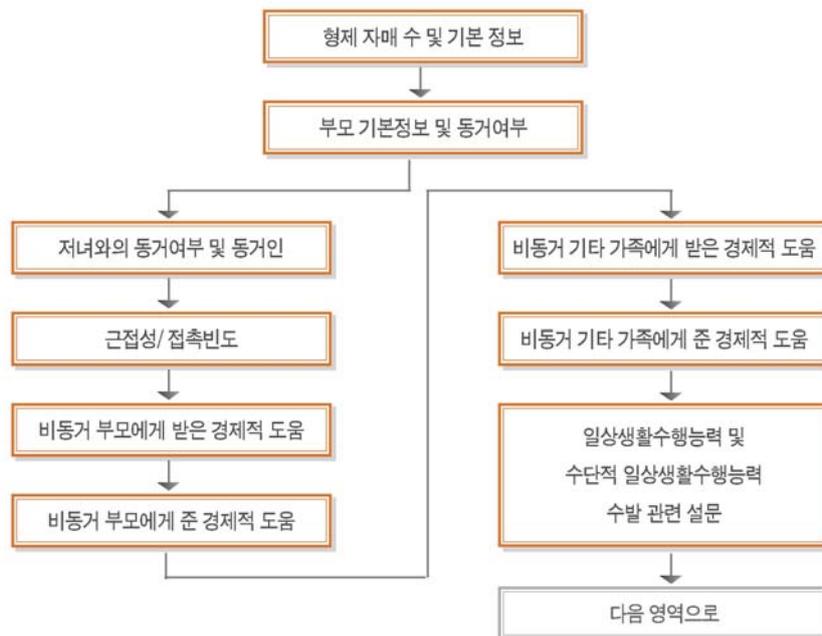
- 현재 응답자와 동거하는 배우자가 45세 이상이면 조사대상자가 되기 때문에 별도로 배우자 기본정보를 묻지 않았다.
- 그러나 응답자의 배우자가 사망, 실종, 별거중인 경우와 배우자 연령이 45세 미만인 경우 조사대상자에서 제외되기 때문에 이러한 경우에 해당하는 응답자에게만 배우자 정보(생년월일, 학력, 고용상태)를 별도로 물었다.

2) 가족영역

[그림 III-2] 가족영역(자녀 및 손자녀) 설문구조



[그림 III-3] 가족영역(형제자매, 부모, 기타 가족) 설문구조



◆ 가족영역에서 주의해야 할 사항 및 주요 개념

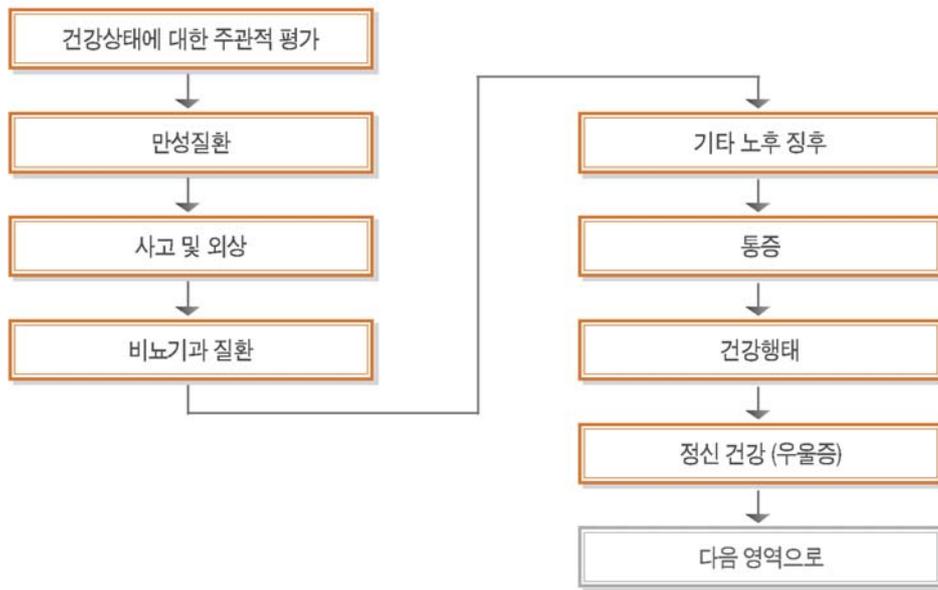
- **가족관계에서 응답자의 주관적 판단 존중** : 가족 관계는 친자, 혈연, 호적 여부와 상관없이 응답자가 생각하는 관계를 그대로 존중하였다. 예를 들어, 혼인 여부 질문에서 “배우자 있음”이란 사실혼도 포함하며, “별거”는 불화로 인한 별거뿐만 아니라 장기수용시설, 장기해외체류 등 다른 이유의 별거를 포함시키되 응답자의 주관적인 판단을 기준으로 한다. 응답자의 아버지가 생부가 아니더라도 응답자 본인이 “아버지”라고 여기면 그대로 아버지로 인정하며, 자녀도 입양자녀, 직접 낳은 자녀, 재혼을 통해 얻은 자녀 상관없이 자신의 주관적인 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 드리거나, 김치나 반찬해 드리기와 같은 서비스내용도 비금전적인 지원에 포함시키다. 다만, 손자녀 돌보기, 수발돌기 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’ 붙이기과 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 III-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.

[그림 III-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 실제로 자신직접 지출한 금액을 중심으로 물었다.

[그림 III-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.

[그림 III-8] 신체기능측정 설문 구조와 도구



◆ 신체계측 설문 안내

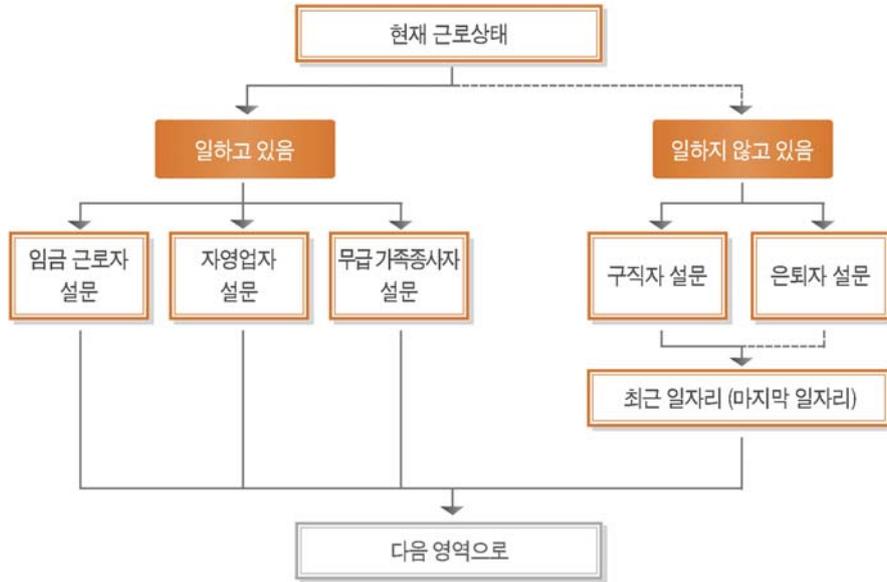
- 노화와 관련한 신체계측은 여러 가지가 있으나, 2006년 고령화연구패널조사 기본조사에서는 악력만을 측정하였다.
- 외국의 고령자패널조사에서는 공간이 확보가 된다면 걷기 속도를 측정하기도 하고, 한 발로 서서 균형잡기에 서부터 혈액이나 타액을 채취하여 검사하기도 한다. 우리나라의 주택구조와 면접원들이 계측해 올 수 있는 현실적인 방법을 고려하여 제1차 기본조사에서는 악력을 측정하였다.
- 악력 측정은 측정을 할 수 있는 상태의 응답자인지 아닌지를 파악하고, 본인인 원하지 않는거나 현재 한쪽 손이 다치거나 아픈 상황이면 측정하지 않는 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, 앞으로 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.

[그림 III-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 유무형 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 따라 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

[그림 III-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 응답하도록 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 1일부터 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

[그림 III-11] 자산영역의 설문구조



◆ 자산영역에서의 주의사항

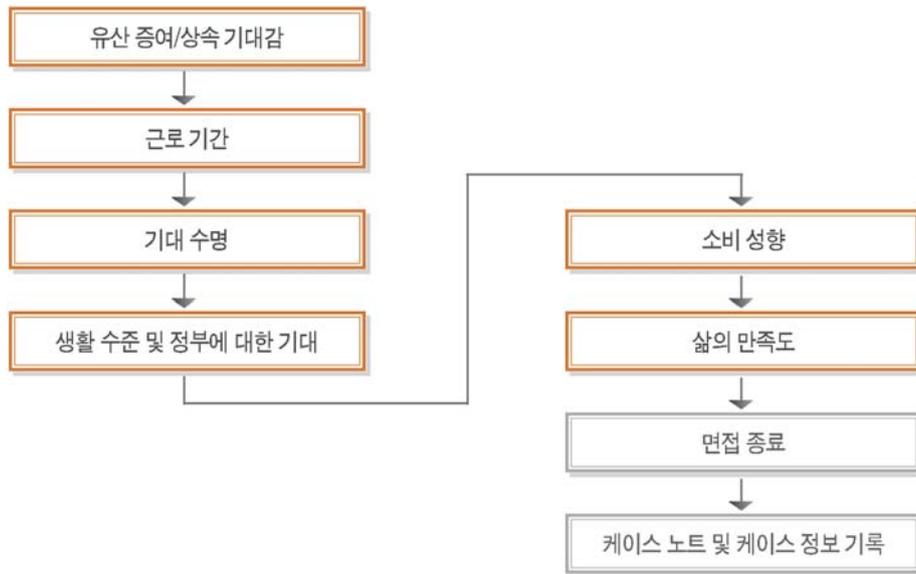
- 자산영역에서 조사하는 자산은 부동산, 사업체, 금융, 상속/증여, 부채이며, **‘명의’를 기준으로 자신의 명의로 된 자산만을 응답**하도록 하였다. 명의가 공동명의로 되어 있는 경우, 명의자수를 기입하게 함으로써 정확한 자산규모를 파악하고자 하였다.
- 문항번호 ‘F001부터 F048’까지의 현 거주 주택의 자산가치의 설문을 우선 묻고, 그 외의 부동산 자산에 대한 설문이 뒤따라 나오는 구조이므로 부동산 자산은 현재 거주지 자산과 그 외의 부동산 자산을 합해주어야 한다.

◆ 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.

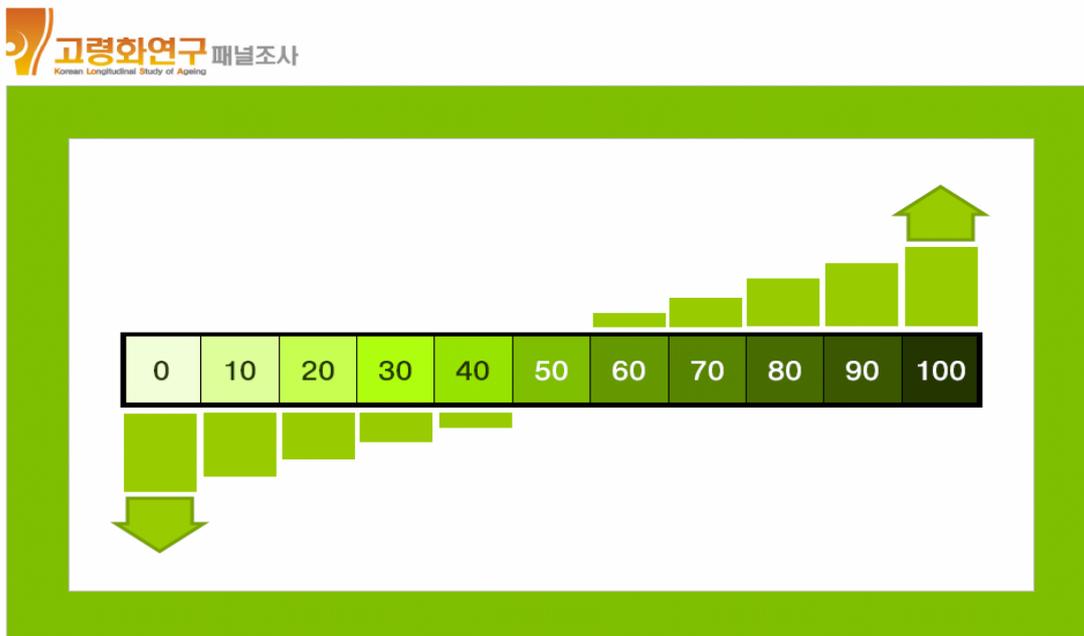
(※ 제IV장을 참고, 더 자세한 내용은 홈페이지 사용자관련자료에서 다운로드 받을 것)

7) 주관적 기대감과 삶의 만족도

[그림 III-12] 주관적 기대감과 삶의 만족도 설문구조 및 사용도구



[그림 III-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 물어 보도기도 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.

4 데이터 변수명 규칙

◆ 변수명은 기본적으로 ‘W01+설문문항번호’의 형식을 가지고 있다.

◆ 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. ____님께서 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: 따로 말씀하시는 분들은 따로 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 자녀에 대한 대답을 하면 첫 번째 루프가 돌게 되고
 >> 변수명은 W01Ba02_01, W01Ba03_01, W01Ba04_01이 된다.
- 응답자가 두 번째 자녀에 대한 대답을 하면 두 번째 루프가 돌게 되고
 >> 변수는 W01Ba02_02, W01Ba03_02, W01Ba04_02가 된다.
- 응답자가 세 번째 자녀에 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

항목 무응답과 보정방법 (Imputation)

※ 항목무응답에 대한 다중대체 보정방법(Multiple Imputation) 관련 자세한 내용은 홈페이지 사용자안내서란 안에 “다중대체 보정방법 리포트”를 참고.

현실에서 만나는 거의 모든 자료는 결측값을 포함하고 있다. 연구자가 변수들을 제어할 수 있는 실험과 달리 설문 조사 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 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 프로그래밍화하여 여러 가지 조사 연구에 적용해 왔고 우수한 결과를 보여 온 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 특징이 발견되었고 그에 따라 대체 방법이 적절히 변형되었다.

- 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 있

었다. 각각의 경우 사용된 대체 방법에 대한 자세한 설명은 KLoSA multiple imputation 보고서에서 자세히 설명되었다.

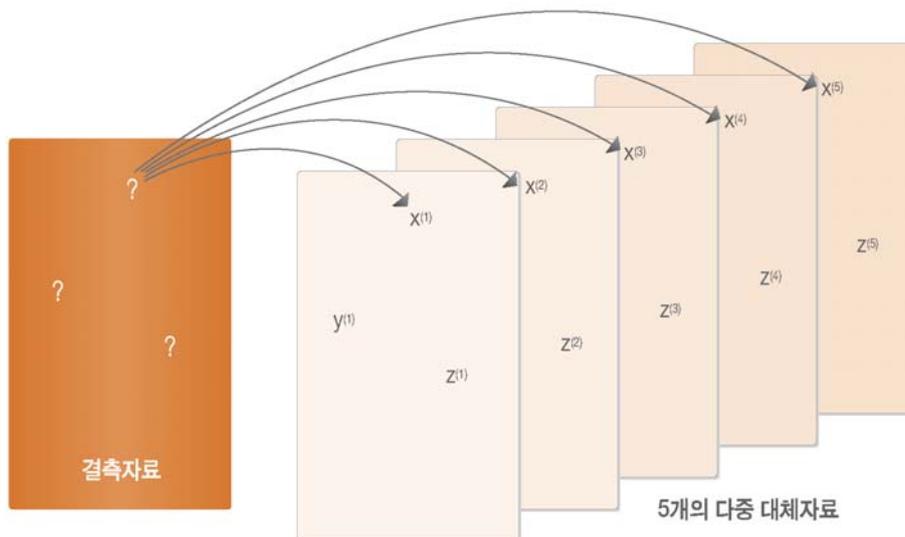
2 결측값이 대체된 자료의 형태 및 구분

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

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.

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



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 w01_v10k이고 대체된 자료는 w01_i1_v10k부터 w01_i5_v10k까지 5개의 자료이므로 각 대체된 자료를 원 자료와 결합하여 대체된 변수를 포함한 전체 자료를 만들어 분석을 시행하면 된다. 이와 같은 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).

```
* 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 실시 하길 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 다중대체(multiple imputation)된 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 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ 이라 하자. 또한, 이 모수의 추정된 분산을 각각 W_1, W_2, \dots, W_m 이라 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 W_i 가 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

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

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

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

으로 구할 수 있다.

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

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

여기서, t-분포의 자유도 ν 는 $\nu = (\nu_0^{-1} + \hat{\nu}_{\text{obs}}^{-1})^{-1}$ 로 계산되는데 ν_0 는 $\nu_0 = (m-1) \left(1 + \frac{1}{m+1} \frac{\bar{W}_m}{B_m} \right)^2$ 을 나타내고 $\hat{\nu}_{\text{obs}}$ 에

서 ν_{com} 이 결측값이 없을 때 모수 θ 에 대한 추정의 자유도를 나타낼 때 $\hat{\nu}_{\text{obs}} = (1 - \hat{\gamma}_D) \left(\frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \right) \nu_{\text{com}}$ 을 나타낸다.

또한, 여기서 $\hat{\gamma}_m$ 은 $\hat{\gamma}_m = (1 + 1/m) B_m / T_m$ 으로서 결측에 의하여 손실된 모수 θ 에 대한 정보의 부분(fraction of information about θ 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 등에서의 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.html>에서 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 W_m 의 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
-----Variance-----
Parameter      Between      Within      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      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 나타나고 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
총화 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
-----Variance-----
Parameter          Between      Within      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      t for H0: Theta0 = 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 나타나고 표준편차는 59.73이다. 금융자산의 표본 설계 가중치를 이용한 평균에 대한 95% 신뢰구간은 (927.19, 1161.34)로 계산되어 지며 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-----			DF
	Between	Within	Total	
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	t for H0:		
			Theta0	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

생성변수 목록

※ 생성된 변수와 관련한 자세한 사항은 홈페이지 '사용자안내서'란에 "생성변수설명" 파일을 참고.

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. 각 영역별 설문문항 변수 마지막부분에 해당영역에서 만들어준 생성변수를 넣었고 그 리스트는 다음 표와 같고 생성변수가 만들어진 과정은 '생성변수 설명' 파일에 기록하고 있으므로 자세한 내용은 이 파일을 참조하여 사용하여야 한다.

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
w01hhsz	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 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_m	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_m	가구 내 자산 대표응답자(결측치 대체)

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

데이터 이용방법

고령화연구패널조사와 관련된 모든 자료는 인터넷 홈페이지를 통해 공개된다. 고령화연구패널조사의 인터넷 홈페이지 <http://klosa.kli.re.kr> 또는 한국노동연구원 홈페이지 <http://www.kli.re.kr>에 들어와서 “고령화연구패널조사” 배너를 통해 들어오는 방법이 있다.

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. 원자료를 다운로드 받는 한국노동연구원 ‘통계자료시스템’에서도 설문지와 코드북을 다운로드 받을 수 있는데, 이 파일은 홈페이지에서 다운받는 파일과 동일하다. 관련 자료들은 별도로 인쇄를 하거나 CD 등의 형태로 배포하지 않으므로 연구자 각자가 인터넷을 이용하여 다운로드 받고 본인의 편의에 맞게 출력해서 사용하여야 한다.

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.

◆ 제공하는 SPSS 데이터는 12.0 버전에서 SAS는 9.0버전에서 작업하였으므로, 그 이하의 버전프로그램에선는 변수명 등에서 오류가 발생할 수도 있다.

◆ 데이터 다운로드 실행 순서

- 홈페이지에서 회원가입을 하고 아이디(등록한 이메일)와 비밀번호로 로그인 후 ‘데이터 다운로드’ 버튼을 누르면 ‘통계자료시스템’으로 자동연결되어 데이터를 다운 받을 수 있다.
- **주의사항!!**
 사용자들은 자신의 컴퓨터에서 ‘팝업 차단 해지’를 해 주어야 ‘통계자료시스템’으로 들어갈 수 있다.
- [그림 VI-1]은 데이터를 다운로드 받기 위하여 로그인을 하면 팝업창이 뜨면서 노동연구원 ‘통계자료시스템’으로 위의 그림과 같이 자동연결이 된다. 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 박스쳐진 영역을 3번으로 규정하고 각각을 설명하면 다음과 같다.
- 데이터는 ‘전체 자료 다운받기’와 ‘부분 자료 다운받기’로 나누어 자료를 다운로드 받을 수 있다. 데이터 전체를 다운받기 원하면 [그림 VI-1] “3번 전체 자료 다운로드 받기”를 선택하고, 영역별 혹은 관심분야별 키워드를 이용하여 부분적으로 데이터를 받길 원하면 “2번 부분자료 다운받기”에서 원하는 자료를 추출할 수 있다.

[그림 VI-1] 통계자료시스템의 구성



◆ 전체자료 내려받기

- 전체자료다운로드를 받기를 원하는 사용자는 [그림 VI-1] 3번 “전체자료 내려받기”에 해당하는 부분에서 데이터(원자료)와 설문지, 코드북을 모두 다운로드 받을수 있다. 이때 원자료, 설문지, 코드북은 모두 국문과 영문으로 각각 zip 파일 형식으로 올려져 있으므로 원하는 언어에 따라 선택할 수 있다.

• 전체자료 내려받기 중 원자료 내려받기 안내!

‘전체 자료 다운받기’에서 제공되는 데이터는 SPSS 와 SAS 형식으로 각각 국문, 영문 데이터로 모두 4가지 형식(국문 SPSS, 영문 SPSS, 국문 SAS, 영문 SAS)으로 제공되고 파일명은 다음 표의 1번부터 4번까지에 해당된다. 이 데이터는 응답자의 응답을 그대로 반영된 것으로 모름이나 응답거절과 같은 응답을 그대로 담고 있다. 그리고 설문문항에서 나온 변수들을 서로 조합하여 “생성변수”를 만들었는데 각 영역별 마지막 문항 뒷부분에 새로 생성한 변수를 넣어주었다. 번호 5번부터 8번에 해당하는 데이터는 응답자가 ‘모르겠음’이나 ‘응답거절’로 표명했던 무응답에 대한 보정이 다중대체(Multiple Imputation)된 데이터로 기본설문문항의 변수를 포함하지 않고 주로 자산이나 소득영역에서 무응답한 변수들을 중심으로 그 값이 보정된 데이터 셋(set)으로 구성되었다. (* 다중대체된 데이터 사용방법은 부록 참조)

번호	데이터명	데이터 설명
1	w01_v1.0k(SAS).zip	1차 기본조사 1.0버전(A영역~G영역)_국문 SAS 데이터
2	w01_v1.0k(SPSS).zip	1차 기본조사 1.0버전(A영역~G영역)_국문 SPSS 데이터
3	w01_v1.0e(SAS).zip	1차 기본조사 1.0버전(A영역~G영역)_영문 SAS 데이터
4	w01_v1.0e(SPSS).zip	1차 기본조사 1.0버전(A영역~G영역)_영문 SPSS 데이터
5	w01_i_v1.0k(SAS).zip	1차 기본조사 1.0버전(무응답 대체)_국문 SAS 데이터
6	w01_i_v1.0k(SPSS).zip	1차 기본조사 1.0버전(무응답 대체)_국문 SPSS 데이터
7	w01_i_v1.0e(SAS).zip	1차 기본조사 1.0버전(무응답 대체)_영문 SAS 데이터
8	w01_i_v1.0e(SPSS).zip	1차 기본조사 1.0버전(무응답 대체)_영문 SPSS 데이터

◆ 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.

◆ 2006년 고령화연구패널 1차 기본조사 데이터 1.0버전의 사용과 관련하여 추가적인 내용이 발생하면 홈페이지 사용자안내서 메뉴를 통해 파일을 올려드릴 예정입니다.

KLoSA

Korean Longitudinal Study of Ageing

2006년 고령화연구패널조사 - 1차 기본조사 사용자 안내서