데이터 질의에 대한 답변

KIHASA
한국보건사회연구원
Korea Institute for Health and Social Affairs

질문1. 가구원정보와 자녀(엄마)번호 연결하기

Q 1.

- 데이터 구조

```
/*************************************************************************************;
/* 가구용 형태의 데이터셋 부를 1번, 모를 2번, 첫째자녀 3번, 둘째자녀 4번 ~ 일곱째자녀 9번  */
/* 가구주(남성) + 배우자(남성) => 1번 */
/* 가구주(여성) + 배우자(여성) => 2번 */
/* 자녀1 => 3번, 자녀2 => 4번, … 자녀7 => 9번 */
/*************************************************************************************/
```

가 구					
1번 가구원	2번 가구원	3번 가구원	4번 가구원	…	9번 가구원
가구주(남성) + 배우자(남성)	가구주(여성) + 배우자(여성)	자녀1	자녀2	…	자녀7
1번 변수명 변경	2번 변수명 변경	3번 변수명 변경	4번 변수명 변경	…	9번 변수명 변경

질문1. 가구원정보와 자녀(엄마)번호 연결하기

Q 1.

- 가구주와의 관계가 가구주(남성), 가구주(여성), 배우자(남성), 배우자(여성), 자녀1~7까지 데이터 생성

```
data gaguju_m gaguju_f spouse_m spouse_f child1 child2 child3 child4 child5 child6 child7;
    set hpda12_r;
    if h12_g2=10 and h12_g3=1 then output gaguju_m; /* 가구주(남성) */
    else if h12_g2=10 and h12_g3=2 then output gaguju_f; /* 가구주(여성) */
    else if h12_g2=20 and h12_g3=1 then output spouse_m; /* 배우자(남성) */
    else if h12_g2=20 and h12_g3=2 then output spouse_f; /* 배우자(여성) */
    else if h12_g2=11 then output child1; /* 첫째자녀 */
    else if h12_g2=12 then output child2; /* 둘째자녀 */
    else if h12_g2=13 then output child3; /* 셋째자녀 */
    else if h12_g2=14 then output child4; /* 넷째자녀 */
    else if h12_g2=15 then output child5; /* 다섯째자녀 */
    else if h12_g2=16 then output child6; /* 여섯째자녀 */
    else if h12_g2=17 then output child7; /* 일곱째자녀 */
    else delete;
run;
```

질문1. 가구원정보와 자녀(엄마)번호 연결하기

Q 1.

- 1번 가구원 데이터셋

1번 가구원

가구주(남성) + 배우자(남성)

1번 변수명 변경

```
proc sort data=gaguju_m; by h12_merkey; run;
proc sort data=spouse_m; by h12_merkey; run;

data father;
    set gaguju_m spouse_m;
    by h12_merkey;
    rename h12_pind = f_h12_pind
           h12_pid = f_h12_pid
           h12_g1 = f_h12_g1
           h12_g2 = f_h12_g2
           h12_g3 = f_h12_g3
           h12_g4 = f_h12_g4
           h12_g6 = f_h12_g6
           h12_g7 = f_h12_g7
           h12_g8 = f_h12_g8
           h12_g9 = f_h12_g9
           h12_g10 = f_h12_g10
           h12_g11 = f_h12_g11
           h12_g12 = f_h12_g12;
run;
```

- 2번 가구원 데이터셋

2번 가구원

가구주(여성) + 배우자(여성)

2번 변수명 변경

```
proc sort data=gaguju_f; by h12_merkey; run;
proc sort data=spouse_f; by h12_merkey; run;

data mother;
    set gaguju_f spouse_f;
    by h12_merkey;
    rename h12_pind = m_h12_pind
           h12_pid = m_h12_pid
           h12_g1 = m_h12_g1
           h12_g2 = m_h12_g2
           h12_g3 = m_h12_g3
           h12_g4 = m_h12_g4
           h12_g6 = m_h12_g6
           h12_g7 = m_h12_g7
           h12_g8 = m_h12_g8
           h12_g9 = m_h12_g9
           h12_g10 = m_h12_g10
           h12_g11 = m_h12_g11
           h12_g12 = m_h12_g12;
run;
```

질문1. 가구원정보와 자녀(엄마)번호 연결하기

Q 1.

– 3번 가구원 데이터셋

3번 가구원

자녀1

3번 변수명 변경

```
data n_child1;
    set child1;
    rename h12_pind = ch1_h12_pind
           h12_pid = ch1_h12_pid
           h12_g1 = ch1_h12_g1
           h12_g2 = ch1_h12_g2
           h12_g3 = ch1_h12_g3
           h12_g4 = ch1_h12_g4
           h12_g6 = ch1_h12_g6
           h12_g7 = ch1_h12_g7
           h12_g8 = ch1_h12_g8
           h12_g9 = ch1_h12_g9
           h12_g10 = ch1_h12_g10
           h12_g11 = ch1_h12_g11
           h12_g12 = ch1_h12_g12;
    ch1=1;
run;
```

– 9번 가구원 데이터셋

9번 가구원

자녀7

9번 변수명 변경

```
data n_child7;
    set child7;
    rename h12_pind = ch7_h12_pind
           h12_pid = ch7_h12_pid
           h12_g1 = ch7_h12_g1
           h12_g2 = ch7_h12_g2
           h12_g3 = ch7_h12_g3
           h12_g4 = ch7_h12_g4
           h12_g6 = ch7_h12_g6
           h12_g7 = ch7_h12_g7
           h12_g8 = ch7_h12_g8
           h12_g9 = ch7_h12_g9
           h12_g10 = ch7_h12_g10
           h12_g11 = ch7_h12_g11
           h12_g12 = ch7_h12_g12;
    ch7=1;
run;
```

질문1. 가구원정보와 자녀(엄마)번호 연결하기

Q 1.

- 1번~9번 데이터 셋을 key변수인 가구머지키로 가로결합

```
proc sort data=father; by h12_merkey; run;
proc sort data=mother; by h12_merkey; run;
proc sort data=n_child1; by h12_merkey; run;
proc sort data=n_child2; by h12_merkey; run;
proc sort data=n_child3; by h12_merkey; run;
proc sort data=n_child4; by h12_merkey; run;
proc sort data=n_child5; by h12_merkey; run;
proc sort data=n_child6; by h12_merkey; run;
proc sort data=n_child7; by h12_merkey; run;

data new_w12;
    merge father mother n_child1 n_child2 n_child3 n_child4 n_child5 n_child6 n_child7;
    by h12_merkey;
run;
```

```
/* 자녀 수 구하기 */
data new_w12_childsum;
    set new_w12;
    array c(7) ch1-ch7;
    childsum1 = sum(of c(*));
    childsum2 = sum(of ch1-ch7);
    childsum3 = sum(ch1, ch2, ch3, ch4, ch5, ch6, ch7);

    array x(3) childsum1-childsum3;
        do i=1 to 3;
        if x(i) = . then x(i) = 0;
        end;
    drop i;
run;

proc freq data=new_w12_childsum; tables childsum1 childsum2 childsum3 / missprint; run;
```

childsum1	빈도	백분율	누적 빈도	누적 백분율
0	1892	41.89	1892	41.89
1	1266	28.03	3158	69.91
2	1101	24.37	4259	94.29
3	225	4.98	4484	99.27
4	29	0.64	4513	99.91
5	4	0.09	4517	100.00

childsum2	빈도	백분율	누적 빈도	누적 백분율
0	1892	41.89	1892	41.89
1	1266	28.03	3158	69.91
2	1101	24.37	4259	94.29
3	225	4.98	4484	99.27
4	29	0.64	4513	99.91
5	4	0.09	4517	100.00

childsum3	빈도	백분율	누적 빈도	누적 백분율
0	1892	41.89	1892	41.89
1	1266	28.03	3158	69.91
2	1101	24.37	4259	94.29
3	225	4.98	4484	99.27
4	29	0.64	4513	99.91
5	4	0.09	4517	100.00

질문2. 15세 이상 가구원 수, 15세 미만 가구원 수 구하기

Q 2.

- 가구용 데이터 사용

```
/* 나이변수 생성(age1~age9) */
data h12_age;
    set h12_r;
    age1=2016-h1201_5;
    age2=2016-h1201_17;
    age3=2016-h1201_29;
    age4=2016-h1201_41;
    age5=2016-h1201_53;
    age6=2016-h1201_65;
    age7=2016-h1201_77;
    age8=2016-h1201_89;
    age9=2016-h1201_101;
    keep h12_id--h1201_110 age1-age9;
run;

proc means data=h12_age min max; var age1-age9; run;
```

변수	최솟값	최댓값
age1	18.0000000	97.0000000
age2	1.0000000	99.0000000
age3	0	109.0000000
age4	0	95.0000000
age5	0	94.0000000
age6	0	92.0000000
age7	1.0000000	93.0000000
age8	4.0000000	21.0000000
age9	2.0000000	2.0000000

질문2. 15세 이상 가구원 수, 15세 미만 가구원 수 구하기

Q 2.

```
/* 방법1 */
/* 15세 이상 가구원 수 */
data h12_age15_up;
    set h12_age;
    array x[9] age1-age9;
    array y[9] age15_u1-age15_u9;
        do i=1 to 9;
            if 15 <= x[i] <= 109 then y[i]=1;
            else y[i]=0;
        end;
    drop i ;
run;

data h12_age15_up_sum;
    set h12_age15_up;
    age15_usum=sum(of age15_u1-age15_u9);
run;

proc freq data=h12_age15_up_sum; tables age15_usum; run;

/* 15세 미만 가구원 수 */
data h12_age15_down;
    set h12_age;
    array x[9] age1-age9;
    array y[9] age15_d1-age15_d9;
        do i=1 to 9;
            if 0 <= x[i] < 15 then y[i]=1;
            else y[i]=0;
        end;
    drop i;
run;

data h12_age15_down_sum;
    set h12_age15_down;
    age15_dsum=sum(of age15_d1-age15_d9);
run;

proc freq data=h12_age15_down_sum; tables age15_dsum; run;
```

```
/* 방법2 */
data age15_up_down;
    set h12_age;
    /* 15세 이상 */
    if 15 <= age1 <=109 then age15_up1=1; else age15_up1=0;
    if 15 <= age2 <=109 then age15_up2=1; else age15_up2=0;
    if 15 <= age3 <=109 then age15_up3=1; else age15_up3=0;
    if 15 <= age4 <=109 then age15_up4=1; else age15_up4=0;
    if 15 <= age5 <=109 then age15_up5=1; else age15_up5=0;
    if 15 <= age6 <=109 then age15_up6=1; else age15_up6=0;
    if 15 <= age7 <=109 then age15_up7=1; else age15_up7=0;
    if 15 <= age8 <=109 then age15_up8=1; else age15_up8=0;
    if 15 <= age9 <=109 then age15_up9=1; else age15_up9=0;
    age15_up_sum=sum(of age15_up1-age15_up9);
    /* 15세 미만 */
    if 0 <= age1 < 15 then age15_down1=1; else age15_down1=0;
    if 0 <= age2 < 15 then age15_down2=1; else age15_down2=0;
    if 0 <= age3 < 15 then age15_down3=1; else age15_down3=0;
    if 0 <= age4 < 15 then age15_down4=1; else age15_down4=0;
    if 0 <= age5 < 15 then age15_down5=1; else age15_down5=0;
    if 0 <= age6 < 15 then age15_down6=1; else age15_down6=0;
    if 0 <= age7 < 15 then age15_down7=1; else age15_down7=0;
    if 0 <= age8 < 15 then age15_down8=1; else age15_down8=0;
    if 0 <= age9 < 15 then age15_down9=1; else age15_down9=0;
    age15_down_sum=sum(of age15_down1-age15_down9);
run;

proc freq data=age15_up_down; tables age15_up_sum age15_down_sum; run;
```

질문2. 15세 이상 가구원 수, 15세 미만 가구원 수 구하기

Q 2.

- 머지용 데이터 사용

```
/* 나이변수 생성(age) */
data hpda12_age;
    set hpda12_r;
    age = 2016-h12_g4;
    if 15 <= age <= 109 then age15_up=1;
    else age15_up=0;

    if 0 <= age < 15 then age15_down=1;
    else age15_down=0;
    keep h12_id--h1201_110 age age15_up age15_down;
run;
```

```
/* 15세 이상 */
proc sort data=hpda12_age; by h12_merkey h12_pid; run;
```

```
data hpda12_age15_up;
    set hpda12_age;
    by h12_merkey;
    if first.h12_merkey then age15_up_sum=0;
        age15_up_sum+age15_up;
    if last.h12_merkey;
run;
```

```
proc freq data=hpda12_age15_up; tables age15_up_sum; run;
```

```
/* 15세 미만 */
proc sort data=hpda12_age; by h12_merkey h12_pid; run;
```

```
data hpda12_age15_down;
    set hpda12_age;
    by h12_merkey;
    if first.h12_merkey then age15_down_sum=0;
        age15_down_sum+age15_down;
    if last.h12_merkey;
run;
```

```
proc freq data=hpda12_age15_down; tables age15_down_sum; run;
```

질문2. 15세 이상 가구원 수, 15세 미만 가구원 수 구하기

Q 2.

- 가구용 방법1 결과

age15_up_sum	빈도	백분율	누적 빈도	누적 백분율
1	2072	31.48	2072	31.48
2	2870	43.61	4942	75.09
3	1001	15.21	5943	90.31
4	515	7.83	6458	98.13
5	105	1.60	6563	99.73
6	17	0.26	6580	99.98
8	1	0.02	6581	100.00

age15_dsum	빈도	백분율	누적 빈도	누적 백분율
0	5398	82.02	5398	82.02
1	562	8.54	5960	90.56
2	525	7.98	6485	98.54
3	86	1.31	6571	99.85
4	8	0.12	6579	99.97
5	2	0.03	6581	100.00

- 가구용 방법2 결과

age15_up_sum	빈도	백분율	누적 빈도	누적 백분율
1	2072	31.48	2072	31.48
2	2870	43.61	4942	75.09
3	1001	15.21	5943	90.31
4	515	7.83	6458	98.13
5	105	1.60	6563	99.73
6	17	0.26	6580	99.98
8	1	0.02	6581	100.00

age15_down_sum	빈도	백분율	누적 빈도	누적 백분율
0	5398	82.02	5398	82.02
1	562	8.54	5960	90.56
2	525	7.98	6485	98.54
3	86	1.31	6571	99.85
4	8	0.12	6579	99.97
5	2	0.03	6581	100.00

- 머지용 방법1 결과

age15_up_sum	빈도	백분율	누적 빈도	누적 백분율
1	2072	31.48	2072	31.48
2	2870	43.61	4942	75.09
3	1001	15.21	5943	90.31
4	515	7.83	6458	98.13
5	105	1.60	6563	99.73
6	17	0.26	6580	99.98
8	1	0.02	6581	100.00

age15_down_sum	빈도	백분율	누적 빈도	누적 백분율
0	5398	82.02	5398	82.02
1	562	8.54	5960	90.56
2	525	7.98	6485	98.54
3	86	1.31	6571	99.85
4	8	0.12	6579	99.97
5	2	0.03	6581	100.00

질문3. 필요한 변수만을 남기고 싶을 때

Q 3.

```
/* keep, drop 외에는 방법이 없음 */
/* 다만 sas에서는 proc sql을 사용하여 데이터의 변수 순서까지 변경이 가능 */
proc sql;
    create table hpda12_new as
    select h12_pid, h12_pind, h12_merkey, h_new, h12_cobf, p12_wgl, p12_wsl, p12_wgc, p12_wsc, h12_reg5, h12_reg7, h12_cin,
           h12_din, h12_hc, nh1201_1, nh1201_2, h1201_1, h1201_110, h12_g1, h12_g2, h12_g3, h12_g4, h12_g6, h12_g7, h12_g8,
           h12_g9, h12_g10, h12_g11, h12_g12
    from w12.Koweps_hpda12_2017_beta1;
quit;
```

```
/* SPSS 변환 */
PROC EXPORT DATA= w12.Koweps_h12_2017_beta1 /* 변환 하고자 하는 data명 */
            OUTFILE= "D:₩Kihasa₩Koweps₩panel data(1~12)₩2017년 12차 한국복지패널조사(koweps) 데이터 및 조사설계서(beta1)₩
                     (2017년 12차 한국복지패널조사) 데이터(beta1)_spss₩Koweps_h12_2017_beta1"
            /* "저장하고자 하는 경로 및 데이터명 지정" */
            DBMS=SPSS REPLACE;
            /* 변환하고자 하는 패키지 */
RUN;
```

```
/* STATA 변환 */
PROC EXPORT DATA= w12.Koweps_h12_2017_beta1
            OUTFILE= "D:₩Kihasa₩Koweps₩panel data(1~12)₩2017년 12차 한국복지패널조사(koweps) 데이터 및 조사설계서(beta1)₩
                     (2017년 12차 한국복지패널조사) 데이터(beta1)_stata₩Koweps_h12_2017_beta1"
            DBMS=STATA REPLACE;
RUN;
```

질문4. 가중치 적용

Q 4.

□ 가중치란?

○ 모집단에서 추출확률 p로 표본을 추출했다면 w=1/p만큼 대표하도록 함.

- 예) 1/100로 표본을 추출한 경우 1개의 단위가 모집단에 있는 100개의 단위를 대표함.

□ 가중치에 고려된 사항

○ 불균등확률 추출에 대한 보정

○ 단위무응답(unit nonreponse)에 대한 보정

○ 특정 변수(성별, 연령 등)에 대한 가중표본의 사후 조정(모집단 정보를 알 때)

- 표본추정치의 정도 개선
- 포괄성 및 무응답 조정

☞ **표본의 모집단에 대한 대표성(representativeness) 및 포괄성(coverage) 확보를 통한 편향 감소 효과**

질문4. 가중치 적용

□ 가구조사(Household Survey)

○ 복합표본추출(complex sampling) 또는 2단계 층화 집락추출(two-stage stratified cluster sampling)

- 1단계 추출단위(PSU: primary sampling units) : 조사구(구획 또는 블록)
- 2단계 추출단위(SSU : secondary sampling units): 가구

☞ 1단계 : 50개중 20개 조사구를 임의로 추출한 경우로 20개 조사구는 50/20=2.5의 가중치를 부여받음.

☞ 2단계 : 조사구당 8가구 중에서 2가구씩을 추출함으로 표본조사구내 2가구는 8/2=4로 조사구당 4가구를 대표함.

☞ 결과적으로 조사구추출확률의 역수: 2.5*조사구내 가구추출확률의 역수: 4=10으로서 표본가구당 10가구씩을 대표하게 됨.

☞ 모집단 총 가구수 ; 400 가구 → 20(표본 PSU)*2(표본 SSU)*10(=W)=400

질문4. 가중치 적용

Q 4.

□ 종단가중치

○ 종단 가중치(Longitudinal weight)는 조사시점을 기준으로 조사차수별로 발생하는 무응답자에 대해서 무응답 보정 가중치를 연속적으로 적용하는 것으로, 차수 간 순수변동의 비교분석을 목적으로 사용

○ 패널탈락으로 인한 무응답 보정을 위해, t차웨이브 종단면 응답여부 변수와 t-1차 웨이브 변수들의 관계를 로지스틱회귀모형을 이용하여 응답확률을 추정

- 로지스틱회귀모형의 설명변수로 응답자의성별, 연령, 지역, 경제활동상태 사용
- t차 웨이브 종단면 응답여부 변수는 t차와 t-1차 웨이브 모두 응답한 경우1, 그렇지 않은 경우는 0을 부여

○ t차 웨이브의 개인 종단면 기본가중치에 로지스틱회귀모형으로 추정된 응답확률의 역수를 곱하여, t차 웨이브의 개인 종단면 가중치를 조정

○ 개인별 변동상황에 따라 t차년도 종단면 가중치를 조정
(신규가구원의 경우 개인별종단면가중치 0값 존재)

질문4. 가중치 적용

□ STATA Weight의 종류

가중치 종류	설명	예시
Frequency Weight : *fw*	• 도수 분포표 상의 각 계급별 도수를 의미(반드시 정수) • 빈도표 작성후 각 계급별 도수를 가중치로 적용함.	regress y x1 x2 [fweight=pop]
Sampling Weight or Probability Weight :*pw*	• 복합표본추출과정에서의 추출확률을 가중치로 적용 • svy 명령문에서 주로 사용됨. • 패널 가중치는 pw 임.	regress y x1 x2 [pweight=pop]
Analytical Weight : *aw*	• 분산 안정화를 목적으로 분산의 역수를 가중치로 적용	regress y x [aweight=pop]
Importance Weight : *iw*	• 특정효과나 통제를 목적으로 활용	

질문4. 가중치 적용

Q 4.

□ 통계분석 프로그램별 가중값 적용여부 비교

통계패키지	분석프로그램	가중값 적용여부
SAS	• PROC GENMOD • PROC GLM • PROC MIXED	종단 가중값 적용가능
STATA	• xtgls • xtreg • xtmixed	가중값 적용 불가능 (frequency weight 허용)
SPSS	• Mixed-Linear	residual weight 적용 가능
HLM	• Multilevel Model	weight 적용 가능

질문5. 무응답 대체방법

□ 무응답의 의미

자료를 수집하는 과정에서 일부 항목이 측정되지 않으면 그 항목에 대한 응답이 발생하지 않았다는 의미로 무응답(nonresponse)이 발생했다고 하거나,
또는 그 항목의 값이 관측되지 않고 빠져있다는 의미로 결측값(missing value)

□ 여러 가지 무응답 분석 방법

1. 완전히 응답한 개체를 이용한 분석 (Complete-case Analysis)
2. 가중값 보정방법 (Weighting Adjustment)
3. 이용 가능한 개체 분석 (Available-case Analysis)
4. **대체방법 (Imputation Methods)**
5. 우도함수(likelihood function)를 근거로 한 무응답 자료 분석법

질문5. 무응답 대체방법

□ 단일 대체(single imputation)

대체는 결측값의 예측분포(predictive distribution)의 평균 또는 추출값(draw)을 사용

응답자료로부터 무응답의 예측분포를 만드는 방법이 필요

일반적으로 무응답의 예측분포는 명백한 모형 또는 함축적인 모형을 통하여 만들어짐.

1. 명백한 모형을 근거로 한 대체방법

(a) 평균 대체: 응답자의 평균을 이용하여 무응답값을 대체하는 방법

(b) 회귀 대체: 회귀식을 응답자의 자료로 추정한 후 무응답값을 적합된 회귀식으로부터 예측하여 대체하는 방법

(c) 확률적 회귀 대체: 위의 회귀대체에서 설명한 회귀예측값에 예측값의 불확실성을 고려하는 잔차를 더하여 무응답을 대체하는 방법

질문5. 무응답 대체방법

Q 5.

□ 단일 대체(single imputation)

대체는 결측값의 예측분포(predictive distribution)의 평균 또는 추출값(draw)을 사용

응답자료로부터 무응답의 예측분포를 만드는 방법이 필요

일반적으로 무응답의 예측분포는 명백한 모형 또는 함축적인 모형을 통하여 만들어짐.

2. 함축적 모형을 근거로 한 대체방법

(a) 핫덱 대체(hot deck imputation): 무응답을 현재 진행 중인 연구에서 “비슷한” 성향을 가진 응답자의 자료로 대체하는 방법(표본조사에서 흔히 사용)

(b) 콜드덱 대체: 핫덱과 비슷하나 대체할 자료를 현재 진행 중인 연구에서 얻는 것이 아니라 외부출처 또는 이전의 비슷한 연구에서 가져오는 방법

(c) 혼합방법: 몇 가지 다른 방법을 혼합하는 방법(예를들어, 회귀대체를 이용하여 예측값을 얻고 핫덱방법을 이용하여 잔차를 얻어 두 값을 더하는 경우 등)

질문5. 무응답 대체방법

□ **패널조사 항목 무응답 대체**

패널조사의 최대 장점은 패널에 대하여 **반복적으로 동일한 설문이 측정**되므로 특정 시점에서 무응답이 발생한다면 이 패널의 **이전 시점 조사 자료를 활용**하여 대체를 실시한다면 보다 정확한 대체를 실시할 수 있음.

(a) 평균대체(mean imputation)

- 응답한 개체들만을 가지고 구한 평균값으로 무응답을 대체하는 방법

(b) 핫덱대체(hot deck imputation)

- 응답자의 값으로 무응답을 대체하는 방법
- 일반적으로 무응답과 특성이 비슷한 개체를 찾기 위해서 여러 관련 변수들을 가지고 대체군을 형성한 다음에, 그 대체군 내에서 기증자를 무작위로 선택하여 기증자의 응답값으로 대체를 실시

질문5. 무응답 대체방법

□ **패널조사 항목 무응답 대체**

(c) 비대체(ratio imputation)

- 소득 활동과 같이 매년 일정 비율로 증가하는 변수의 경우 직전 시점의 응답값에 일정한 비율을 곱하여 대체를 실시하는 방법
- 비율은 대체 대상 시점과 이전 시점이 모두 측정된 패널들을 대상으로 추정하는 것이 일반적임

(d) 중위수대체(median imputation)

- 무응답이 발생 빈도가 매우 낮은 경우 프로그램화하기 상대적으로 복잡한 핫덱대체 대신 중위수(median)를 가지고 대체하는 방법

질문5. 무응답 대체방법

□ 패널조사 항목 무응답 대체

(e) 회귀대체(regression imputation)

- 종단면 자료의 대체에 흔히 사용되는 회귀대체를 패널 자료에 적용하는 것이 가능
- 특정 시점의 특정 항목의 무응답을 예측하기 위하여 이전 시점의 동일 항목의 값 외에 다른 연관된 변수*들을 포함하여 회귀식을 형성하고 예측값에 근거하여 대체를 실시하는 방법

  * 설명력(R^2)이 높은 변수를 선택하기 위한 변수 선택 단계를 거쳐 대체 모형을 결정

- 비대체는 회귀대체의 특별한 예로 볼 수 있음.
- 실제 자료는 회귀선(regression line)에 존재하지 않고 오차를 가지고 측정되므로 이를 반영하는 확률적 회귀대체를 실시하는 것이 바람직한 것으로 알려져 있음 (Little and Rubin, 2002)

질문5. 무응답 대체방법

□ 패널조사 항목 무응답 대체

(f) LOCF(Last Observation Carried Forward) 방법

- 특정 조사 시점에서 무응답이 발생한 경우 해당 패널의 이전 시점의 조사 자료들 중 가장 최근에 조사된 응답값을 가지고 무응답을 대체하는 방법
- 임상실험의 결측값을 대체하기 위해 흔히 사용되는 방법
- 시간이 지남에 따라 응답값이 변화(예를 들어, 소득, 지출 등)하는 경우, 편향을 가져올 수 있다는 단점을 지님

이 외, 논리적 대체(logical imputation), 최빈값 대체(mode imputation), 확률에 근거한 대체(imputation based on the probability), 임의대체(random imputation) 등이 있음.

질문5. 무응답 대체방법

□ **(참고) 국외 노동 관련 패널조사의 무응답 대체**

○ 독일 German Socio-Economic Panel (GSOEP)에서는 무응답률이 낮은 변수는 **평균대체**를 사용(Frick·Grabka, 2004)

○ 미국 Survey of Income and Program Participation (SIPP)은 지난 차수 자료를 이용하여 유사한 특성을 가진 기증자의 값으로 대체하는 순차적인 **핫덱대체** (sequential hot deck imputation)를 사용(U.S. Census Bureau, 2016).

○ 캐나다 Survey of Labor and Income Dynamics(SLID)에서도 소득의 횡단면 대체에서는 **최근접이웃 핫덱대체**(nearest neighbor hotdeck imputation)를 사용하는데 매칭(matching)을 위한 변수들을 설정하여 최근접이웃을 찾아낸 후 이 응답자의 값으로 대체를 실시

질문5. 무응답 대체방법

□ (참고) 국외 노동 관련 패널조사의 무응답 대체

○ 영국 British Household Panel Survey (BHPS)에서는 범주형 소득관련 변수를 대체하는 방법 중 하나로 **핫덱대체**를 사용

○ 또한 무응답 대체 중에서 cross-wave imputation 방법이 있는데 대체하고자 하는 변수의 지난 차수의 값에 랜덤하게 선택한 비슷한 특성을 가지는 케이스의 변화율을 적용하는 방법으로 **비대체**의 일종이라고 볼 수 있음(Taylor 등, 2010)

○ 미국 Panel Study of Income Dynamics (PSID)에서 사용된 대체 방법 중에서 **전년도 자료 이월(carryover) 방법**은 지난해의 소득이 존재하는 경우 전년도 자료에 소득 증가분을 고려하여 이월하는 방법으로 비대체라고 볼 수 있음(Duffy, 2011).

참고문헌

손창균(2016). 복지패널분석에서 가중치 활용. 제9회 한국복지패널 데이터설명회 특강자료.

송주원·안형진(2009). 무응답 자료 처리 및 분석. 통계청 통계교육원.

이혜정·송주원(2017). 패널자료에서의 항목무응답 대체 방법 비교. 응용통계연구 30(3), pp. 377-390

Duffy, D. (2011). 2007 PSID Income and Wage Imputation Methodology, Survey Research Center- Institute for Social Research Technical Series Paper #11-03, University of Michigan, Michigan.

Frick, J. R. and Grabka, M. M. (2004). Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income Distribution, DIW Discussion Papers No. 376, DIW Berlin.

Little, R. J. A. and Rubin, D. B. (2002), Statistical Analysis With Missing Data , Wiley: New York.

Taylor, M. F., Brice, J., Buck, N., and Prentice-Lane, E. (2010). British Household Panel Survey User Manual Volume A-Introduction (Technical Report and Appendices), University Essex, Colchester.

U.S. Census Bureau (2016). Survey of Income and Program Participation 2014 Panel Users' Guide, U.S. Department of Commerce Economic and Statistics Administration U.S. Census Bureau.

감사합니다

KiHASA
한국보건사회연구원
Korea Institute for Health and Social Affairs