※ 이 강연의 내용은 한국보건사회연구원과 복지패널 연구진의 입장과 무관한 강연자의 개인 의견임을 밝힙니다.

2021년 한국복지패널 데이터 설명회

한국복지패널조사 데이터의 구조와 활용

한국보건사회연구원 이원진

한국보건사회연구원

목차

가구·개인 데이터의 일반 구조

횡단 데이터
Cross-sectional

개인ID	연도	성별	연령	취업
1	2019	여	34	취업
2	2019	남	16	취업
3	2019	남	49	비취업
4	2019	여	78	비취업
…	…	…	…	…

가상데이터

- 2019년 남성과 여성의 취업률 차이는 얼마인가?

반복횡단 데이터
Repeated Cross-sectional

개인ID	연도	성별	연령	취업
1	2019	여	34	취업
2	2019	남	19	취업
3	2019	남	49	비취업
4	2019	여	78	비취업
…	…	…	…	…
101	2020	남	56	비취업
102	2020	여	34	취업
103	2020	남	24	취업
104	2020	여	18	비취업
…	…	…	…	…

가상데이터

- 2019년과 2020년 사이에 성별 취업률 격차는 얼마나 감소하였는가?

패널 데이터
Panel, Longitudinal

개인ID	연도	성별	연령	취업
1	2019	여	34	취업
2	2019	남	19	취업
3	2019	남	49	비취업
4	2019	여	78	비취업
…	…	…	…	…
1	2020	여	35	비취업
2	2020	남	20	취업
3	2020	남	50	취업
4	2020	여	79	비취업
…	…	…	…	…

가상데이터

- 2019~2020년 취업→비취업 이행확률이 성별로 다른가?

가구 데이터

- 이천만 가구 중 가구주가 여성인 가구의 비율은 얼마인가?

가구ID	가구원수	가구소득	가구원1 (가구주) 성별	가구원1 (가구주) 연령	가구원2 성별	가구원2 연령	가구원3 성별	가구원3 연령	…
1	2	102	남	46	여	49			…
2	1	403	여	27					…
3	3	262	여	52	여	32	남	13	…
4	2	39	남	48	남	14			…
…	…	…	…	…	…	…	…	…	…

가상 데이터

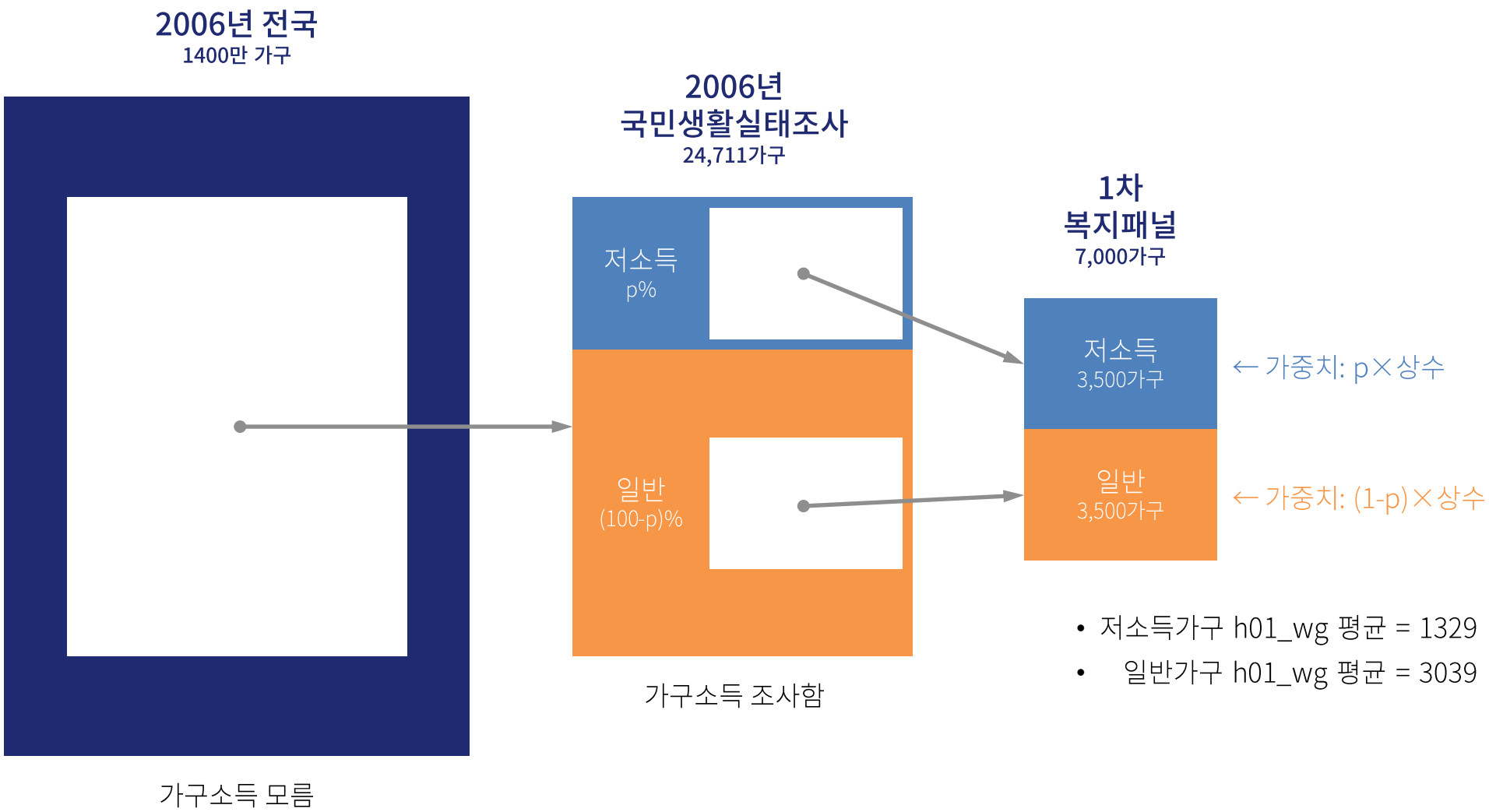
개인 데이터

- 오천만 개인 중 여성 비율은 얼마인가?
- 오천만 개인 중 가구주가 여성인 가구에 속한 개인의 비율은 얼마인가?

가구ID	개인ID	가구원수	가구소득	가구주 성별	가구주 연령	개인 성별	개인 연령
1	11	2	102	남	46	남	46
1	12	2	102	남	46	여	49
2	21	1	403	여	27	여	27
3	31	3	262	여	52	여	52
3	32	3	262	여	52	여	32
3	33	3	262	여	52	남	13
4	41	2	39	남	48	남	48
4	42	2	39	남	48	남	14
…	…	…	…	…	…	…	…

가상 데이터

1차 복지패널 설계


2006년 전국
1400만 가구
가구소득 모름
2006년
국민생활실태조사
24,711가구
저소득
p%
일반
(100-p)%
가구소득 조사함
1차
복지패널
7,000가구
저소득
3,500가구
← 가중치: p×상수
일반
3,500가구
← 가중치: (1-p)×상수
• 저소득가구 h01_wg 평균 = 1329
• 일반가구 h01_wg 평균 = 3039

1차 가구 데이터
koweps_h01_2006

- 7,072가구
- 가구당 평균 가구원수 = 2.666명
- 가구당 평균 가구가중치(h01_wg) = 2246.483
- 가구가중치 총합 = 7,072가구 × 2246.483 = 15,887,128가구
  → 전국 전체 가구 대표

1차 가구-가구원 머지데이터
koweps_hpc01_2006

- 7,072가구에 속한 모든 개인
- 18,856명 = 7,072가구 × 2.666명
- 개인당 평균 개인가중치(p01_wg) = 2494.773
- 개인가중치 총합 = 18,856명 × 2494.773 = 47,041,434명
  → 아동 포함한 전국 전체 개인 대표

1차 가구원 데이터
koweps_p01_2006

- 18,856명 중 가구원용 조사표 응답 대상인 개인 14,463명
- 대상: 만15세 이상 가구원(중고등학생 제외)
- 가구-가구원 머지데이터와 동일한 개인가중치 부여
  → 15세 미만과 중고등학생 제외한 전국 전체 개인 대표
  → 단, 조사 미완 269명 존재하므로 실제 분석의 대표성에 일정한 한계

1차 가구 데이터 + 1차 가구-가구원 머지데이터

koweps_h01_2006 koweps_hpc01_2006

h01_id 가구패널ID	h01_ind 가구생성차수	h01_sn 가구분리일련번호	h01_merkey 가구머지키	h01_wg 가구가중치	p01_wg 개인가중치	h01_pind 가구원진입차수	h01_pid 개인ID
1	1	1	10101	1000	1011	1	101
1	1	1	10101	1000	1011	1	102
2	1	1	20101	1500	1517	1	201
3	1	1	30101	800	809	1	301
3	1	1	30101	800	809	1	302
3	1	1	30101	800	809	1	303
…	…	…	…	…	…	…	…

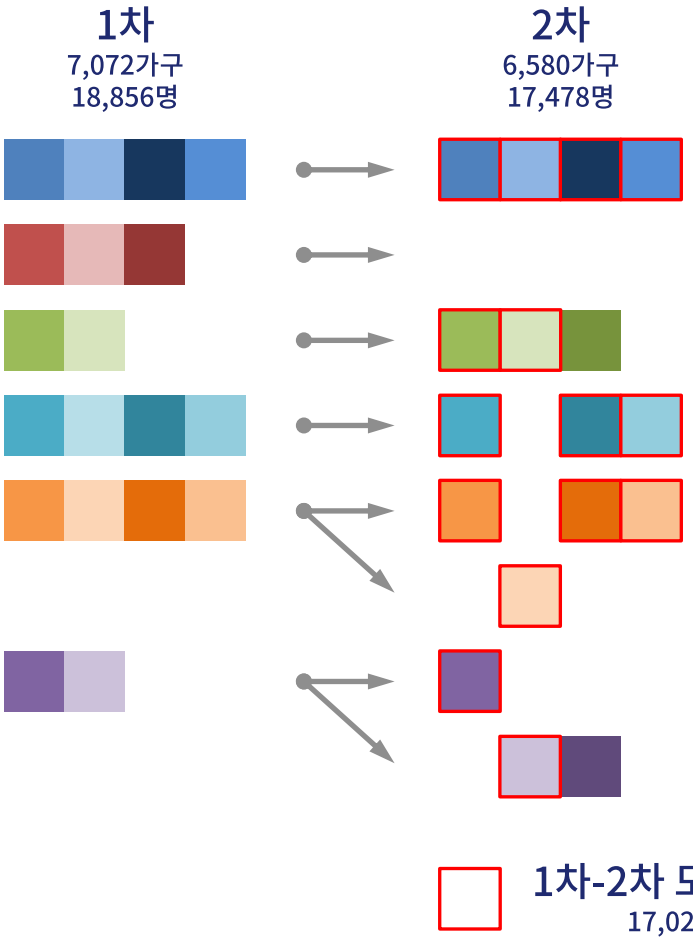
가상 데이터

- 1차는 가구생성차수, 가구분리일련번호, 가구원진입차수 모두 1
- 가구머지키: 가구패널ID, 가구생성차수, 가구분리일련번호의 조합으로 구성

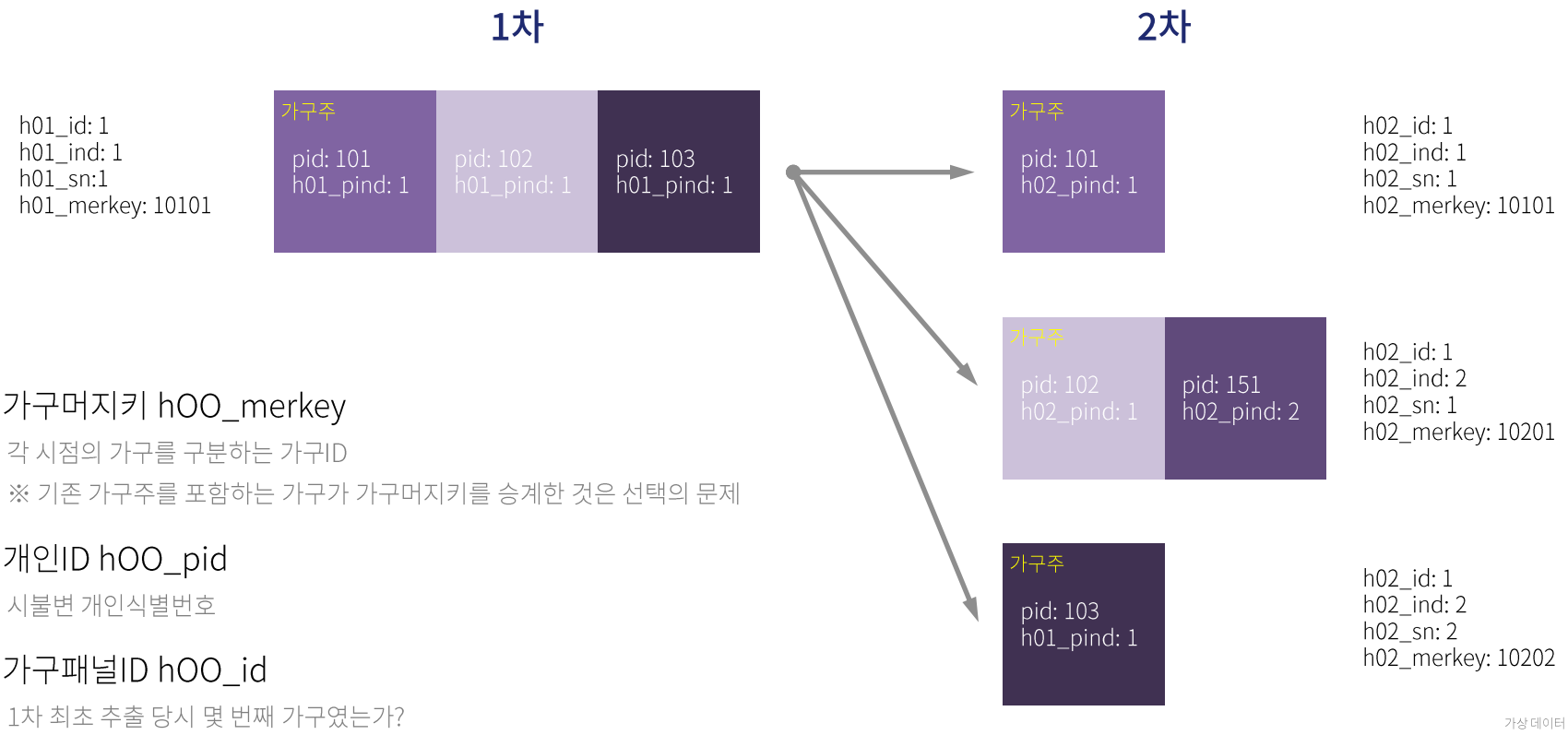
  h01_merkey = h01_id×10000 + h01_ind×100 + h01_sn
- 1차 가구가중치와 1차 개인가중치는 기본적으로 동일하되, 가구모수와 인구모수에 맞추어 scale 조정
- 가구가중치가 w인 가구는 전국에 w가구 존재하므로, 해당 가구에 속한 개인도 기본적으로 전국에 w명 존재

  p01_wg = h01_wg×1.011215

2차 복지패널 설계



- 가구 구성이 그대로 유지된 채로 조사되기도 하고
- 가구가 통째로 탈락하기도 하고
- 결혼 등으로 가구원이 추가되기도 하고
- 사망 등으로 가구원이 사라지기도 하고
- 가구원이 분가하기도 하고
- 분가한 가구에서 가구원이 추가되기도 한다.



- 가구머지키 hOO_merkey

  각 시점의 가구를 구분하는 가구ID

  ※ 기존 가구주를 포함하는 가구가 가구머지키를 승계한 것은 선택의 문제

- 개인ID hOO_pid

  시불변 개인식별번호

- 가구패널ID hOO_id

  1차 최초 추출 당시 몇 번째 가구였는가?

- 가구생성차수 hOO_ind

  몇 번째 차수에 생성된 가구인가?

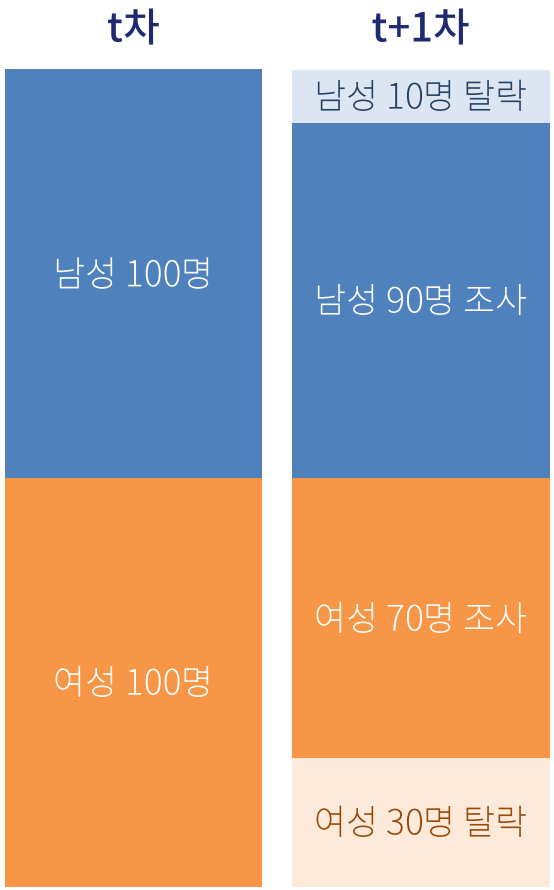
- 가구원진입차수 hOO_pind

  몇 번째 차수에 진입한 가구원인가?

- 가구분리일련번호 hOO_sn

  원가구로부터 몇 번째로 분리된 가구인가?

종단가중치의 기본 개념



- 1차 표본 여성 비율
- 1차 횡단가중치 적용 = 51.76%

  tabulate h01_g3 [aweight=p01_wg]

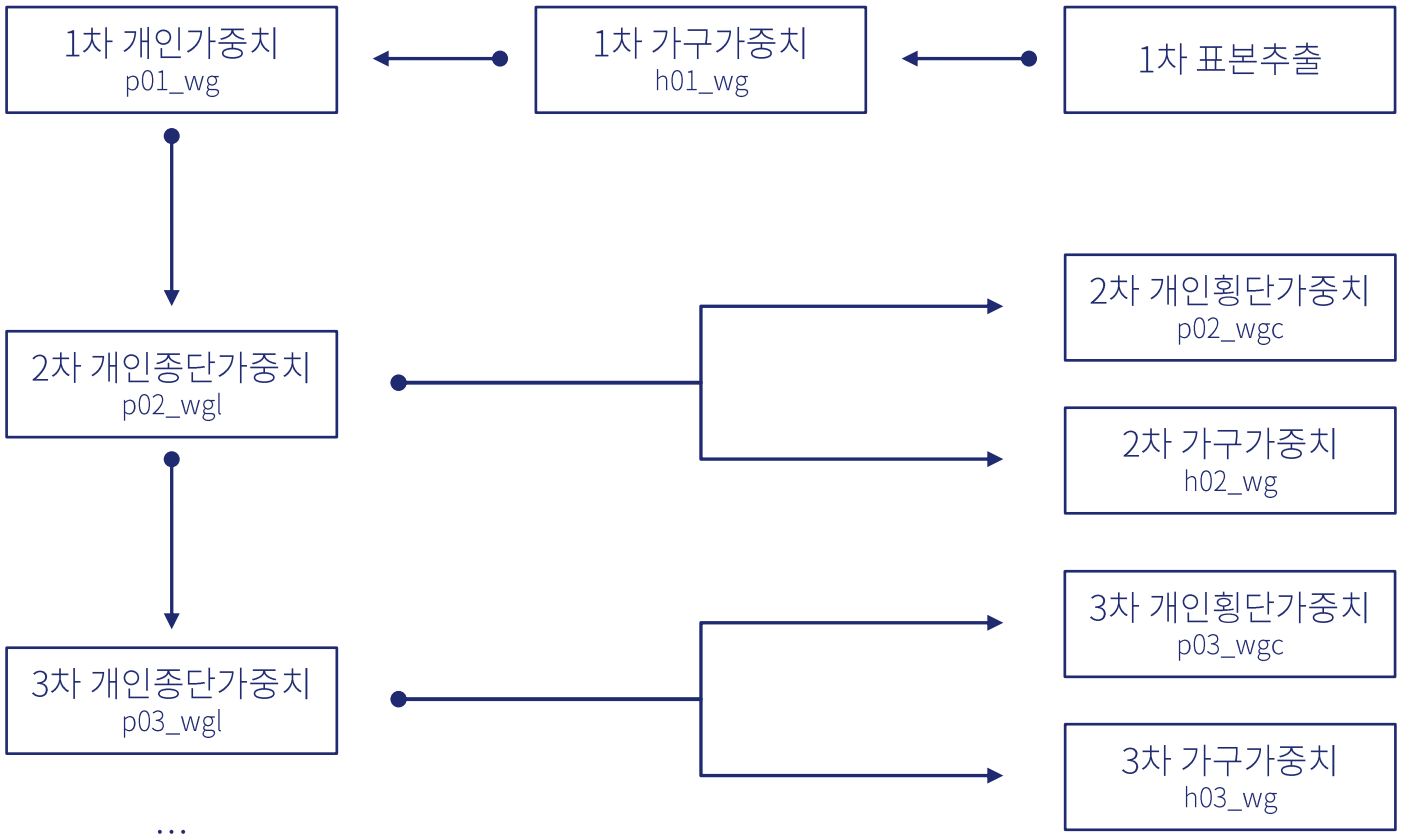
- 1차-2차 균형패널 표본 여성 비율
- 1차 횡단가중치 적용 = 52.01%

  tabulate h02_g3 [aweight=p01_wg] if _merge==3

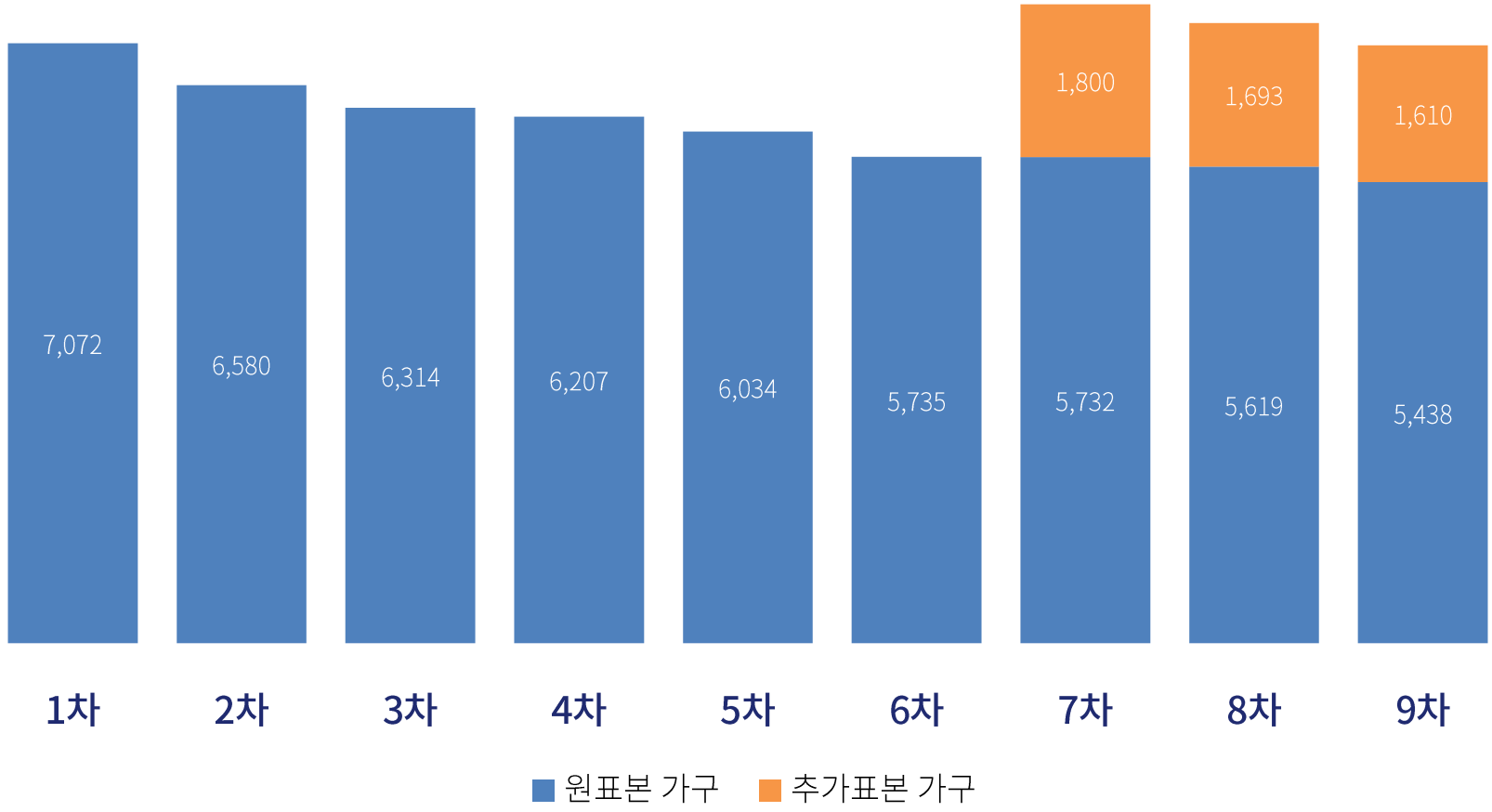
- 1차-2차 균형패널 표본 여성 비율
- 2차 종단가중치 적용 = 51.73%

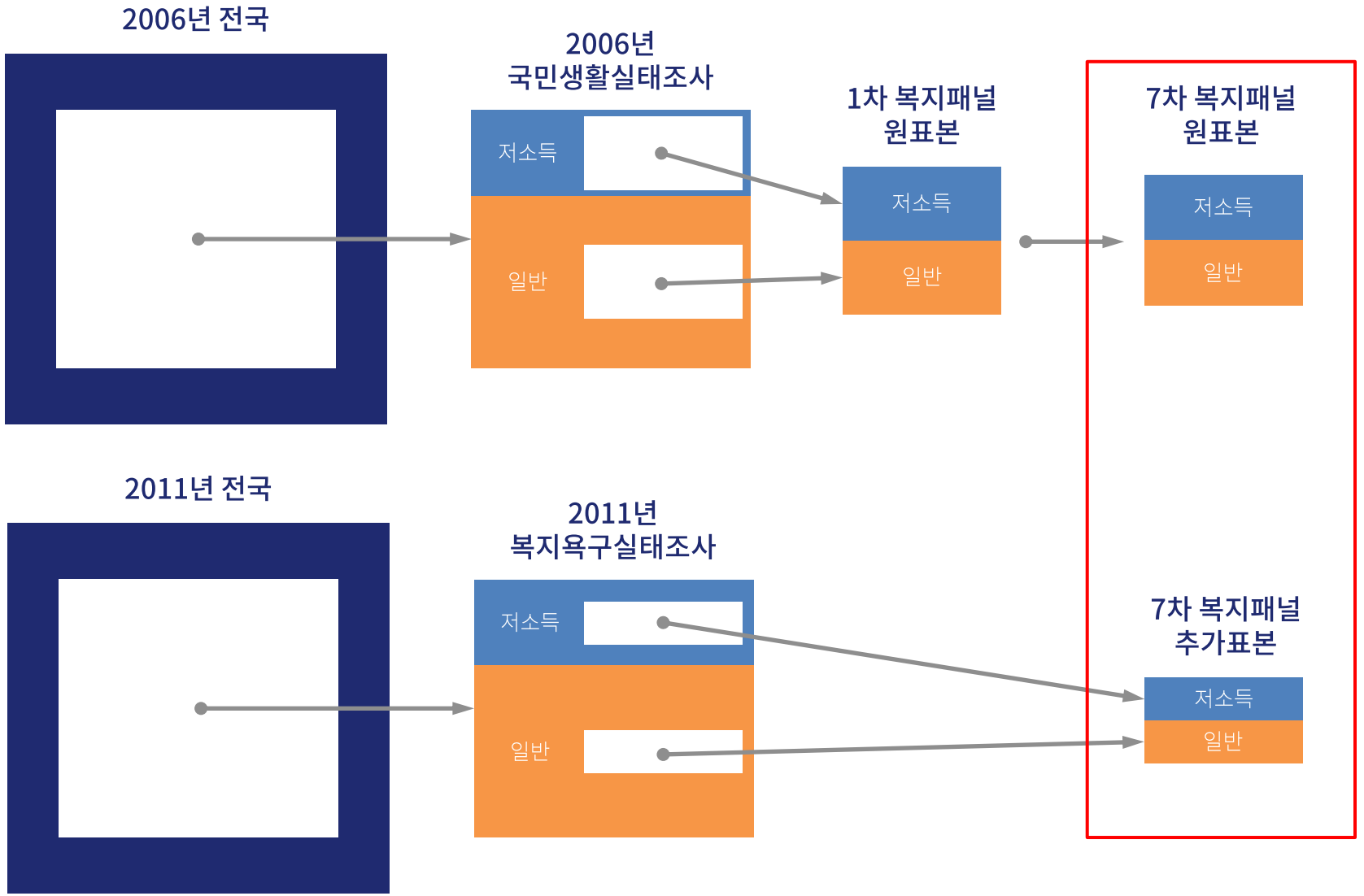
  tabulate h02_g3 [aweight=p02_wgl] if _merge==3

복지패널 가중치 부여 체계



7차 복지패널 추가표본 설계




2006년 전국
2006년
국민생활실태조사
저소득
일반
1차 복지패널
원표본
저소득
일반
7차 복지패널
원표본
저소득
일반
2011년 전국
2011년
복지욕구실태조사
저소득
일반
7차 복지패널
추가표본
저소득
일반

7차 가구 데이터 + 7차 가구-가구원 머지데이터

koweps_h07_2012 koweps_hpc07_2012

h_new 추가표본	h07_id 가구패널ID	h07_ind 가구 생성차수	h07_merkey 가구머지키	h07_wg 가구원표본 가중치	h07_wg_all 가구통합표본 가중치	h07_pind 가구원 진입차수	h07_pid 개인ID	p07_wgc 개인원표본 횡단가중치	p07_wgl 개인원표본 종단가중치	p07_wgc_all 개인통합표본 횡단가중치	p07_wgl_all 개인통합표본 종단가중치
0	1	1	10101	600	700	1	101	500	450	650	450
0	2	1	20101	3000	2500	1	201	2400	2300	800	2300
0	2	1	20101	3000	2500	1	202	2300	2200	850	2200
0	2	6	20601	2800	2900	1	203	1800	1900	1200	1900
0	2	6	20601	2800	2900	6	251	1900	2000	1400	2000
…	…	…	…	…	…	…	…	…	…	…	…
1	8001	7	80010701		2000	7	800101			2200	0
1	8001	7	80010701		2000	7	800102			2100	0
1	8001	7	80010701		2000	7	800103			1900	0
1	8002	7	80020701		3200	7	800201			3000	0
…	…	…	…	…	…	…	…	…	…	…	…

가상 데이터

- 통합표본 = 원표본+추가표본
- 통합표본 분석 시 통합표본가중치 사용, 원표본 분석 시 원표본가중치 사용
- 7차는 추가표본의 최초 조사차수이므로, 추가표본의 종단가중치 없고, 개인원표본종단가중치=개인통합표본종단가중치

- 개인원표본횡단가중치 총합 = 49,779,441명
- 원표본 여성 비율, 개인원표본횡단가중치 적용 = 49.90%

tabulate h07_g3 [aweight=p07_wgc] if h_new==0

- 개인통합표본횡단가중치 총합 = 49,779,439명
- 통합표본 여성 비율, 개인통합표본횡단가중치 적용 = 49.91%

tabulate h07_g3 [aweight=p07_wgc_all]

가중치 활용의 필요성

기술 분석의 경우

- 가중치를 적용해야 함.
- 특히, 복지패널은 저소득층을 과대표집하였으므로 가중치를 적용해야 할 필요성이 매우 큼.
- 15차 가구 통합표본 가처분소득 평균, 가중치 미적용 = 4025만

  summarize h15_din
- 15차 가구 통합표본 가처분소득 평균, 가중치 적용 = 5043만

  summarize h15_din [aweight=h15_wg_all]

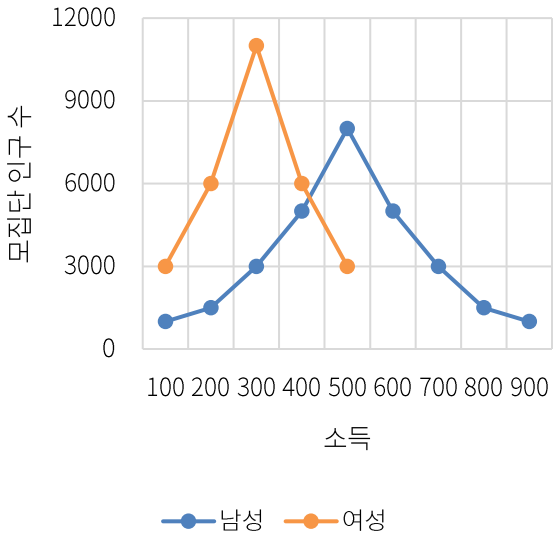
변수 간 인과관계 분석의 경우

- 분석 모델이 정확하면, 가중치를 적용하지 않는 것이 괜찮거나 오히려 더 효율적일 수 있다는 주장이 있음.
- 하지만 분석 모델이 정확하지 않다면?
- 게다가, 분석 모델이 정확하더라도 저소득층 과대표집 때문에 변수 간 관계에 편의가 발생하는 경우가 있음.

성별	교육수준	가중치	소득
남성	저학력	500	100
남성	고학력	3500	130
여성	저학력	1800	50
여성	고학력	1200	60

가상데이터

가중치 미적용		가중치 적용
• 여성 비율 = 50% • 여성의 고학력 비율 = 50% • 남성의 고학력 비율 = 50%		• 여성 비율 = 43% • 여성의 고학력 비율 = 40% • 남성의 고학력 비율 = 88%
• 남성-저학력 소득 평균 = 100 • 남성-고학력 소득 평균 = 130 • 여성-저학력 소득 평균 = 50 • 여성-고학력 소득 평균 = 60	**정확한 분석모델**	• 남성-저학력 소득 평균 = 100 • 남성-고학력 소득 평균 = 130 • 여성-저학력 소득 평균 = 50 • 여성-고학력 소득 평균 = 60
• 남성 소득 평균 = 115 • 여성 소득 평균 = 55	**정확하지 않은 분석모델**	• 남성 소득 평균 = 126 • 여성 소득 평균 = 54



성별	소득	모집단	여성 과대표집 표본			저소득층 과대표집 표본		
		인구	표본 수	추출확률	가중치	표본 수	추출확률	가중치
남성	100	1000	10	1%	100	50	5%	20
남성	200	1500	15	1%	100	75	5%	20
남성	300	3000	30	1%	100	150	5%	20
남성	400	5000	50	1%	100	50	1%	100
남성	500	8000	80	1%	100	80	1%	100
남성	600	5000	50	1%	100	50	1%	100
남성	700	3000	30	1%	100	30	1%	100
남성	800	1500	15	1%	100	15	1%	100
남성	900	1000	10	1%	100	10	1%	100
여성	100	3000	150	5%	20	150	5%	20
여성	200	6000	300	5%	20	300	5%	20
여성	300	11000	550	5%	20	550	5%	20
여성	400	6000	300	5%	20	60	1%	100
여성	500	3000	150	5%	20	30	1%	100

가상 데이터

모집단

- 소득=500-200×여성

여성 과대표집
가중치 적용

- 소득=500-200×여성

여성 과대표집
가중치 미적용

- 소득=500-200×여성

저소득층 과대표집
가중치 적용

- 소득=500-200×여성

저소득층 과대표집
가중치 미적용

- 소득=386-130×여성

데이터 구조와 연구문제

횡단 가구 데이터

- 2019년 남성 가구주 가구와 여성 가구주 가구의 빈곤율 격차는 얼마인가?

  해당 시점 가구가중치 적용

  표본 추가 이후 통합표본이라면, 가구통합표본가중치 적용

  robust standard error 사용 고려

연도	가구 머지키	가구주 성별	빈곤
2019	10101	여성	빈곤
2019	20101	남성	비빈곤
2019	30101	남성	빈곤
2019	40101	여성	비빈곤
…	…	…	…

가상 데이터

횡단 개인 데이터

- 2019년 남성 가구주 가구에 속한 개인과 여성 가구주 가구에 속한 개인의 빈곤율 격차는 얼마인가?
- 2019년 남성 개인과 여성 개인의 빈곤율 격차는 얼마인가?
- 2019년 남성 가구주 가구에 속한 개인과 여성 가구주 가구에 속한 개인의 우울 격차는 얼마인가?
- 2019년 남성 개인과 여성 개인의 우울 격차는 얼마인가?

  해당 시점 개인횡단가중치 적용

  표본 추가 이후 통합표본이라면, 개인통합표본횡단가중치 적용

  가구 단위 cluster-robust standard error 사용 고려

연도	가구 머지키	개인ID	가구주 성별	개인 성별	빈곤	우울
2019	10101	101	여성	여성	빈곤	5
2019	20101	201	남성	남성	비빈곤	2
2019	20101	202	남성	여성	비빈곤	4
2019	30101	301	남성	남성	빈곤	3
2019	30101	302	남성	여성	빈곤	6
2019	30101	303	남성	여성	빈곤	2
2019	40101	401	여성	여성	비빈곤	1
…	…	…	…	…	…	…

가상 데이터

pooled 가구 또는 개인 데이터

- 2019년과 2020년 사이에 빈곤율이 얼마나 변화하였는가?
- 2019년과 2020년 사이에 가구주 성별 또는 개인 성별 빈곤율 격차가 얼마나 변화하였는가?

각 시점별 횡단가중치 적용

표본 추가 이후 통합표본이라면, 통합표본횡단가중치 적용

패널데이터를 반복횡단데이터처럼 활용한 trend 분석의 시계열 안정성에 대해서는 주의 필요

특히 표본 추가 전 원표본과 표본 추가 이후 통합표본의 시계열 연속성 주의

연도별 통계치가 아니라 여러 연도를 결합한 통계치를 구할 경우 가중치의 scale에 대한 고민 필요

가구패널ID 단위 cluster-robust standard error 사용 고려, 최소한 개인ID 단위 cluster-robust standard error 사용 필요

연도	가구 패널ID	가구 머지키	개인ID	가구주 성별	개인 성별	빈곤	우울
2019	1	10101	101	여성	여성	빈곤	5
2019	2	20101	201	남성	남성	비빈곤	2
2019	2	20101	202	남성	여성	비빈곤	4
…	…	…	…	…	…	…	…
2020	1	10101	101	여성	여성	빈곤	8
2020	2	20101	201	남성	남성	비빈곤	1
2020	2	21501	202	여성	여성	빈곤	6
2020	2	21501	251	여성	남성	빈곤	5
…	…	…	…	…	…	…	…

가상 데이터

종단 가구 데이터 wide form

- t년 빈곤 가구의 t+n년 빈곤탈출확률은 얼마인가?
- t~t+n년 빈곤→비빈곤 이행확률의 가구주 성별 격차가 존재하는가?

  표본 추가 이후 시기만을 대상으로 한다면, 통합표본가중치 적용

  표본 추가 전 시기와 표본 추가 이후 시기를 포함한다면, 원표본으로 구성된 균형패널이므로 원표본가중치 적용

  원칙적으로 해당 상황에 정확히 부합하는 가중치를 제공하고 있지 않고, 대체로 마지막 차수 가구가중치 적용 고려

  robust standard error 사용 고려

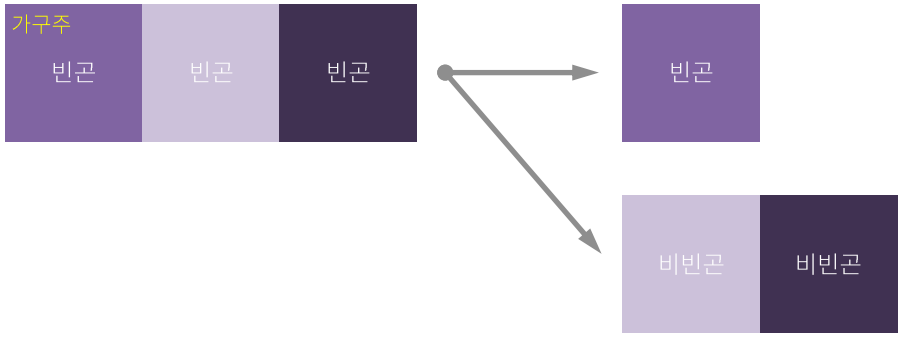
가구 머지키	t년 가구주 성별	t년 빈곤	t+n년 가구주 성별	t+n년 빈곤
10101	여성	빈곤	여성	빈곤
20101	남성	비빈곤	여성	빈곤
30101	남성	빈곤	남성	비빈곤
40101	여성	비빈곤	여성	비빈곤
…	…	…	…	…

가상 데이터

※ 가구 단위 종단 분석이 가능한가?

기존 가구주를 포함하는 가구가 가구머지키를 승계한 것은 선택의 문제

가구 단위 종단 분석에서는 분가 가구를 제외



종단 개인 데이터 wide form

- t년 빈곤 인구의 t+n년 빈곤탈출확률은 얼마인가?
- t~t+n년 빈곤→비빈곤 이행확률의 가구주 성별 또는 개인 성별 격차가 존재하는가?

표본 추가 이후 시기만을 대상으로 한다면, 통합표본가중치 적용

표본 추가 전 시기와 표본 추가 이후 시기를 포함한다면, 원표본으로 구성된 균형패널이므로 원표본가중치 적용

원칙적으로 해당 상황에 정확히 부합하는 가중치 미제공

1~t차 균형패널이라면, 대체로 마지막 차수 개인종단가중치 적용 고려

t~t+n차 균형패널이라면, t차 표본의 t+n차 조사확률 추정하여 직접 가중치 구성 고려

가구패널ID 단위 cluster-robust standard error 사용 고려

가구 패널ID	개인ID	t년 가구 머지키	t년 가구주 성별	t년 개인 성별	t년 빈곤	t+n년 가구 머지키	t+n년 가구주 성별	t+n년 개인 성별	t+n년 빈곤
1	101	10101	여성	여성	빈곤	10101	여성	여성	빈곤
2	201	20101	남성	남성	비빈곤	20101	남성	남성	빈곤
2	202	20101	남성	여성	비빈곤	20101	남성	여성	빈곤
3	301	30101	남성	남성	빈곤	30101	남성	남성	빈곤
3	302	30101	남성	여성	빈곤	30501	여성	여성	비빈곤
3	303	30101	남성	여성	빈곤	30501	여성	여성	비빈곤
4	401	40101	여성	여성	비빈곤	40101	여성	여성	빈곤
…	…	…	…	…	…	…	…	…	…

가상 데이터

종단 가구 또는 개인 데이터 long form

- 고정효과, 확률효과 등 패널 회귀분석
- 관찰되지 않은 시불변 이질성을 통제할 때, 빈곤이 우울에 영향을 미치는가?

  대체로 개인별 마지막 차수에 관찰된 개인종단가중치를 활용하다는 것이 적절하다는 견해가 있으나, 불분명

  표본 추가 전 원표본과 표본 추가 이후 통합표본 혼재 여부, 균형패널 또는 불균형패널 여부 등에 따라서도 가중치 선택 고민

  가구패널ID 단위 cluster-robust standard error 사용 고려, 최소한 개인ID 단위 cluster-robust standard error 사용 필요

연도	가구 패널ID	가구 머지키	개인ID	가구주 성별	개인 성별	빈곤	우울
2019	1	10101	101	여성	여성	빈곤	5
2019	2	20101	201	남성	남성	비빈곤	2
2019	2	20101	202	남성	여성	비빈곤	4
…	…	…	…	…	…	…	…
2020	1	10101	101	여성	여성	빈곤	8
2020	2	20101	201	남성	남성	비빈곤	1
2020	2	21501	202	여성	여성	빈곤	6
2020	2	21501	251	여성	남성	빈곤	5
…	…	…	…	…	…	…	…

가상 데이터

감사합니다