

한국복지패널(KoWePs) 데이터처리 및 머지 방법

한국보건사회연구원

데이터 머지방법 소개에 앞서

KoWePs 1차 데이터 Beta14버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 머지용
데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 2차 데이터 Beta14버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 머지용
데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 3차 데이터 Beta12버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 머지용
데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 4차 데이터 Beta11버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 머지용
데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 5차 데이터 Beta10버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 머지용
데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 6차 데이터 Beta8버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 장애인,
머지용 데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 7차 데이터 Beta8버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 머지용
데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 8차 데이터 Beta6버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 머지용
데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 9차 데이터 Beta4버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 가구원용,
장애인용, 머지용 데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 10차 데이터 Beta3버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 가구원용,
아동용, 머지용 데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 11차 데이터 Beta2버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

1

1~12차 종단에러체크 인한 가구용, 머지용
데이터 수정

데이터 머지방법 소개에 앞서

KoWePs 12차 데이터 Beta1버전 공개

- 현재 koweps.re.kr에서 다운 가능(3월 31일 업로드 완료)

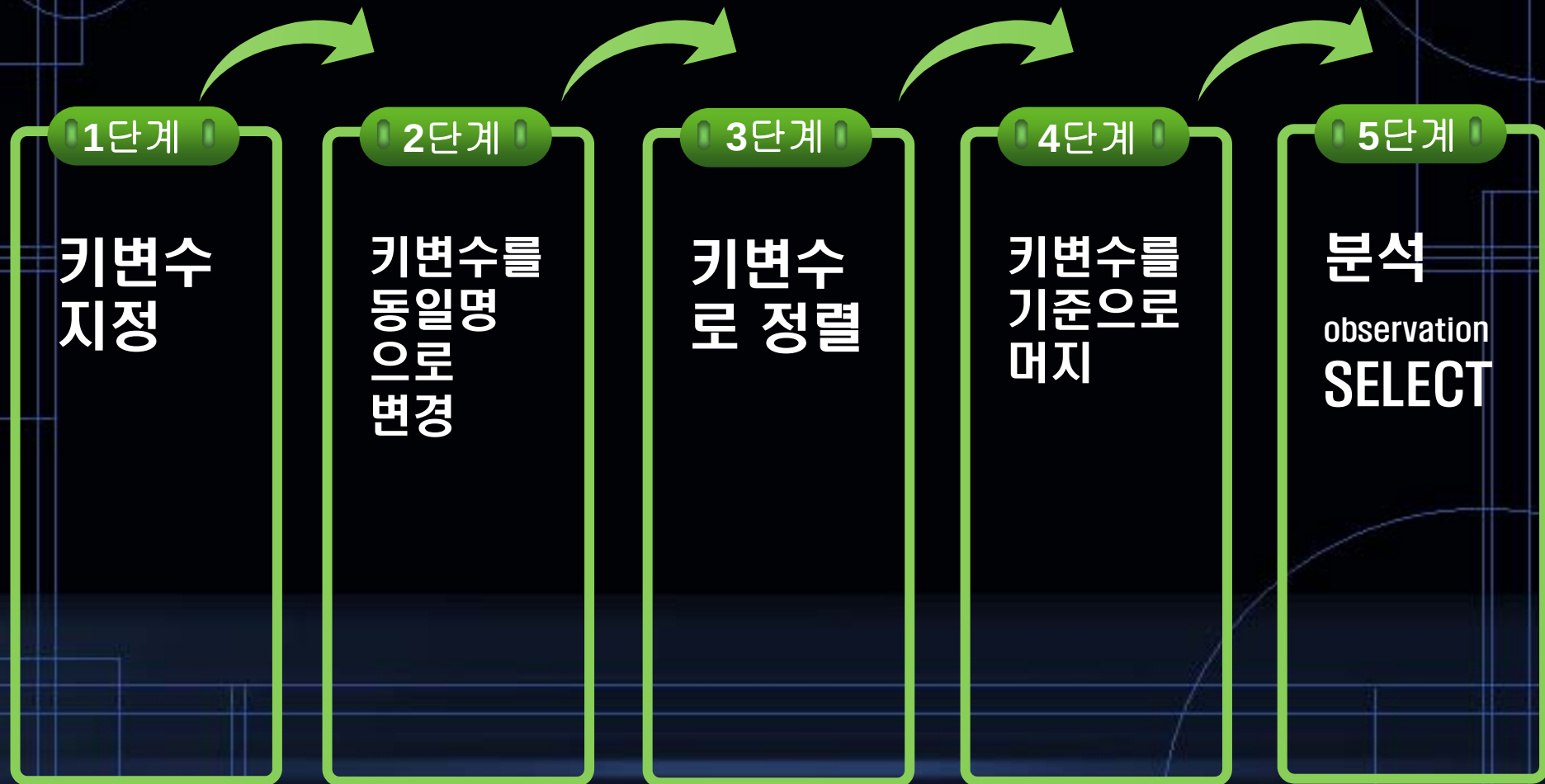
1

가구용, 가구원용, 장애인용, 머지용 데이터 공개

데이터 머지 방법 종류

- **개체 결합방식(Case Merge)**
 - 동일한 패널 ID에 대해서 관측치들을 옆으로 결합하는 방식
- **덧붙임 결합방식(Append Merge)**
 - 새로운 패널 ID의 관측치들을 아래로 추가하는 형태로 결합하는 방식

개체 결합방식의 단계



변수명 구조

(가구용(h), 가구원용(p), 부가조사(c, wc, da))

+ (패널차수) + (조사영역별 일련번호) _

+ (각 조사영역에 속한 문항의 일련번호)

→ h0102_15, h0202_15, h0302_15, h0202_aq1, h0302_3aq1
h0402_4aq1, h0503_5aq1, h0708_7aq1, h0808_8aq1,
h0903_9aq1, h1003_10aq1, h1103_11aq1, h1208_12aq1

※ 1,2,3,4,5,6,7,8,9,10,11,12차의 각 조사영역에 속한 문항의 일련번호는 동일한 문항에 대해서 동일한 번호를 부여하였음.

※ 2차조사에서 추가하였거나 많이 변형된 문항은 aq를 붙였음.

※ 3차: 3aq, 4차: 4aq, 5차: 5aq, 6차: 6aq, 7차: 7aq, 8차: 8aq,
9차: 9aq, 10차: 10aq, 11차: 11aq, 12차: 12aq를 붙였음

1~12차 데이터 종류 및 구성

1. 1차, 4차, 7차, 10차 데이터 종류

– 가구용, 가구원용, 아동부가조사

2. 2차, 5차, 8차, 11차 데이터 종류

– 가구용, 가구원용, 복지인식부가조사

3. 3차, 6차, 9차, 12차 데이터 종류

– 가구용, 가구원용, 장애인부가조사

1~12차 데이터 종류 및 구성

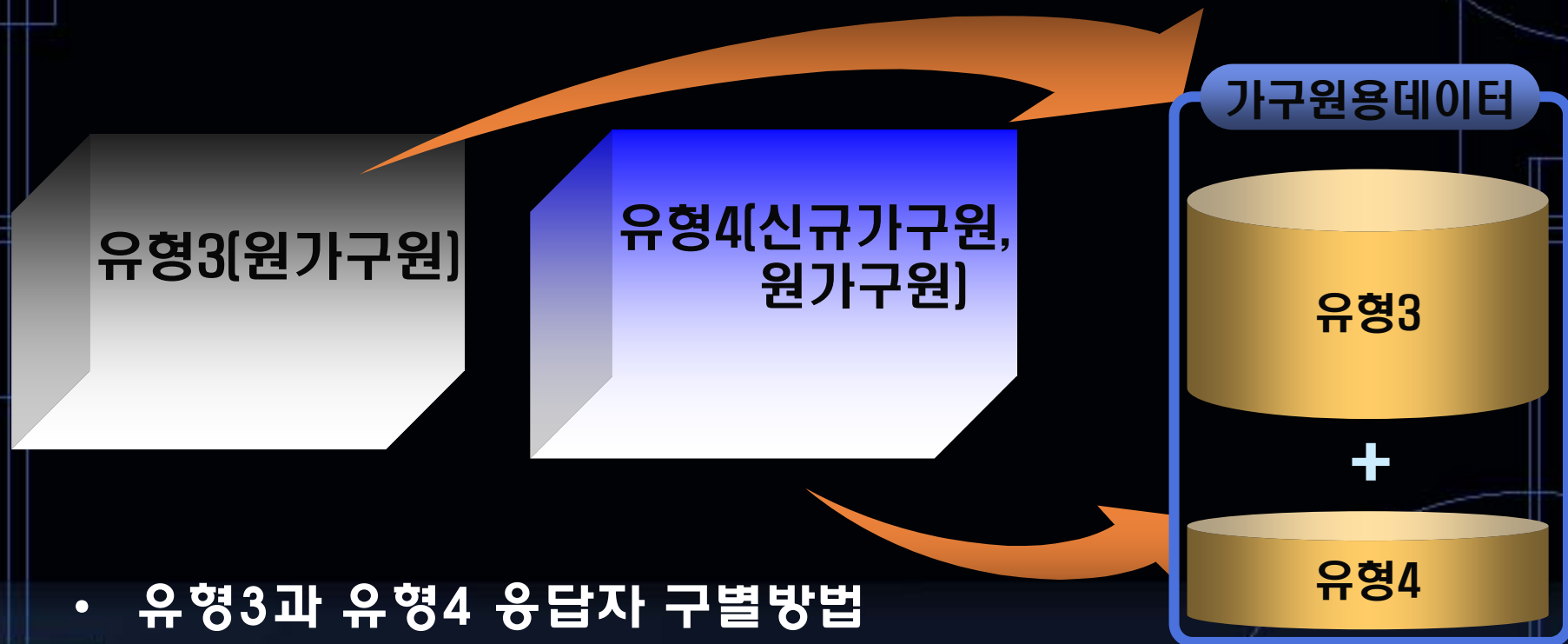
12차 가구용 데이터 구성



- 유형1(원가구)와 유형2(분가가구) 구별방법
 - 가구생성차수(h12_ind)
 - : 유형1=> 원가구(1), 기존분가(2~11)
 - 유형2=> 당해년도 분가가구(12)

1~12차 데이터 종류 및 구성

12차 가구원용 데이터 구성



- **유형3과 유형4 응답자 구별방법**
 - 설문유형구분변수(p12_tq) : 3(유형3), 4(유형4)
- **원가구원과 신규가구원 구별방법(가구원진입차수:p12_pind)**
 - 유형3=> 원가구원(1), 기존 신규가구원(2~11)
 - 유형4=> 원가구원(1), 당해년도 신규가구원(12)

가구 및 개인 식별 변수

- 가구식별변수

- 가구패널 ID, 가구생성차수, 가구분리일련번호

- 가구머지키변수(h01_merkey, h02_merkey, h03_merkey
h04_merkey, h05_merkey, h06_merkey
h07_merkey, h08_merkey, h09_merkey
h10_merkey, h11_merkey, h12_merkey)

- 개인식별변수

- 개인패널 ID(h01_pid, h02_pid, h03_pid, h04_pid,
h05_pid, h06_pid, h07_pid, h08_pid,
h09_pid, h10_pid, h11_pid, h12_pid)

횡단면 분석시 다양한 형태의 머지방법

1. 가구용(개인단위변수) + 개인용(가구원용, 부가조사)

- Key Variable: 개인패널 ID(h01_pid, h02_pid, h03_pid, h04_pid, h05_pid, h06_pid, h07_pid, h08_pid, h09_pid, h10_pid, h11_pid, h12_pid)

※ 가구용 개인단위변수들(Wide Type)의 Long Type 변환절차

절차1: 가구용 데이터를 가구원번호(1~9)로 나누어 9개의 분리데이터를 생성

절차2: 가구원번호에 해당하는 변수에 대해서 식별(머지)변수명을 부여하고
덧붙임 결합방식으로 결합

2. 가구원용 + 부가조사

- Key Variable: 개인패널 ID(h01_pid, h02_pid, h03_pid, h04_pid, h05_pid, h06_pid, h07_pid, h08_pid, h09_pid, h10_pid, h11_pid, h12_pid)

가구용 데이터 → 가구원용으로 Reshaping

표지	가구일반사항																		
	1번 가구원							2번 가구원							3번 가구원				
	가구패널 ID	가구원 진입차수	개인패널 ID	가구원 번호	가구주와의 관계	성 별	태어난 연도	가구원 진입차수	개인패널 ID	가구원 번호	가구주와의 관계	성 별	태어난 연도	가구원 진입차수	개인패널 ID	가구원 번호	가구주와의 관계	성 별	태어난 연도
변수명	h12_id	h12_pind1	h12_pid1	h1201_2	h1201_3	h1201_4	h1201_5	h12_pind2	h12_pid2	h1201_14	h1201_15	h1201_16	h1201_17	h12_pind3	h12_pid3	h1201_26	h1201_27	h1201_28	h1201_29
관측	1000	1	10001	1	010	1	1950	1	10002	2	020	2	1952	2	10003	3	011	1	1975
	2000	1	20001	1	010	1	1952	1	20002	2	020	2	1954						
	3000	1	30001	1	010	1	1948	1	30002	2	020	2	1950	2	30003	3	012	2	1973

↓

1번 가구원 데이터파일	하위 관측치	가구패널ID	가구원 진입차수	개인 패널 ID	가구원번호	가구주와의 관계	성별	태어난 연도	...
	식별변수명	h12_id	h12_pind	h12_pid	h12_g1	h12_g2	h12_g3	h12_g4	...
	관측	100	1	10001	1	010	1	1950	...
		200	1	20001	1	010	1	1952	...
		300	1	30001	1	010	1	1948	...
+									
2번 가구원 데이터파일	하위 관측치	가구패널ID	가구원 진입차수	개인 패널 ID	가구원번호	가구주와의 관계	성별	태어난 연도	...
	식별변수명	h12_id	h12_pind	h12_pid	h12_g1	h12_g2	h12_g3	h12_g4	...
	관측	100	1	10002	2	020	2	1952	...
		200	1	20002	2	020	2	1954	...
		300	1	30001	2	020	2	1950	...
+									
⋮									
+									
9번 가구원 데이터파일	하위 관측치	가구패널ID	가구원 진입차수	개인 패널 ID	가구원번호	가구주와의 관계	성별	태어난 연도	...
	식별변수명	h12_id	h12_pind	h12_pid	h12_g1	h12_g2	h12_g3	h12_g4	...
	관측	100	1	10009	9	015	1	1980	...
		300	1	20009	9	016	2	1985	...

가구용 데이터 → 가구원용으로 Reshaping

```
/* 1번 가구원 데이터셋 */  
data a1;  
  set h12_m.t;  
  if h1201_2=1;  
  rename h12_pind1=h12_pind  
h12_pid1=h12_pid  
h1201_2=h12_g1  
h1201_3=h12_g2  
h1201_4=h12_g3  
h1201_5=h12_g4  
h1201_6=h12_g6  
h1201_7=h12_g7  
h1201_8=h12_g8  
h1201_9=h12_g9  
h1201_11=h12_g10  
h1201_12=h12_g11  
h1201_13=h12_g12  
  
h1202_1=h12_med1  
h1202_2=h12_med2  
h1202_3=h12_med3  
h1202_4=h12_med4  
h1202_5=h12_med5  
h1202_6=h12_med6  
h1202_7=h12_med7  
h1202_8=h12_med8  
h1202_aq1=h12_g9_1  
h1202_9=h12_med9  
h1202_3aq1=h12_med10
```

- 가구원번호가 1인것만 추출

- 설문지에서 매트릭스 형태(1.가구일반사항, 2.건강및의료A, 3.경제활동, 4.사회보험, 8.소득)를 머지용 데이터 변수명으로 바꿈
- 위의 작업을 가구원번호 1~9번까지 실행

가구용 데이터 → 가구원용으로 Reshaping

```
data h12_gagu;  
  set a1 a2 a3 a4 a5 a6 a7 a8 a9;  
  drop h12_pind1 h12_pid1 h1201_2  
        h12_pind2 h12_pid2 h1201_14  
        h12_pind3 h12_pid3 h1201_26  
        h12_pind4 h12_pid4 h1201_38  
        h12_pind5 h12_pid5 h1201_50  
        h12_pind6 h12_pid6 h1201_62  
        h12_pind7 h12_pid7 h1201_74  
        h12_pind8 h12_pid8 h1201_86  
        h12_pind9 h12_pid9 h1201_98  
  
        h1202_1 h1202_2 h1202_3  
        h1202_10 h1202_11 h1202_12  
        h1202_19 h1202_20 h1202_21  
        h1202_28 h1202_29 h1202_30  
        h1202_37 h1202_38 h1202_39  
        h1202_46 h1202_47 h1202_48  
        h1202_55 h1202_56 h1202_57  
        h1202_64 h1202_65 h1202_66  
        h1202_73 h1202_74 h1202_75  
  
        h1203_1 h1203_2 h1203_3  
        h1203_11 h1203_12 h1203_13  
        h1203_21 h1203_22 h1203_23  
        h1203_31 h1203_32 h1203_33  
        h1203_41 h1203_42 h1203_43  
        h1203_51 h1203_52 h1203_53  
        h1203_61 h1203_62 h1203_63  
        h1203_71 h1203_72 h1203_73  
        h1203_81 h1203_82 h1203_83
```

- 9개의 가구원별 데이터셋들을 세로결합

- 기존의 매트릭스 형태의 변수들을 삭제
- 이것으로 가구용 데이터가 개인단위로 바뀐 데이터셋 구축

가구용(개인단위조사) + 가구원용

```
proc sort data=h12_gagu;  
  by h12_pid;  
run;  
proc sort data=p12;  
  by h12_pid;  
run;  
data koweps_hp12;  
  merge h12_gagu p12;  
  by h12_pid;  
run;
```

```
/* (가구용 + 가구원용) + 부가조사 */  
proc sort data=koweps_hp12;  
  by h12_pid;  
run;  
proc sort data=da12;  
  by h12_pid;  
run;  
data koweps_hpda12;  
  merge koweps_hp12 da12;  
  by h12_pid;  
run;
```

- 가구용(개인단위조사) + 가구원용을 개체결합 하기에 앞서 Key variable인 개인패널 ID로 정렬 후에 결합

- 가구용(개인단위조사)+가구원용 결합한 데이터를 부가조사와 개체결합
- 개체결합하기에 앞서 Key variable인 개인패널 ID로 정렬 후 결합

- 가구원용 + 부가조사도 동일한 방법
- 가구용 + 부가조사도 가구용+가구원용과 동일한 방법

종단면 분석시 다양한 형태의 머지방법

1. 1차 가구용 + 2차 가구용 + 3차 가구용 + 4차 가구용 + 5차 가구용 +
6차 가구용 + 7차 가구용 + 8차 가구용 + 9차 가구용 + 10차 가구용 +
11차 가구용 + 12차 가구용

– Key Variable: 가구머지키변수(h01_merkey, h02_merkey, h03_merkey,
h04_merkey, h05_merkey, h06_merkey, h07_merkey, h08_merkey
h09_merkey, h10_mekrey, h11_merkey, h12_merkey)

2. 1차 가구원용 + 2차 가구원용 + 3차 가구원용 + 4차 가구원용
+ 5차 가구원용 + 6차 가구원용 + 7차 가구원용 + 8차 가구원용
+ 9차 가구원용 + 10차 가구원용 + 11차 가구원용 + 12차 가구원용

– Key Variable: 개인패널 ID(h01_pid, h02_pid, h03_pid, h04_pid, h05_pid,
h06_pid, h07_pid, h08_pid, h09_pid, h10_pid, h11_pid, h12_pid)

종단면 분석시 다양한 형태의 머지방법

3. (1차 가구용(개인단위조사)+가구원용+아동) +
(2차 가구용(개인단위조사)+가구원용+복지인식) +
(3차 가구용(개인단위조사)+가구원용+장애인) +
:
(10차 가구용(개인단위조사)+가구원용+아동) +
(11차 가구용(개인단위조사)+가구원용+복지인식) +
(12차 가구용(개인단위조사)+가구원용+장애인)

— Key Variable: 개인패널 ID(h01_pid,h02_pid,h03_pid,h04_pid, h05_pid,
h06_pid,h07_pid,h08_pid,h09_pid,h10_pid,h11_pid,h12_pid)

1~12차 가구용을 종단면 데이터로 구축(균형패널)

1. 1차가구용 + 2차가구용 + 3차가구용 + 4차가구용 + 5차가구용 +
 ... + 9차가구용 + 10차가구용 + 11차가구용 + 12차가구용
 - 가구일반사항(경상소득, 가처분소득, 지역변수(5개권역, 7개권역), 가구구분,
 가구원수, 가구형태, 경찰상태)을 종단면 데이터로 구축

```
data panel_h01;
  set w1.Koweps_h01_2006_beta13;
  keep h_merkey h01_merkey h01_reg5 h01_reg7 h01_dir
        h0101_110 h0103_4 h01_ws h01_wg;
  h_merkey=h01_merkey;
run;
```

```
proc sort data=panel_h01; by h_merkey; run;
proc sort data=panel_h02; by h_merkey; run;
proc sort data=panel_h03; by h_merkey; run;
proc sort data=panel_h04; by h_merkey; run;
proc sort data=panel_h05; by h_merkey; run;
proc sort data=panel_h06; by h_merkey; run;
proc sort data=panel_h07; by h_merkey; run;
proc sort data=panel_h08; by h_merkey; run;
proc sort data=panel_h09; by h_merkey; run;
proc sort data=panel_h10; by h_merkey; run;
proc sort data=panel_h11; by h_merkey; run;
proc sort data=panel_h12; by h_merkey; run;
/* 12차 기준으로 1~12 모두 조사된 가구: 4309 */
data k_panel_b_h0;
  merge panel_h01 panel_h02 panel_h03 panel_h04 panel_h05
        panel_h06 panel_h07 panel_h08 panel_h09 panel_h10
        panel_h11 panel_h12;
  by h_merkey;
  if h01_merkey=h02_merkey=h03_merkey=h04_merkey=h05_merkey
     h06_merkey=h07_merkey=h08_merkey=h09_merkey=h10_merkey
     h11_merkey=h12_merkey;
run;
```

- 1차,2차,3차,4차,5차,6차,7차,8차,9차,
10차,11차,12차 가구용 데이터셋에서
분석할 변수만 남김

- h_merkey 생성(key variable)

- 1차,2차,3차,4차,5차,6차,7차,8차,9차
10차,11차,12차를 개체결합 방식으로
가로결합

- 1~12차에서 모두 조사된 가구만 남김

1~12차 가구용을 종단면 데이터로 구축(균형패널)

```
data k_panel_b_h;  
  set k_panel_b_h0;  
  hid=_n_;
```

```
  time=1;  
  reg5=h01_reg5;  
  reg7=h01_reg7;  
  cin=h01_cin;  
  din=h01_din;  
  h_hc=h01_hc;  
  f_num=h0101_1;  
  f_form=h0101_110;  
  work=h0103_4;  
  h_ws=h01_ws;  
  h_wg=h01_wg;  
  output;
```

```
  time=2;  
  reg5=h02_reg5;  
  reg7=h02_reg7;  
  cin=h02_cin;  
  din=h02_din;  
  h_hc=h02_hc;  
  f_num=h0201_1;  
  f_form=h0201_110;  
  work=h0203_4;  
  h_ws=h02_ws;  
  h_wg=h02_wg;  
  output;
```

- time(시차)변수 생성(1차=1, 2차=2, 3차=3, 4차=4, 5차=5, 6차=6, 7차=7, 8차=8, 9차=9, 10차=10, 11차=11, 12차=12)

- hid(종단패널 ID) 생성(순서대로 재 생성)

- 새로운 공통변수 생성(각 차수에 있는 기존 변수를 이용)
ex) 경상소득(cin)은 1차~12차(h01_cin~h12_cin)

- 1~12차의 가로결합된 가구데이터를 다시 세로로 구성

1~12차 가구용을 종단면 데이터로 구축 (불균형패널)

```
/* 불균형 패널 */
/* 12차 기준으로 1~12 모두 조사된 가구: 9879(1차 7072 + 7차 추가 1800 + 1~12차 분가 1007) */
data k_panel_ub_h0;
  merge panel_h01 panel_h02 panel_h03 panel_h04 panel_h05
        panel_h06 panel_h07 panel_h08 panel_h09 panel_h10
        panel_h11 panel_h12;
  by h_merkey;
run;
```

- 균형패널과 달리
조건 없이 머지

```
data k_panel_ub_h1;
  set k_panel_ub_h0;
  hid=_n_;
  time=1;
  reg5=h01_reg5;
  reg7=h01_reg7;
  cin=h01_cin;
  din=h01_din;
  h_hc=h01_hc;
  f_num=h0101_1;
  f_form=h0101_110;
  work=h0103_4;
  h_ws=h01_ws;
  h_wg=h01_wg;
  output;

  time=2;
  reg5=h02_reg5;
  reg7=h02_reg7;
  cin=h02_cin;
  din=h02_din;
  h_hc=h02_hc;
  f_num=h0201_1;
  f_form=h0201_110;
  work=h0203_4;
  h_ws=h02_ws;
  h_wg=h02_wg;
  output;
```

- 균형패널과 동일하게 종단데이터 구축
- 단, 마지막에 각 차수별 조사가 안된
가구를 삭제해야함(9879×12)

```
/* 해당차수에 조사가 안된 가구 삭제 */
data k_panel_ub_h2;
  set k_panel_ub_h1;
  if h_hc=. then delete;
run;
```


1~12차 가구용을 종단면 데이터로 구축 (불균형패널)

```
/* 가구에서 1~12차 동안 조사된 횟수 */  
data hid_count0;  
    set k_panel_ub_h2;  
    a=1;  
run;  
  
proc sort data=hid_count0; by h_merkey time; run;  
  
data hid_count;  
    set hid_count0;  
    by h_merkey;  
    hid_count+a;  
    if last.h_merkey then do;  
        output; hid_count=0;  
    end;  
    keep h_merkey hid_count;  
run;  
  
proc sort data=k_panel_ub_h2; by h_merkey; run;  
proc sort data=hid_count; by h_merkey; run;  
  
data k_panel_ub_h;  
    merge k_panel_ub_h2 hid_count;  
    by h_merkey;  
run;
```

- 1~12차 동안 조사된 횟수 변수 생성(누적합)
- 가구데이터와 머지용데이터는 1~12차 결합용(개체 및 덧붙임) 데이터(가구, 머지)를 제공함