패널 데이터 품질 개선을 위한 항목무응답 대체 방법[1)]

Imputation of Item Nonreponse for Quality Improvement in Panel Data

이혜정 | 한국보건사회연구원 부연구위원

수집한 패널 데이터의 사후적인 품질 관리 방법으로는 데이터 정제를 통한 정합성 관리, 패널이탈에 대한 지속적인 검토를 통한 표본 대표성 확보, 가중치 작성, 항목무응답에 대한 처리 등이 있다. 이 글에서는 한국복지패널조사 및 한국의료패널조사에서의 항목무응답 대체 방법을 소개하고 최신 통계 방법인 기계학습 통계기법을 기반으로 한 항목무응답 대체 방법의 효과를 파악하여 적절한 대체 방법을 제안하고자 한다.

이에 대한 종합적인 결과와 활용 방안은 세 가지로 요약할 수 있다. 첫째, 패널 자료에서의 항목무응답 대체 시 기계학습 통계기법을 적용할 수 있는 가능성을 확인하였다. 특히 랜덤 포레스트 대체 방법은 편향뿐만 아니라 다른 평가지표도 우수한 결과를 보여 실무에서 활용해 볼 수 있다고 생각한다. 둘째, 대체군 활용 여부에 따라 대체 효과가 확연히 달라지는 결과를 통해 무응답 대체 시 대체군 활용의 중요성을 다시 한번 확인하였다. 셋째, 보조변수로 활용하는 설명변수 개수 증가에 따른 대체 효과를 확인하였다. 복잡하고 포괄적인 모형보다는 무응답 대체 대상 변수와 연관성이 큰 설명변수를 탐색하고 선정하는 것이 무응답 대체 효과 향상에 효과적이라고 생각한다. 바람직한 대체 방법은 통계적 추론에서 발생할 수 있는 무응답 편의가 감소하고 모집단 분포로부터 표본 분포가 왜곡되지 않고 비슷하게 유지될 수 있어야 한다. 이 점을 인지한다면 더욱 정확하고 신뢰성 있는 무응답 대체 자료를 제공할 수 있을 것이다.

1) 이 글은 이혜정, 지희정, 이지혜. (2019).『보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로』에 수록된 내용을 바탕으로 재구성하여 작성하였다.

1. 들어가며

패널 데이터는 동일한 개체를 시간에 따라 반복적으로 측정하여 수집한 자료로, 개체 간 차이 및 시간에 따른 개체 변화를 고려한 동태적 분석이 가능하다. 현재 국내 패널조사는 21개로 규모가 작지 않다. 패널 자료 활용도 측면에서 보면, 한국복지패널조사를 활용한 연구 결과물은 2008년 30편에서 2016년 148편으로 5배 정도 대폭 증가하였다(이현주 외, 2017, p. 143). 이는 품질 높은 패널 데이터를 생산하고 있다는 결과이며 앞으로도 지속적으로 유지해 나가야 할 방증으로 볼 수 있다. 수집한 패널 데이터의 사후적인 품질 관리 방법으로는 데이터 정제(data cleaning)를 통한 정합성 관리, 패널이탈(attrition)에 대한 지속적인 검토를 통한 표본 대표성 확보, 가중치(weighting) 작성, 항목무응답에 대한 처리 등이 있다.

이 글에서는 패널 데이터 품질 관리 중에서 항목무응답을 처리하는 방법에 대해 살펴보고자 한다. 항목무응답은 설문조사에서 응답자가 일부 문항에 대해 정확하게 응답하지 않고 모름이나 응답 거절을 선택한 것을 의미한다. 패널조사에서도 항목무응답이 발생하는데, 초기 조사 차수에서는 응답자와 면접자 간 유대감 형성 부족, 이후 조사 차수에서는 응답자의 패널 피로도 증가가 원인인 것으로 볼 수 있다. 이렇게 수집된 패널 데이터를 사용하여 통계 분석을 하면 추정 시 편향이 발생하고, 목표 모집단에 대한 추정 및 검정에서의 오류로 인해 잘못된 연구 결과를 도출하게 된다. 이를 해결하기 위해서는 최적의 대체 방법을 이용하여 적절한 값으로 채워 주는 항목무응답 대체를 실시하는 것이 바람직하다고 볼 수 있다.

한국보건사회연구원에서 공동으로 주관하고 있는 한국복지패널조사 및 한국의료패널조사의 항목무응답 대체 방법을 소개하고, 최신 통계기법인 기계학습 통계기법을 패널 데이터의 항목무응답 대체에 적용해 본 다음 기존의 대체 방법과 비교하여 활용 가능성을 가늠해 보고자 한다.

2. 보건복지 분야 패널조사의 항목무응답 대체 방법 소개

현재 국내 패널조사 중에서 한국복지패널, 한국의료패널, 한국노동패널, 한국고령화패널은 항목무응답 대체를 하고 있다. 한국복지패널조사 및 한국의료패널조사는 항목무응답 비율이 높지 않으며, 주요 변수에 한해 대체를 실시하여 대체 대상 변수 개수가 많지 않은 편이다. 2개 패널 데이터에서의 항목무응답 처리 과정을 살펴보면 다음과 같다.

가. 한국복지패널조사

한국복지패널조사는 2006년 1차 원표본 7072가구를 구축하였으며, 2012년 신규 표본 1800가구를 추가 구축하였다. 14차 조사(2019년)에서는 총 6612가구를 대상으로 하여 6331가구에

대한 조사를 완료하였다(여유진 외, 2019, p. 23).

한국복지패널조사는 가구의 경상소득 및 가처분소득에 대해 항목무응답 대체를 실시하고 있다(한국보건사회연구원, 2019, p. 94). 해당 가구의 소득변수가 결측인 경우 가구 구분(일반·저소득층 가구)을 할 수 없기 때문에 필수적으로 대체하고 있다. 가구의 경상소득 및 가처분소득에 대한 무응답 비율은 2차 1.29%(226개), 3차 0.31%(52개), 6차 0.04%(6개)로 매우 낮은 편이며, 이를 제외한 나머지 조사 차수에서는 무응답이 없는 것으로 나타났다. 무응답을 포함하는 2, 3, 6차는 무응답이 모두 대체값으로 대체되어 있다.

가구의 경상소득 및 가처분소득에 대한 대체 방법은 다음과 같다. 무응답 변수와 연관성이 있는 변수를 선택한 후 모형 적합(model fitting) 과정을 거쳐 다중대체를 실시하는 것이다. 보통 기본 5개의 대체 자료 세트가 생성되는데, 이용자의 편의를 위하여 5개의 대체 자료 세트 중에서 소득변수를 가장 적절하게 설명하는 대체 자료 세트 1개를 선택한다(김미곤 외, 2008, pp. 119-120). 선택된 대체 자료 세트의 무응답 대체 변수를 배포용 자료에 붙여 제공한다. 이때 무응답 대체 변수에 대한 대체 적용 여부를 나타내는 플래그(flag) 변수가 포함되어 있어 응답값과 대체값을 구분할 수 있다.

나. 한국의료패널조사

한국의료패널조사는 2008년에 1차 조사를 시작한 후 2019년 14차 조사를 마지막으로 12년간 제1기 조사를 운영하였다(박은자 외, 2019, p. 23). 2008년 1차 원표본 7866가구를 구축하고 2012년 신규 표본 2222가구를 추가 구축하였다. 마지막 14차 조사(2019년)에서는 6493가구에 대한 조사를 완료하였다(박은자 외, 2019, pp. 23-24).

한국의료패널조사의 항목무응답 대체 변수는 보건의료서비스의 경우 의료비, 교통비, 요양비, 간병비 등이고 의약품의 경우 처방약값, 의약품 구매 등이며 보건의료용품 및 기구의 경우 약국에서의 일반 의약외품 구매, 의료기기 구매·임대·수리 등이다. 이 변수들의 무응답 비율은 현저히 낮은 편이다. 이 변수들의 무응답 대체는 의료비 생성 변수를 위한 항목무응답 처리라고 볼 수 있으며, 의료비 생성 변수는 모두 값을 가진다. 의료비 생성 변수는 5개의 가구 지출 의료비와 2개의 개인 지출 의료비이며, 이를 배포용 자료에 포함하여 제공하고 있다. 무응답 대체 방법은 주로 평균대체이며 필요에 따라 0 또는 결측치 처리를 한다.

3. 보건복지 분야 패널조사의 항목무응답 대체 방법 모의실험

항목무응답 변수를 대체하기 위해서는 다양한 대체 방법 중에서 적절한 방법을 선정해야 한다. 대체 방법 선정 전에 대체 대상 변수의 설문 구조, 무응답 비율과 분포 특성, 활용 가능한 보조변수 정보 등을 두루 검토하는 과정이 우선되어

야 한다. 이 장에서는 무응답 비율에 따른 보건복지 분야 패널 데이터의 항목무응답 대체 방법의 효과를 평가하고자 한다. 무응답 비율은 5%, 10%, 20%, 30%, 50%로 다양하게 구성한다. 한국복지패널 및 한국의료패널 자료의 경우 무응답 비율이 낮은 편이어서 실제 상황을 반영하기 위해 5%로 하였고, 점진적으로 무응답 비율을 높여 이에 따른 무응답 대체 방법의 효과를 모의실험을 통해 살펴보고자 한다. 고려한 대체 방법은 최근 다양한 분야에서 사용되고 있는 기계학습 통계기법을 기반으로 하여 항목무응답 대체에 적용함으로써 활용 가능성을 가늠해 보고자 한다. 항목무응답 대체 시 많이 활용하는 평균, 핫덱, 비대체 방법도 함께 비교한다. 다양한 패널조사에 따른 무응답 대체 효과도 파악하기 위해 2개의 패널 데이터(한국복지패널, 한국의료패널)를 사용하여 모의실험을 하고자 한다. 먼저 한국복지패널 데이터를 이용한 모의실험을 살펴본다. 내용은 모의실험의 자료 설계 과정, 적용한 대체 방법의 종류, 대체의 효과를 판단할 때 사용한 평가지표, 모의실험 결과로 구성한다.

가. 한국복지패널 데이터를 이용한 모의실험

1) 자료 설계

모의실험을 위한 자료로는 2019년에 배포된 한국복지패널 12~13차(2017~2018년) 자료를 사용하였고, 13차 조사 때 조사한 항목인 '작년 한 해 연간 경상소득(이하 경상소득)'을 항목무응답 대체 대상 변수로 정하였다. 경상소득은 실제로 무응답 대체를 하고 있는 변수이기도 하다. 연구 대상 모집단은 6360가구로, 12~13차 경상소득 및 항목무응답 대체 시 설명변수로 사용하는 변수에 대해 모두 응답한 가구만을 대상으로 한정하였다.

결측 자료 메커니즘이 임의결측(MAR: Missing At Random)을 따른다는 가정하에 경상소득과 연관성이 있는 '가구주의 배우자 유무'와 '가구주의 학력'을 가지고 4개의 대체군 집단(이하 집단)으로 나누었다(표 1). 집단별 경상소득 평균은 집단 1이 1615만 원, 집단 2가 3359만 원, 집단 3이 3156만 원, 집단 4가 7054만 원이었다. 유의수준 5%하에서 (집단 1), (집단 2, 집단 3), (집단 4)로 묶이는 것으로 나타났다.

배우자가 있고 학력이 고졸 이상인 경우 무응답 가능성이 높다는 점을 반영하여 집단별 무응

표 1. 가구주의 배우자 유무 및 학력에 따른 집단 구분(한국복지패널)

	고졸 미만	고졸 이상
배우자 없음	집단 1	집단 2
배우자 있음	집단 3	집단 4

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 84 〈표 4-1〉.

표 2. 집단별 무응답 발생 개수의 비율(한국복지패널)

(단위: %)

	고졸 미만	고졸 이상
배우자 없음	10	18
배우자 있음	22	50

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 85 〈표 4-2〉.

답 발생 개수의 비율을 다르게 구성하였다(표 2). 예를 들어 무응답 비율이 10%라고 가정할 때 집단 1은 집단 1 전체 개수의 1%(=10×0.1), 집단 2는 집단 2 전체 개수의 1.8%(=10×0.18), 집단 3은 집단 3 전체 개수의 2.2%(=10×0.22), 집단 4는 집단 4 전체 개수의 5%(=10×0.5)가 집단별로 발생시켜야 할 무응답 비율이다.

또, 경상소득은 매년 조사되어 무응답이 이전 차수의 경상소득에 의존하는 것이 가능하므로 경상소득이 높을수록 무응답이 많이 발생한다고 가정하였다. 이에 따라 4개 대체군 집단에서 중위수를 기준으로 중위수 미만에서는 집단별 무응답 가구 수의 30%, 중위수 이상에서는 70%의 무응답이 발생한다고 가정하였다. 임의결측 가정을 만족시키기 위해 12차 경상소득은 무응답 없이 모두 값을 가지도록 하고 13차 경상소득은 무응답으로 처리하였다. 이는 비대체에서 비(ratio)를 구할 때 비의 분모가 결측되지 않는 효과도 있다.

무응답 비율별 발생 개수를 살펴보면, 무응답 비율 5%는 318가구, 10%는 636가구, 20%는 1272가구, 30%는 1908가구, 50%는 3108가구이다(부록 참조).

연구 대상 모집단 자료에서 무응답 가구를 단순임의추출방법(simple random sampling)으로 추출하는 과정을 100번 반복하여 최종 모의실험용 무응답 자료 100개를 생성하였다.

2) 대체 방법

모의실험에서 사용한 대체 방법은 평균대체, 핫덱대체, 비대체, K-최근접 이웃 대체, 랜덤 포레스트 대체, 서포트 벡터 머신 대체, 신경망 대체로 총 7가지이다. 이 중에서 K-최근접 이웃 대체, 랜덤 포레스트 대체, 서포트 벡터 머신 대체, 신경망 대체 방법은 기계학습 통계기법에 기반한 것이다. 평균대체, 핫덱대체, 비대체는 대체군 활용 여부에 따라 2개의 대체값을 생성하였다. 대체 방법을 자세히 살펴보면 다음과 같다.

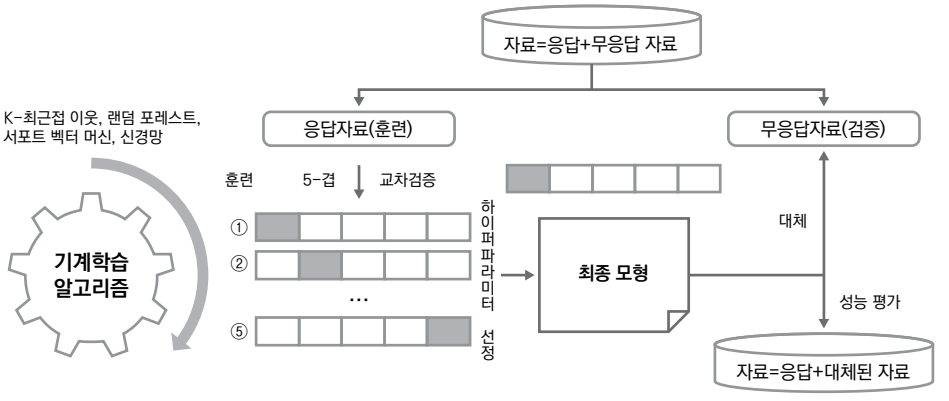
평균대체는 대체군을 사용하지 않는 경우 13차 경상소득에 대해 응답한 값만을 가지고 구한 평균값으로 무응답을 대체하였다. 대체군을 사용한 경우는 먼저 대체군을 형성하고, 각 대체군 내에서의 13차 경상소득 평균값으로 무응답을 대체하였다. 대체군 형성 시 고려한 변수는 학력 범주 2개(고졸미만·이상), 배우자 유무 범주 2개(배우자 있음·없음), 12차 경상소득 범주 2개(중위수 미만·이상)이며 총 8개로 구분하였다.

핫덱대체는 대체군을 사용하지 않는 경우 13차 경상소득에 대해 응답한 모든 값을 기증자 후보로 사용하여 기증자 중에서 무작위로 추출한 후 기증자의 응답값으로 무응답을 대체하였다. 대체군을 사용한 경우는 대체군 내에서 응답한 모든 값을 기증자 후보로 사용하여 기증자 중에서 무작위로 추출한 후 기증자의 응답값으로 무응답을 대체하였다. 이때 사용한 대체군은 평균대체와 동일하고, 무응답 대체 시 한 번 사용한 기증자는 이후 대체 시 사용하지 않는 것을 원칙으로 하였다. 단, 무응답 비율 30%와 50%의 경우에는 해당 대체군 내의 기증자 부족으로 무응답에 대해 제대로 채울 수 없어서 기증자를 중복으로 사용하였다.

비대체는 대체군을 사용하지 않는 경우 이전 차수와 해당 차수에 대해 모두 경상소득 응답값을 가지고 있는 개체만을 대상으로 하여 비를 계산한 후 비의 평균값을 구하였다. 무응답은 이전 차수의 응답값에 해당 비를 곱한 값으로 대체하였다. 대체군을 사용하는 경우 학력 및 배우자 유무 범주로 4개의 대체군을 생성한 후 대체군 내에서의 비의 평균을 계산하였다. 무응답은 이전 차수의 응답값에 해당 대체군 내의 비를 곱한 값으로 대체하였다.

마지막은 기계학습 통계기법에 기반한 대체 방법으로, [그림 1]은 대체 방법 절차를 나타낸다. 자료를 응답 자료와 무응답 자료로 구분한 후 응답 자료는 훈련 자료로, 무응답 자료는 검증 자료로 간주한다. 다음으로 최적의 하이퍼파라미터를 찾기 위해 훈련 자료에 기계학습 통계기법(K-최근접 이웃, 랜덤 포레스트, 서포트 벡터 머신, 신경망)을 적용한다. 이때 하이퍼파라미터 튜닝(parameter tuning)에는 5겹 교차검증(5-fold cross-validation) 또는 10겹 교차검증을 사용

그림 1. 기계학습 방법론에 기반한 대체 방법 절차



자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 75 [그림 3-4].

표 3. 무응답 대체 시 사용한 설명변수(한국복지패널)

설명변수	변수명
3개	가구주의 학력 및 배우자 유무, 경상소득(12차)
6개	가구주의 학력, 배우자 유무, 경상소득(12차), 생활비, 가구 내 근로자 수, 가구원 수

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 91 〈표 4-8〉.

한다. 교차검증은 하이퍼파라미터를 결정할 때 많이 활용하는 방법 중 하나이며, 이 모의실험에서는 5겹 교차검증을 사용한다. 하이퍼파라미터는 모형 설계 시 필요한 파라미터이며, 보통 훈련 전에 결정한다. 설정해야 할 하이퍼파라미터의 개수가 늘어남에 따라 복잡도가 증가하고 최적의 하이퍼파라미터를 찾는 것도 어려워진다. 5겹 교차검증을 사용하여 하이퍼파라미터를 결정하는 과정은 다음과 같다. 5겹 교차검증을 실시하기 위해 우선 훈련 자료를 5겹 자료로 나눈다. 첫 번째 시도에서 첫째부터 넷째 겹 자료를 가지고 적합 과정을 거친 다음에 다섯째 겹 자료로 모형을 평가한다. 두 번째 시도에서는 첫째부터 셋째, 다섯째 겹 자료로 적합 과정을 거친 후에 넷째 겹 자료로 모형을 평가한다. 이러한 방식으로 총 5번의 시도를 하여 5개 결과에 대한 평균을 계산한다. 설정한 하이퍼파라미터 조합에 대한 결과값 중에서 가장 작은 값을 가지는 하이퍼파라미터를 최종 선택한다. 선정된 하이퍼파라미터를 사용하여 전체 훈련 자료에 다시 적합 과정을 거쳐 최종 모형을 구축한다. 마지막으로 검증 자료에 최종 모형을 적용하여 예측값을 구하며, 무응답은 이 예측값으로 대체하여 최종 생성한다.

기계학습 통계기법에 기반한 대체 방법은 평균, 핫덱, 비대체 방법에서 활용한 보조변수 3개를 동일하게 사용하였다. 이는 무응답을 대체한 방법들의 효과를 비교할 때 동일한 조건을 만족시키기기 위해서이다. 또한 더 많은 정보를 사용하여 무응답을 대체한 경우에 대한 효과를 보기 위해 부가적인 모의실험을 하였으며, 기존의 보조변수 3개에서 6개로 늘려 결과를 살펴보았다(표 3).

3) 평가지표

각 대체 방법에 따라 생성한 대체값의 효과를 5가지 평가지표로 평가하였다. 5가지 평가지표는 편향(bias), 제곱근평균제곱오차(RMSE: Root Mean Square Error), 포함률(coverage rate), 예측의 정확성, 그리고 추정의 정확성이다. 예측의 정확성에 대해서는 절대거리를 사용하는 평균절대편차(MAD: Mean Absolute Deviation)와 제곱거리를 사용하는 제곱근평균제곱편차(RMSD: Root Mean Square Deviation)를 구하였다. 추정의 정확성은 상대평균의 절대 차이(Absolute relative difference in means), 변동계수의 절대차이(Absolute difference in the coefficient of variation), 표준화된 3차 적률(왜도)의 절대차이, 표준화된 4차 적률(첨도)의 절대 차이를 사

표 4. 항목무응답 대체 방법 평가 기준

평가지표	표기	설명
편향	BIAS	평균 추정치와 참값의 일치 정도
제곱근평균제곱오차	RMSE	평균 추정치를 얼마나 일관되게 추정하는지 평가
포함률	COV	평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률
예측의 정확성	-	개별 개체의 대체값과 실제값 간의 유사 정도
추정의 정확성	-	대체된 자료의 적률이 실제 자료의 적률을 잘 유지하는지 평가

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 요약 5 〈요약표 1〉.

용하여 평가하였다.

대체 방법의 평가지표에 대한 표기와 설명은 〈표 4〉와 같다. 평가지표를 간단하게 표기하면 'BIAS'는 편향, 'RMSE'는 제곱근평균제곱오차, COV는 참값의 포함률이다. 예측의 정확성 지표의 경우 'MAD'는 평균절대편차, 'RMSD'는 제곱근평균제곱편차이고 추정의 정확성 지표의 경우 'ADB'는 상대평균의 절대 차이-, 'ADV'는 변동계수의 절대 차이, 'ADS'는 표준화된 왜도의 절대 차이, 'ADK'는 표준화된 첨도의 절대 차이를 의미한다.

포함률을 제외한 나머지 평가지표(편향, 제곱근평균제곱오차, 예측의 정확성 및 추정의 정확성)의 값은 작을수록 우수한 효과를 가지는 대체 방법으로 볼 수 있다.

4) 모의실험 결과

무응답 비율별 대체 방법에 따라 생성한 대체값에 대한 5가지 평가지표를 계산하였을 뿐만 아니라 히스토그램 및 상자그림을 통해 대체된 자료와 실제 자료 간 분포 비교도 하였다(이혜정, 지희정, 이지혜, 2019, pp. 92-114). 이를 바탕으로 각각의 무응답 비율에서 평가지표별로 가장 우수한 효과를 보이는 대체 방법에 대해 해당 무응답 비율을 표기하여 모의실험 결과를 〈표 5〉와 같이 정리하였다. 대체 방법의 표기는 다음과 같이 정의하였다. 'mean'은 평균대체, 'hotdeck'은 핫덱대체, 'ratio'는 비대체, 'knn'은 K-최근접 이웃 기반 대체, 'rf'는 랜덤 포레스트 기반 대체, 'svm'은 서포트 벡터 머신 기반 대체, 'mlp'는 인공 신경망 기반 대체 방법이다. 설명변수의 개수에 따라 대체 방법을 구분하기 위해, knn을 예로 들면 설명변수가 3개인 경우 knn_3으로, 6개인 경우에는 knn_6으로 표기하였다. 나머지 대체 방법도 동일하게 적용하였다.

3개의 설명변수를 사용한 랜덤 포레스트 대체 방법(rf_3)의 효과가 가장 우수한 것으로 나타났다. 무응답 비율과 상관없이 경상소득의 평균 추정량은 참값과의 차이(편향)가 가장 작았으며, 평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률(포함률)도 높은 편이었다. 또한 무응답 비율에 따라 약간 차이가 있지만 대부분 자료의 분포도 잘 유지(추정의 정확성)하는 편에 속하였다.

표 5. 평가지표별 가장 우수한 효과를 가지는 대체 방법-1(한국복지패널)

(단위: %)

대체 방법	BIAS	RMSE	COV	예측의 정확성		추정의 정확성			
				MAD	RMSD	ADB	ADV	ADS	ADK
mean									
hotdeck									
ratio			5, 10, 20, 30				20, 30, 50	50	50
mean_3			5, 10						
hotdeck_3			5, 10						
ratio_3			5, 10, 20						
knn_3		5, 10	5, 10, 20, 30		10	5, 10		10, 30	
rf_3	5, 10, 20, 30, 50	20, 30, 50	5, 10, 20, 30			30, 50	5, 10	5, 20	5, 10, 20, 30
svm_3			5, 10, 20, 30	5, 10, 20, 30, 50	5, 20, 30, 50	20			
mlp_3			5, 10, 20, 30						

주: BIAS – 편향, RMSE – 제곱근평균제곱오차, COV – 참값의 포함률, MAD – 평균절대편차, ADB – 상대평균의 절대 차이, ADV – 변동계수의 절대 차이, ADS – 표준화된 왜도의 절대 차이, ADK – 표준화된 첨도의 절대 차이.
자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 116 〈표 4-12〉.

다음으로 3개 설명변수를 사용한 K-최근접 이웃(knn_3) 및 서포트 벡터 머신(svm_3) 대체 방법의 결과도 좋은 편이었다. 특히 3개의 설명변수를 사용한 서포트 벡터 머신 대체 방법은 무응답 비율에 상관없이 개별 개체의 대체값과 실제값의 유사 정도(예측의 정확성)가 가장 우수하였다. 이렇듯 기계학습 통계기법을 이용한 대체 방법의 효과는 좋은 편이었다.

한편 대체군을 활용하지 않은 비대체(ratio)도 무응답 비율 30% 이하에서는 포함률이 높아 효과가 좋은 편이었다. 특히 무응답 비율 50%의 경우 편향, 포함률과 추정의 정확성은 K-최근접 이웃 및 신경망 대체 방법보다 효과가 우수하였다. 비대체는 대체 대상 변수의 이전·해당 차수 간 변동이 크지 않을 때 효과적인 대체 방법이므로, 실제 패널 자료의 항목무응답 대체 시 많이 사용하는 대체 방법 중 하나이다.

평균 및 핫덱 대체 방법의 경우 무응답 대체 시 대체군 형성 유무에 따라 대체 효과에 큰 차이를 보였다. 이는 선행연구와 동일한 결과로, 대체군을 활용한 무응답 처리의 중요성을 다시 한번 확인하였다. 더불어 대체 대상 변수의 특징을 고려하여 대체군을 더욱 세밀하게 구분한다면 대체 효과도 좀 더 향상될 수 있을 것이다.

다음은 추가로 실시한 모의실험의 결과를 정리한 것이다. 무응답 대체 시 무응답 대체 대상

표 6. 평가지표별 가장 우수한 효과를 가지는 대체 방법-2(한국복지패널)

(단위: %)

대체 방법	BIAS	RMSE	COV	예측의 정확성		추정의 정확성			
				MAD	RMSD	ADB	ADV	ADS	ADK
knn_6			5, 10, 20				10		
rf_6	10, 20, 30, 50	10, 20, 30, 50	5, 10, 20, 30	50	30, 50	5, 10, 20, 30, 50	5, 20, 30, 50	5, 10, 50	5, 10, 20, 30, 50
svm_6		5	5, 10, 20, 30	5, 10, 20, 30	5, 10, 20				
mlp_6	5		5, 10, 20, 30					20, 30	

주: BIAS - 편향, RMSE - 제곱근평균제곱오차, COV - 참값의 포함률, MAD - 평균절대편차, ADB - 상대평균의 절대 차이, ADV - 변동계수의 절대 차이, ADS - 표준화된 왜도의 절대 차이, ADK - 표준화된 첨도의 절대 차이.
자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 118 〈표 4-13〉.

변수와 관련 있는 여러 개의 설명변수를 더 포함하여 무응답 처리 효과를 살펴보았다. 〈표 5〉와 동일한 방법으로 4개의 대체 방법 중 평가지표별로 가장 우수한 효과를 나타내는 대체 방법의 무응답 비율을 표기하였다(표 6).

6개 설명변수를 사용한 랜덤 포레스트 대체 방법(rf_6)은 6개 설명변수를 사용한 대체 방법 중에서 무응답 비율과 상관없이 효과가 가장 우수하였다. 또한 3개의 설명변수를 사용한 랜덤 포레스트 대체 방법과 비교해 보면, 추정의 정확성은 유사하였으나 나머지 평가지표(편향, 포함률, 예측의 정확성)는 소폭 향상된 결과를 보였다.

한편 6개 설명변수를 사용한 K-최근접 이웃 대체 방법(knn_6)은 3개 설명변수를 사용했을 때보다 효과가 좋지 않은 것으로 나타났다. 이는 설명변수의 개수가 증가함에 따라 나타나는 차원의 저주로 인한 과적합의 영향이라고 볼 수 있다.

6개 설명변수를 사용한 랜덤 포레스트 및 K-최근접 이웃 대체 방법의 결과를 통해 알 수 있는 무응답 대체 효과 향상 요인은 다음과 같다. 무응답 대체 대상 변수와 관련 있는 설명변수를 탐색하여 선택하고 적절한 설명변수 개수를 선정하는 등의 과정을 필수적으로 실시하는 것이다.

마지막으로 6개 설명변수를 사용한 신경망 대체 방법(mlp_6)의 효과는 3개일 때보다 향상된 결과를 보였으며, 무응답 비율이 증가할수록 효과는 더 좋아지는 것으로 나타났다.

이 추가적인 모의실험은 기계학습 방법에 기반한 대체 방법에만 적용하였다. 평균, 핫덱, 비대체의 경우에는 대체군으로 활용될 변수의 개수가 증가하게 되면 대체군 내 개체의 개수가 부족한 현상이 발생하여 대체하기 어려울 수 있다는 한계가 있기 때문이다. 실제로 무응답 비율 30%와 50%에서 핫덱대체는 기증자 부족으로 비복원 추출 대신 복원 추출을 하였으며, 비대체의 경우 비(ratio)의 변동이 커지게 되는 현상을 발견하였다.

나. 한국의료패널 데이터를 이용한 모의실험

1) 자료 설계

모의실험을 위한 자료로는 2018년에 배포된 한국의료패널 8~9차(2014~2015년) 자료를 사용하였고, 9차 조사 때 조사한 항목인 '작년 한 해 월평균 생활비(이하 생활비)'를 항목무응답 대체 대상 변수로 선정하였다. 연구 대상 모집단은 6473가구로, 8~9차 생활비 및 항목무응답 대체 시 설명변수로 사용하는 변수에 대해 모두 응답한 가구를 대상으로 확정하였다.

결측 자료 메커니즘이 임의결측을 따른다는 가정하에 생활비와 연관성이 있는 '가구주의 연령'(이하 연령)과 '가구주의 학력'(이하 학력)을 가지고 4개의 대체군 집단(이하 집단)으로 나누었다(표 7). 집단별 생활비 평균은 집단 1이 197만 원, 집단 2가 305만 원, 집단 3이 115만 원, 집단 4가 198만 원이었다. 유의 수준 5%하에서 (집단 3), (집단 1, 집단 4), (집단 2)로 묶이는 것으로 나타났다.

생활비가 많을수록 응답하지 않을 가능성이 높다는 점을 반영하여 집단별 무응답 발생 개수의 비율을 달리 구성하였다(표 8).

또, 생활비는 매년 조사되어 무응답이 이전 차수의 생활비에 의존하는 것이 가능하므로, 생활비가 높을수록 무응답이 많이 발생한다고 가정하였다. 이에 따라 4개 대체군 집단에서 중위수를 기준으로 중위수 미만에서는 집단별 무응답 가구수의 30%, 중위수 이상에서는 70%의 무응답이 발생한다고 가정하였다. 임의결측 가정을 만족시키기 위해 8차 생활비는 무응답 없이 모두 값을 가지도록 하고 9차 생활비는 무응답으로 처리하였다.

무응답 비율별 발생 개수는 무응답 비율 5%가 324가구, 10%가 649가구, 20%가 1296가구, 30%가 1942가구, 50%가 3236가구이다(부록 참조).

표 7. 연령 및 학력에 따른 집단 구분(한국의료패널)

	고졸 미만	고졸 이상
60세 미만	집단 1	집단 2
60세 이상	집단 3	집단 4

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 120 〈표 4-14〉.

표 8. 집단별 무응답 발생 개수의 비율(한국의료패널)

(단위: %)

	대졸 미만	대졸 이상
60세 미만	6	58
60세 이상	12	24

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 120 〈표 4-15〉.

연구 대상 모집단 자료에서 무응답 가구를 생성하는 과정을 반복적으로 실시하여 모의실험용 무응답 패널 자료를 100개 생성하였다. 무응답 가구 생성에는 단순임의추출 방법을 사용하였다.

2) 대체 방법

모의실험에서 사용한 대체 방법은 평균대체, 핫덱대체, 비대체, K-최근접 이웃 대체, 랜덤 포레스트 대체, 서포트 벡터 머신 대체, 신경망 대체로 총 7가지이다. 이 중에서 K-최근접 이웃 대체, 랜덤 포레스트 대체, 서포트 벡터 머신 대체, 신경망 대체 방법은 기계학습 통계기법에 기반한 것이다. 평균대체, 핫덱대체, 비대체는 대체군 활용 여부에 따라 2개의 대체값을 생성하였다. 대체 방법을 자세히 살펴보면 다음과 같다.

평균대체는 대체군을 사용하지 않는 경우 9차 생활비에 대해 응답한 값만을 가지고 구한 평균값으로 무응답을 대체하였다. 대체군을 사용한 경우는 먼저 대체군을 형성하고, 각 대체군 내에서의 9차 생활비 평균값으로 무응답을 대체하였다. 대체군 형성 시 고려한 변수는 연령 범주 2개(60세 미만·이상), 학력 범주 2개(고졸 미만·이상), 8차 생활비 범주 2개(중위수 미만·이상)이며 총 8개로 구분하였다.

핫덱대체는 대체군을 사용하지 않는 경우 9차 생활비에 대해 응답한 모든 값을 기증자 후보로 사용하여 기증자 중에서 무작위로 추출한 후 기증자의 응답값으로 무응답을 대체하였다. 대체군을 사용한 경우는 대체군 내에서 응답한 모든 값을 기증자 후보로 사용하여 기증자 중에서 무작위로 추출한 후 기증자의 응답값으로 무응답을 대체하였다. 이때 사용한 대체군은 평균대체와 동일하고, 무응답 대체 시 한 번 사용한 기증자는 이후 대체 시 사용하지 않는 것을 원칙으로 하였다. 단, 무응답 비율 30%와 50%의 경우에는 해당 대체군 내에서의 기증자 부족으로 무응답에 대해 제대로 채울 수 없어서 기증자를 중복으로 사용하였다.

비대체는 대체군을 사용하지 않는 경우 이전 차수와 해당 차수에 대해 모두 생활비 응답값을 가지고 있는 개체만을 대상으로 하여 비를 계산한 후 비의 평균값을 구하였다. 무응답은 이전 차수의 응답값에 해당 비를 곱한 값으로 대체하였다. 대체군을 사용하는 경우 연령 및 학력 범주로 4개의 대체군을 형성한 후 대체군 내에서의 비의 평균을 계산하였다. 무응답은 이전 차수의 응답값에 해당 대체군 내의 비를 곱한 값으로 대체하였다.

기계학습 통계기법에 기반한 대체 방법은 '가. 한국복지패널 데이터를 이용한 모의실험'에서 자세하게 설명하였으며, 동일한 과정으로 대체를 실시하였다.

기계학습 통계기법에 기반한 대체 방법은 평균, 핫덱, 비대체 방법에서 활용한 보조변수 3개를 동일하게 사용하였다. 이는 동일한 조건하에서 무응답을 대체한 방법들의 효과를 비교하기 위해서이다. 한국복지패널 데이터를 이용한 모의실험에서처럼 추가로 모의실험을 하여 더 많은

표 9. 무응답 대체 시 사용한 설명변수(한국의료패널)

설명변수	변수명
3개	가구주의 연령 및 학력, 생활비(8차)
10개	가구주의 학력, 연령, 배우자의 유무 및 근로 여부, 생활비(8차), 가구 근로소득, 만성질환 여부, 저축 여부, 가구 내 근로자 수, 가구원 수

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 126 〈표 4-21〉.

정보를 사용한 경우(보조변수 3개 → 10개 증가)에 대한 무응답 대체 효과를 살펴보았다(표 9).

3) 평가지표

평가지표는 한국복지패널 데이터를 이용한 모의실험에서 사용한 5가지 평가지표(편향, 제곱근평균제곱오차, 포함률, 예측의 정확성, 추정의 정확성)를 동일하게 사용하여 각 대체 방법에 따라 생성한 대체값에 대한 효과를 평가하였다(표 4).

4) 모의실험 결과

무응답 비율별 각 대체 방법에 따라 생성한 대체값에 대한 5가지 평가지표를 계산하였을 뿐만

표 10. 평가지표별 가장 우수한 효과를 가지는 대체 방법-1(한국의료패널)

(단위: %)

대체 방법	BIAS	RMSE	COV	예측의 정확성		추정의 정확성			
				MAD	RMSD	ADB	ADV	ADS	ADK
mean									
hotdeck								5	
ratio			5, 10, 20				20, 30, 50	20, 30, 50	50
mean_3			5, 10						
hotdeck_3			5, 10						
ratio_3			5, 10				5, 10	10	20, 30
knn_3	5	5, 10, 20, 30	5, 10, 20, 30	5		5, 10, 20, 30			
rf_3	5, 10, 20, 30, 50	5, 10, 20, 30, 50	5, 10, 20, 30			50			
svm_3			5, 10, 20	5, 10, 20, 30, 50	5, 10, 20, 30, 50				10
mlp_3			5, 10, 20, 30						5

주: BIAS - 편향, RMSE - 제곱근평균제곱오차, COV - 참값의 포함률, MAD - 평균절대편차, ADB - 상대평균의 절대 차이, ADV - 변동계수의 절대 차이, ADS - 표준화된 왜도의 절대 차이, ADK - 표준화된 첨도의 절대 차이.
자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 151 〈표 4-25〉.

아니라 히스토그램 및 상자그림을 통해 대체된 자료와 실제 자료 간 분포 비교도 하였다(이혜정 외, 2019, pp. 128-149). 각각의 무응답 비율에서 평가지표별 효과가 가장 우수한 대체 방법에 대해 해당 무응답 비율을 표기하여 〈표 10〉과 같이 정리하였다.

3개의 설명변수를 사용한 랜덤 포레스트 대체 방법(rf_3)의 효과가 우수한 것으로 나타났다. 무응답 비율과 상관없이 생활비에 대한 평균 추정량은 참값과의 차이(편향)와 평균 추정치를 얼마나 일관되게 추정하는지 평가하는 제곱근평균제곱오차가 가장 작았고, 평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률(포함률)도 높았다.

다음으로 3개의 설명변수를 사용한 K-최근접 이웃 대체 방법(knn_3)의 제곱근평균제곱오차와 포함률의 결과가 좋은 편이었다. 한편 3개의 설명변수를 사용한 서포트 벡터 머신 대체 방법(svm_3)은 무응답 비율에 상관없이 개별 개체의 대체값과 실제값 사이의 유사 정도(예측의 정확성)가 가장 우수하였다.

대체군을 활용하지 않은 비대체 방법(ratio)도 기계학습 통계기법의 결과와 유사하였으며, 특히 무응답 비율 20% 이상에서는 자료의 분포를 잘 유지(추정의 정확성)하는 측면에서 우수한 결과를 나타냈다.

대체군을 활용한 비대체(ratio_3)의 효과도 무응답 비율 10% 이하에서는 포함률과 추정의 정확성이 우수하게 나타났다. 그러나 무응답 비율 20%에서부터는 대체군 내의 개체 개수가 적어져서 효과가 좋지 못한 결과를 보였다. 이는 무응답 대체 시 정보 부족이 발생한 것이므로 대체군 내의 개체 개수는 충분히 보장되어야 할 것이다.

대체군을 사용하지 않은 평균 및 핫덱 대체는 무응답 비율이 낮은 5%에서도 나머지 대체 방법에 비해 저조한 결과를 나타냈으며, 한국복지패널 데이터를 이용한 모의실험과 동일하였다. 대체군을 사용하여 평균 및 핫덱대체를 실시하면 효과가 개선될 수 있다는 점을 다시 한번 확인하였다.

다음은 부가적으로 실시한 모의실험으로, 무응답 대체 시 무응답 대체 대상 변수와 관련 있는 여러 개의 변수를 더 포함했을 때의 무응답 처리 효과를 살펴보았다. 〈표 11〉은 4개의 대체 방법 중에서 평가지표별 가장 우수한 효과를 보이는 대체 방법에 대해 무응답 비율을 표기한 것이다.

4개의 대체 방법 중에서 10개의 설명변수를 사용한 랜덤 포레스트 대체 방법(rf_10)이 편향도 작고 포함률도 높은 편이며 참값에 더 가깝게 대체하였다. 이와 더불어 10개의 설명변수를 사용한 랜덤 포레스트 대체 방법은 무응답 비율과 상관없이 3개의 설명변수를 사용할 때보다 예측의 정확성이 향상되었다.

한편 10개의 설명변수를 사용한 신경망 대체 방법(mlp_10)은 무응답 비율과 상관없이 편향이 작게 추정되어 실제값과 유사하다고 볼 수 있다. 또한 무응답 비율이 증가할수록 3개의 설명변수를 사용한 경우보다 효과가 향상되었다. 그러나 무응답 비율이 50%인 경우 추정의 정확성

표 11. 평가지표별 가장 우수한 효과를 가지는 대체 방법-2(한국의료패널)

(단위: %)

대체 방법	BIAS	RMSE	COV	예측의 정확성		추정의 정확성			
				MAD	RMSD	ADB	ADV	ADS	ADK
knn_10			5, 10, 20					5, 10, 20, 30, 50	5, 10, 20, 30, 50
rf_10	10, 20, 30	5, 10, 20, 30	5, 10, 20, 30	5, 10, 50	5, 10, 20, 30, 50	5, 10, 20	5, 10, 50		
svm_10	5		5, 10, 20	5, 10, 20, 30, 50					
mlp_10	5, 10, 30, 50	5, 20, 30, 50	5, 10, 20, 30			30, 50	20, 30		

주: BIAS – 편향, RMSE – 제곱근평균제곱오차, COV – 참값의 포함률, MAD – 평균절대편차, ADB – 상대평균의 절대 차이, ADV – 변동계수의 절대 차이, ADS – 표준화된 왜도의 절대 차이, ADK – 표준화된 첨도의 절대 차이.
자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구–항목무응답 대체 방법을 중심으로. p. 153 〈표 4-26〉.

에 해당하는 표준화된 왜도의 절대 차이와 표준화된 첨도의 절대 차이가 일부 대체된 자료에서 큰 값을 가지므로 무응답 대체 시 주의가 필요하다.

10개의 설명변수를 사용한 K-최근접 이웃 대체 방법(knn_10)은 4가지 대체 방법 중에서 효과가 가장 저조했을 뿐만 아니라 3개의 설명변수를 사용했을 때보다 평가지표의 결과도 좋지 않은 것으로 나타났다.

10개의 설명변수를 사용한 서포트 벡터 머신 대체 방법(svm_10)의 경우 예측의 정확성은 무응답 비율과 상관없이 우수한 편이었으나, 편향은 무응답 비율 10%에서부터 큰 값을 가지는 것으로 나타났다.

4. 나가며

패널 데이터의 품질 향상을 위한 방안 중 하나로 항목무응답에 대한 처리가 있다. 항목무응답이 적절한 값으로 대체된 자료를 사용한다면 표본의 대표성을 확보할 수 있으며, 통계적 분석 결과에서 편향이 감소하게 되고 목표 모집단에 대한 통계적 추론인 추정과 검정에서의 오류를 최소화할 수 있다. 이 글은 한국보건사회연구원에서 공동으로 주관하고 있는 한국복지패널조사와 한국의료패널조사에서 향후 발생할 수 있는 무응답 패턴 변화에 대비하기 위해 작성하였다. 무응답 발생 패턴의 변화로는 무응답 대체 대상 변수의 추가 발생, 무응답 비율 변화 등이 있다. 한국복지패널조사 및 한국의료패널조사에서의 항목무응답 대체 방법을 살펴보았고, 기계학습 통계기법을 기반으로 한 항목무응답 대체 방법의 효과를 파악하여 적절한 대체 방법을 제안하고자 하였다. 이에 따른 종합적인 결과와 활용 방안은 세 가지로 요약할 수 있다.

첫째, 패널 자료에서의 항목무응답 대체 시 기계학습 통계기법을 적용할 수 있는 가능성을 확

인하였다. 대체 방법 선정은 어떠한 평가지표에 중점을 두는지에 따라 달라질 수 있다. 무응답 비율에 상관없이 개별 개체의 대체값과 실제값 간의 유사 정도(예측의 정확성)에 초점을 둔다면 서포트 벡터 머신 대체 방법을, 평균 추정치와 참값의 일치 정도(편향)에 초점을 둔다면 랜덤 포레스트 대체 방법을 추천한다. 특히 랜덤 포레스트 대체 방법은 편향뿐만 아니라 다른 평가지표도 우수한 결과를 보여 실무에서 활용해 볼 수 있다고 생각한다.

둘째, 대체군 활용 여부에 따라 대체 효과가 확연히 달라지는 결과를 통해 무응답 대체 시 대체군 활용의 중요성을 다시 한번 확인하였다. 평균, 핫덱, 비대체 시 대체군은 무응답 대체 변수와 연관성이 높은 설명변수로 형성할 것을 추천한다. 또한 대체군을 더욱 견고하게 형성할수록 대체 효과는 더 향상될 수 있다고 생각한다.

셋째, 보조변수로 활용 하는 설명변수의 개수 증가에 따른 대체 효과를 확인하였다. 부가적으로 실시한 모의실험에서 K-최근접 이웃 대체 방법은 효과가 좋지 않은 것으로 나타났다. 이는 설명변수 개수가 증가함에 따라 나타나는 차원의 저주로 인한 과적합의 영향으로 볼 수 있다. 차원이 증가하면 같은 점들 사이의 값이라도 최소 길이가 점차 길어져 평균 길이와 비슷해지므로 최소 길이라는 의미가 사라지기 때문이다. 실무에서 활용할 때 차원의 저주 문제는 정규화(regularization), 군집화(clustering), 차원 축소, 설명변수 선택 방법 등으로 해결하는 것을 추천한다. K-최근접 이웃, 랜덤 포레스트 대체 방법의 결과를 통해 더 많은 설명변수를 추가하면 더 많은 정보를 포함할 수 있으나, 무응답 대체 효과가 항상 향상된다고 볼 수는 없었다. 이렇듯 복잡하고 포괄적인 모형보다는 무응답 대체 대상 변수와 연관성이 높은 설명변수를 탐색하고 선정하는 것이 무응답 대체 효과 향상에 효과적이라고 생각한다.

대체 방법은 패널 자료의 무응답 특징, 범위 등을 자료의 형태에 따라 달리 고려해야 하므로 모의실험 결과를 일반화할 수 없다. 그렇기 때문에 항목무응답을 대체하기 전에 대체 대상 변수의 무응답 비율, 분포, 특성, 대체 대상 변수와의 연관 변수 등을 탐색하는 과정이 우선되어야 한다. 이 과정을 바탕으로 최적의 대체 방법을 선정해야 한다. 바람직한 대체 방법은 통계적 추론에서 발생할 수 있는 무응답 편의가 감소해야 하고 모집단 분포로부터 표본 분포가 왜곡되지 않고 비슷하게 유지될 수 있어야 한다. 이 점을 인지한다면 더욱 정확하고 신뢰성 있는 무응답 대체 자료를 제공할 수 있을 것이다.

부록

부표 1. 한국복지패널에서의 집단별 무응답 발생 개수

(단위: 가구)

	무응답 비율 5%			무응답 비율 10%			무응답 비율 20%			무응답 비율 30%			무응답 비율 50%		
	경상소득 중위수		전체	경상소득 중위수		전체	경상소득 중위수		전체	경상소득 중위수		전체	경상소득 중위수		전체
	미만	이상		미만	이상		미만	이상		미만	이상		미만	이상	
집단 1	10	22	32	19	45	64	38	89	127	57	134	191	95	223	318
집단 2	17	40	57	34	80	114	69	160	229	103	240	343	172	400	572
집단 3	21	49	70	42	98	140	84	196	280	126	294	420	210	490	700
집단 4	48	111	159	95	223	318	191	445	636	286	668	954	477	1113	1590
전체	96	222	318	190	446	636	382	890	1272	572	1336	1908	954	2226	3180

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 86 〈표 4-3〉, 〈표 4-4〉 p. 87 〈표 4-5〉, 〈표 4-6〉, 〈표 4-7〉.

부표 2. 한국의료패널에서의 집단별 무응답 발생 개수

(단위: 가구)

	무응답 비율 5%			무응답 비율 10%			무응답 비율 20%			무응답 비율 30%			무응답 비율 50%		
	생활비 중위수		전체	생활비 중위수		전체	생활비 중위수		전체	생활비 중위수		전체	생활비 중위수		전체
	미만	이상		미만	이상		미만	이상		미만	이상		미만	이상	
집단 1	6	13	19	12	27	39	23	55	78	35	82	117	58	136	194
집단 2	56	132	188	113	263	376	225	526	751	338	788	1126	563	1314	1877
집단 3	12	27	39	23	55	78	47	109	156	70	163	233	116	272	388
집단 4	23	55	78	47	109	156	93	218	311	140	326	466	233	544	777
전체	97	227	324	195	454	649	388	908	1296	583	1359	1942	970	2266	3236

자료: 이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. p. 121 〈표 4-16〉, p. 122 〈표 4-17〉, 〈표 4-18〉, p. 123 〈표 4-19〉, 〈표 4-20〉.

참고문헌

김미곤, 손창균, 여유진, 김계연, 유현상, 오지현,…김혜진. (2008). 2008년 한국복지패널 기초분석 보고서. 서울: 한국보건사회연구원.

박은자, 정연, 서제희, 배정은, 이나경, 김은주,…박현아. (2019). 제2기 한국의료패널 구축·운영을 위한 연구. 서울: 한국보건사회연구원.

여유진, 오미애, 이병재, 최준영, 이주미, 김근혜,…김정욱. (2019). 2019년 한국복지패널 기초분석보고서. 세종: 한국보건사회연구원.

이현주, 오미애, 정은희, 정해식, 김현경, 손창균,…박형존. (2017). 한국복지패널의 진단과 향후 개선 과제. 세종: 한국보건사회연구원.

이혜정, 지희정, 이지혜. (2019). 보건복지 분야 패널자료 품질 개선 연구-항목무응답 대체 방법을 중심으로. 세종: 한국보건사회연구원.

한국보건사회연구원. (2019). 한국복지패널 14차년도 조사자료 User's Guide. https://www.koweps.re.kr:442/data/guide/list.do에서 2020. 6. 22. 인출.