

다문화가족종단연구 콜로키움3

패널조사의 데이터 에러와 자료관리 방안

한국노동연구원 이상호(shlee@kli.re.kr)
최효미(chm@kli.re.kr)

콜로키움 순서 및 목차

1차 : 패널조사 기획 및 전략수립의 주요쟁점들

2차 : 초기 패널구축과정에서의 조사성공률 유지방안

3차 : 데이터 에러 및 자료관리방안

- 패널조사에서 데이터 에러 및 무응답관리의 중요성
- 패널조사에 특수한 비표본오차들
- 통계적 방법을 이용한 데이터 에러와 무응답 관리
- 설문지 설계과정에서의 비표본오차의 관리
- 실사과정에서의 비표본 오차의 관리 전략
- 데이터 클리닝 과정
- 데이터 가공 과정
- 데이터 분석과정
- 학술대회

1. 패널조사의 데이터 에러와 무응답관리의 중요성

- 더 많은 정보 = 더 많은 에러!
- 패널조사는 조사 자체가 복잡하고 문항의 수 자체가 절대적으로 많기 때문에 데이터 에러의 발생가능성이 큼
- 무엇보다 “동일한 조사대상을 반복적으로 추적조사하기 때문에 발생하는 오류”를 어떻게 다룰 것인가가 관건
- 더욱이 CAPI의 도입으로 코딩과 에디팅 과정이 사전적 프로그래밍에 의해서 탑재되기 때문에 클리닝의 성격이 “사후적인 ”과정에서 어느 정도는 “사전적” 과정으로 파악됨
- 데이터 에러의 성격을 이해하기 위해서 패널조사에서 발생하는 비표본오차 (nonsampling errors)의 특성을 이해하는 것이 중요

2. 패널조사의 비표본오차

2.1 비표본오차란 무엇인가

- 총조사오차(total survey errors)
 - 모집단 평균과 표본조사에 기초한 추정치의 평균 사이에 발생하는 차이
- 표본오차 (sampling errors)
 - 표본선택이 모집단을 대표하지 못했을 때 발생하는 에러
 - 정확한 확률표집 => 표본오차의 크기는 표본수에 의해서 결정됨
 - 가중치 부여를 통해서 해결
- 비표본오차(nonsampling errors)
 - 표본오차를 제외한 모든 종류의 오차를 총칭하는 용어
 - 잘못된 표본틀 설정, 설문지 설계, 코딩, 에디팅, 등 전 과정에서 발생

2. 패널조사의 비표본오차

2.1 비표본오차란 무엇인가

- 명세화오차(specification errors) : 측정되어야 할 개념과 설문을 통해 실제로 측정된 개념간의 차이
- 표본틀/포함오차(frame/coverage errors) : 잘못된 표집틀로 인해서 조사단위의 중복, 누락, 불완전 포함 등이 발생하는 것
- 무응답오차(nonresponse errors) : 단위 무응답과 항목 무응답, 불완전 응답 등
- 측정오차(measurement errors) : 모호한 설문, 지시문 혼란, 잘못된 용어 사용 등으로 인해 발생하는 설문지 오차
- 처리오차(processing errors) : 데이터 에디팅, 코딩, 가중치 부여 등의 과정에서 발생하는 오차

2. 패널조사의 비표본오차

2.1 비표본오차란 무엇인가?

- Biemer and Lyberg(2003) : 표집오차보다 비표집오차에서 체계적오차의 위험도가 더 큼
- Assael and Keon(1982)의 실험연구 : 비표본오차가 총조사 오차의 95%를 차지함
- 비표본오차를 통계적인 개념으로만 협소하게 생각하는 것은 문제가 있음
=> 즉, 가중치 부여나 임putation과 같은 통계적 해법으로 해결
- 오히려 비표본 오차는 조사 전과정에서 발생하므로 조사전략 수립 시점부터 조사의 목적, 모집단 설정, 조사방법, 문헌 검토 등을 통한 체계적인 오차 관리가 필요

2. 패널조사의 비표본오차

2.2 패널조사의 비표본오차

- 응답여부에 따른 구분 : 응답 오차 vs. 무응답 오차
- 편의발생 여부에 따른 구분 : 체계적 오차 vs. 비체계적 오차
- 시점에 따른 구분 : 횡단면 오차 vs. 종단면 오차
- 주체에 따른 구분 : 면접원 오차 vs. 응답자 오차
 - 표본이탈(2차 콜로키움)
 - 회고오차 (recall errors)
 - 돌출효과 (seam effects)
 - 패널 컨디셔닝 효과(panel conditioning effects)
 - 모드효과 (mode effects)

2. 패널조사의 비표본오차

2.3 표본이탈과 데이터 에러와의 관련성

- 표본이탈은 자료의 대표성 뿐만 아니라, 종단면 에러의 핵심적인 원천이 되기도 함
- 첫째. 설문구조의 설계가 조사 누락기간 동안의 정보를 파악하는데 부적절하거나 불충분함으로써 발생
- 둘째. 표본이탈로 인한 우측절단 문제의 발생. 이에 대한 정확한 처리와 식별
- 셋째. 이탈후 재복귀자의 경우 회고기간이 상대적으로 길기 때문에 회고설문의 경우 자료의 정확성이 떨어질 수 있는 가능성

=> 뒤에서 다룰 패널조사의 응답 및무응답 오차와 연관

2. 패널조사의 비표본오차

2.4. 회고오차

- 응답자가 과거의 사건이나 상황에 대해 정확하게 응답하지 않음으로 인해서 발생하는 측정오차
- 회고오차의 종류
 - 망각(forgetting), 텔레스코핑(telescoping), 돌출효과(seam effects)
- 회고오차에 영향을 미치는 요인
 - 시간효과(time effects) : 오래된 기억일 수록 편의(bias)가 증가
 - 명료성(saliency) : 기억을 복구해야 할 사건이나 행동이 명료할 경우 특히 긍정적 사건이 더 정확하게 회고됨 (혼인력 > 이혼력)
 - 개인사의 복잡성 차이 : 실업기간이 길거나, 고용-실업 상태의 반복이 잦은 경우 회고오차 발생 가능성 증대
- 응답자 특성차이 : 여성, 고학력자, 인종 등

2. 패널조사의 비표본오차

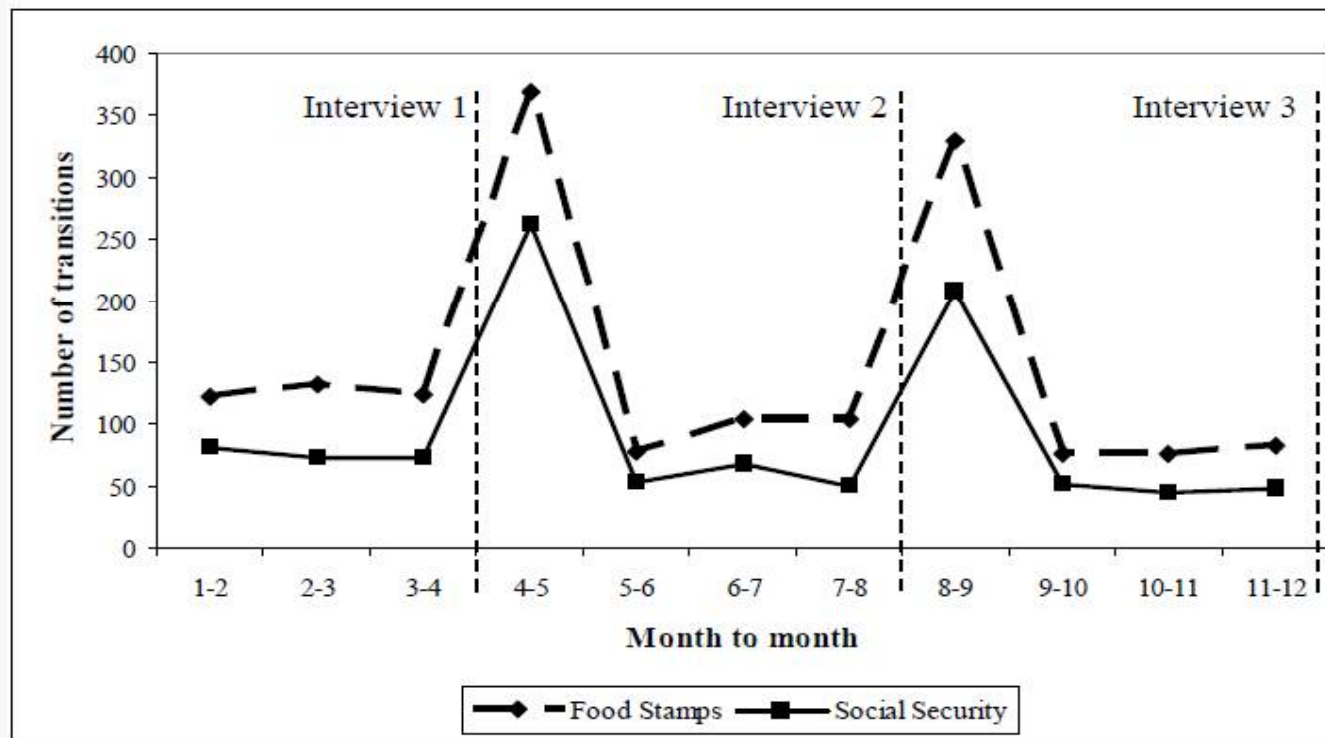
2.5 돌출효과(seam effects)

- 동일한 조사항목에 대한 응답 값이 조사차수가 변화하는 시점에 급격하게 변하는 것을 말함
- 미국 센서스국에서 실시하는 SIPP(Survey of Income and Program Participation)조사에서 그 심각성이 최초로 제기
- Jackle(2008)에 따르면 정부보조금 수령 여부의 상태변화 중 22% 정도가 ‘돌출부’에서 발생
- 돌출효과의 유형
 - 생략(Omission or Under-reporting)
 - 오분류(misclassification or re-definition)
 - 잘못된 시점보고(misplacement of events in time)

2. 패널조사의 비표본오차

2.5 돌출효과(seam effects)

<SIIP 소득문항에서 나타난 돌출효과>



Source: Burkhead and Coder (1985), Table 2 and Table 5. Based on waves 1 to 3 of the 1983/4 Survey of Income and Program Participation.

2. 패널조사의 비표본오차

2.6 패널조건화(panel conditioning)

- 응답자가 이전 조사를 한번 이상 해보았기 때문에 발생하는 응답상의 변화(편의)를 말함
- 생활만족도와 같은 응답자의 주관적 태도와 관련된 문항에서 빈번하며, 조사 차수간의 간격이 짧을 때 더 확연하게 나타남
- 다문화가족 종단연구의 경우 주관적 태도에 관한 문항이 많으므로 더욱 주의가 필요
- 인지적 고무가설(cognitive stimulus hypothesis)
 - 정치설문조사 의 투표율 변화를 설명하는 이론을 패널조사에 적용
- 모름/무응답 증가, 사회적 선망 편의(social desirability bias), 묵종(acquiescence), 중간반응 편의(centrality bias), 양극단 반응(yes and no saying), 면식에 의한 편의(auspices)

2. 패널조사의 비표본오차

2.7 모드효과(mode effects)와 데이터 에러

- 모드효과란, 조사방식의 변화로 인해서 발생하는 표본이탈, 무응답오차, 측정오차 등을 종합적으로 일컫는 용어임
- 동일한 조사 차수내에서 면접/전화/자기기입식 조사를 병행하는 multiple mode의 문제
=> CAPI의 경우 유치조사가 불가능하므로, 종이설문과 병행시 문제점
- CAPI과정에서 도입에서 주의해야할 데이터 에러
 - 데이터 모듈간의 조합
 - 변수 맵핑, 데이터 산출과정상의 에러
 - missing 값의 처리(9999, OR -1, -2...)
 - 최대/최소값/변수값의 범위/소수점 자리의 처리
 - 문항의 조합(주소의 입력과 합산, 연도변수의 구분과 합산) 등
 - 전년도 정보의 누적(주소, 연락처 등 수정시 정보의 저장)

3. 비표본오차의 관리 방안

3.1 비표본 오차 관리의 개요

- 데이터 에러를 감소시키기 위한 전략은 크게 3가지로 구분가능
 - 설문지 설계(CAPI 시스템)을 통한 사전적 통제
 - 조사후 데이터 클리닝 과정
 - 무응답 대체(imputation)를 활용한 통계적 처리
- 설문지설계/ 데이터 클리닝 과정을 통한 에러 감소가 우선적.
- 국내 패널조사는 외국에 비해 항목무응답이 많지 않고, 무응답 대체에 많은 비용과 전문인력, 시간이 소요됨. 무응답대체의 종류로는...
 - 논리적 대체/최근값 대체/과거값을 이용한 비보정 대체
 - 결측치의 숫자가 미미한 경우 중위값 대체
 - 다음 차수 정보를 이용한 대체
 - 회귀모형을 이용한 대체 Multiple imputation

3. 비표본오차의 관리 방안

3.2 설문지설계과정에서의 비표본 오차 관리방안

- 응답자에게 질문의 형태를 가급적 상세하게 구술함
- 회고 오차, 특히 돌출 효과를 감소시키기 위한 설문 설계에 주의가 필요
- 종속형 설문(dependent interviewing : DI)의 활용
 - 사전적 형태 : 산업/직업 분류 문항의 조사차수간 과잉변화 감소
 - 사후적 형태 : 가구소득 등에서 생략을 줄여줌
- 영국 BHPS의 사례 : 2006년부터 체계적으로 DI 도입을 시도. 이를 위해 2003년 '소득 및 고용관련 조사척도의 개선'에 대한 실험연구를 실시
 - - 사전적 형태의 DI는 고용 스펬의 오류를 14.2% 감소
 - - 사후적 형태의 DI는 비근로소득의 오류를 감소
 - - 경제활동상태, 고용관련 스펬자료, 금융 분야 문항에 전면

3. 비표본오차의 관리 방안

3.3 DI 설계시 유의할 점

- 시점의 일치
 - 1차년도 조사당시 **조사시점으로부터 ‘지난 1년 동안’**으로 질문

마1 작년 한해 동안 _____님 닥 가족들 중 근로소득이 있었던 분이 계셨습니까?

H**2101

1	2	3	4	5	6	7	8	9	10	11
---	---	---	---	---	---	---	---	---	----	----

▶ 근로소득이란, 근로(일)의 대가로 벌어들인 수입을 말하며, 직장 또는 일자리에서 받은 임금이나 자영업자의 소득 등이 포함됩니다.

(1) 있었다
(2) 없었다 → 마2로 가십시오.

마1-1 있었다면, 작년 한해 동안 _____님 닥 전체의 연간 총 근로소득은 얼마 되었습니까?

H**2102

1	2	3	4	5	6	7	8	9	10	11
---	---	---	---	---	---	---	---	---	----	----

■ 작년 근로소득 : 연간 총소득 _____만원

3. 비표본오차의 관리 방안

3.3 DI 설계시 유의할 점

- Histroy 설문시 유의할 점 : 혼인력 설문의 사례(KLIPS 1,2차)

혼1 최초로 결혼하신 날짜는 언제입니까?
(날짜를 잘 모르시면 결혼하신 나이를 기입하십시오)

■ 최초 결혼년도 : 19 ____ 년 ____ 월 (만 ____ 세)

	년	월	만나이
변수명	P**5541	P**5542	P**5543
응답차수	1 2 3 4 5 6 7 8 9 10 11		

혼2 최초로 결혼한 분과 지금도 같이 살고 계십니까?

P**5544 1 2 3 4 5 6 7 8 9 10 11

(1) 그렇다 → 아 1로 가십시오
(2) 그렇지 않다

혼2-1 그렇지 않다면, 다음 중 어디에 해당되니까? 또, 그 시기는 언제부터입니까?

3. 비표본오차의 관리 방안

3.3 DI 설계시 유의할 점

- History 설문시 유의할 점 : 혼인력 설문의 사례(KLIPS 3차년도 이후)

사 1 현재 혼인상태는 어디에 해당되니까?(사실혼의 경우에는 기혼인 것으로 간주합니다. 또한, 주말부부이거나 요양 등으로 인해 함께 살고 있지 않을 경우에는 별거가 아닙니다.)

- (1) 미혼이다 (4) 이혼하였다
(2) 기혼이며 배우자가 있다 (5) 배우자가 먼저 사망하였다
(3) 별거중이다

사 2 지난 조사 이후 _____님의 혼인상태에 변화(결혼, 별거, 이혼, 사별 등)가 있었습니까?

- (1) 변화가 없었다 → 아 1로 가십시오.
(2) 변화가 있었다

사2-1 언제 어떤 변화가 있었습니까? 다음의 <보기>를 참조하여 변화가 일어난 순서대로 모두 응답하여 주십시오.

《보기》 최초로 결혼한 시기가 2006년 7월 4일이고, 2006년 8월에 이혼한 후 2007년 3월에 재혼한 경우

변 화 내 용						변 화 시 점	
1.	1.처음 결혼함○	2.재혼()	3.별거()	4.이혼()	5.사별()	6.기타()	2006년 7월
2.		2.재혼()	3.별거()	4.이혼○	5.사별()	6.기타()	2006년 8월
3.		2.재혼○	3.별거()	4.이혼()	5.사별()	6.기타()	2007년 3월

3. 비표본오차의 관리 방안

3.4 실사과정에서의 데이터 관리

CAPI Interviewer 로그인

로그인

2008년 제11차 한국노동패널 조사
(KLIPS : Korea Labor & Income Panel Study)

• 사용자ID
• 비밀번호

한국노동연구원 Hankook Research

조사 진행 대상 가구 선택

2008년 제11차 한국노동패널 조사
조사대상가구선택

가구번호(0000) 중 일부를 주체로 검색

• 가구번호
• 가구명
• 주소
• 전화번호

주요조사 진행현황의 확인사항에 따라 검색조건을 설정하십시오.

가구의 조사 진행 현황을 표시해 주십시오

☐ 전체가구 ☐ 조사가진행중인가구 ☐ 미조사가구 ☐ 무응답가구 ☐ 기타

검색 시작하기

설문 진행

[가구원 1]

■ 아래 가구원 정보 중 변경이 되었거나 잘못된 부분이 있으면 해당 사항을 선택해 주십시오.

가구원 일련번호 : 1

1. 성명 : 김명성
2. 성별 : 남
3. 생/월/일 : 88
4. 생년월일 : 1988년 10월 08일

1. 성명
2. 성별
3. 생/월/일
4. 생년월일
5. 수정사항 없음

기본정보확인 4

3. 비표본오차의 관리 방안

3.4 IRP(Intelligent Review Program)를 이용한 설문지 리뷰

Page 3 of 29

2. 가구원별 학력

가구번호	문가여부	가구원번호	이름	생년월일	만나이	응답유형
1554		1	윤병선	1956년 11월 20일	52	설문 시작

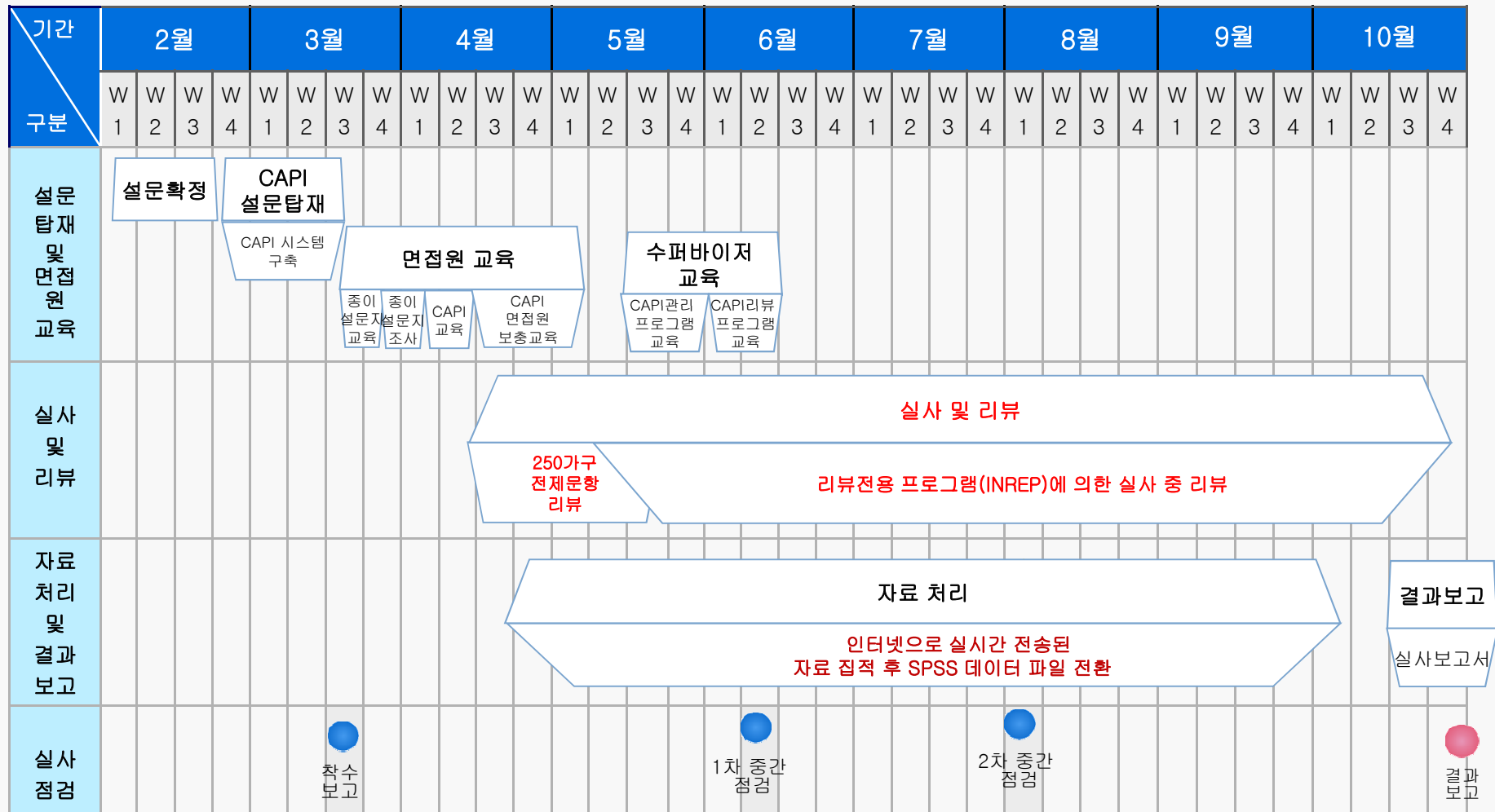
확인내용	가구용			개인용					
	문항	4년제 대학	메모내용	취업	미취업		신규용		
최종학력	문10-1	4년제 대학		문44		문39		문항	
대학종류				문56		문51		아1/아13	
대학유형				문57		문52		아16	
대학유형_기타								아17	
변경이유									
변경이유2									
이수여부	문10-2	졸업		재학여부		재학여부		재학여부	
변경이유									
변경이유2									
졸/수/중 시기	문10-3	1981년 2월						아3/아15	
변경이유									
변경이유2									
부가교육									
중학교 재학경험			문1	다녔다					
검정고시로 중학교 졸업자격 획득여부			문1-2						
졸업한 중학교의 입학년도			문1-3	1969					
졸업한 중학교의 졸업년도			문1-3	1972					
중퇴한 중학교의 입학년도			문6-1						
중퇴한 중학교의 중퇴년도			문6-1						
고등학교 재학경험			문7	다녔다					
검정고시로 고등학교 졸업자격 획득여부			문7-2						
졸업한 고등학교의 입학년도			문7-3	1972					
졸업한 고등학교의 졸업년도			문7-3	1975					
중퇴한 고등학교의 입학년도			문12-1						
중퇴한 고등학교의 중퇴년도			문12-1						
대학교 재학경험			문22	다닌 적이 있으며 졸업(수료)하였다					

HID : 1554[ver. 5]

3 / 29

3. 비표본오차의 관리 방안

3.4 실사과정에서의 데이터 관리



3. 비표본오차의 관리 방안

3.5 데이터 클리닝

- 데이터 클리닝이란?
 - 데이터 클리닝은 조사설계와 현장조사, 그리고 데이터 코딩 및 가공과정에서 발생한 오차들을 관리하는 체계적인 과정을 의미
- 데이터 클리닝을 위한 준비들
 - 설문구조의 숙지
 - 핵심 변수 체크리스트 작성 : ID, 성별, 연령, 학력 등 기본적인 인적정보 및 조사에서 가장 중요한 핵심변수들을 선정하여 체크 리스트 작성
 - Cross Checking의 전략수립 : 상식, 이론, 선행연구 등에 의거하여 문항간 논리적 연관성을 사전적으로 체크하여 집중 클리닝 변수 선정

3. 비표본오차의 관리 방안

3.5 데이터 클리닝

1) 횡단면 클리닝

- 같은 차수 안에서의 오류를 바로 잡는 작업
- KEY 변수 클리닝 : 기본정보 내의 ID, 이름, 성별, 연령, 학력 등의 오류를 바로 잡는 작업

=> 자료 사용자들이 발견하기 쉬운 에러이기 때문에 집중적 관리가 필요

=> 조사업체의 1차 원자료 산출과정에서 이루어져야 할 사항임

- Cross-Checking : 소득과 지출, 정부보조금 지급 등 서로 관련된 문항을 선정하여 논리적 타당성 여부를 체크 및 관리

=> 데이터 오류 여부를 판단하는 것이 쉽지 않으며 전문지식을 필요로 함.

=> 연구진의 자발적 노력이 필요 , 분석과정과 연계

3. 비표본오차의 관리 방안

3.5 데이터 클리닝

9차 발송전 데이터클리닝 작업내용 최종(070511).xls [호환 모드]

	A	B	C	D
1	에러 유형	구분	진행상황	사연정리 Sheet
5	가구주 관계 맞추기			
6	배우자 관계의 성별 비교	DCP	완료	N/A
7	가구주가 없는 집	DCP	완료	N/A
8	가구주 관계 중복	DCP	완료	N/A
9	가구주 관계의 "10" 가구주와 가구주 성명 미매칭	DCP	완료	N/A
10	합가 찾기			
11	신규 가구원이면서 유형 설문지 진행자 리스트업	DCP	완료	9차합가리스트
12	차수별 이름 중복 리스트 리스트업	DCP	완료	
13	기본 정보(성명, 성별, 양음, 생년월일) 확인	DCP	완료	기본정보확인_성명 기본정보확인_성별 기본정보확인_생년월일
14	기본 정보(성명, 성별, 양음, 생년월일) 미입력 데이터 확인	DCP	완료	N/A
15	학력 에러 (vs. 지난차수)			
16	학력 낮아짐	DCP	완료	학력
17	학력 동일, 학년 낮아짐	DCP	완료	학력

9차 데이터클리닝 작업내용

G	H	I	J	K	L	M	N	O	P	Q	R
개인근로소득합	가구 문23-1. 총근로소득 (만)	확인내용									
6000	600	예러아님	02가구원만	600만원 소득있음.							
3600	1950	예러아님	05가구원이	11월에 신규(재혼)로 들어옴. 05가구원의 2005년 10월까지의 소득은 가구용에 들어옴.							
1500	1800	예러아님		설문지에 그대로 적혀있음.							
3650	4100	예러아님		가구근로소득에는 세전소득을 적음.							
4800	6120	예러아님	04가구원이	12월에 분가했음. 분가 가구원의 근로소득인 1320만원이 차이남.							
1080	2680	예러아님		개인 근로소득의 합이 2680이 맞음.							
1330	1150	예러아님		가구용 근로소득은 유형의 임금을 계산하였으며, 개인용은 취업의 설문에서 응답자가 1330이라							
2400	3720	예러아님		6월에 분가한 04가구원의 2005년 근로소득이 가구용에서 잡혔음. 그래서 1320만원의 차이가 남.							
0	2760	예러아님	03가구원이	12월 분가하였으므로 개인소득에 2005년 근로소득이 없음.							
1788	2328	예러아님		설문지에 그대로 적혀있음.							
1180	1380	예러아님		설문지에 그대로 적혀있음.							
5280	7260	예러아님	04가구원이	12월에 분가했음. 분가 가구원의 근로소득인 1980만원 차이남.							
0	1200	예러아님		개인근로소득의 합이 잘못 계산되었음.							
2760	5010	예러아님	03가구원이	11월에 분가했음. 분가 가구원의 근로소득인 2250만원 차이남.							
3520	3960	예러아님		개인 근로소득의 합이 3960이 맞음.							
0	300	예러아님		설문지에 그대로 적혀있음.							
1800	2400	예러아님		2006년 5월에 분가한 03가구원의 근로소득 600까지 포함됨.							
0	1560	예러아님		2006년 3월에 분가한 03가구원의 근로소득 1560만원까지 포함됨.							
0	1800	예러아님		2006년 8월에 분가한 04가구원의 근로소득 1800만원까지 포함됨.							
5700	7300	예러아님		9월에 분가한 03가구원의 소득 1600만원까지 포함됨.							
1800	2150	예러아님		설문지에 그대로 적혀있음.							
0	2000	예러아님	01가구원이	2006년 1월 사망하였으므로 그 전의 근로소득 2000만원이 포함됨.							
0	200	예러아님		설문지에 그대로 적혀있음.							
0	344	예러아님		5월에 분가한 04가구원의 근로소득 344만원이 포함됨.							
1176	2300	예러아님		12월에 분가한 05가구원의 근로소득 1124만원이 포함됨.							

3. 비표본오차의 관리 방안

3.5 데이터 클리닝

2) 종단 클리닝

- 서로 다른 차수들 간의 비교를 통해 오류를 바로 잡는 작업으로 일반적으로 횡단 클리닝의 후행과정임
- 클리닝 대상 변수가 많지는 않음 (ID, 성별, 생년월일, 학력 등)
- 그러나 오류를 찾는데 상당한 시간이 소요될 뿐만 아니라, 이전조사까지 클리닝 된다는 점에서 데이터 확인 및 수정에도 많은 시간이 소요됨
- 조사 차수가 늘어날 수록 종단 클리닝에 대한 부담은 가중됨
- 지난 차수 정보를 이용한 종단면 클리닝 : DI와 연계 가능

3. 비표본오차의 관리 방안

3.5 데이터 클리닝

D	E	F	G	H	I	J	K	L	M
prjob1	prjob2	prjob3	prjob4	prjob5	prjob6	prjob8	현재 주로하는일	에러유형	
801	영등포노회 생활협동조합	유기농산물판매	위탁관리	상용직	관리자	임금	물품판매	지난조사 임금->이번조사 비임금	
301	청한섬유	보세웨타	재단	상용직	주임	임금	편직 관리	지난조사 임금->이번조사 비임금	
801	미싱	의류 제조	미싱	상용직	직원	임금	미싱	지난조사 임금->이번조사 비임금	
801	약국	의약품 제조판매	약제 조판매	상용직	약사	임금	약제 조판매	지난조사 임금->이번조사 비임금	
601	학생과외	과외	과외	임시직	아르바이트	임금	과외	지난조사 임금->이번조사 비임금	
601	과외지도	수학과외	수학과외	임시직	아르바이트	임금	수학과외	지난조사 임금->이번조사 비임금	
801	부동산 사무실	부동산 알선	부동산 알선	상용직	공인중개사	임금	부동산 알선	지난조사 임금->이번조사 비임금	
601	조선선재	운송업	운송업	상용직	운전기사	임금	트레일러로 목재를 실어	지난조사 임금->이번조사 비임금	
802	씨앤디	부동산매매 중개	부동산중개	상용직	없음	임금	부동산중개	지난조사 임금->이번조사 비임금	
801	필 부동산 중개사무소	부동산중개업	중개업	상용직	중개인	임금	중개업	지난조사 임금->이번조사 비임금	
701	삼성화재	보험회사	보험설계업	상용직	사원	임금	보험설계업	지난조사 임금->이번조사 비임금	
801	F.M.C상사	자동차판매	중고자동차	상용직	사원/직원	임금	중고자동차매매	지난조사 임금->이번조사 비임금	
601	붉은장미	유흥업소	주방	상용직	주방	임금	주방	지난조사 임금->이번조사 비임금	
601	대한생명 정읍지점	보험업	보험세일즈	상용직	팀장	임금	보험세일즈	지난조사 임금->이번조사 비임금	
301	금성출판사	책출판업	책판매	상용직	영업직	임금	책판매	지난조사 임금->이번조사 비임금	
701	개인과외	과외	과외	임시직	가정교사	임금	과외	지난조사 임금->이번조사 비임금	
601	아트피아 창원점	화장품판매	화장품판매	상용직	사원/직원	임금	화장품판매	지난조사 임금->이번조사 비임금	
801	웃가게	의류도매	의류판매	임시직	직원	임금	의류판매	지난조사 임금->이번조사 비임금	
801	국선도	기체조	사범	자영업자/고용		비임금	사범	지난조사비임금->이번조사 임금	
801	윤선생영어교실 대리점	영어학습지도	영어방문지	자영업자/고용		비임금	영어방문지도	지난조사비임금->이번조사 임금	
801	목공예	목공예제작	목공예제작	자영업자/고용		비임금	목공예제작	지난조사비임금->이번조사 임금	
801	없음	전기내선공사	전기내선공	자영업자/고용		비임금	전기내선공사	지난조사비임금->이번조사 임금	
801	대진통상	삼성전자에납품	운전직	자영업자/고용		비임금	운전직	지난조사비임금->이번조사 임금	
801	서울우유(구포대리점)	우유판매회사	서울우유 기	자영업자/고용		비임금	서울우유 가정배달	지난조사비임금->이번조사 임금	
801	아이엔지학원	교육	강의	가족종사자		비임금	강의	지난조사비임금->이번조사 임금	
802	토속식당	토속식당	식당에서 홀	가족종사자		비임금	식당에서 홀서빙 및 설것	지난조사비임금->이번조사 임금	
801	연세우유 경주대리점	우유배달	우유배달	자영업자/고용		비임금	우유배달	지난조사비임금->이번조사 임금	
802	매일우유 포항대리점	우유배달	우유배달	자영업자/고용		비임금	우유배달	지난조사비임금->이번조사 임금	
801	신홍택시	택시	운전	자영업자/고용		비임금	운전	지난조사비임금->이번조사 임금	
801	오성운수	운수업	운전	자영업자/고용		비임금	운전	지난조사비임금->이번조사 임금	
801	소백건축자재	건축자재 판매시공	건축미장	자영업자/고용		비임금	건축미장	지난조사비임금->이번조사 임금	
802	웰렘아트	지점토교재 문구	영업담당	가족종사자		비임금	영업담당	지난조사비임금->이번조사 임금	

3. 비표본오차의 관리 방안

3.5 데이터 클리닝

3) 종단클리닝 후 횡단클리닝(연계 클리닝)

- 종단 클리닝이 완료되면, 종단면 클리닝으로 인해 횡단면 일치성이 깨지지 않는지 재확인 작업 실시

< KLIPS 학력 변수의 사례 >

- a. CAPI 프로그래밍(학력 확인, 지난 차수, 다른 모듈의 설문과 연계)
- b. IRP를 통한 조사업체 리뷰과정
- c. 조사업체의 자체 클리닝
- d. 연구진의 횡단면 에러 체크 : 학력-연령, 학력-사교육문항, 가구학력-개인학력
- e. 연구진의 종단면 에러 체크 : 지난 차수 정보와 매칭, 에러 여부 분석
- f. 클리닝된 원자료를 조사업체에서 재송부
- g. 연구진 클리닝 완료 여부 재확인
- h. 연구진 클리닝된 변수와 다른 변수들간의 충돌 여부 재확인
- i. 클리닝으로 인해 응답 누락 문항 발생시 (전화) 재조사 실시 -> d과정부터 재실시

4. 데이터 가공

4.1 기본 개념

- 데이터 가공의 정의

- 조사업체에서 산출한 원자료를 사용자가 보다 편리하게 사용할 수 있도록 데이터 구조 및 변수를 재구성, 탐색하는 과정

- 데이터 가공과정은 데이터 클리닝 과정과 연계

- 필요한 경우 외부데이터와 매칭

- 직업능력개발원의 교육고용패널 : 학생 성적 매칭

- 한국노동연구원의 사업체패널 : 한국신용평가정보의 기업정보 매칭

- => 다문화가족종단연구의 경우 조사를 통해서 구축되지 않았지만, 학교나 학급정보의 매칭이 가능

4. 데이터 가공

4.2 데이터 매핑(mapping)

- 데이터 매핑이란, 실사업체에서 온 원자료를 유저용 데이터로 만들기 위해 변수를 매칭하는 과정을 일컬음
- 매핑(mapping)은 데이터 가공의 첫번째 단계이자 가장 중요한 단계임

11차 릴리즈 버전 레이블	11차 릴리즈 버전 변수명	11차 HRC 변수명	11차 HRC 레이블
성별 11	h110251	Q1327A11	문7.2007성별- 11
성별 12	h110252	Q1327A12	문7.2007성별- 12
성별 13	h110253	Q1327A13	문7.2007성별- 13
성별 14	h110254	Q1327A14	문7.2007성별- 14
성별 15	h110255	Q1327A15	문7.2007성별- 15
가구주와의 관계 1	h110261	Q1328A1	문8.2008가구주와의 관계-1
가구주와의 관계 2	h110262	Q1328A2	문8.2008가구주와의 관계-2
가구주와의 관계 3	h110263	Q1328A3	문8.2008가구주와의 관계-3
가구주와의 관계 4	h110264	Q1328A4	문8.2008가구주와의 관계-4

4. 데이터 가공

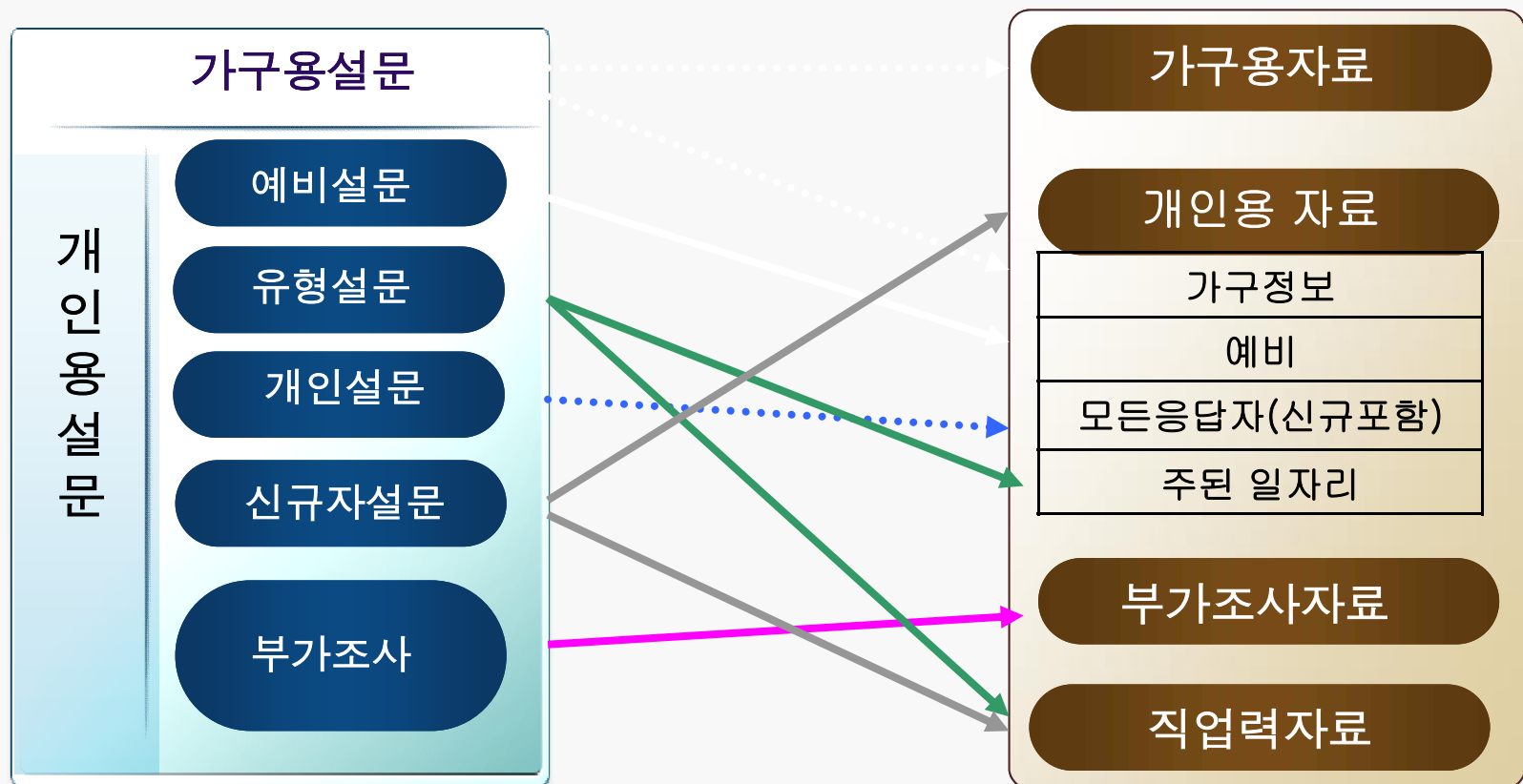
4.3 데이터 가공(restructure)

- 데이터 가공은 크게 데이터파일 가공과 변수가공으로 구분됨
 - 데이터파일 가공 : 가구용, 개인용과 같이 데이터파일의 제공 형태를 결정 하는 과정 (*아래의 예시는 노동패널의 경우를 기준으로 함)
 - a. 데이터파일은 몇 개로 하며, 어떻게 구성할 것인가를 결정 : 노동패널의 경우 가구-개인-신규-직업력-부가용 데이터로 제공되다가, 9차년도부터 개인용데이터와 신규용 데이터를 통합하여 제공함
 - b. 각 데이터파일에는 어떤 대상을 포함시킬 것인지 결정 : 노동패널 가구용 데이터의 경우 원표본 전체를 가공하는 것을 원칙으로 함. 이에 따라 당해년도 조사에 실패 하거나 가구가 소멸된 경우도 매해 데이터 가공에 포함됨
 - c. 각 데이터파일에 어떤 정보를 포함시킬 것인가 결정 : 노동패널 개인용 데이터에는 가구용 설문 의 일부 내용과 직업력 데이터의 주된 일자리 정보가 포함됨
- => 이처럼 데이터파일1(개인용)에 다른 파일(직업력)의 내용을 붙여서 제공 하면, 유저들의 편의성은 증진되지만 가공뿐 아니라 클리닝의 분량도 2배로 늘어나기 때문에 연구진의 수고가 가중되므로 주의.

4. 데이터 가공

4.3 데이터 가공(restructure)

- 노동패널의 설문과 데이터파일의 연결 구조



4. 데이터 가공

4.3 데이터 가공(Data Restructuring)

- 변수 가공
 - a. 원자료의 변수를 가공변수로 rename만 하면 되는 경우 (예 : 성별, 연령 등)
 - b. 항목번호를 조정해서 가공하는 경우
 - c. 연구진이 동일차수 내의 원자료를 이용하여, 변수 자체를 생성하여 만들어주는 경우 (예 : 가구원수 등)
 - d. 과거 데이터의 변수를 끌어다 붙여주는 경우 (예 : 일자리 시작시기 등) 등이 있음
- 노동패널의 경우 10차년도 공개용 자료부터 통합 변수명으로 제공
 - 9차년도까지는 매해 횡단자료의 순서 순에 따라 변수명을 부여하여, 차수 간 변수명이 조금씩 차이가 있었음
 - 10차년도에 변수명 부여규칙을 재정비하고, 차수간 동일한 변수를 갖도록 데이터를 재정비함. (*차수를 나타내는 자리의 변수는 다름)
 - 10차년도 변수명 재정비는 1년이 넘는 기간에 걸친 작업이었음. 조사초기 부터 변수명과 관련한 명확한 기준을 정하여 적용하는 것이 좋음

4. 데이터 가공

4.3 데이터 가공(Data Restructuring)

- 이전 ‘개인응답자의 종사상지위’ 변수

1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
p01026	p02351	p03363	p04746	p05748	p06741	p07736	p08736	p09736	p10736

- 10차년도 release 자료부터는 이들 변수명이 모두 P**0314로 일치되어 제공
단, 중간에 **에는 해당 차수가 들어가므로, 1차년도와 5차년도 및 10차년도
개인응답자의 종사상 지위는 각각 p010314, p050314, p100314로 됨
- 변수명 부여 규칙

<u>H</u>	<u>10</u>	<u>01</u>	<u>41</u>
설문종류	조사차수	SET번호	변수번호

4. 데이터 가공

4.3 데이터 가공(Data Restructuring)

- 변수 명 부여 규칙
 - 설문종류 : 가구용의 경우 H, 개인용의 경우 P, 부가조사의 경우 A.
(직업력 데이터는 기존과 동일하며, J로 시작)
 - 조사차수 : 설문종류 다음에 오는 두 자리는 해당하는 조사 차수를 의미
 - set 번호 : 여러 개의 관련 문항을 하나의 set으로 구성하여, 관련 문항들은 동일한 set 번호를 부여. 예를 들어 가구소득과 관련된 문항은 set 번호가 21 ~ 22번으로 부여.
 - 변수번호 : 해당하는 변수의 일련번호
- 9차년도 가구 소득관련 문항을 찾고 싶으면 h0921**을 찾으면 되고, 차수별로 앞 두 자리만 변동하므로, 이용자들의 변수색인이 훨씬 쉽고 프로그램 작성 시에도 편리성이 확보됨
- 연구진 입장에서는 차수별로 나오던 코드북/설문지 등의 물리적 양을 줄일 수 있으나, 가공단계에서는 변수명 일치/항목조정 등을 해야 하므로 약간의 수고가 더 필요

[illegible]

5. 코드북, 유저용 설문지 작성

5.2 유저용 설문 작성

- 데이터 이용자들이 코드북만으로는 설문지 내용과 의도를 제대로 파악하기 어렵기 때문에 유저용 설문지를 별도로 작성하여 제공
 - 유저용 설문지에는 해당 변수명을 표기해서 제공

문 17 _____님께서 학교를 졸업하거나 그만 둔 이후 지금까지 2개월 이상 계속 취업을 하신 적이 있습니까?

A**0341

1	2	3	4	5	6	7	8	9	10	11
---	---	---	---	---	---	---	---	---	----	----

(1) 취업을 한 적이 없다 → 문 35로 가십시오.

(2) 취업을 한 적이 있다



문 17-1 졸업 이후 현재까지 2달 이상 계속 일한 일자리는 모두 몇 군데나 됩니까?(현재까지 그 직장(일자리)에 취업하고 계시다면 1개의 일자리입니다)

(A**0342

1	2	3	4	5	6	7	8	9	10	11
---	---	---	---	---	---	---	---	---	----	----

)군데

6. 데이터 관리-사소하지만 잊지말아야 할 것들

6.1 key-file 관리

- Key-file은 다음 차수 조사를 위한 기본 정보로, 한번이라도 조사된 적이 있는 가구 및 개인의 핵심적인 이력 정보를 관리하는 것임.
 - 성별, 연령, 연락처, 일자리정보, 혼인상태 등으로, 다음 차수 info-sheet에 제공되는 정보 위주로 DB를 관리.
 - info-sheet 종단면 데이터의 일치성을 확보하기 위해, 전년도 정보를 담은 일종의 참고자료(*CAPI의 경우 프로그램 로직으로 처리 가능)
- ID변수에 대한 개념 정리는 처음부터 해두자!
 - 노동패널의 경우 5차년도까지 원자료에 가구원번호를 개인고유번호(PID)가 아닌 가구원 번호로 넣어와서, PID오류가 상당수 있었음

6. 데이터 관리-사소하지만 잊지 말아야 할 것들

6.2 ID 관리

- ID변수에 대한 개념 정리는 처음부터 !
 - ‘당해년도 개인번호’로 자료가 제공되면 안됨 => 가구번호나 개인고유번호는 해마다 바뀌지 않아야 하며, 원자료에 처음부터 기록하도록 해야 함 !!
 - 데이터의 종단면 연결에 필요한 ID변수를 미리 정의하고, 원자료 작성 단계에서부터 관리해야 함
 - 노동패널의 경우 가구번호(HHID)의 관리는 비교적 잘 이루어졌으나, 개인고유번호(PID)는 5차년도까지 가구용 원자료에 붙여오지 않았음
 - 이로 인해 PID오류가 상당수 있었을 뿐 아니라, 게다가 연구진이 가공하는 과정에서도 PID를 맞추느라 상당한 노력이 소진되었음
 - 6차년도 이후로는 원자료상 가구원번호에 개인고유번호(PID) 전체를 붙이게 하여 문제를 해결

6. 데이터 관리-사소하지만 잊지 말아야 할 것들

6.3 비공개 데이터 관리

- 연구목적상 필요한 비공개 데이터는 연구진이 주도적으로 관리.체크!!
- 예) KLIPS혼인상태의 종단면 에러가 빈번
 - 자료의 일치성을 위해 인포시트에 정확한 정보가 제공되면 응답 거부 가능성이 증가 (사실혼 혹은 이혼이력)
 - => 해법: KLIPS 연구진이 별도로 비공개 혼인력 리스트 작성
 - => 인포시트 상으로는 응답자들의 거짓 응답내용을 출력
 - => 그러나,공개용 데이터에는 클리닝된 내용을 제공함
 - 실사과정에서는 응답내용이 지속적으로 거짓 응답되므로, 이는 연구진이 의도적으로 다른 정보로 교체하여 자료를 제공하는 것임

6. 데이터 관리-사소하지만 잊지 말아야 할 것들

6.4 실사업체와 연구진의 데이터 일치

- 데이터 클리닝이 이뤄지는 과정에서 실사업체와 연구진의 데이터가 달라지는 일이 발생할 수 있으므로 주의
 - 패널조사는 종단면 조사이므로, 실사업체와 연구진이 동일한 데이터를 가지고 있는 것은 매우 중요
- 클리닝의 분량이 방대하기 때문에, 단순한 오류에 대해서는 연구진이 임의적으로 자료를 수정할 수 있는 가능성 존재
 - => 차기 조사와의 일관성을 위해서 실사업체의 데이터도 반드시 동시에 수정이 이루어져야 함
- 종단면 데이터의 클리닝을 누가 할 것인지도 미리 정해두어야 함
 - => 현재 노동패널의 경우 5차 이전은 연구진이 데이터를 업데이트, 6차년도 이후로는 실사업체가 데이터를 업데이트하고 있음
 - => 데이터 클리닝을 통한 원자료 수정 횟수 만큼 재가공 과정 반복(10-20회)

6. 데이터 관리-사소하지만 잊지말아야 할 것들

6.5 표준코드의 사용

- 다른 연구 자료 등과의 호환성을 고려해서, 되도록 표준 코드를 사용
 - 지역코드는 되도록 우편번호 시스템으로..
 - 표준직업분류, 산업분류도 표준코드로 맞추면 연구의 편의가 증진됨
 - 표준 코드는 아니지만, 생년월일의 경우는 (양/음)으로 구분되므로, 프로그래밍을 통해서 양력으로 전환한 생년월일을 제공하면 편리