연구보고서(수시) 2019-08

보건복지 분야 패널자료 품질 개선 연구

- 항목무응답 대체 방법을 중심으로

이혜정
지희정 · 이지혜

KIHASA
한국보건사회연구원
Korea Institute for Health and Social Affairs

【책임연구자】
이혜정 한국보건사회연구원 부연구위원

【주요 저서】
제20차(2017) 한국노동패널 기초분석보고서
한국노동연구원, 2018(공저)
패널자료 품질 개선 연구(Ⅷ)
한국노동연구원, 2018(공저)

【공동연구진】
지희정 고려대학교 의과대학 의학통계학교실 연구교수
이지혜 한국보건사회연구원 연구원

연구보고서(수시) 2019-08

보건복지 분야 패널자료 품질 개선 연구
- 항목무응답 대체 방법을 중심으로

발 행 일 2019년 11월
저 자 이 혜 정
발 행 인 조 흥 식
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1994년 7월 1일(제8-142호)
인 쇄 처 (사)아름다운사람들복지회

ISBN 978-89-6827-611-8 93510

발간사

패널자료는 개체 간 차이 및 시간에 따른 개체의 변화를 고려한 동태적 분석이 가능하여 사회과학, 자연과학, 보건 및 의학 등의 다양한 분야에서 널리 활용되고 있다. 이와 같은 상황에서 신뢰성 있는 통계적 분석 결과를 얻기 위해서는 패널자료의 질이 담보되어야 한다. 패널자료의 질을 결정짓는 하나의 요소로 항목무응답에 대한 처리를 들 수 있다. 항목무응답이 포함된 자료의 경우 무응답을 통계적 방법에 기반하여 적절한 값으로 대체해 완전한 자료를 만든 뒤 사용하는 것이 바람직하다고 볼 수 있다.

이 연구는 우리 연구원에서 공동으로 주관하고 있는 한국복지패널조사 및 한국의료패널조사 자료의 품질 개선을 위한 연구로 항목무응답 대체 방법을 중심으로 살펴보았다. 이를 통해 항목무응답 대체에 대한 새로운 기법을 제안하고 지속 가능한 무응답 대체 자료를 제공하고자 하였다. 먼저 기계학습과 딥러닝이 여러 분야에서 활발하게 사용되고 있는 시기에 맞추어 최신 통계방법을 패널자료의 항목무응답 대체에 적용해 보았다. 이때 기존의 대체 방법과 함께 비교하여 적절한 대체 방법을 제시하였다. 다음으로 두 개 패널자료에서 향후 항목무응답 발생 패턴의 변화에 대비하기 위해 국내외 패널조사의 선행연구를 통하여 벤치마킹 가능성을 모색하고자 하였다. 또한 다양한 조건을 고려한 모의실험을 실시하여 상황에 따른 최적의 방법을 제시하였다.

이 연구는 우리 연구원의 이혜정 부연구위원의 책임하에 이지혜 연구원이 참여하였다. 외부 연구진으로 참여한 지희정 연구교수에게 감사의 마음을 전한다. 또한 이 연구와 관련하여 유익한 조언을 해 주신 원내 정

영철 정보통계연구실장, 오미애 빅데이터정보연구센터장, 고경환 선임연구위원 그리고 원외 고려대학교 송주원 교수, 코리아크레딧뷰로 김은경 박사를 비롯하여 익명의 검독위원들에게도 감사의 마음을 전한다. 이 보고서의 결과는 우리 연구원의 공식적인 견해가 아니라 연구진의 의견임을 밝혀 둔다.

2019년 11월

한국보건사회연구원 원장

조 흥 식

목 차

표 목차

그림 목차

Abstract

A Study of Quality Improvement in Health and Welfare Panel Data – Focusing on Imputation of Item Nonresponse

Project Head: Lee, Hyejung

Panel data is widely used in social, natural, healthcare and medical sciences, as it enables dynamic analysis of inter-subject differences and changes over time in the same subject. As in other forms of data, nonresponse in panel surveys occurs when the observed values for some variables are not measured. Nonresponse can be attributed to lack of bond at the beginning of the survey and increased panel fatigue as the investigation progresses. Imputation of missing responses is necessary to improve data quality.

This study is about bringing quality improvement to KOWEPS and KHP data, especially in terms of imputation of item nonresponse. We examined how Korea and other countries have been handling missing reponses in panel data. Also, machine learning and deep learning techniques were examined as a potential alternative to imputation of missing responses in panel data. Our simulation results show that imputation methods based on machine learning and deep learning generally

Co-Researchers: Jee, Hee-Jung · Lee, Jihye

outperform mean and hot-deck imputation. Specifically, we propose an imputation method based on random forest. It has been found that a large number of explanatory variables do not necessarily improve performance. Exploring and selecting variables that are highly related to the target imputation variables perform better than using a complex and comprehensive model.

Panel data is widely used as data for policy diagnosis and establishment. Imputation of item nonresponse is an important part of data quality management and must be managed continuously for statically reliable data production.

*Keywords: panel data, imputation method, machine learning

요 약

1. 연구의 배경 및 목적

패널자료는 개체 간 차이 및 시간에 따른 개체의 변화를 고려한 동태적 분석이 가능하다. 이 같은 이유로 정책입안자나 연구자가 정책을 마련하거나 효과를 평가할 때 활발히 사용하고 있다. 패널자료의 활용도가 높은 상황에서 신뢰성 있는 통계적 분석 결과를 얻기 위해서는 패널자료의 질이 담보되어야 한다. 패널자료의 질을 결정짓는 하나의 요소로 항목무응답에 대한 처리를 들 수 있다.

항목무응답이란 설문조사 시 응답자에게 가능한 한 정확한 응답을 받는 것이 좋지만 여러 가지 제약으로 불가피하게 일부 문항에 대해 응답을 받지 못하는 경우를 의미한다. 패널조사에서 항목무응답은 조사 초기의 경우 응답자와 면접자 간의 유대감 형성 부족으로, 조사 차수가 거듭됨에 따라서는 응답자의 패널피로도 증가 또는 개인 사정 등으로 인해 나타날 수 있다. 이렇듯 항목무응답 발생 가능성이 상존하므로 패널자료의 품질 향상을 위해서는 무응답 처리 기법과 관련한 심층 연구가 필요하다.

현재 국내외 패널자료를 생산하고 있는 여러 기관에서는 사용자 편의를 위해서 항목무응답 변수를 적절한 값으로 대체하여 제공하고 있다. 한국노동패널조사와 한국고령화연구패널조사는 기본 자료 제공 이외에 추가로 무응답 대체 자료를 배포하고 있다. 우리 연구원에서 공동으로 주관하고 있는 한국복지패널조사와 한국의료패널조사도 일부 변수에 대해 무응답을 대체하고 있다. 두 개 패널자료의 경우 항목무응답 비율은 높지 않은 편이나 향후 항목무응답 발생 패턴의 변화에 대비할 필요가 있다.

한편 최근 빅데이터 및 정보 산업혁명 시대가 도래함에 따라 인공지능(AI: artificial intelligence)에 대한 관심이 높아지고 있으며 이와 관련된 연구도 빠르게 진행되고 있다. 인공지능은 기계학습(machine learning)과 딥러닝(deep learning)을 포함하고 있는 가장 큰 영역이다. 이 연구에서도 기계학습과 딥러닝이 여러 분야에서 널리 사용되고 있는 상황에 맞추어 이러한 최신 통계방법을 패널자료의 항목무응답 대체에 적용해 보고자 한다. 이때 기존의 대체 방법과 비교함으로써 적절한 대체 방법을 제시하고자 한다.

2. 주요 연구 결과

먼저 국내외 패널자료에서 실시하고 있는 대체 방법의 처리 현황을 보면 패널조사는 특화된 조사 목적을 가지고 있지만 인구학적 배경, 소득, 자산 등과 같이 유사한 문항을 공통적으로 포함하고 있다. 대부분의 패널조사는 이러한 유사한 문항에 대해 무응답 대체를 실시하고 있다. 성별, 연령 등과 같은 인구학적 변수의 무응답률은 초반 조사 차수에서는 높은 편이나 점점 낮아지는 경향을 나타내므로 무응답 처리는 이전 및 이후 차수 간 관계를 고려하여 대체하고 있다. 주요 무응답 대체 변수는 소득, 자산, 생활비 등과 같은 금액과 관련된 것이었다. 또한 이러한 주요 무응답 대상 변수 대체 시 사용되는 보조변수가 무응답인 경우에는 필요에 따라 함께 대체하였다. 무응답 대체는 대체 대상 변수 또는 개체에 대해 한 가지 대체 방법만이 아닌 여러 가지 대체 방법을 사용하여 대체하기도 하였다. 이는 설문구조, 자료의 분포, 변수의 특성 등에 따라 사용할 수 있는 정보가 다르기 때문이며, 활용 가능한 정보를 최대한 사용하여 대체하고 있는 것으로 나타났다. 그리고 대체할 때 사용되는 보조변수는 패널자료

의 장점이라고 할 수 있는 이전 차수의 정보를 사용하기도 하였으나, 대부분 해당 차수의 정보를 활용하였다.

다음은 한국복지패널 및 한국의료패널 자료에 근거하여 모의실험을 실시한 결과이다. 대체 대상 변수는 한국복지패널자료의 경우 '경상소득'이었고, 한국의료패널자료는 '생활비'로 선정하였다. 무응답 비율을 최저 5%에서 최고 50%까지 극대화하여 무응답 비율의 변화에 따른 대체 방법의 효과를 살펴보았다. 또한 빅데이터 분석기법 중 하나인 기계학습 통계방법을 무응답 대체에 적용해 봄으로써 활용 가능성도 가늠해 보았다. 대체 방법의 효과는 〈요약표 1〉을 기준으로 평가하였다.

〈요약표 1〉 대체 방법 평가 기준

평가지표	표기	설명
편향	BIAS	평균 추정치와 참값과의 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

첫 번째 한국복지패널자료의 결과를 보면 무응답 비율과 상관없이 3개의 설명변수를 사용한 랜덤 포레스트 대체 방법(rf_3)의 효과가 가장 우수한 것으로 나타났다(〈요약표 2〉 참조). 즉, 경상소득에 대한 평균 추정량의 편향도 작고 포함률도 높은 편이며 추정의 정확성도 잘 유지하는 편이었다. 3개의 설명변수를 사용한 K-최근접 이웃 대체 방법(knn_3)의 결과도 좋은 편이었다. 그리고 3개의 설명변수를 사용한 서포트 벡터 머신 대체 방법(svm_3)은 무응답 비율에 상관없이 예측의 정확성이 가장 우수하였다. 이렇듯 전반적으로 기계학습 통계방법에 기반한 대체 방법

의 결과는 좋은 것으로 나타났다.

〈요약표 2〉 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: COV는 91% 이상의 포함률을 만족하는 경우임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

무응답 대체 대상 변수와 관련 있는 정보를 더 포함하였을 때 대체 방법의 효과를 평가하기 위하여 부가적으로 모의실험을 하였다. 설명변수를 더 포함(총 6개)하여 대체한 결과는 무응답 비율에 상관없이 랜덤 포레

스트 대체 방법(rf_6)이 예측의 정확성을 제외한 나머지 지표의 결과에서 모두 우수하였다(〈요약표 3〉 참조). 신경망 대체 방법(mlp_6)은 기존보다 결과가 향상되었으며, 무응답 비율이 증가할수록 대체 효과도 점점 향상되었다. 한편 K-최근접 이웃 대체 방법(knn_6)은 기존보다 효과가 좋지 않은 것으로 나타났다.

〈요약표 3〉 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: COV는 91% 이상의 포함률을 만족하는 경우임.
자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

두 번째 한국의료패널자료의 모의실험 결과를 보면 무응답 비율과 상관없이 3개의 설명변수를 사용한 랜덤 포레스트 대체 방법(rf_3)의 효과가 우수한 것으로 나타났다(〈요약표 4〉 참조). 즉, 생활비에 대한 평균 추정량의 편향이 가장 작고 포함률도 높았다. 다음으로 K-최근접 이웃 대체 방법(knn_3) 및 서포트 벡터 머신 대체 방법(svm_3)도 편향이 작고 포함률이 높은 것으로 나타났다. 서포트 벡터 머신 대체 방법의 경우 무

응답 비율과 상관없이 예측의 정확성이 아주 우수하였다.

대체군을 활용하지 않은 비대체 방법(ratio)도 기계학습 방법의 결과와 유사하게 나타났으며, 특히 추정의 정확성이 우수하였다.

〈요약표 4〉 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: COV는 91% 이상의 포함률을 만족하는 경우임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

설명변수를 더 포함(총 10개)하여 대체한 결과는 랜덤 포레스트 대체 방법(rf_10)이 편향도 작은 편이고 포함률도 높은 편이며 참값에 더 가깝게 대체하여 4가지 대체 방법 중에서 가장 우수하였다(〈요약표 5〉 참조). 신

경망 대체 방법(mlp_10)은 작은 편향을 가져서 참값과 유사하고 포함률도 높은 편으로 나타났다. 단, 무응답 비율이 50%일 때 추정의 정확성에 해당하는 표준화된 왜도의 절대 차이 및 표준화된 첨도의 절대 차이가 일부 대체된 자료에서 큰 값을 가지는 것으로 나타나, 무응답 대체 시 주의가 필요하다. 서포트 벡터 머신 대체 방법(svm_10)은 예측의 정확성은 우수한 편이었으나, 무응답 비율 10%에서부터 편향이 큰 편에 속하였다. K-최근접 이웃 대체 방법(knn_10)은 기존보다 효과가 좋지 않은 것으로 나타났다.

〈요약표 5〉 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: COV는 91% 이상의 포함률을 만족하는 경우임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

3. 결론 및 시사점

두 개의 모의실험 결과를 통해 패널자료에서의 항목무응답 대체 시 기계학습 통계방법을 적용할 수 있다는 점을 확인하였다. 랜덤 포레스트 대체 방법은 우수한 결과를 보여 실무에서도 활용할 수 있을 것으로 보인다.

부가적으로 실시한 모의실험에서 K-최근접 이웃 대체 방법의 효과는 좋지 않은 편이었다. 이는 설명변수의 개수가 증가함에 따라 나타나는 차원의 저주로 인한 과적합 영향으로 볼 수 있다. 차원의 저주는 정규화(regularization), 군집화(clustering), 차원 축소, 설명변수 선택 방법 등으로 해결할 수 있다. K-최근접 이웃, 랜덤 포레스트 대체 방법의 결과를 통하여 더 많은 설명변수를 추가하면 더 많은 정보를 포함할 수 있으나, 무응답 대체 효과가 항상 향상된다고 볼 수는 없다. 복잡하고 포괄적인 모형보다는 무응답 대체 대상 변수와 연관성이 높은 설명변수를 탐색하고 선정하는 것이 무응답 대체 효과를 향상하는 데 효과적이라고 생각된다.

*주요 용어: 패널자료, 항목무응답 대체 방법, 기계학습

제 1 장 서 론

1 서 론

제1절 연구 배경 및 목적

패널자료는 개체 간 차이 및 시간에 따른 개체의 변화를 고려한 동태적 분석이 가능하다. 이 같은 이유로 정책입안자나 연구자가 정책을 마련하거나 효과를 평가할 때 활발히 사용하고 있다. 일례로 한국복지패널조사를 활용한 연구결과물[2]은 2008년 30편에서 2016년 148편으로 5배 정도 대폭 증가하였다(이현주 외, 2017, p.143). 이처럼 패널자료의 활용도가 높은 상황에서 신뢰성 있는 통계적 분석 결과를 얻기 위해서는 패널자료의 질이 담보되어야 한다. 패널자료의 질을 결정짓는 하나의 요소로 항목무응답에 대한 처리를 들 수 있다.

항목무응답이란 설문조사 시 응답자에게 가능한 한 정확한 응답을 받는 것이 좋지만 여러 가지 제약으로 불가피하게 일부 문항에 대해서 응답을 받지 못하는 경우를 의미한다. 패널조사에서 항목무응답은 조사 초기의 경우 응답자와 면접자 간 유대감 형성 부족으로, 조사 차수가 거듭됨에 따라서는 응답자의 패널피로도 증가 또는 개인 사정 등으로 인해 나타

2) 2008년, 2012년, 2016년 한국복지패널 활용 연구 성과

(단위: 편)

구분	2008년	2012년	2016년	합계
보고서 및 기타	5	8	6	19
학술지논문	19	44	105	168
학위논문	6	34	37	77
합계	30	86	148	264

자료: 이현주 외. (2017). 한국복지패널의 진단과 향후 개선과제. p.143 〈표 5-1〉 인용.

날 수 있다. 이렇듯 항목무응답 발생 가능성이 상존하므로 패널자료의 품질 향상을 위해서는 이와 같은 무응답 처리 기법과 관련한 심층 연구가 필요하다.

패널자료에 항목무응답이 존재함으로써 발생할 수 있는 문제점을 예를 들어 설명하면 다음과 같다. 정책입안자가 복지정책의 수혜 대상자를 선정할 때 중위소득을 기준으로 선정한다고 가정하자. 실제로 이러한 기준은 더 복잡하나 이해가 쉽도록 단순하게 중위소득이라고 정의한다. 중위소득 산정 시 필요한 분석 자료의 소득 관련 정보에 무응답을 포함하고 있어 완전한 자료가 아니었다. 보통 무응답은 고소득 또는 저소득 가구에서 발생하는 편이다. 먼저 고소득 가구의 무응답이 많은 자료에서 이를 적절한 값으로 대체하지 않고 분석하면 중위소득이 실제(참값)보다 낮게 추정될 것이다. 이렇게 되면 복지 혜택을 받아야 할 대상자가 누락되는 문제가 발생하게 된다. 이와 반대로 저소득 가구의 무응답이 많은 자료를 사용하면 중위소득이 실제보다 높게 추정될 것이다. 이와 같은 상황은 복지 혜택을 받지 않아야 하는 대상자가 추가로 혜택을 받게 되어 예산을 제대로 사용하지 못하는 결과를 초래할 수 있다. 이러한 문제점을 해결하기 위한 방안으로 무응답 대체는 필수 불가결하다고 볼 수 있다. 적절한 무응답 대체는 표본의 대표성을 확보할 수 있으며, 통계적 분석 결과에서 편향이 감소하게 되고, 목표 모집단에 대한 통계적 추론인 추정과 검정에서의 오류를 최소화할 것이다. 따라서 이 연구는 보건복지 분야 정책 연구를 하는 데 사용하는 패널자료의 품질 제고를 위해 필요한 기초연구라고 볼 수 있다.

이 연구는 무응답 대체에 대한 새로운 기법 제안 및 지속 가능한 무응답 대체 자료를 제공하는 데 목적을 두고 있다. 먼저 최근 빅데이터 및 정보 산업혁명 시대가 도래함에 따라 인공지능(AI)에 대한 관심이 높아지고

있으며 이와 관련된 연구도 빠르게 진행되고 있다. 인공지능은 기계학습과 딥러닝을 포함하고 있는 가장 큰 영역이다. 기계학습과 딥러닝이 여러 분야에서 활발하게 사용되고 있는 상황에 부합하여 이러한 최신 통계방법을 패널자료의 항목무응답 대체에 적용해 보고자 한다. 이때 기존의 대체 방법과 비교함으로써 적절한 대체 방법을 제시하고자 한다.

다음으로 현재 국내외 패널자료를 생산하고 있는 여러 기관에서는 사용자 편의를 위해서 항목무응답 변수를 적절한 값으로 대체하여 제공하고 있다. 한국노동패널조사와 한국고령화연구패널조사는 기본 자료 제공 이외에 추가로 무응답 대체 자료를 배포하고 있다. 우리 연구원에서 공동으로 주관하고 있는 한국복지패널조사와 한국의료패널조사도 일부 변수에 대해 무응답을 대체하고 있다. 두 개 패널자료의 경우 항목무응답 비율은 높지 않으나, 향후 항목무응답 발생 패턴의 변화(예: 무응답 대체 대상 변수의 추가적인 발생, 무응답 비율의 변화 등)에 대비할 필요가 있다. 이를 위해 국내 및 해외 패널조사의 선행연구를 바탕으로 벤치마킹 가능성을 모색하고자 한다. 또한 다양한 조건을 고려한 모의실험을 실시하여 상황에 따른 최적의 방법을 찾아보고자 한다.

제2절 연구 내용 및 방법

1. 연구 내용

이 연구는 보건복지 분야 패널자료에서의 항목무응답 대체 방법에 대해서 살펴보며, 각 장에서의 연구 내용은 다음과 같다.

제2장에서는 먼저 국내외 패널자료에서 실시하고 있는 항목무응답 대체 방법에 대한 선행연구를 하고자 한다. 국내는 한국복지패널조사, 한국의료패널조사, 한국노동패널조사, 한국고령화패널조사이다. 외국은 영국의 가구패널조사(BHPS: British Household Panel Survey), 미국의 소득역동성패널연구(PSID: Panel Study of Income Dynamics), 독일의 사회경제패널(SOEP: Socio Economic Panel) 그리고 미국의 의료비지출패널조사(MEPS: Medical Expenditure Panel Survey) 등이다.

제3장에서는 패널자료에서의 무응답 구조를 파악하기 위해 필요한 요소인 결측자료 메커니즘과 무응답 패턴에 대해 설명하고자 한다. 그리고 항목무응답 대체 방법에 대해 소개하며 그에 대한 장단점을 알아보고 대체 방법의 효과를 평가할 때 기준이 되는 평가지표도 소개하고자 한다.

제4장에서는 보건복지 분야 패널자료인 한국복지패널자료 및 한국의료패널자료에 근거하여 모의실험을 실시하고자 한다. 대체 대상 변수는 한국복지패널자료의 경우 '경상소득'이며 한국의료패널자료의 경우 '생활비'이다. 모의실험의 목적은 다양한 무응답 비율(5%, 10%, 20%, 30%, 50%)에 따른 항목무응답 대체 방법별 효과를 평가하는 것이다. 항목무응답 대체 방법은 평균대체, 핫덱대체, 비대체, K-최근접 이웃 대체, 랜덤포레스트 대체, 서포트 벡터 머신 대체, 신경망 대체로 총 7가지를 고려한다. 평균대체, 핫덱대체, 비대체의 경우 대체군 활용 여부에 따라 2가지

의 무응답 대체값을 생성한다. K-최근접 이웃 대체, 랜덤 포레스트 대체, 서포트 벡터 머신 대체, 신경망 대체의 경우 앞선 대체 방법에서 대체군을 형성할 때 고려한 보조변수를 사용하여 대체한다. 또한 부가적인 모의실험으로 대체 대상 변수와 관련 있는 정보를 더 포함한 경우에 대한 무응답 대체 효과 효과도 살펴보고자 한다.

선정된 무응답 대체 대상 변수는 대체 방법을 통해 대체하며, 여러 가지 평가지표에 근거하여 대체 방법별 효과를 비교·평가한다. 대체 방법의 효과는 평균 추정치와 참값의 일치 정도를 확인할 수 있는 '편향' 및 평균 추정치를 일관되게 추정하는지 알 수 있는 '제곱근평균제곱오차', 평균 추정치에 대한 95% 신뢰구간에서의 참값의 '포함률', 개별 개체의 대체값과 실제값 간의 유사 정도를 보여 주는 '예측의 정확성', 대체된 자료의 적률이 실제 자료의 적률을 잘 유지하는지 볼 수 있는 '추정의 정확성'으로 평가하였다. 그리고 대체된 자료와 실제 자료 간 분포 비교를 위하여 히스토그램과 상자그림으로 살펴보고자 한다.

제5장에서는 연구 결과를 요약하고 향후 과제를 제안한다.

2. 연구 방법

국내외 패널자료에서 실시하고 있는 항목무응답 대체 방법에 대한 현황을 파악해 본다. 이를 비롯하여 국내외 선행연구, 자료 분석, 전문가 자문회의 등을 활용하여 패널자료에서의 무응답, 항목무응답 대체 방법 그리고 기계학습 통계방법에 대해 살펴보고자 한다. 모의실험에서 사용한 자료는 한국복지패널조사(2019년 공개 자료) 및 한국의료패널조사(2018년 공개 자료)이며, 통계패키지는 SAS와 R을 사용한다.

제 2 장

국내외 패널조사에서의 항목무응답 처리 사례 연구

2 국내외 패널조사에서의 항목무응답 처리 사례 연구

제1절 국내 패널조사

1. 한국복지패널조사

한국복지패널조사는 연령, 소득계층, 경제활동 상태 등에 따른 다양한 인구집단별 생활 실태, 복지 욕구 등을 동태적으로 파악하여 정책 마련과 제도 개선을 위해 활용되고 있다. 2006년 1차 원표본 7072가구를 구축하였으며 2012년 신규 표본 1800가구를 추가 구축하였다. 2018년(13차 조사)에 총 6474가구를 완료하여 원표본유지율은 60.3%이고 신규 표본(7차 조사) 가구유지율은 77.3%인 것으로 나타났다(김태완 외, 2018, p.23, 36).

주요 설문 내용은 가구일반사항, 건강 및 의료, 경제활동 상태, 사회보험, 주거, 생활비, 소득, 가구별 복지서비스 이용 등 다양한 항목별로 구성되어 있다.

가. 한국복지패널조사의 무응답 대체 방법

소득에 대한 대체 방법은 다중대체를 실시하였으며 대체 과정은 다음과 같다(김미곤 외, 2008, p.119). 무응답 변수와 연관성이 있는 변수를 선택한 후 모형을 적합하여 다중대체를 실시하였다. 필요에 따라서는 정규성을 만족하도록 변수변환하며, 대체 후 역변환을 하여 원래 값으로 변

환하였다. 또한 대체된 값이 이산치인 경우에는 반올림하였다. 보통 기본 5개의 대체자료세트가 생성되는데, 한국복지패널자료의 경우 이용자의 편의를 위하여 5개의 대체자료세트 중에서 소득변수를 가장 적절하게 설명하는 대체자료세트 한 개를 선택하였다(김미곤 외, 2008, pp.119-120). 이 선택된 대체자료세트의 무응답 대체 변수를 추출하여 배포용 자료에 붙여서 제공하고 있다.

나. 한국복지패널조사의 무응답률 및 무응답 대체 변수

항목무응답 변수에 대한 무응답률은 따로 제공하고 있지 않다.

무응답 대체 변수는 가구의 경상소득 및 가처분소득이며, 해당 가구의 소득변수가 결측인 경우 가구구분(일반·저소득층 가구)을 할 수 없기 때문에 필수적으로 대체하고 있다(한국보건사회연구원, 2019b, p.94). 대체값과 응답값을 구분해 주는 플래그(flag) 변수가 함께 제공되고 있다.

2. 한국의료패널조사

한국의료패널조사는 국가 보건의료체계의 대응성 및 접근성 향상과 효율화를 위한 정책을 마련하는 데 기초 자료로 활용되고 있다. 1차 조사는 2008년 상반기에 7866가구를 구축한 후 지속적인 패널 유지를 위하여 2012년 2520가구를 신규 유치하였다. 2018년(13차 조사) 6497가구에 대한 조사를 완료하였으며 1차 원표본 유지율은 53.8%이고 2차 원표본 유지율은 72.3%로 나타났다(김남순 외, 2018, p.33).

주요 설문 내용은 가구 및 가구원 정보, 질환, 민간의료보험, 의료 이용, 의약품 복용, 의료비 지출, 건강 관련 인식 및 행태 등으로 구성되어 있다.

가. 한국의료패널조사의 무응답 대체 방법

의료비 관련 생성변수를 만들 때 구성되는 항목 변수에 대해 무응답 대체 방법을 적용한다. 무응답 대체 방법은 평균대체이며 필요에 따라 0 또는 결측치 처리를 한다.

나. 한국의료패널조사의 무응답률 및 무응답 대체 변수

유효값, 모름/무응답, 해당 사항 없음(설문 문항의 응답 대상이 아님) 등과 같이 자세하게 항목을 구분하고 있으며 코드북을 보면 이 항목을 기준으로 변수별 빈도와 비율을 제공하고 있다. 변수별 무응답률은 모름/무응답 항목을 통해 파악 가능하다.

무응답 대체 변수는 보건의료 서비스의 경우 교통비, 요양비, 간병비 등이 있고 의약품의 경우 처방약값, 의약품 구매 등이 있다. 이러한 변수를 대체한 다음에 의료비 변수를 생성한다. 의료비 생성변수는 5개의 가구지출 의료비와 2개의 개인지출 의료비이며, 이를 배포용 자료에 포함하여 제공하고 있다. 각각의 가구·개인지출 의료비에 대한 변수별 세부 항목은 다르게 구성되어 있으며, 자세한 내용은 코드북에 정리되어 있다.

3. 한국노동패널조사

한국노동패널조사는 노동시장 관련 양질의 조사자료를 생산하여 노동시장연구의 활성화를 통해 보다 합리적이고 정확한 노동시장 및 고용정책의 수립과 평가에 기여하는 데 있다. 1998년 총 5000가구를 구축하였으며 패널마모를 보완하기 위해서 2번의 표본 추가가 있었다. 구축된 추

가표본 가구 수는 2009년의 경우 1415가구이고, 2017년의 경우 신규패널 표본구축을 위한 예비표본 6034가구이다(이혜정, 2018a, p.1). 2017년(20차)에 98표본 기준으로 5761가구를 완료하여 원표본유지율은 67.1%이고, 통합표본은 6685가구를 완료하여 원표본유지율은 84.1%로 나타났다(김유빈 외, 2018, p.5, 30).

주요 설문 내용은 가구의 특성과 가구원들의 경제활동 및 노동시장 이동, 소득, 자산 및 소비, 교육 및 직업훈련, 사회생활 등으로 다양한 항목으로 구성되어 있다.

가. 한국노동패널조사의 무응답 대체 방법

무응답 대체 방법은 변수의 특성에 따라 다양한 방법을 사용하고 있다. 가장 간단한 평균대체, 중위수대체, 최빈값대체에서부터 임의대체, 확률기반 대체, 핫덱대체, 회귀대체, 비대체 방법을 이용하고 있다(송주원, 2017, p.3). 대체의 정확성을 높이기 위해서 무응답 대체 변수와 관련된 보조변수를 활용하여 대체군을 생성한 후 대체군 내에서 각각 대체를 실시하였다. 또한 무응답 변수에 대해 대체를 실시할 때 하나의 대체 방법만을 사용하는 것이 아니라 조건에 따라 적합한 여러 가지 대체 방법을 활용하여 대체하였다.

예를 들어 이자소득 대체 과정을 설명하면 다음과 같다. 먼저 은행예금 총액을 응답한 경우와 아닌 경우로 구분한다. 은행예금 총액을 응답한 경우는 해당 연도 이자소득/해당 연도 금융자산총액의 비(ratio)를 이용하여 비대체를 실시한다. 은행예금 총액을 응답하지 않은 경우에는 다시 3가지로 구분하여 대체한다. 첫 번째는 작년도 이자소득이 존재하는 경우로 해당 연도 이자소득/이전 연도 이자소득에 대한 비를 계산하여 비대체

를 실시한다. 두 번째는 작년도 이자소득이 존재하지 않으나 지난 한 달 금융소득이 존재하는 경우에는 지난 한 달 금융소득×12의 값으로 대체한다. 위의 방법들로 무응답이 대체되지 않은 값들은 핫덱대체를 실시한다.

나. 한국노동패널조사의 무응답률 및 무응답 대체 변수

무응답 대체는 주로 소득, 자산, 생활비 등과 같은 금액 변수이며, 금액 변수들을 대체하는 데 사용되는 보조 변수가 무응답인 경우도 함께 대체하였다. 무응답 대체 자료를 기본 자료 제공 이외에 추가로 배포하고 있으며, 무응답 대체 자료에는 플래그 변수가 포함되어 있다. 또한 무응답 대체 자료에 대한 코드북을 제공하고 있으며, 코드북에는 무응답 대체 변수에 대한 대체 방법 정보가 정리되어 있다.

4. 고령화연구패널조사

고령화연구패널 조사의 목적은 고령자의 사회, 경제, 심리, 인구학적 형성 및 건강 상태 등을 파악하여 사회·경제 정책 수립 시 활용될 기초 자료를 생산하는 데 있다(한국고용정보원, 2019a, p.1). 고령화연구패널은 2006년 1만 254명을 대상으로 시작되어 격년 주기로 추적조사하고 있다(한국고용정보원, 2019a, p.1). 5차 조사에서 모집단을 반영하기 위해 기존 패널 외 1962년생과 1963년생을 추가로 추출하였다. 최근 조사된 6차 조사는 7490명이 참여하였고 표본유지율은 78%이다(한국고용정보원, 2019a, p.1).

고령화연구패널의 설문 내용은 가구 배경, 인적 속성, 가족, 건강, 고

용, 소득과 소비, 자산, 주관적 기대감과 삶의 질, 사망자 설문으로 구성되어 있다(한국고용정보원, 2019a, p.1).

가. 고령화연구패널조사의 무응답 대체 방법

고령화연구패널의 무응답 대체 방법은 순차적 회귀(sequential regression) 대체이다(한국고용정보원, 2019b, p.1). 자료 이용자들을 위해 다중 대체 가이드를 제공하고 있으며 대체 방법이 간단히 기술되어 있다. 연속형 변수의 경우 5번의 다중 대체를 수행하였고, 범주형 변수의 경우 단일 대체를 하였다(한국고용정보원, 2019b, p.2).

나. 고령화연구패널조사의 무응답률 및 무응답 대체 변수

고령화연구패널은 범주형 변수에 대해 빈도수를 제공하고 있다. 따라서 '모르겠음' 또는 '응답거부'의 빈도수에 대한 정보는 알 수 있으나, 무응답률은 따로 제공하지 않는다.

인구통계학적 변수, 자산, 소득 등에서 무응답 비율이 있는 변수에 대해 대체를 수행하고 5개의 다중 대체 데이터 세트를 제공하고 있다. 대체된 데이터 세트에는 다중대체 번호를 나타내는 변수와 대체값에 대한 플래그 변수가 포함되어 있다.

5. 소결

현재 국내에서 실시되고 있는 패널조사는 21개 정도이며 이 중에서 한국복지패널, 한국의료패널, 한국노동패널, 한국고령화패널은 항목무응답

에 대한 대체를 실시하고 있다. 4개 패널조사를 제외한 나머지 패널조사는 무응답 대체를 실시하고 있지 않거나 고려 중인 것으로 파악되었다. 이 절에서는 4개의 국내 패널조사에서 실시하고 있는 항목무응답 대체 현황을 살펴보았다.

다음 〈표 2-1〉은 앞서 살펴본 국내 패널조사별 무응답 대체 변수 및 대체 방법을 저자가 다시 정리한 내용이다. 한국노동패널과 한국고령화패널의 경우 변수 항목별 다양한 대체 방법을 활용하여 대체를 실시하고 있으며, 기본 배포용 자료 외에 대체 자료를 추가로 제공하고 있다. 한국복지패널과 한국의료패널의 경우에는 대체 변수 개수가 많지 않으며, 대체 변수를 기본 배포용 자료에 포함하여 함께 제공하고 있다.

패널조사별 무응답 대체 시 대체 변수별 또는 개체별 다양한 대체 방법을 사용하고 있다. 이는 대체 대상 변수에 따른 무응답률, 설문구조, 분포 특성 등이 다르고 개체마다 활용할 수 있는 정보가 각각 다르기 때문이다. 그리고 대체할 때 사용되는 보조변수는 패널자료의 장점이라고 할 수 있는 이전 차수의 정보를 사용하기도 하였으나, 주로 해당 차수의 정보를 활용하였다.

패널자료는 조사의 목적에 따라 특화된 설문 문항으로 구성되어 있어서 다르지만, 기본적이거나 유사한 문항(예: 인구학적 변수, 소득, 자산 등)을 공통적으로 포함하고 있다. 보통 이 문항에 대해 대체하고 있는 것으로 나타났다. 추후 한국복지패널이나 한국의료패널에서 추가적으로 무응답 대체 변수를 처리해야 할 때 참고자료로 활용할 수 있을 것이다.

〈표 2-1〉 국내 패널조사 무응답 대체 현황

패널조사	대체 변수	대체 방법
한국복지패널	경상소득 가처분소득	- 다중대체
한국의료패널	가구지출 의료비 개인지출 의료비	- 평균대체 - 0 대체 - 결측치 처리
한국노동패널	가구 소득 (근로소득, 금융소득, 이전소득 등)	- 개인 근로소득(다른 항목에서 조사된 변수) 사용 - 비대체 - 핫덱대체 - 중위수대체
	생활비 및 저축	- 비대체 - 중위수대체
	자산 및 부채	- 비대체 - 핫덱대체 - 중위수대체
	기타 (주거, 교육비)	- 회귀대체 - 비대체 - 중위수대체 - 논리적 대체
	개인 근로소득	- 회귀대체 - 비대체 - 트리를 활용한 핫덱대체 - 중위수대체
	종업원 수	- 음이항 회귀대체 - 확률적 EM 방법을 활용한 대체 - 비대체 - 중위수대체 - 논리적 대체 - 0 대체
	비정규직 여부	- 트리를 활용한 최빈값 대체
	취업·퇴직 시기	- 트리를 활용한 평균대체 - 분포추정 후 무작위 추출(random sample generating)
한국고령화연구패널	인구학적 변수 자산, 소득 등	- 순차적 회귀대체 - 다중대체

자료: 1) 김미곤 외. (2008). 2008년 한국복지패널 기초분석 보고서. 서울: 한국보건사회연구원.
2) 한국보건사회연구원. (2019a). 2008-2016 연간데이터 한국의료패널 코드북(버전 1.5) 및 변수표. https://www.khp.re.kr:444/web/data/board/view.do?bbsid=55&seq=2092에서 2019. 6. 7. 인출.
3) 한국노동패널조사. (2019). 항목 무응답 flag 변수 정보(20190418). https://www.kli.re.kr/klips/selectBbsNttList.do?bbsNo=96&key=497에서 2019. 9. 16. 인출.
4) 한국고용정보원. (2019b). 고령화연구패널 1-6차 자료 무응답 대체값 산출을 위한 다중대체 안내서. https://survey.keis.or.kr/klosa/klosaguide/List.jsp에서 2019. 9. 16. 인출.

제2절 해외 패널조사

이 절에서는 영국, 미국, 유럽 등 해외 주요 패널에서의 무응답 대체 사례를 조사한다. 해외 주요 패널로는 영국의 가구패널조사(BHPS: British Household Panel Survey), 미국의 소득역동성패널연구(PSID: Panel Study of Income Dynamics), 독일의 사회경제패널(SOEP: Socio Economic Panel), 그리고 미국의 의료비지출패널조사(MEPS: Medical Expenditure Panel Survey)를 중심으로 살펴보고자 한다.

1. 영국의 British Household Panel Survey(BHPS)[3)]

영국의 British Household Panel Survey(BHPS)는 영국 국민과 가계의 사회·경제적 변화를 파악하기 위해 ESRC-UK Longitudinal Centre와 Essex 대학의 사회경제연구원(Institute for Social and Economic Research)이 수행하는 패널조사이다.

BHPS는 16세 이상의 성인을 대상으로 매년 대표성이 있는 표본 5000가구 이상, 대략 1만 명의 개인(개체)을 조사하도록 설계되었다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A2-1). 가구와 개인에 대해 매년 연속적으로 자료를 수집하며 개인이 새로운 가계로 편입될 경우 그 가계의 구성원들도 면접의 대상이 된다. 이로써 동태적 모집단을 대표하는 표본을 설계하였다. 4차 연도부터 11~15세 가구원을 대상으로 조사를 추가하였으며, 1997년과 1999년에 새로운 하위 표본을

3) Marcia Freed, T(ed)., Brice, J., Buck, N., & Prentice-Lane, E. (2018). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. Colchester: University of Essex. Retrieved from http://www.iser.essex.ac.uk/bhps/documentation 2019. 10. 12.를 참고하여 정리함.

추가하였다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A2-1).

패널 설문조사의 첫 3년차의 주요 주제는 가구 구성, 노동 시장, 소득 및 자산, 주거, 건강 그리고 사회·경제적 가치에 국한되어 조사하였지만 이후 건강의 변화, 소득에 따른 라이프 사이클의 변동, 교육 등 폭넓은 항목을 조사하고 있다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A2-1). 현재 18차(2018년)의 데이터가 공개된 가장 최신 자료이다.

가. BHPS의 무응답 대체 방법

BHPS에서는 핫덱대체(hotdeck imputation)와 회귀대체(regression imputation) 두 가지 방법론이 주로 사용되었다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A5-23).

핫덱대체는 범주형 금액 변수(categorical money variables)에 사용하였다. 여기서 범주형 금액 변수란 대리인이 밝힌 개인의 소득 및 배당과 투자로 인한 구간 소득(banded income)을 의미한다. 또한 '복지 혜택으로 인한 소득'처럼 회귀 대체로 그 값을 예측하는 것이 적절하지 않은 변수들에 대해 핫덱대체를 활용하였다. 핫덱대체의 신뢰도는 그 기증자(donor)와 수증자(recipient)가 유사한 특성을 보유할 때 높아진다. 따라서 각 개체들을 정밀하게 그룹화해야 하는데, BHPS에서는 이를 위해 CHAID 기법을 사용하였다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A5-23).

회귀대체는 정량형 금액 변수(amount money variable)에 사용된다. 여기서 회귀대체란 완전히 응답한 개체들을 대상으로 회귀분석을 실시하

고, 모형적합이 가장 잘된 모형을 활용하여 무응답을 대체하는 방법을 말한다. BHPS에서는 대체에 사용한 회귀모형의 모형 적합도(R^2 등)를 제시하여 모형적합에 대한 정보를 보고한다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A5-23).

한편 BHPS는 시점 간 대체(cross-wave imputation)도 수행하였고, 그 과정에 핫덱과 회귀대체 방법을 활용하였다. 시점 간 대체는 시점 내 대체에서 고려하지 않는 변수의 변화에 대해서도 고려해야 한다. 시점 간 변화가 크지 않은 변수들에 대한 가장 좋은 예측치(best predictor)는 동일 개체의 다른 시점의 값이라 할 수 있다. 그러나 시점 간 변화가 상당한 변수에 대해서는 동일개체라 하더라도 위 방식을 사용하기 어렵다. 따라서 이 패널에서는 동일시점에서 그 개체와 가장 유사한 기증자의 값과, 동일 개체의 다른 시점의 값 모두를 비교 분석하여 시점 간 대체를 수행하였다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A5-24).

나. BHPS의 무응답률 및 무응답 대체 변수

BHPS는 각 연차의 응답률(response rates)과 응답 상태(response status)들을 보고하고 있다. 응답률은 1차 연도만 보고하고 있으며 전체 대상자 중 유효한 표본의 비율로 정의하였다. 1차 연도 모든 인터뷰에 응답한 개인은 92%였다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A4-28). 또한 2차 연도부터는 응답 상태를 보고하고 있다. 응답 상태는 전년 대비 금년의 응답자 비율을 뜻한다. BHPS의 2018년도 응답 상태는 92%로 2017년도에 모든 인터뷰에 응답한 7755명 중 7124명(92%)이 2018년에 응답하였다(Marcia Freed, Brice, Buck & Prentice-Lane, 2018, p.A4-45).

항목무응답의 경우 모든 차수의 변수들에 대한 코드북과 빈도수를 BHPS 홈페이지에서 온라인상으로 확인할 수 있다.

무응답은 주로 소득과 주택비용처럼 개인에게 민감할 수 있는 변수에서 주로 발생하였다. 따라서 BHPS는 소득과 같은 변수에 집중적으로 수행되었다. 참고로 대체된 자료를 사용하는 것을 원하지 않는 사용자를 위하여 플래그 변수를 달아 두었다. 차수별로 대체된 변수와 해당 변수의 플래그 변수는 BHPS 매뉴얼에 기술되어 있다.

2. 미국의 Panel Study of Income Dynamics(PSID)[4)]

미국의 통계청은 린든 존슨 대통령의 "가난과의 전쟁(War on Poverty)" 정책을 평가하기 위해 1996년 약 3만 가구를 대상으로 경제기회조사(SEO: Survey of Economic Opportunity)를 수행하였다. 이후 미시간 대학 조사연구센터(SRC: Survey Research Center)에서 미국의 전체 모집단을 대표할 수 있는 표본 5000가구에 대한 추적조사를 수행하였는데 이 조사가 Panel Study of Income Dynamics(이하 PSID)이다.

PSID는 총 4802가구, 1만 8233명에서 시작되었으며, 성인이 되어 원가구로부터 독립한 가구원들도 추적조사를 함으로써 시간이 지남에 따라 대표성을 유지할 수 있도록 하였다. 이에 따라 최근 조사된 2017년에는 총 9607가구, 2만 6445명을 대상으로 조사를 하였다. 수집되는 자료는 고용, 소득, 자산, 지출, 건강 행위와 건강 상태, 기초 자료 등이 있다.

4) University of Michigan(Institute for social research, Survey research center). (2019). *Panel Study of Income Dynamics(PSID)*. Retrieved from https://psidonline.isr.umich.edu/default.aspx 2019. 10. 12. 참고하여 정리함.

가. PSID의 무응답 대체 방법[5)]

1) 자산(wealth)

PSID의 자산은 사업(또는 농장) 자산, 거래계좌, 담보대출, 부동산 자산, 주식 자산 등 9가지 항목으로 질문하였다. 이들 중 8가지 항목에 대해서는 핫덱으로 대체하였다. 나머지 1가지 항목은 주택의 순가치로 주택가치에서 주택대출액을 빼어 계산한다. 주택의 순가치는 두 변수로 계산해야 하므로 두 변수의 결측 여부의 조합에 따라 그룹을 나누었고 대체방법 적용 여부에 따라 순차적으로 대체 방법을 적용하였다. 우선 이전 차수 값과 브래킷(bracket) 정보를 이용하여 대체한다. 이때 브래킷 정보는 구체적인 금액을 답할 수 없는 경우, 금액에 대하여 범위로 응답한 변수이다. 만약 이전 차수 값과 브래킷 정보를 이용할 수 없는 개체에 대해서는 이전 차수의 이전 값을 이용하여 대체한다. 이 경우도 불가능한 개체에 대해서는 대출과 주택의 가치의 관계를 가정하여 계산하였다. 예를 들어 주택의 가치는 응답하였으나 대출 변수가 무응답인 경우 주택의 가격에 0.6을 곱하여 대체하였다.

2) 소비와 지출(Consumption and expenditures)

소비와 지출에는 식료품 지출, 집 유지 관리비, 가구 설비와 장비, 의료비, 여행, 여가, 건강관리비, 주거비, 교통비, 교육비, 양육비가 조사되었다. 소비와 지출에서는 크게 두 가지 대체 방법으로 무응답 대체를 수행

5) Duffy, D. (2011). *2007 PSID income and wage imputation methodology*. In PSID Technical Series Paper# 11-03. Survey Research Center-Institute for Social Research, University of Michigan Ann Arbor.

하였다.

첫 번째, 열린 브래킷(unfolding bracket) 정보가 조사된 경우 금액의 구간을 의미하는 브래킷을 조건부로 하여 조건부 평균대체를 수행하였다.

두 번째, 브래킷 정보가 없는 경우는 회귀대체 모형을 사용하였다. 대체모형은 지출 카테고리별로 나누어 최소제곱법으로 적합하였다. 이때 연령을 3차 다항식으로, 가구 수는 비제한적 스플라인(unrestricted spline)으로 모형에 포함하였다.

3) 소득(Income)

소득에서 가구 총소득변수는 소득의 항목들과 근로시간, 임금변수들을 이용하여 생성한다. 생성변수를 만드는 데 관여된 변수들을 대체하였으며 소득에 대한 대체 방법은 2011년에 출간된 2007 PSID 소득과 임금의 대체 방법 문서에 자세히 기술되어 있다.

소득에서 무응답 대체는 각 항목의 특성에 따라서 대체 방법을 달리하였는데 아래와 같은 다양한 대체 방법들을 사용하였다.

- 다른 변수 이용: PSID에서 수집된 다른 변수의 값을 이용. 예) 총급여 또는 임금은 개인의 급여 또는 임금으로 계산
- 이전 차수의 값을 이용: 이전 소득에 인플레이션을 보정하여 현 차년도의 소득을 대체
- 그룹 평균 대체: 같은 직업 또는 산업 종사자들의 관측값의 평균으로 대체
- 핫덱대체: 정의된 관측 데이터로부터 임의로 선택한 값을 대체
- 중위수대체: 관측값들의 중위수를 대체

소득 항목의 많은 변수들은 중위수대체를 수행하였다. 그러나 비슷한 질문이 있는 경우 또는 이전 차수에 비해 큰 변화가 없을 것으로 간주되는 소득변수들에 대해서는 최대한 비슷한 변수의 값 또는 이전 값을 이용하여 대체하였으며, 이러한 값들도 무응답인 개체에 대해서는 중위수로 대체하는 등 순차적으로 대체하였다.

소득 항목에서는 전반적으로 중위수대체를 하였는데 사업 또는 농장의 순수익 항목에서는 중위수대체가 아닌 핫덱으로 대체하였다. 그 이유는 순수익의 값 중 -99만 9997달러는 순손실이지만 금액을 알 수 없는 경우를 의미한다. 따라서 순수익 변수는 핫덱대체로 음의 소득값을 갖는 기증자의 값을 대체할 수 있도록 하였다.

마지막으로 근로시간과 관련된 변수들에 대한 대체는 3단계 대체 과정을 거쳤다. 1단계는 대체 전 자료 에디팅이며, 2단계는 대체 단계로 일정한 상수 값을 대체하거나 평균대체 등을 하였다. 3단계는 대체된 값이 시간에 대한 변수이므로 대체값이 합당한 값인지 검토하여 수정하는 단계이다.

이처럼 PSID의 무응답 대체는 하나의 대체 방법을 사용하지 않았다. 변수의 특성에 맞는 대체 방법으로 대체하였다. 또한 하나의 변수에 대해서도 동일한 대체 방법을 적용하는 것이 아니라 정보 사용이 가능한 개체에 대해서는 정보를 사용하여 대체하고 불가능한 경우 중위수대체를 하였다.

나. PSID의 무응답률 및 무응답 대체 변수

PSID의 응답률은 전 차수에서 조사를 완료했던 가구 중 당해 차수에서 조사를 완료했던 가구의 백분율로 계산되었다. 최근 조사된 2017년 기준

전체 응답률은 88.8%이며, 원표본(main PSID)의 응답률은 90.1%, 이주 표본(immigrant sample)이 83.6%, 2017년 이주 표본은 75.7%였다(Institute for Social Research, University of Michigan, 2019).

PSID에서 가장 중요한 질문(자산, 소득, 소비)에 대해 각 변수의 대체 방법과 대체된 개체의 빈도수 등을 이용자 가이드 부록으로 제공하고 있다. 최근 공개된 이용자 가이드는 2017년 자료에 대한 것이며, 이 중 소득 항목에서 대체할 필요가 없는 가구는 71.91%이며, 하나의 변수가 대체된 경우는 16.47%, 8개의 변수가 대체된 가구는 0.04%였다(Institute for Social Research, University of Michigan, 2019). 모든 변수 각각에 대한 항목무응답률은 PSID의 웹페이지에서 찾아볼 수 있다.

PSID는 자료 이용자들의 편의를 위해 조사된 원래의 변수들을 이용하여 주요 생성변수 - 소득, 근로시간, 임금, 재산, 소비와 지출 등의 항목 - 를 만들어 제공한다. 이때 생성변수를 만드는 데 사용되는 변수들을 대체하였다.

각 변수들을 대체할 때 변수의 특성에 따라 대체 방법을 다르게 적용하였고 개체(record)별로도 적용하는 방법이 불가능할 수 있으므로 순차적으로 대체 방법을 달리 적용하였다.

3. 독일의 Socio-Economic Panel(SOEP)[6]

독일의 Socio-Economic Panel(SOEP)는 독일 가구에 대한 대표적인 종단면(longitudinal)연구로 “사회 정책의 미시적 분석의 근간(Microanalytic Foundations of Social Policy)”을 목적으로 1984년

6) German Institute for Economic Research(DIW Berlin). (2019). *Socio-Economic Panel (SOEP)*. Retrieved from https://www.diw.de/en/soep 2019. 10. 12. 참고하여 정리함.

에 시작되었다. 독일의 경제연구기관(IDW Berlin)이 주관하고 있다. 현재는 대략 1만 5000가구와 3만 명의 가구원이 SOEP 조사에 참여하고 있다. 자료 수집의 목적은 가계 구성, 직업, 고용, 수입, 건강 및 삶의 만족도 등 독일 국민들의 전반적인 삶의 질을 이해하는 데 있다(Goebel et al., 2019).

SOEP의 표본은 독일에 거주하는 모집단을 대표하기 위해 비교적 복잡하게 설계되었다. SOEP 표본은 1984년 서독 거주자 대상인 표본 A(4528가구), 외국인 가구 대상인 표본 B(1393가구)에서 1990년 동독 지역으로 표본을 확대하는 등 1990년 통일 이전의 독일민주공화국(GDR)에 대한 자료와 1995년 이후 독일 사회에서 일어나는 인구의 변화를 설명하기 위한 이민자 정보까지 담고 있다. 한편 패널 이탈(panel attrition)을 보완하기 위해 추가적으로 표본을 추출하는데 현재 표본N까지 조사하여 총 3만 1384가구를 대상으로 조사하고 있다(Kara, Zimmermann, & SOEP Group, 2018).

가. SOEP의 무응답 대체 방법[7)]

SOEP에서 무응답은 단위 무응답(unit non-response)과 응답의 논리적 오류가 원인인 경우도 있었으며, 일반적인 형태는 항목무응답이었다. 무응답 대체에 앞서 단위 무응답과 논리적 오류는 각각 가중치 보정법(weighting procedures)과 에디팅(editing) 과정으로 해결하였다. 여기서 에디팅이란, 응답한 값에 논리적 오류가 있을 때 적절한 값으로 변환하는 작업을 의미한다. 이후 항목무응답을 대체하였다.

---

7) Grabka, M. M., & Westermeier, C. (2015). *Editing and multiple imputation of item non-response in the wealth module of the German Socio-Economic Panel*. SOEP Survey Papers Series C, 272.

자산 항목은 먼저 필터변수(filter variables)로 질문을 한 후 자산 또는 부채가 있는 경우, 해당 하위 질문으로 세부적인 자산과 부채에 대한 질문을 하였다. 따라서 무응답 대체도 우선 필터변수를 대체한 후 대체된 필터변수의 결과에 따라 하위질문인 수치형 변수들을 대체하였다.

1) 필터변수의 대체 방법

필터변수에 대한 대체는 로지스틱 회귀를 이용하여 단일 대체(single imputation)를 수행하였다. 항목마다 항목무응답과 부분 단위 무응답(partial unit non-response)을 구분하여 모형적합을 하였다. 모형에는 개인 수준의 연령과 성별 그리고 가구 수준에서 광범위한 공변량을 포함하였으며, 모형별로 공변량은 조금씩 상이하다. 예측값이 0.5보다 작으면 자산을 소유하지 않은 것으로 가정하여 필터변수를 대체하였다.

2) 하위 문항의 대체 방법

하위 문항의 수치형 변수(metric variables)들은 다중 대체(multiple imputation) 방법으로 대체되었다. 다중 대체는 크게 두 가지 방식으로 진행하였다. 기본 방법인 행과 열 방법과 연쇄방정식에 의한 다중 대체(MICE: multiple imputation by chained equations) 방법이다.

행과 열 방법(row-and-column method)은 패널조사에서 항목무응답 대체 기법으로 Little & Su(1989)가 제안하였다. SOEP에서 이 방법은 대체하려는 개체의 다른 차수 자료의 정보를 이용해 대체를 수행할 때 사용하였다. 이때 행의 효과는 전체 개체에 대해 모든 기간에서의 자료를 이용하는 것이며 열의 효과는 시간의 흐름에 따른 정보를 포함한다.

만약 이전 차수의 자료를 이용할 수 없는 경우 대안으로 동일 차수 내에서는 연쇄방정식에 의한 다중 대체를 사용하였다. 또한 결측값이 존재하는 변수도 예측 방정식에 포함하는 'Fully conditional specification' 방법을 사용하였다. 행과 열 방법은 자산 관련 변수의 대체 방법 중 50%를 차지하고 있다.

위와 같은 방법으로 항목무응답을 대체하였을 때 2012년 순가치의 지니계수를 산출한 결과, 대체하지 않은 자료(유효 관측치 수 =11168)의 지니계수보다 대체 후 자료(대체 후 관측치 수=18536)의 지니계수가 8.2% 포인트 더 작은 것으로 보고되었다(Grabka & Westermeier, 2015).

나. SOEP의 무응답률 및 무응답 대체 변수

SOEP는 개인 수준의 자산(wealth)에 대한 자료를 2002년부터 수집하기 시작하였으며 2007년과 2012년에 추가적으로 정보를 수집하였다. 특히 2002년부터 고소득자 표본을 추가적으로 추출하여 조사하였다. SOEP에서 자산은 민감한 정보로 연령, 성별, 이민 여부 등과 같은 인구학적 정보보다 더 높은 무응답률을 보인다(Grabka & Westermeier, 2015). 자산과 관련된 항목 중 무응답률은 2002, 2007, 2012년이 비슷한 패턴을 보인다. 세 개의 조사 차수 모두 개인보험 항목의 무응답률이 13%가량으로 가장 높았고 금융자산이 약 8%, 그 외는 5% 이하였다(Grabka & Westermeier, 2015).

무응답률이 가장 높은 항목인 자산과 소득에서 다중 대체를 하였고, 플래그 변수를 포함한 데이터 세트를 제공하고 있다. 자산 항목은 자산과 부채 부문으로 나누어져 있다. 자산에 관한 질문은 주택의 시장가치, 다

른 소유물, 금융자산, 빌딩론(building-loan) 계약, 개인보험, 사업체 자산, 유형자산이 있다. 부채에 관한 질문은 주택의 시장가치 부채, 다른 소유물에 대한 부채, 신용 대출로 이루어져 있다. 반면 키와 몸무게 등은 무응답률이 높지 않고, 지난 차수의 값을 이용한 대체값을 제공하고 있다.

4. 미국의 Medical Expenditure Panel Survey(MEPS)[8)]

Medical Expenditure Panel Survey(MEPS)는 1996년부터 미국 보건복지부(HHS: Department of Health and Human Service) 산하 의료관리품질조사국(AHRQ: Agency for Healthcare Research and Quality)이 수행하고 있다. MEPS는 표본 추출된 국민들을 대상으로 보건의료지출, 건강보험급여, 국민 건강 상태 그리고 인구학적 변수 등을 조사하고 있다.

MEPS의 조사 대상자(패널)는 전년도 National Health Interview Survey(NHIS) 조사의 응답자로 한다. 따라서 매년 패널을 추가하며, 각 패널은 2년간 총 5회의 조사를 받는다. 예를 들어 1996년부터 조사가 시작되었으므로 그해에 추출한 첫 패널(패널1)은 2년간 조사하고 1997년부터 패널2 표본을 추가로 추출하여 2년간 조사하는 등 매년 표본을 추가한다. 현재 패널21에 대한 조사까지 진행되었다. 한 해에 표본을 오버랩하면서 표본 증가에 따라 통계적 검정력(power)이 높아진다. 1996년 8588가구(패널1)를 조사하였고 1997년에는 1만 2986가구(패널1과 패널2)를 조사하였으며, 최근 공표된 2016년은 1만 3492가구(패널20과 패널21)를 조사하였다.

---

8) Agency for Healthcare Research and Quality(AHRQ). (2019a). *Medical Expenditure Panel Survey(MEPS).* Retrieved from https://www.meps.ahrq.gov/ 2019. 10. 12. 참고하여 정리함.

MEPS는 총 4가지 영역으로 나누어 조사하며 가구 영역(HC: Household Component), 의료제공자 영역(MPC: Medical Provider Component), 보험 영역(IC: Insurance Component), 너싱홈 영역(NHC: Nursing Home Component)으로 구성되어 있다.

- 가구 영역(HC)은 가구 및 가구원을 대상으로 조사하고 있으며 모든 정보의 핵심 영역이다. 가구 영역에서 얻은 정보를 토대로 의료제공자 및 보험 영역으로부터의 자료를 보완한다.
- 의료제공자 영역(MPC)은 대표성 있는 표본을 추출한 것은 아니며 가구 영역을 에디팅하고 대체하기 위한 목적으로 조사되었다. 가구 영역의 응답자가 이용한 의료기관 및 약국에 대해 전화와 우편으로 조사한다. 조사의 내용은 의료기관의 기본정보, 의료이용자의 입원기간, 상병, 의료비 지출 등이 있다. 의료제공자 영역의 자료는 가구 영역과 통합하여 제공하고 있다.
- 보험 영역(IC)은 민간과 공공 부문의 고용주를 추출하여 고용주와 직원에 대한 보험 정보를 조사하였다. 보험 영역의 자료는 요약표 형태로 제공하고 있으며 원자료는 제공하지 않는다.
- 너싱홈 영역(NHC)은 너싱홈에 대한 조사로 1996년에만 실시하였다.

가. MEPS의 무응답 대체 방법[9)]

의료비 지출은 의료 이용자가 직접 지불하는 금액뿐만 아니라 제3의 의료비 지급자(예, 민간보험, 메디케어, 메디케이드 등)에 의해 지불되는 금액이 있다. 따라서 의료비 지출지(source of payment)를 10가지로 구분하였고, 모든 의료비 지출지에서 지출된 금액을 합산하여 총지출 금액을 정의하였다. 여기서 10개의 의료비 지출지는 의료이용자, 민간보험(private insurance), 메디케어(medicare), 메디케이드(madicaid), 제대 군인 의료처(Veteran's Administration) 등이 있다.

이처럼 의료비 지출을 상세하게 파악하기 위해선 가구 또는 가구 구성원이 장기간에 걸쳐 광범위한 기록을 유지해야 하는데, 이것이 쉽지 않기 때문에 의료비 지출에서 무응답이 흔히 발생한다. 또한 의료비 지출의 거래 행태가 의료제공자와 제3의 의료비 지급자 사이에서만 나타날 때 가구 영역에서는 그에 관한 정보를 알지 못하는 경우도 있다.

따라서 의료비 지출을 추정할 때는 가구 영역(HC) 자료뿐만 아니라 의료제공자 영역(MPC)의 자료를 이용하며 만약 무응답으로 인해 자료를 이용할 수 없는 경우에는 무응답 대체를 수행한다. 무응답 대체는 가중 축차 핫덱(weighted sequential hotdeck)대체 방법을 사용하였다. 우선 기증자와 수증자 간 매칭을 위해서 보험 종류, 입원 일수, 병원 방문 이유, 응급실 방문 여부, 지역 등의 변수를 활용하였다.

MEPS 대체 방법은 (1) 모든 지불의 자료원이 전체 결측인지 또는 (2) 부분 결측인지와 (3) 의료이용에 대한 총비용이 보고되었는지에 따라 달

9) Machlin, S. R. and Dougherty, D. D. (2007). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. *Method-ology Report 19*. Agency for Healthcare Research and Quality, Rockville, Md. Retrieved from http://www.meps.gov/mepsweb/data_files/pub-lications/mr19/mr19.pdf.

라질 수 있다. MEPS에서 의료비 지출에 대한 항목무응답은 완전(full) 무응답과 부분(partial) 무응답 두 가지 형태로 구분한다. MEPS가 발간한 무응답 관련 매뉴얼에서는 다음의 세 가지 예를 활용하여 핫덱대체의 이해를 돕고 있다(Machlin & Dougherty, 2007).

먼저 완전 무응답의 경우이다. 어느 환자가 입원하였고, 그 환자는 메디케어와 민간보험을 활용하였을 때다. 이 상황은 의료비 거래 행위가 의료비를 지급하는 제3자와 의료제공자 간에 일어나므로 총의료비 지출에 대해 응답할 수 없는 경우이다. 입원 환자 중 해당 환자와 비슷한 기증자를 선별한 후 그 기증자의 의료비 지출 내용을 그대로 대체한다([그림 2-1] 참조).

[그림 2-1] 예시 1. 완전 무응답 대체

Example 1. Complete imputation

Payment source	Donor	Recipient (pre-imputation)	Recipient (post-imputation)
Medicare	$1,840	Missing	$1,840
Private insurance	$792	Missing	$792
Total expenses	$2,632	--	$2,632

자료: Machlin, S. R., & Dougherty, D. D. (2007). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. *Methodology Report 19*. Agency for Healthcare Research and Quality, Rockville, Md. Retrieved from http://www.meps.gov/mepsweb/data_files/publications/mr19/mr19.pdf. p.5, Example 1 표 인용.

다음의 두 예는 부분 무응답의 경우다. 먼저 자신이 지출한 의료비용(5달러)에 대해서는 응답하였으나 보험사가 지급한 비용은 알 수 없는 경우이다. 이 환자와 가장 유사한 기증자를 선택하고, 그 기증자의 총의료비 지출(997달러)을 이용한다. 즉, 기증자의 총의료비 지출에서 해당 환자 부담분(5달러)을 차감하여 보험사의 부담분(992달러)을 대체한다([그림 2-2] 참조).

[그림 2-2] 예시 2. 부분 무응답 대체

Example 2. Partial imputation

Payment source	Donor	Recipient (pre-imputation)	Recipient (post-imputation)
Out of pocket	$26	$5	$5
Private insurance	$971	Missing	$992
Total expenses	$997	--	$997

자료: Machlin, S. R., & Dougherty, D. D. (2007). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. *Methodology Report 19*. Agency for Healthcare Research and Quality, Rockville, Md. Retrieved from http://www.meps.gov/mepsweb/data_files/publications/mr19/mr19.pdf. p.5, Example 2 표 인용.

그다음으로 의료기관이 청구(charge)한 금액은 알고 있으나, 다른 비용에 대해서는 무응답이 있을 때 기증자의 청구금액과 의료비 지출의 비로 결측을 대체할 수 있다([그림 2-3] 참조).

[그림 2-3] 예시 3. 총비용을 이용한 대체

Example 3. Imputation using total charge

Payment source	Donor	Recipient (pre-imputation)	Recipient (post-imputation)
Total charges	$5,171	$4,173	$4,173
Total expenses	$4,248	missing	$3,421
Medicare	$3,411	missing	$2,737
Private insurance	$837	missing	$684

자료: Machlin, S. R., & Dougherty, D. D. (2007). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. *Methodology Report 19*. Agency for Healthcare Research and Quality, Rockville, Md. Retrieved from http://www.meps.gov/mepsweb/data_files/publications/mr19/mr19.pdf. p.5, Example 3 표 인용.

또한 가구 영역(HC)에서 인구통계학 변수들, 소득, 고용 등이 가중 축차 핫덱대체 방법으로 대체되었다. 대체 시 보조변수는 연령, 교육수준, 고용 상태, 인종, 성별, 종교 등이 사용되었는데 설문의 구조와 특성에 따

라 보조변수와 대체 과정은 상이하였다. 자세한 내용은 가구 영역(HC) 자료에 대한 문서에 기술되어 있다(Agency for Healthcare Research and Quality, 2019b).

나. MEPS의 무응답률 및 무응답 대체 변수

아래 [그림 2-4]는 MEPS의 Methodology Report #19(Machlin & Dougherty, 2007)에 제시된 표를 발췌한 것이다. 2001년 서비스 타입별 의료이용에 대한 응답률을 나타낸다. 응급실 방문의 경우 가구 영역(HC)의 응답률은 8.1%에 불과하였고 의료제공자 영역(MPC)의 응답률은 47.9%였다(Machlin & Dougherty, 2007).

[그림 2-4] 2001년 MEPS에서 서비스 타입별 의료이용에 대한 응답률

Table 1. Distribution of source of expenditure data for survey-reported health care events, by type of service, 2001 MEPS

		Hospital events				
	Office visits	Outpatient visits	Emergency room visits	Inpatient stays	Dental visits[1]	Home health[2]
Number of events	142, 793	15, 763	5, 904	3, 405	26,438	3,155
	Percent distribution by source of data[3]					
Total	100.0	100.0	100.0	100.0	100.0	100.0
MPC	27.9	46.7	47.9	61.4	--	42.3
HC	17.5	6.2	8.1	3.7	47.1	9.4
Imputed: Partial[4]	19.2	8.2	9.7	4.9	11.8	0.1
Imputed: Full	35.3	38.9	34.3	30.0	41.1	48.2

[1]Dental care providers are not surveyed in the MEPS Medical Provider Component, so MPC category is not applicable.
[2]Expense data for home health are collected on a monthly rather than a per visit basis.
[3]Percentages for office visits do not add to exactly 100.0 due to rounding.
[4]Includes events where expense information was imputed for some but not all payment sources.

자료: Machlin, S. R., & Dougherty, D. D. (2007). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. *Methodology Report 19*. Agency for Healthcare Research and Quality, Rockville, Md. Retrieved from http://www.meps.gov/mepsweb/data_files/publications/mr19/mr19.pdf. p.4, Table 1 인용.

나이, 소득 등의 변수에서 결측값이 발생했을 때 에디팅 또는 핫덱대체를 수행하였다. MEPS에서는 특히 가구 영역 이벤트 파일을 제공하고 있다. 이것은 의료이용 시 조사된 자료이며 처방전에 대한 파일, 치과 방문, 외래 방문, 응급실 방문 등 의료이용 형태에 따라 파일을 구분하여 제공하였다. 각 파일에서 의료비는 민감한 자료이므로 무응답률이 다른 변수보다 비교적 높은 편이다. MEPS는 의료비의 무응답을 대체하여 대체값과 대체 방법에 대한 플래그 변수를 데이터 세트에 함께 제공하고 있다.

6. 소결

이 절에서는 해외 주요 패널조사인 BHPS, PSID, SOEP, MEPS의 항목무응답 대체 방법 사례를 조사하였다. 해외의 주요 패널자료를 보면 공통적으로 인구학적 변수들과 같이 덜 민감한 자료는 항목무응답률이 낮았다. 더욱이 연령, 인종, 성별 등과 같이 변동이 없는 변수들은 무응답 대체에 앞서 이전 차년도 자료를 이용하여 에디팅이 가능하였다. 그러나 소득, 자산, 지출과 같은 민감한 변수는 항목무응답률이 높았다. 따라서 이러한 민감한 변수에 대한 무응답 대체를 수행하였다.

소득, 자산, 지출, 비용 등의 금액 변수는 치우친 분포이므로 내재적 모형을 이용하는 핫덱대체 또는 중위수대체 방법을 많이 사용하였다. 회귀모형 등의 명시적 모형을 사용하는 경우는 변환하여 모형에 적합하였다. PSID는 지출에서 조건부 평균대체를 하기도 하였는데, 이 경우는 브래킷 정보가 주어진 경우로 평균대체 시 이상치의 영향이 덜하다.

앞서 살펴본 해외 패널조사 무응답 대체 방법을 저자가 다시 정리하면 〈표 2-2〉와 같다. 이와 같이 패널조사에 따라 다양한 방법으로 무응답을 대체하였다. 더구나 같은 조사 자료에서도 변수에 따라 다른 방법을 적용

하기도 하였다. 이러한 이유는 설문의 설계, 구조 또는 변수의 특성에 따라 대체 방법이 다르기 때문이다. 또한 모든 레코드에 반드시 동일한 방법을 적용해 무응답을 대체하는 것이 아니며 레코드마다 이용 가능한 정보가 다를 수 있으므로 최대한 가능한 정보를 이용하도록 하였다. PSID에서 소득은 다른 항목에서 비슷하게 조사된 변수를 이용하거나 이전 차수의 값을 이용하는 등 최대한 정보를 고려하여 대체하였다. 만약 이러한 정보도 이용할 수 없는 레코드는 중위수 대체를 하였다. MEPS는 가구응답자에서 의료비의 값이 무응답인 경우 의료제공자의 자료를 이용하거나 이를 이용할 수 없는 가구응답자는 의료비의 무응답을 가중 축차 핫덱대체 방법으로 대체하였다.

이처럼 항목무응답 대체는 설문의 구조와 패널조사의 설계, 자료의 분포 등을 고려하여 각 상황에 맞는 적절한 대체 방법을 선정한다. 따라서 무응답 대체에 앞서 자료의 분포 등을 파악하고 패널조사 자료를 이해하는 과정이 필요하다.

〈표 2-2〉 해외 패널조사 무응답 대체 현황

패널조사	국가	대체 변수	대체 방법
BHPS	영국	소득과 주택비용	- CHAID를 이용한 가중 축차 핫덱대체 - 확률적 회귀대체
PSID	미국	자산	- 핫덱대체 - 이전 차년도 값 이용
		소비와 지출	- 조건부 평균대체 - 회귀대체(다항식, 스플라인 항 포함)
		소득	- 다른 항목에서 조사된 변수 이용 - 이전 차년도 값 이용 - 그룹 평균대체 - 핫덱대체 - 중위수대체
SOEP	독일	자산과 소득 (필터 변수)	- 로지스틱 회귀를 이용한 단일 대체
		자산과 소득 (연속형 변수)	- 기본적인 방법: 행과열 방법을 이용한 다중 대체 - 대안의 방법: 연쇄방정식에 의한 다중 대체(MCMC)
MEPS	미국	인구통계학 변수들, 소득, 고용	- 가중 축차 핫덱대체
		의료비	- 가중 축차 핫덱대체

자료: 1) Marcia Freed, T(ed)., Brice, J., Buck, N., & Prentice-Lane, E. (2018). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices*. Colchester: University of Essex. Retrieved from http://www.iser.essex.ac.uk/bhps/documentation 2019. 10. 12.

2) University of Michigan Institute for social research, (Survey research center). (2019). *Panel Study of Income Dynamics(PSID)*. Retrieved from https://psidonline.isr.umich.edu/default.aspx 2019. 10. 12.

3) German Institute for Economic Research(DIW Berlin). (2019). *Socio-Economic Panel (SOEP)*. Retrieved from https://www.diw.de/en/soep 2019. 10. 12.

4) Agency for Healthcare Research and Quality(AHRQ). (2019a). *Medical Expenditure Panel Survey(MEPS)*. Retrieved from https://www.meps.ahrq.gov/ 2019. 10. 12.

5) Machlin, S. R., & Dougherty, D. D. (2007). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. *Methodology Report 19*. Agency for Healthcare Research and Quality, Rockville, Md. Retrieved from http://www.meps.gov/mepsweb/data_files/publications/mr19/mr19.pdf.

제 3 장

결측자료 메커니즘과 항목무응답 대체 방법

3 결측자료 메커니즘과 항목무응답 대체 방법

제1절 무응답 분류 및 패턴, 결측자료 메커니즘[1)]

패널조사에서도 일반적인 설문조사와 마찬가지로 무응답이 발생하기 마련이다. 패널응답자의 접촉 부재, 개인 사정 및 조사 참여 거절 등으로 조사를 실패하는 경우, 조사에는 참여하였으나 일부 설문 문항(소득과 자산 등 금액 관련 질문 또는 민감한 질문)에 대해 응답거절이나 모르겠음이라고 하는 경우가 해당된다. 이러한 상황으로 인하여 특정 조사 시점에서의 변수값이 측정되지 않아서 알 수 없는 경우를 총칭하여 패널자료에서의 무응답이라고 한다.

이 절에서는 패널자료의 무응답 구조를 파악하기 위하여 무응답 분류와 패턴, 결측자료 메커니즘(missing data mechanisms)에 대해 살펴보고자 한다.

1. 패널자료에서의 무응답 분류 및 패턴

패널자료에서의 무응답은 단위무응답(unit nonresponse)과 항목무응답(item nonresponse)으로 나뉜다. 단위무응답은 어느 조사 시점에서 응답자가 조사에 참여하지 않아서 자료 전체에 대한 정보가 모두 누락된 것이다. 항목무응답은 응답자가 조사에는 참여하여 그 조사 시점의 자료는 존재하나 전체 설문 문항 중 일부 문항 값에 대한 정보가 누락된 것

1) 이혜정(2018b), pp.8-10를 인용하여 재구성하였음.

을 의미한다. 보통 값이 있는 자료만을 가지고 분석하는데, 무응답 분포가 응답 분포와 같지 않은 경우 분석 결과에 편향(bias)이 발생하여 잘못된 연구 결과를 초래할 수 있다(이혜정, 2018b, p.8). 단위무응답은 가중치로 조정할 수 있으며, 항목무응답은 적합한 대체 방법을 사용하여 무응답을 대체값으로 채워 넣음으로써 완전한 형태의 자료로 구성하여 해결할 수 있다(이혜정, 2018b, p.8).

패널자료의 무응답 패턴(pattern)에는 단조(monotone) 패턴과 비단조(nonmonotone) 패턴이 있다. 단조 패턴은 패널 응답자가 조사 참여를 중단한 이후 다시는 참여하지 않는 경우를 의미한다(이혜정, 2018b, p.9). 하지만 비단조 패턴은 단위무응답이 간헐적으로 발생하는 형태로, 즉 현 조사 차수에 참여하지 않았더라도 다음 조사 차수에서는 참여하는 것을 의미한다(이혜정, 2018b, p.9).

2. 패널자료에서의 결측자료 메커니즘

무응답 자료를 이해하기 위해서는 결측자료 메커니즘 구조를 파악해야 한다. 결측자료 메커니즘은 무응답과 자료에 있는 변수와의 관계를 의미한다. 패널자료에서의 결측자료 메커니즘은 횡단면 자료의 결측자료 메커니즘과 비슷한 완전임의결측, 임의결측, 비임의결측으로 구분할 수 있다(Little & Rubin, 2002; 이혜정, 2018b, p.9 재인용). 3가지 결측자료 메커니즘에 대해 자세하게 살펴보았다.

결측자료 메커니즘은 결측이 발생할 확률을 모형화한 식 $f(R|Y,X,\phi)$로 표현할 수 있다. 여기서 $Y=(y_{ij})$는 $(N\times p)$ 행렬 자료이며 Y_{obs}는 Y의 응답값들, Y_{mis}는 Y의 무응답이라고 정의한다(이혜정, 2018b, p.9). y_{ij}는 i번째 개체의 j번째 응답값이라고 하면

$y_i = (y_{obs,i}, y_{mis,i})$로 구성한다(이혜정, 2018b, p.9). $R = (r_{ij})$은 무응답지시행렬(missing-data indicator matrix)이고, R은 y_{ij}가 무응답이면 $r_{ij} = 1$이고 y_{ij}가 응답값이면 $r_{ij} = 0$이다. 자료에서 관측된 다른 변수는 $X = (X_{ik})$는 $(N \times q)$ 행렬로, ϕ는 모수들(parameters)이라고 정의한다.

완전임의결측(MCAR: Missing Completely At Random) 가정하에서 Y의 결측은 Y에는 의존하지 않으나 X에는 의존한다. 아주 강한 가정이며,

$$f(R|Y,X,\phi) = f(R|X,\phi), \text{ 모든 } Y,\ X,\ \phi$$

로 표현할 수 있다(이혜정, 2018b, p.9).

보통 소득 수준이 높으면 소득 문항에 대한 무응답 확률이 높은 편이다. 그러나 소득 수준과 상관없이 소득 문항이 무응답일 확률이 모든 응답자들 사이에서 동일하다면 자료는 완전임의결측 가정을 만족한다(이혜정, 2018b, p.10). 즉, 소득 응답자와 무응답자 간에 본질적인 차이가 없기 때문에 두 집단 간 소득 분포는 동일하다는 가정이다(이혜정, 2018b, p.10).

임의결측(MAR: Missing At Random) 가정하에서 Y의 결측은 Y_{obs}와 X에는 의존하나, Y_{mis}에는 의존하지 않는다(이혜정, 2018b, p.10). 앞서 설명한 완전임의결측보다는 완화된 가정으로,

$$f(R|Y,X,\phi) = f(R|Y_{obs},X,\phi), \text{ 모든 } Y_{obs},\ X,\ \phi$$

로 표현할 수 있다.

모든 응답자들에 대한 소득세 정보가 있다고 가정했을 때 소득 문항이 무응답일 확률은 응답자들의 소득세에 따라 다르다고 하자(이혜정, 2018b, p.10). 그러나 같은 소득세를 내는 응답자들의 소득이 다르지 않으면 자료는 임의결측 가정을 만족한다. 즉, 소득세 정보를 알고 있다면 소득문항의 무응답은 소득과는 상관없이 임의적(Random)이라고 가정할 수 있다.

비임의결측(NMAR: Not Missing At Random) 가정하에서 Y의 결측은 Y_{obs}, Y_{mis}, X 에 모두 의존한다(이혜정, 2018b, p.10). 이는 가장 약한 가정으로 완전임의결측이나 임의결측이 아니면 비임의결측으로 볼 수 있으며,

$$f(R|\,Y, X, \phi) = f(R|\,Y_{obs}, Y_{mis}, X, \phi),\ \text{모든}\ Y,\ X,\ \phi$$

로 표현할 수 있다(이혜정, 2018b, p.10).

소득 문항의 무응답은 소득 자체와 관련이 있다. 소득세에 대한 정보를 알고 있더라도 이와 관련없이 소득이 높은 응답자가 응답하지 않을 확률이 더 높다면 자료는 비임의결측 가정을 만족한다(이혜정, 2018b, p.10).

제2절 평균, 핫덱, 비대체 방법

이 절에서는 실제 항목무응답 대체 시 자주 사용되고 있는 평균, 핫덱, 비대체를 중심으로 살펴보았다.

1. 평균대체

평균대체는 응답한 값들만으로 구한 평균값으로 무응답을 대체하는 방법이다. 대체 방법이 쉽고 간단하여 무응답 비율이 낮은 경우 많이 사용되고 있다. 완전임의결측이라면 평균 추정치의 불편성도 만족하는 특징이 있다. 그러나 무응답 대체값이 모두 동일한 평균값을 가지므로 대체된 자료는 표준편차를 과소 추정 그리고 분포를 왜곡할 수 있는 단점이 있다.

2. 핫덱대체

핫덱대체는 응답자의 응답값으로 무응답을 대체하는 방법이다. 무응답과 유사한 특성이 있는 응답값을 찾기 위해서 응답값들을 다양한 변수(보조 정보)를 사용하여 대체군으로 형성한다. 대체군에 있는 응답값들은 기증자라고 하고, 기증자의 응답값을 받는 -즉, 무응답이어서 대체할 대상자- 수증자라고 정의한다. 대체할 무응답과 동일한 특성을 가지는 대체군 내에 있는 기증자를 무작위로 추출한 후, 선택된 기증자의 응답값으로 무응답을 대체한다. 자료의 분포를 잘 유지하며 변수의 형태에 따른 영향을 받지 않아서 가장 많이 활용되고 있는 대체 방법 중 하나이다.

3. 비대체

비대체는 일정한 증감 비율을 가지는 변수에 대해 적용할 수 있는 대체 방법으로 소득, 자산 등 금액 관련 변수를 대체할 때 많이 활용되고 있다. 회귀대체의 특별한 경우로 다음과 같이 대체한다.

먼저 평균 증감 비율

$$ratio = \frac{1}{n}(\sum_{i \in R} \frac{y_{it}}{y_{it-1}})$$

을 구한다(이후 $ratio$). 비율은 대체 대상 시점과 이전 시점이 모두 응답된 개체들만을 대상으로 하여 추정한다. 여기서 R은 해당 시점(t차수)과 이전 시점($t-1$차수)에서 모두 응답한 개체들만의 집합이다. y_{it}는 i번째 개체의 t차수에서의 응답값이고 y_{it-1}는 i번째 개체의 $t-1$차수에서의 응답값이다. n은 집합 R에 속하는 원소들의 전체 개수를 의미한다.

해당 t차수에서의 무응답은 이전 $t-1$차수의 응답값에 $ratio$를 곱한 값으로 대체한다.

$$y_{it} = y_{it-1} \times ratio$$

이전 및 해당 시점 간 비율을 계산할 수 있으면 비교적 대체하기 간단한 방법이다. 개체의 특성이 다르다면 대체군을 형성한 다음 각 대체군 내의 비율을 사용하는 것이 대체의 정확도가 향상되는 방법 중 하나라고 볼 수 있다.

제3절 기계학습 통계방법을 이용한 대체 방법

이 절에서는 항목무응답 처리 시 '기계학습과 딥러닝(이하 기계학습)' 통계방법을 적용하여 대체하는 방법에 대해 살펴보고자 한다. 최근 기계학습 통계방법에 기반을 둔 무응답 대체 관련 연구가 많아지고 있다. Silva-Ramírez, Pino-Mejías, López-Coello & Cubiles-de-la-Vega(2011)

는 무응답 비율 5%에서 범주형 변수만으로 구성된 자료에 대해 신경망 대체 방법으로 대체한 결과가 평균, 회귀, 핫덱대체 방법보다 우수하다는 점을 밝혔다. Jerez et al.(2010)은 유방암 환자 자료의 무응답 변수에 대해 기계학습 통계방법(다층 퍼셉트론, K-최근접 이웃 등), 평균, 핫덱, 다중 대체 방법을 사용하여 대체하였다. 대체 자료로 유방암 진단 여부의 정확성을 확인한 결과, 기계학습 통계방법 대체 방법에서 정확성이 유의하게 향상되어 우수한 대체 방법으로 나타났다. 손세림(2019)은 무응답 비율 30%와 50%에서 순서형 범주 무응답 변수(3, 5개)를 순서형 로지스틱 모델과 기계학습 통계방법(순서형 의사결정나무, 랜덤 포레스트)을 이용하여 대체하였다. 절편이 모형에 미치는 영향에 따라 다른 결과를 보였는데, 절편의 영향이 작으면 범주의 개수와 무응답 비율에 상관없이 기계학습 통계방법을 이용한 대체 방법의 사용을 추천하였다.

기계학습 통계방법에 대해 개략적으로 살펴본 후 무응답 처리 시 기계학습 통계방법을 적용하여 대체하는 과정을 설명한다. 이 연구의 모의실험에서 사용할 K-최근접 이웃, 랜덤 포레스트, 서포트 벡터 머신, 신경망을 중심으로 소개하고자 한다.

1. 기계학습 통계방법론 개요

인공지능이란 인간의 지적 능력을 인공적으로 구현한 것을 의미한다. 최근 들어 인공지능 관련 연구와 산업에서의 적용이 활발하게 이루어지고 있어 일반인에게는 새로운 분야로 인식될 수 있다. 하지만 이미 1952년부터 학자들은 다양한 영역에서 인공적인 두뇌의 가능성을 논의하였다. 1956년에 이르러 미국 다트머스에서 개최된 한 학회에서 존 매카시가 이 용어를 사용하면서부터 학문의 분야로 들어섰다. 최근 이 분야가

폭발적으로 성장하게 된 배경은 GPU 등 컴퓨터 계산 효과의 비약적인 발전, 방대한 데이터 축적, 모델 개선 등이라고 볼 수 있다. 특히 방대한 양의 데이터를 저장하고 처리하는 기술의 발전은 인공지능 분야의 한 단계 도약을 이끌었다. 소프트뱅크의 회장 손정의는 인공지능이 모든 산업을 다시 정의할 것이라고 언급했으며 인공지능 분야의 잠재적 성장력과 산업에 미치는 영향을 평가하였다.

기계학습은 인공지능의 한 분야다. 'Machine learning'의 저자 Mitchell, T. M.(1997)에 따르면 기계학습은 어떠한 작업에 관한 꾸준한 경험을 통하여 그에 대한 효과를 높이는 것이라고 정의하였다. 컴퓨터가 꾸준한 경험을 하도록 하는 것이 학습 알고리즘이며, 이 알고리즘과 기술을 연구하는 것이 기계학습의 핵심이라고 볼 수 있다. 예를 들면 고양이 사진을 식별하는 작업을 하기 위해 수많은 고양이 사진과 고양이가 아닌 사진을 비교하는 경험(학습)을 한다. 그리고 새로운 사진에 대해 고양이인지 아닌지 식별하는 효과를 최대화하는 학습 알고리즘을 찾는 것이 기계학습에 관한 일련의 과정이다.

기계학습은 학습의 종류에 따라 지도 학습, 비지도 학습, 반지도 학습으로 구분할 수 있다. 지도학습은 입력과 레이블(출력)이 있는 자료를 통해 인과관계 혹은 상관관계를 찾아 학습하는 것이다. 대표적으로 회귀분석, 의사결정나무, 신경망 등이 있으며 레이블이 이산형이면 분류이고 연속형이면 회귀라고 한다. 비지도 학습은 레이블이 없는 자료만을 사용하여 학습하는 것으로 군집화, 이상치 탐지, 연관규칙분석 등이 해당된다. 반지도 학습은 레이블이 있는 자료와 없는 자료를 모두 활용하여 학습하는 것이다. 일반적으로 다수의 레이블이 없는 자료를 소수의 레이블이 있는 자료로 보충하여 학습한다. 이 외에도 모델 업데이트 방식이나 최종 목적에 따라 배치학습, 온라인학습, 강화학습으로 구분하기도 한다.

2. K-최근접 이웃[2)]

K-최근접 이웃은 기계학습 중 메모리 기반 학습기법의 가장 간단한 형태이다. 훈련자료를 있는 그대로 저장하는 것이 모델을 만드는 과정의 전부이며, 분류와 회귀에 모두 사용할 수 있다. 이름에서 유추할 수 있듯이 새로운 자료에 대한 레이블을 예측할 때 훈련자료에서 가장 가까운 K개의 자료를 찾는 것이다. 모델 구축을 특별히 할 필요가 없고 새로운 자료가 들어오면 자료 사이의 거리를 측정하면서 학습하므로 게으른 모델(lazy Model) 또는 사례기반학습(instance-based learning)이라고 명명하기도 한다. 모델 훈련 과정에서 분석자가 직접 설정해야 하는 변수를 하이퍼파라미터(hyper-parameter)라고 한다. 하이퍼파라미터 설정은 모델 효과(과소적합 또는 과대적합 가능성)에 가장 큰 영향을 주기 때문에, 최적화 기법 관련 연구를 많이 하고 있다. 보통 교차검증을 활용하며 K-최근접 이웃을 구현하기 위한 하이퍼파라미터는 최근접 이웃 개수(K) 및 거리측정방식이 해당된다. K의 개수가 커질수록 과소적합 가능성이 커지고, 개수가 작아질수록 과대적합 경향이 커진다. 거리 측정 방식은 주로 유클리디언 거리(euclidean distance)와 맨해튼 거리(manhattan distance)가 사용되며 식은 다음과 같다. p차원 공간에서 두 점을 $X_1 = (x_{11}, x_{12}, \cdots, x_{1p})$과 $X_2 = (x_{21}, x_{22}, \cdots, x_{2p})$라고 하자.

2) Hastie, T., Tibshirani, R., & Friedman, J.(2009), James, G., Witten, D., Hastie, T., & Tibshirani, R.(2013), 박창이, 김용대, 김진석, 송종우, 최호식.(2018) 등을 참고하여 정리함.

• 유클리디언 거리

$$d(X_1, X_2) = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2}$$

• 맨해튼 거리

$$d(X_1, X_2) = \sum_{i=1}^{p}|x_{2i} - x_{1i}|$$

최근접 이웃 개수와 거리 측정 방식이 결정되면 분류 및 회귀 자료에 따라 모델을 구성할 수 있다. 먼저 분류 자료의 경우 $N_k(x)$는 x와 거리가 가까운 K개 훈련자료들의 집합이라고 하면 K개 이웃은 1이고, K개 이외는 -1로 대응한다. 이웃이 더 많은 레이블값이 자료에 대한 예측 레이블값이 된다.

$$\hat{y}(x) = \arg \max_{j \in \{-1, 1\}} \sum_{x_j \in N_k(x)} I(y_j = l)$$

여기서 $I(\bullet)$은 지시함수(indicator function)이다.

회귀 자료의 경우 K개의 이웃 레이블 평균이 자료에 대한 예측 레이블값이 된다.

$$\hat{y}(x) = \frac{1}{K} \sum_{x_j \in N_k(x)} y_j$$

입력자료가 추가되면 모든 자료와의 거리를 계산하므로 자료의 수가 증가하면 계산량이 증가하고 분류 시간 역시 증가한다는 단점이 있다. 훈련자료가 N개이고 예측자료가 M개이면 시간복잡도(time complexity)는 $O(MN^2)$이 된다. 또한 모델을 생성하지 않아서 항상 훈련자료 전체를 저장해야 하는 단점이 있다. 다른 모델에 비해 단순하고 우수한 효과를 가지고 있으나 대용량 자료를 빠르게 계산해야 하는 경우에서는 권장하지 않는다.

3. 랜덤 포레스트[3)]

의사결정트리에서 과적합을 방지하고자 제안된 앙상블은 여러 모델을 연결하여 더 우수한 모델을 만드는 기법이다. 의사결정트리는 직관적으로 이해하기 쉬울 뿐만 아니라 분석도 쉽다는 장점이 있다. 반면에 최적의 의사결정트리를 찾기 위해 설명변수 및 깊이(depth)를 선택해야 하며, 이 과정에서 새로운 자료에 대해서는 과적합되기 쉽다. 이러한 단점을 개선하는 앙상블은 평균을 취하여 편의를 줄이고, 하나의 모형이 아닌 다양한 모형을 결합하여 분산을 감소시킨다. 그러나 예측력이 높은 모델이나 해석의 경우 어려움이 있고 단일 모델에 비해 예측 시간이 많이 걸리는 단점이 있다. 앙상블 모형 중 하나인 배깅(bagging)은 bootstrap aggregation의 약자로 자료 집합으로부터 크기가 같은 표본을 여러 번 무작위로 추출하여 여러 개의 예측모형을 만든 후 결과를 조합하는 방법이다. 배깅에서 bootstrap은 크기가 같은 표본을 무작위로 복원추출하는 것이다.

랜덤 포레스트에서 각각의 트리는 비교적 잘 예측하나 자료의 일부에 과적합하는 경우 서로 다른 방향으로 과적합된 트리들을 평균하여 과대적

3) Hastie et al. (2009). James et al. (2013), 박창이 외(2018) 등을 참고하여 정리함.

합을 줄이는 것이다. 트리의 예측효과는 유지하면서 과대적합은 줄어드는 것인데 이는 이론적으로도 증명되었다. 특히 입력 변수가 많으면 배깅과 같거나 더 우수한 효과를 가진다. 하이퍼파라미터 튜닝을 하지 않아도 잘 작동하고 자료의 스케일에 민감하지 않아 실무에서 많이 활용되고 있다. 좋은 트리는 레이블을 잘 예측하면서 다른 트리와는 구별되는 특성을 가져야 한다. 그래서 트리 생성 시 무작위성을 도입하는데, 트리 생성 시 데이터포인트를 무작위로 선택하는 방법 또는 분할 테스트 시 특성을 무작위로 선택하는 방법이 있다. 다음은 랜덤 포레스트 구축 과정이다.

(단계1) 생성할 트리 개수 결정
(단계2) 전체 자료에서 정해진 수만큼 복원추출하여 붓스트랩 표본 생성
(단계3) 각 자료로부터 결정트리 생성
(단계4) 생성된 결정트리를 가지고 분류의 경우는 약한 투표전략, 회귀의 경우는 평균하여 최종 학습기 생성

입력변수가 많은 경우 중요한 변수를 찾는 데 유용하다고 볼 수 있다. 자료의 개수가 많아지면 모델을 구현하는 데 시간이 오래 걸리는 단점이 있으나, 모델 구축 과정에서 병렬화(parallelize)가 가능하여 CPU 코어를 늘릴수록 작업시간은 비례해서 짧아진다.

4. 서포트 벡터 머신[4)]

서포트 벡터 머신은 1995년 Cortes & Vapnik이 제안한 모형으로 기계학습 커뮤니티에서 큰 반향을 일으켰다. 효과가 좋을 뿐만 아니라 기본적인 접근 방식이 고전적인 분류 방법과 다르기 때문이었다. 약간의 위반(violation)을 허용하면서 자료를 가능한 한 잘 나누는 초평면(hyperplane)을 찾는 방법으로, 로지스틱이나 선형판별분석(linear discriminant analysis)과는 다른 접근 방법으로 볼 수 있다.

어떤 모형이 다른 모형에 비해 좋다는 기준을 마련하는 것이 모델링의 핵심인데, 서포트 벡터 머신은 마진(margin)이라는 개념을 도입하여 마진이 크면 클수록 좋은 모형이라고 판단한다. 또한 특성 공간을 확장하기 위해 커널(kernel)이라는 개념을 사용하는 것도 서포트 벡터 머신의 차별화된 특성 중 하나이다. 커널은 입력 공간(input space)에서 비선형 문제를 고차원의 특성 공간(feature space)에서의 선형 문제로 대응해 주는 역할을 한다. 또 다른 특징 중 하나로 분류의 경우, 고전적인 방법인 로지스틱 회귀모형은 데이터에 대한 출력값으로 조건부 확률을 추정하는 반면 서포트 벡터 머신은 확률을 추정하지 않고 분류에 대한 결과만 예측한다. 따라서 확률이 필요한 문제에서는 활용하지 않는다.

가. 서포트 벡터 머신 분류

$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N),\ x_i \in R^p,\ y_i \in \{-1, 1\}$을 N개의 훈련자료라고 할 때 초평면은 다음 식과 같이 정의한다.

4) Hastie et al. (2009). James et al. (2013), 박창이 외(2018) 등을 참고하여 정리함.

$$\{x : f(x) = \beta^T x + \beta_0 = 0\}$$

여기서 β는 단위벡터(unit vector)이며 $f(x)$에 따른 분류 규칙은 다음과 같다.

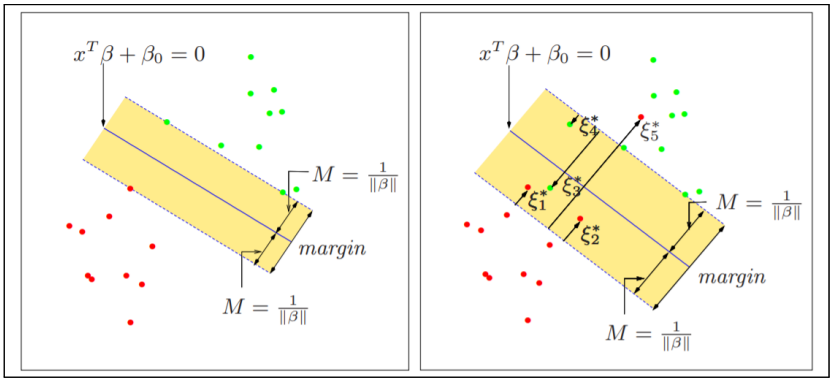
$$G(x) = \begin{cases} 1 & \beta^T x + \beta_0 > 0 \\ -1 & \text{그 외} \end{cases}$$

[그림 3-1]에서 왼쪽 그림은 두 개의 그룹으로 분리가 가능한 경우이다. 실선은 결정경계(decision boundary)를 나타내며 점선은 최대 마진으로 넓이가 $2M = 2/\|\beta\|$이다. 오른쪽 그림은 분리가 불가능하며 겹치는 경우이고, ξ_j^*는 잘못 분류된 경우이다. 이렇듯 마진을 정할 때 구간 내 여유변수(slack variable)를 두고 여유변수에 따른 오분류는 인정하고 분류경계면만 결정하는 것을 소프트마진(soft margin)이라고 한다. 실무에서는 완전히 분리 가능한 초평면을 찾기 어렵기 때문에 보통 이러한 여유를 고려하여 마진의 크기를 최대화하는 초평면을 찾으며 식은 다음과 같다.

$$\min_{\beta, \beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i, \qquad \xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \ \forall i$$

여기서 C는 비용모수(cost parameter)이며 과적합의 정도를 조절할 수 있다.

〔그림 3-1〕 서포트 벡터 분류



주: Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media. p.418 그림 인용.

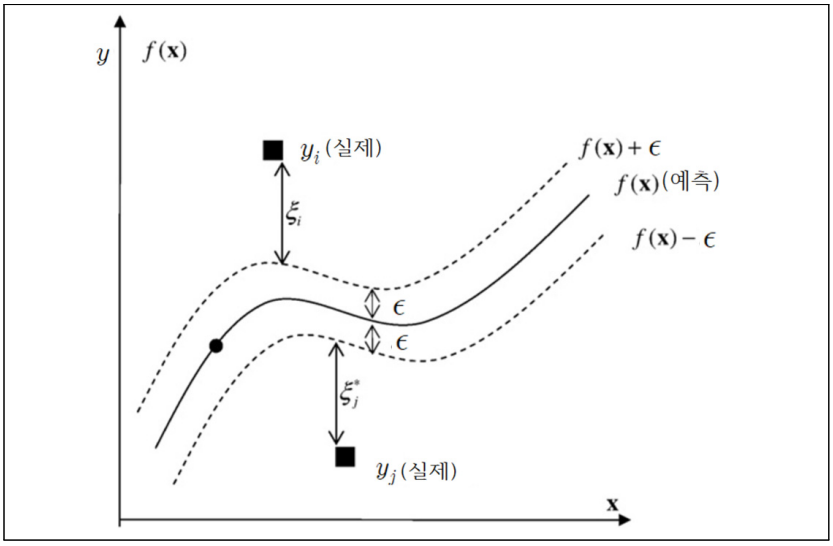
나. 서포트 벡터 머신 회귀

서포트 벡터 머신은 분류에서 많이 사용하고 있으나 회귀에서도 적용 가능하며(Vapnik, 1998), 커널 함수에 근거한 비모수적 방법으로 볼 수 있다. 서포트 벡터 머신 회귀는 아래 식처럼 ϵ-무감도 손실함수(ϵ-insensitive loss function)를 이용하여 임의의 실수값을 예측하는 데 사용할 수 있다.

$$L_\epsilon(x, y, f) = \max(0, |y - f(x)| - \epsilon)$$

ϵ-무감도 손실함수는 실제값과 예측값의 차이($|y-f(x)|$)가 ϵ보다 크면 그 차이값($|y-f(x)|-\epsilon$)을 반환하고 아니면 0을 반환한다. 여기서 ϵ는 $f(x)$ 주변에 위치한 튜브의 반지름을 나타내는 정밀모수(precision parameter)이다([그림 3-2] 참조).

[그림 3-2] 서포트 벡터 머신 회귀에서 사용하는 ϵ-무감도 손실함수



주: Lu, C. J., Lee, T. S., & Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision support systems, 47*(2), 115-125. p.117 그림 인용.

서포트 벡터 머신 회귀의 최적화는 비용함수를 사용하며 식은 아래와 같다. 즉, 실제값(y)과 예측값($f(x)$)이 가능한 한 ϵ 이내를 유지하면서 마진을 최대화하는 것이다.

$$\text{cost}(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\epsilon(x_i, y_i, f(x_i)) + \beta^T\beta$$

여기서 C와 ϵ는 분석자가 직접 설정해야 하는 하이퍼파라미터이다. 이를 여유변수를 사용하여 다시 표현하면 다음 식과 같다.

$$\min \beta^T\beta + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-)$$

여기서 임의의 i에서 $(\beta^T x_i + \beta_0) - y_i \le \epsilon + \xi_i^+, \xi_i^+ \ge 0$이고, 임의의 i에서 $y_i - (\beta^T x_i + \beta_0) \le \epsilon + \xi_i^-, \xi_i^- \ge 0$이다.

다. 서포트 벡터 머신 커널트릭

선형모형은 과적합 방지, 해석의 편리성 등과 같은 특성 때문에 주로 활용되고 있다. 그러나 비선형인 경우 자료에 비선형 특성을 추가하여 고차원 공간으로 사상하는 것이다. 고차원 비선형 매핑(mapping)의 경우 변수 확장에 따른 자료 사용량 및 최적화 난이도의 증가 등과 같은 문제가 있어서 커널트릭을 많이 사용한다. 커널트릭은 자료를 실제로 확장하는 것이 아니라 확장된 특성을 거리 함수에 반영하는 것이다. 일반적으로 많이 사용하는 커널함수는 다항, 가우시안, 시그모이드 함수이다.

- 다항(polynomial) 함수

$$K(x_i, x_j) = (< x_i, x_j > + \kappa)^d, \qquad \kappa > 0$$

- 가우시안(gaussian radial basis) 함수

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}), \qquad \sigma^2 \neq 0$$

• 시그모이드(sigmoid) 함수

$$K(x_i,x_j) = \tanh(\kappa_1 < x_i,x_j > + \kappa_2), \qquad \kappa_1, \kappa_2 > 0$$

분석시간은 자료의 개수에 비례하지만 차원의 크기에는 상대적으로 덜 민감한 편이다. 생물정보학의 마이크로어레이(microarray) 자료와 같이 자료의 개수보다 차원이 훨씬 큰 경우 계산상의 장점을 가진다(박창이 외, 2018). 랜덤 포레스트와 다르게 자료의 전처리가 중요하며 하이퍼파라미터 설정 시 신경을 많이 써야 하는 단점이 있다.

5. 신경망[5)]

인공신경망(artificial neural networks)은 생물학적 뉴런으로부터 영감을 얻어 개발한 학습 알고리즘으로 딥러닝의 핵심 알고리즘이다. 1943년 Warren McCulloch & Walter Pitts는 신경망을 위한 작동모형을 구축하였다. 이후 1958년 Frank Rosenblatt은 간단한 연산을 하는 단층 신경망(single layer perceptron) 모형을 개발하였다. 그러나 컴퓨터의 계산 성능 문제로 30여 년간 사용되지 않다가 오차역전파법(back propagation)이 만들어진 후 활용도가 높아졌다.

인공신경망은 강력하고 확장성이 좋아서 이미지 분류, 음성인식 또는 추천 프로그램, 게임(알파고) 등과 같이 복잡하고 대규모의 기계학습 문제에 적합하다. 결과에 대한 해석이 어렵기 때문에 신용평가처럼 해석이 중요한 분야에서는 잘 사용하지 않는다. 그러나 이와 관련한 연구가 활발히 진행되고 있어 추후 적용 가능성을 배제할 수 없다. 인공신경망은 뉴

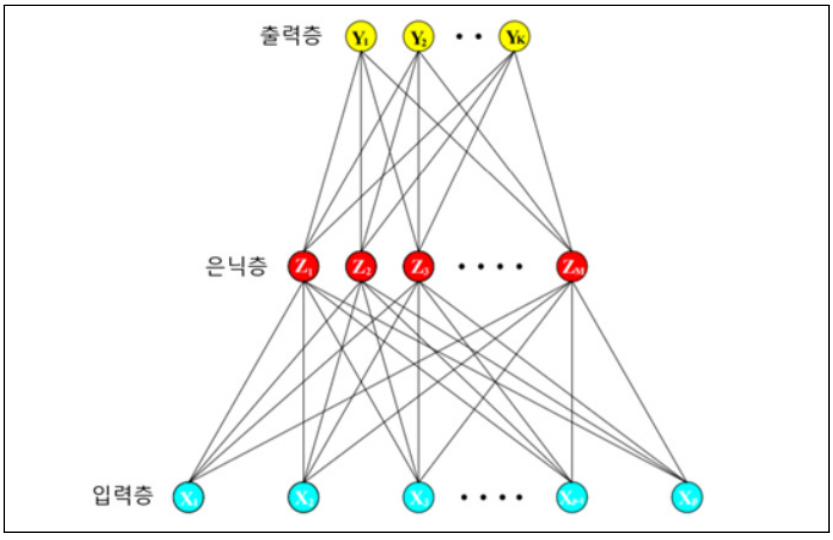
5) Hastie et al. (2009), James et al. (2013), 박창이 외(2018) 등을 참고하여 정리함.

런을 모델화한 유닛을 서로 연결하여 만든 후, 비용(에러)을 최소화하는 모델을 찾는 것이다. 이때 비용은 일반적으로 오차제곱합을 사용한다.

가. 다층 신경망 구조 및 학습 알고리즘

신경망은 입력층(input layer), 은닉층(hidden layer) 그리고 출력층(output layer)으로 구성되어 있다([그림 3-3] 참조). 입력층은 각 입력변수에 대응되는 노드(node)로 구성되며, 개수는 입력 노드의 수와 같다. 이렇게 입력변수는 은닉층을 거쳐서 비선형으로 처리되고 출력층을 통해 출력변수로 대응된다. 은닉층이 2개 이상이면 딥러닝이라고 한다. 분류모형의 경우 분류 클래스의 개수는 출력 노드와 같고, 회귀모형의 경우 출력 노드는 한 개가 된다.

[그림 3-3] 단일 은닉층 피드 포워드 신경망



주: Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media. p.393 그림 인용.

클래스가 K개인 분류 문제에서 출력 노드는 K개의 유닛(unit)으로 구성되는데 $Y_k(k=1, 2, \cdots, K)$는 k번째 클래스에 속할 확률을 모형화한 것이다. 이때 출력변수 Y_k는 자료가 k번째 클래스에 속하는 경우로 k번째 좌표는 1이고 그 외 좌표는 0으로 코딩된다.

은닉 노드 Z_m은 입력 노드의 선형 결합을 활성함수(activation function) $\sigma(v)$를 이용하여 비선형 변환한 것이고, 출력 노드 Y_k는 Z_m의 선형결합 함수로 모형화한 것으로, 식으로 나타내면 다음과 같다.

$$
\begin{aligned}
Z_m &= \sigma(\alpha_{0m} + \alpha_m^T x), && m = 1, 2, \cdots, M \\
T_k &= \beta_{0k} + \beta_k^T Z, && k = 1, 2, \cdots, K \\
f_k(X) &= g_k(T), && k = 1, 2, \cdots, K
\end{aligned}
$$

여기서 $Z=(Z_1, Z_2, \cdots, Z_M)$이고 $T=(T_1, T_2, \cdots, T_K)$이다.

$\sigma(v)$는 이전 노드로부터 전달받은 정보를 다음 노드에 전달하는 역할을 한다. 보통 시그모이드 함수를 사용하며, 가우시안 함수를 사용하기도 하는데 이를 RDF 신경망(radical basis function network)이라고 한다. 단극성 시그모이드 함수 $\sigma(v)=1/(1+e^{-v})$는 증가함수로 출력값이 0과 1 사이에 있다. 양극성 시그모이드 함수 $\sigma(v)=(1-e^{-v})/(1+e^{-v})$는 증가함수로 출력값이 -1과 1 사이에 있으며 $\sigma(0)=0$이다. 시그모이드 함수는 S자 형태의 비선형 미분 가능한 함수로 역전파 학습 알고리즘에서 잘 작동하며, 복잡한 유형의 의사결정 문제에도 효과적으로 적용된다. 또한 입력값이 크더라도 출력값이 급격히 변화하지 않아서 입력값의 작은 변화도 잘 반영되는 장점이 있다.

$g_k(T)$는 출력함수로 정의하며 출력값 T에 대해 최종적으로 비선형 변환을 하는 역할을 한다. K 클래스 분류에서는 항등함수(identity function)를 사용하였으나 최근에는 소프트맥스함수(softmax function) $g_k(T) = e^{T_k} / \sum_{l=1}^{K} e^{T_l}$가 선호되고 있다. 다범주 로지스틱 회귀에서도 사용되는 소프트맥스함수는 항상 양의 값을 가지며, 그 합은 1이라는 특징을 가진다. 회귀에서는 항등함수 $g_k(T) = T_k$가 사용된다.

신경망에서는 훈련데이터를 잘 설명하는 미지의 파라미터인 가중치(weight)를 찾는 것이 훈련 과정이며, 이는 곧 모형을 최적화하는 것과 같다. 미지의 파라미터 α_{0m}, α_m과 β_{0k}, β_k에 대한 벡터를 θ로 나타내면, 회귀에서의 손실함수는 아래 식과 같은 오차제곱합으로 표현할 수 있다.

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2$$

손실함수는 모형에 기반한 추정값과 실제값 간의 차이를 측정하는 것으로, 이를 최소화하는 파라미터를 찾는 것을 훈련 과정으로 볼 수 있다. 일반적으로 손실함수 $R(\theta)$의 전역 최솟값(global minimizer)은 찾기 어려울 뿐만 아니라 과적합일 가능성도 높은 편이다. 따라서 벌점화(penalty term) 또는 알고리즘 조기 종료(early stopping) 등을 추가하여 '괜찮은' 국소 최솟값을 찾는다. $R(\theta)$의 추정치를 찾기 위한 가장 일반적인 방법은 역전파(back-propagation)라고 부르는 기울기 강하(gradient descent) 알고리즘이다. 역전파를 활용한 기울기 강하 알고리즘은 편도함수의 미분을 반복적으로 실시하여 최솟값을 찾아가는 방법이다.

나. 신경망 모형 구축 시 고려 사항

신경망은 대량의 자료에 내재된 정보를 찾아내고, 매우 복잡한 모형을 만들 수 있다는 장점이 있다. 충분한 연산 시간과 자료를 주고 하이퍼파라미터를 세심하게 조정한다면, 다른 기계학습 알고리즘을 능가하는 효과를 가질 수 있다. 그러나 다른 기계학습에 비해 모형 구축 시 고려해야 할 부분이 많은 편이다. 여기서는 입력변수의 척도화, 초기값과 다중 최솟값 문제, 은닉층과 은닉 노드 수에 대해 살펴보았다.

1) 입력변수의 척도화(scaling)

신경망은 매우 복잡한 모형으로 입력변수의 값에 따라 결과가 매우 민감하므로 자료를 처리하는 단계에서 필수적으로 점검이 필요하다. 신경망에 적합한 자료는 범주형의 경우 모든 범주에서 일정한 빈도(frequency) 이상 값들을 가져야 하고, 연속형인 경우 변수값의 범위가 변수별로 차이가 크지 않아야 한다. 예를 들면 연속형 입력 변수의 분포가 평균을 중심으로 대칭이 아닌 경우에는 모형 결과가 좋지 않을 수 있다. 가구소득 분포를 예로 들어 설명하면, 일반적으로 대부분의 가구의 소득은 평균 미만이고 특정한 가구의 소득이 매우 큰 패턴을 가진다. 이에 소득에 대한 로그변환이나 범주화 등과 같은 전처리가 필요하다. 또한 범주형 변수는 가변수(dummy variable)로 처리해야 한다. 그러나 가변수 처리 방법에 따라 결과가 달라질 수 있다. 즉, 남자와 여자에 대한 가변수 처리를 -1과 1로 한 결과와 0과 1로 한 경우는 다를 수 있다. 여러 개의 범주형 변수를 가변수할 때는 같은 범위를 가지도록 처리하는 것이 바람직하다고 볼 수 있다.

2) 초기값과 다중 최솟값 문제

역전파 알고리즘은 초기값에 따라 결과가 많이 다를 수 있으므로 초기값 선택은 매우 중요하다. 가중치가 0에 가까우면 시그모이드 함수는 거의 선형이 되고 신경망 모형은 근사적으로 선형이 된다. 보통 초기값은 0 근처에서 무작위로 선택하므로 초기 모형은 선형 모형에 가까우나, 가중치 값이 증가할수록 비선형 모형이 된다. 그러나 초기값을 0으로 설정하면 반복에 따라 값이 변하지 않을 수 있으며, 반대로 너무 큰 값이면 좋지 않은 추정치를 얻을 수 있다는 점을 주의해야 한다.

보통 신경망에서의 손실함수는 비볼록 함수이고, 여러 개의 국소 최솟값(local minima)을 가진다. 따라서 최종 예측값은 무작위로 선택된 여러 개의 초기값으로 적합하여 구한 추정치 중에서 오차가 가장 작은 것을 선택하는 방법 또는 예측값의 평균이나 최빈값을 선택하는 방법이 있다.

3) 은닉층과 은닉 노드 수

신경망에서는 하이퍼파라미터인 은닉층과 은닉 노드의 수를 설정해야 한다. 은닉층과 은닉 노드가 너무 많으면 가중치들이 많아져서 과대적합이 발생하나, 너무 적으면 과소적합이 있을 수 있으므로 자료의 특징이나 상황에 따라 결정해야 한다. 범용근사자정리(universal approximation theorem)는 하나의 은닉층을 가진 신경망을 이용하여 매끄러운 함수를 근사적으로 표현할 수 있다는 것이다. 그러므로 은닉층은 한 개로 하고, 은닉 노드 수는 적절히 선택한다면 큰 무리가 없을 것이다. 은닉 노드의 수는 교차확인 오차를 사용하여 결정하는 것보다는 적절히 큰 값을 은닉 노드로 놓고 가중치 감소(weight decay)라는 모수에 대한 벌점화를 적용하는 것이 더 좋은 결과를 얻을 수 있다.

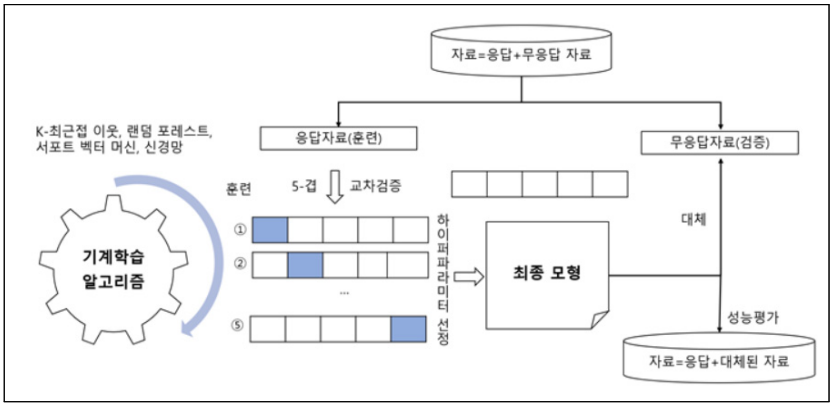
6. 기계학습 통계방법에 기반한 대체 방법

저자는 [그림 3-4]와 같이 기계학습에 기반한 대체 방법의 과정을 구성하였다. 먼저 자료를 응답 자료와 무응답 자료로 구분하며 응답 자료는 훈련자료로, 무응답 자료는 검증자료로 간주한다.

다음으로 최적의 하이퍼파라미터를 찾기 위해 훈련자료에 기계학습을 적용한다. 이때 하이퍼파라미터 튜닝(parameter tuning)은 5-겹 교차검증(5-fold cross-validation) 또는 10-겹 교차검증을 사용한다. 교차검증은 하이퍼파라미터를 결정할 때 많이 활용하는 방법 중 하나이며, 이 연구에서는 5-겹 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. 5-겹 교차검증을 사용하여 하이퍼파라미터를 결정하는 과정을 설명하면 다음과 같다. 5-겹 교차검증을 실시하기 위하여 우선 훈련자료를 5겹 자료로 나눈다. 첫 번째 시도에서 첫 번째부터 4번째 겹 자료를 가지고 적합한 다음에 5번째 겹 자료로 모형을 평가한다. 다음 두 번째 시도에서는 첫 번째부터 3번째 그리고 5번째 겹 자료로 적합한 후에 4번째 겹 자료로 모형을 평가한다. 이러한 방식으로 총 5번의 시도를 할 수 있고 5개 결과에 대한 평균을 계산한다. 설정한 하이퍼파라미터 조합에 대한 결과값 중에서 가장 작은 값을 가지는 하이퍼파라미터를 최종 선택한다.

선정된 하이퍼파라미터를 사용하여 전체 훈련자료에 다시 적합하여 최종 모형을 구축한다. 마지막으로 검증자료에 최종 모형을 적용하여 예측값을 구하며, 무응답은 이 예측값으로 대체하여 최종 생성한다.

[그림 3-4] 기계학습 방법론에 기반한 대체 방법 절차



주: 저자 직접 작성.

제4절 대체 방법 평가 기준

항목무응답을 대체하기 위해서는 2절과 3절에서 소개한 여러 가지 무응답 대체 방법을 사용하여 최적 모형을 적합한 다음, 이를 통해 얻어진 예측값으로 무응답을 대체하면 된다. 무응답 대체 방법에 따라 예측값은 서로 다른 값을 가지므로 효과가 좋은 대체 방법을 선정하는 것이 중요한 과정 중의 하나이다. 최적의 대체 방법 선정은 다양한 평가 기준 지표를 통하여 여러 대체 방법을 비교한 후 대체하고자 하는 항목에 맞는 방법을 최종 선택하는 것이다.

Chambers(2002)는 대체 방법을 평가하는 기준으로 예측의 정확성(predictive accuracy), 순위의 정확성(ranking accuracy), 분포의 정확성(distributional accuracy), 추정의 정확성(estimation accuracy) 그리고 대체의 그럴듯함(imputation plausibility)을 제안하였다. 예측의 정확성은 대체값이 가능한 한 실제값에 가까워야 한다는 의미이다. 순

위의 정확성은 대체 후 대체값의 순위가 실제값의 순위를 유지해야 한다는 것을 나타낸다. 분포의 정확성은 대체된 데이터의 주변(marginal) 및 고차항 분포가 실제값 데이터의 분포를 유지해야 한다는 것이다. 추정의 정확성은 대체된 데이터의 적률이 실제값 데이터의 적률을 유지해야 한다는 의미이다. 마지막 대체의 그럴듯함은 대체값이 그럴듯한 값으로 대체되어야 한다는 것을 나타낸다.

변수를 대체할 때 위의 평가 기준을 모두 만족시켜야 하는 것은 아니다. 예를 들면 순위의 정확성은 변수의 특성이 순서인 경우 고려할 수 있다는 의미이다. 그러므로 평가 기준은 관심 있는 데이터를 구성하는 변수들에 대한 추정된 적률이 유지되는 것에 초점을 맞추는 것으로 완화될 수 있다.

이 연구에서는 평가 기준으로 평균 추정치와 참값과의 편향 및 제곱근 평균제곱오차, 예측의 정확성과 추정의 정확성을 근거로 하여 대체 방법의 효과를 평가하였다. 대체된 자료는 참값과 대체값이 함께 구성되어있다. y_{ij}는 i번째 개체에 대한 j차수 관측값에 대한 참값(true value)으로, $y_{ij,\,imp}$는 대체값(imputation value)으로 정의한다.

$$y_{ij,imp} = \begin{cases} y_{ij} & \text{관찰된경우} \\ y_{ij,\,imp} & \text{무응답인경우} \end{cases}$$

평가지표에 대한 내용은 다음과 같다.

1. 편향

편향(bias)은 대체된 자료의 평균 추정치가 참값과 얼마나 가까운지를 나타내며 식은 다음과 같다.

$$bias = E(\hat{\theta}_{jk,\,imp} - \theta_{jk})$$

여기서, $\hat{\theta}_{jk,\,imp} = \frac{1}{N}\sum_{i=1}^{N} y_{ijk,imp}$ 이고 $\theta_{jk} = \frac{1}{N}\sum_{i=1}^{N} y_{ijk}$ 이고 $k(=1, \cdots, K)$는 모의실험 자료의 개수이다. 평균 추정치가 참값과 얼마나 가까운지를 보여 주는 평가지표로 참값과의 일치 정도를 확인할 수 있다.

2. 제곱근평균제곱오차

제곱근평균제곱오차(RMSE: Root Mean Square Error)는 다음과 같다.

$$RMSE = \sqrt{E(\hat{\theta}_{jk,\,imp} - \theta_{jk})^2}$$

제곱근평균제곱오차를 통해서 정밀도(precision)를 측정할 수 있다. 정밀도는 여러 번 측정하거나 계산한 결과가 서로 얼마나 가까운지에 대한 것이다. 관측의 균질성을 나타내며, 관측된 값의 편차가 적을수록 정밀하다(Wikipedia, 2019). 즉, 평균 추정치가 얼마나 일관되는지(consistent)를 알 수 있는 평가지표이다.

3. 참값의 포함률

평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률(coverage rate)을 계산하였다. 대체된 자료에서의 평균 추정치에 대한 95% 신뢰구간의 식은 다음과 같다.

$$\hat{\theta}_{jk,\,imp} \pm 1.96 \times SE(\hat{\theta}_{jk,\,imp})$$

참값 θ_{jk}가 95% 신뢰구간에 포함하는 비율을 구하면 된다.

4. 예측의 정확성 지표

대체 방법에 대한 예측의 정확성은 개체별 대체값($y_{ij,\,imp}$)과 실제값(y_{ij}) 간의 근사 정도를 평가하는 것으로, 대체 방법이 가능한 실제값을 유지해야 한다는 의미이다(Watson & Starick, 2011). 지표는 대체값과 실제값의 거리를 사용하여 절대거리 및 제곱거리를 측정할 수 있다. 절대거리를 사용하면 평균절대편차(MAD: Mean Absolute Deviation)이고, 제곱거리를 사용하면 제곱근평균제곱편차(RMSD: Root Mean Square Deviation)이며 식은 다음과 같다. MAD는 평균절대편차이고 RMSD는 제곱근평균제곱편차를 의미한다.

$$MAD = E(\frac{1}{N}\sum_{i=1}^{N}|y_{ijk,\,imp} - y_{ijk}|)$$

$$RMSD = E(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{ijk,\,imp} - y_{ijk})^2})$$

5. 추정의 정확성 지표

대체 방법에 대한 추정의 정확성은 실제값 분포의 적률들(Moments)을 얼마나 잘 유지하는지를 평가하는 것이다(Watson & Starick, 2011). Chambers(2000)는 대체값과 실제값 사이의 평균에 대한 절대 차이, 분산, 왜도(Skew), 첨도(Kurtosis)를 지표로 제안하였다. 이 연구에서는 4가지 지표로, 즉 평균 절대 상대 차이(Absolute relative difference in means), 변동계수 절대 차이(Absolute difference in the coefficient of variation), 표준화된 3차 적률(왜도) 절대 차이 그리고 표준화된 4차 적률(첨도) 절대 차이를 사용하여 평가하였다. 각 지표들의 값이 작을수록 좋은 대체 방법이라고 볼 수 있으며 식은 다음과 같다. m_1은 평균 절대 상대 차이이고 m_2는 변동계수 절대 차이이며 m_3는 표준화된 왜도, m_4는 표준화된 첨도를 의미한다. 여기서, $\hat{\sigma}_{jk,imp}$는 대체된 자료의 표준편차($\hat{\sigma}_{jk,imp} = \frac{1}{N-1}\sum_{i=1}^{N}(y_{ijk,\,imp} - \hat{\theta}_{jk})^2$)이고, σ_{jk}는 실제 자료의 표준편차($\sigma_{jk} = \frac{1}{N-1}\sum_{i=1}^{N}(y_{ijk} - \theta_{jk})^2$)이다.

$$m_1 = E(\left|\frac{\hat{\theta}_{jk,\,imp} - \theta_{jk}}{\hat{\theta}_{jk,\,imp}}\right|)$$

$$m_2 = E\left(\left| \frac{\hat{\sigma}_{jk,\,imp}}{\hat{\theta}_{jk,\,imp}} - \frac{\sigma_{jk}}{\theta_{jk}} \right| \right)$$

$$m_3 = E\left(\left| \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(y_{ijk,\,imp} - \hat{\theta}_{jk,\,imp})^3}{\hat{\sigma}^3_{jk,\,imp}} - \frac{(y_{ijk} - \theta_{jk})^3}{\sigma^3_{jk}} \right] \right| \right)$$

$$m_4 = E\left(\left| \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(y_{ijk,\,imp} - \hat{\theta}_{jk,\,imp})^4}{\hat{\sigma}^4_{jk,\,imp}} - \frac{(y_{ijk} - \theta_{jk})^4}{\sigma^4_{jk}} \right] \right| \right)$$

제 4 장 항목무응답 대체 방법 모의실험

4 항목무응답 대체 방법 모의실험

이 장에서는 무응답 비율에 따른 패널자료에서의 항목무응답 대체 방법의 효과를 평가하기 위하여 보건복지 분야 패널자료에 근거하여 모의실험을 실시하였다. 항목무응답 대체 방법은 최근 여러 분야에서 다양하게 사용하고 있는 기계학습 방법론을 기반으로 하여 무응답을 대체한 후 이에 대한 효과를 살펴보았다. 항목무응답 처리 시 자주 활용되고 있는 평균, 핫덱, 비대체 방법과 함께 비교하였다.

1절에서는 한국복지패널자료, 2절에서는 한국의료패널자료에 근거하여 모의실험을 실시하였다.

제1절 한국복지패널자료[6]를 이용한 모의실험

1. 자료 설계

올해 공개된 한국복지패널자료 12~13차(2017~2018년)를 사용하여 모의실험을 위한 자료를 생성하였다. 무응답 대체 대상 변수는 실제로 실시하고 있는 변수인 '경상소득'으로 선정하였다.

모의실험을 위한 연구 대상 모집단은 12~13차에서 경상소득을 모두

6) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr 에서 2019. 5. 14. 인출한 자료 활용.
한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr 에서 2019. 5. 14. 인출한 자료 활용.

응답하고, 무응답 대체 시 설명변수로 사용하는 변수도 모두 응답한 가구만을 추출하여 최종 6360가구로 확정하였다.

무응답 비율은 5%, 10%, 20%, 30%, 50%로 다양하게 설정하였다. 한국복지패널자료는 무응답 비율이 낮은 편이어서 실제 상황을 반영하기 위해 5%를 고려하였다. 그리고 무응답 비율을 점진적으로 늘려 무응답 비율에 따른 무응답 대체 방법의 효과를 살펴보았다.

결측자료 메커니즘이 임의결측을 따른다는 가정하에 모의실험용 무응답 패널자료를 생성한 과정은 다음과 같다. 무응답을 발생시키기 위하여 경상소득과 연관성이 있는 '가구주의 배우자 유무(이하 배우자 유무)'와 '가구주의 학력(이하 학력)'을 가지고 4개의 대체군 집단(이하 집단)으로 나누었다. 배우자 유무 범주는 배우자 없음과 있음인 2개, 학력 범주는 고졸 미만과 이상인 2개로 구분하였다. 집단별 경상소득 평균은 집단 1은 1615만 원, 집단 2는 3359만 원, 집단 3은 3156만 원 그리고 집단 4는 7054만 원이었고, 유의수준 5%하에서 (집단 1), (집단 2, 집단 3), (집단 4)로 묶이는 것으로 나타났다.

〈표 4-1〉 연령 및 학력에 따른 집단 구분(한국복지패널)

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

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

소득이 높을수록 무응답 가능성이 높다는 점을 반영하여 〈표 4-2〉와 같이 셀별로 무응답 비율에 따른 무응답 발생 개수의 비율을 다르게 구성하였다. 예를 들어 무응답 비율이 10%라고 가정했을 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.

〈표 4-2〉 집단별 무응답 발생 개수의 비율(한국복지패널)

(단위: %)

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

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

추가적으로 고려한 가정으로 경상소득은 매해 조사하므로 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 만족하기 위해서 12차 경상소득은 무응답이 없이 모두 값을 가지고 있다고 가정하고 13차 경상소득을 무응답으로 처리하였다. 이는 모의실험에서 비교 대상 중 하나인 비대체에서의 비를 계산할 때 비의 분모가 결측이 되지 않도록 하는 효과도 있다. 사전에 설정한 무응답 비율에 따라 무작위로 추출한 가구의 13차(해당 차수) 경상소득을 무응답으로 처리하였다. 무응답 비율이 5%인 경우 무응답 발생 개수는 318가구, 10%인 경우 636가구, 20%인 경우 1272가구, 30%인 경우 1908가구, 50%인 경우 3108가구이다. 〈표 4-3〉부터 〈표 4-7〉은 무응답 비율별 생성한 무응답 가구 수를 나타낸다.

〈표 4-3〉 무응답 비율 5%에서의 집단별 무응답 발생 개수(한국복지패널)

(단위: 가구)

	경상소득 중위수		전체
	미만	이상	
집단 1	10	22	32
집단 2	17	40	57
집단 3	21	49	70
집단 4	48	111	159
전체	96	222	318

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〈표 4-4〉 무응답 비율 10%에서의 집단별 무응답 발생 개수(한국복지패널)

(단위: 가구)

	경상소득 중위수		전체
	미만	이상	
집단 1	19	45	64
집단 2	34	80	114
집단 3	42	98	140
집단 4	95	223	318
전체	190	446	636

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〈표 4-5〉 무응답 비율 20%에서의 집단별 무응답 발생 개수(한국복지패널)

(단위: 가구)

	경상소득 중위수		전체
	미만	이상	
집단 1	38	89	127
집단 2	69	160	229
집단 3	84	196	280
집단 4	191	445	636
전체	382	890	1272

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〈표 4-6〉 무응답 비율 30%에서의 집단별 무응답 발생 개수(한국복지패널)

(단위: 가구)

	경상소득 중위수		전체
	미만	이상	
집단 1	57	134	191
집단 2	103	240	343
집단 3	126	294	420
집단 4	286	668	954
전체	572	1336	1908

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〈표 4-7〉 무응답 비율 50%에서의 집단별 무응답 발생 개수(한국복지패널)

(단위: 가구)

	경상소득 중위수		전체
	미만	이상	
집단 1	95	223	318
집단 2	172	400	572
집단 3	210	490	700
집단 4	477	1113	1590
전체	954	2226	3180

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

연구 대상 모집단 자료로부터 무응답 가구를 생성하는 과정을 100번 실시하여 모의실험용 무응답 패널자료를 100개 생성하였다. 이때 무응답 가구를 추출하는 방법은 단순임의추출(simple random sampling) 방법을 사용하였다.

2. 대체 방법 및 평가지표

모의실험에서 항목무응답을 대체할 때 사용한 대체 방법과 최적의 대체 방법 선정을 위해 이용한 평가지표에 대해 살펴보았다.

가. 무응답 대체 방법

무응답 대체 방법은 총 7가지로 1) 평균대체, 2) 핫덱대체, 3) 비대체, K-최근접 이웃 대체, 5) 랜덤 포레스트 대체, 6) 서포트 벡터 머신 대체, 7)

신경망 대체이다. 각 대체 방법에 따른 무응답 대체 과정은 다음과 같다.

1) 평균대체

대체군 활용 여부에 따라 2개의 대체값을 생성하였다. 먼저 대체군을 사용하지 않는 경우에는 응답값을 가지고 구한 평균값으로 무응답을 대체하였다.

다음으로 대체군을 형성하여 대체군 내에서의 대체를 실시하였다. 대체군은 학력 범주 2개(고졸 미만·이상), 배우자 유무 범주 2개(배우자 있음·없음), 13차 경상소득 범주 2개(중위수 미만·이상)로 총 8개 대체군으로 구분한 다음, 각 대체군 내에서의 평균값을 구하였다. 무응답은 학력, 배우자 유무, 13차 경상소득 범주에 해당하는 대체군 내의 평균값으로 대체하였다. Welniak & Coder(1980)는 미국의 CPS 자료에 대한 실증분석을 통하여 대체군 형성의 중요성을 확인하였다(김영원, 조선경, 1996, p.147 재인용). 이처럼 대체군을 사용하면 대체된 무응답의 편향(bias)이 감소될 뿐만 아니라 대체의 정확도도 보다 향상될 수 있다. 이 연구에서는 평균대체, 핫덱대체 그리고 비대체의 경우 대체군 활용 여부에 따라 무응답 대체를 실시하였다.

2) 핫덱대체

대체군 활용 여부에 따라 2개의 대체값을 생성하였다. 대체군이 없는 경우의 대체 방법은 응답한 모든 값을 기증자 후보로 사용하여 기증자 중에서 무작위로 추출한 후 기증자의 응답값으로 무응답을 대체하였다.

대체군을 사용한 경우는 대체군 내에서 응답한 모든 값을 기증자 후보

로 사용하여 기증자 중에서 무작위로 추출한 후 기증자의 응답값으로 무응답을 대체하였다. 이때 대체군은 평균대체와 동일하며, 무응답 대체 시 한 번 사용된 기증자는 이후 대체에서는 기증자로 사용하지 않도록 하였다. 단, 무응답 비율 30% 및 50%의 경우 해당 대체군 내에서의 기증자 부족으로 무응답이 제대로 채워지지 않아 기증자를 중복 사용하였다.

3) 비대체

대체군이 없는 경우는 이전 차수와 해당 차수에 대해 모두 응답값을 가지고 있는 개체만을 대상으로 하여 비율을 계산하였다. 무응답은 이전 차수의 응답값에 해당 비율을 곱한 값으로 대체하였다.

대체군을 활용한 경우의 대체 방법은 학력 및 배우자 유무 범주로 4개의 대체군을 생성한 후 대체군 내에서의 비율을 계산하였다. 무응답은 이전 차수의 응답값에 해당 대체군 내의 비율을 곱한 값으로 대체하였다.

4) K-최근접 이웃 기반 대체

평균, 핫덱, 비대체 방법에서 대체군을 형성할 때 보조변수로 사용한 3개 변수를 활용하여 무응답 대체를 실시하였다. 이는 동일한 보조변수를 사용하여 무응답을 대체한 방법들의 효과를 비교하기 위해서이다.

또한 설명변수를 더 포함(3개에서 6개로 증가)하여 무응답 대체를 실시하였다. 더 많은 정보를 사용하여 무응답 대체를 하였을 때의 효과를 살펴보기 위해서이다. 랜덤 포레스트, 서포트 벡터 머신, 신경망 대체 방법에도 동일하게 적용하였으며 다음 〈표 4-8〉은 사용한 설명변수에 대한 목록이다.

〈표 4-8〉 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

주: 경상소득(12차)을 제외한 나머지 설명변수는 무응답 대체 대상 변수와 동일한 13차임.
자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

K-최근접 이웃 대체를 위해 R패키지 caret을 사용하였다. 최적 모형을 찾기 위한 모델 옵션(option)을 조정할 때 파라미터 튜닝을 5-겹 교차검증으로 실시하였다. 10-겹 교차검증도 많이 사용하지만 모형을 적합하는 데 시간이 상당히 오래 걸리는 것으로 나타났다. 무응답 대체 시 시간의 효율성 측면도 중요하다고 생각하여, 이 모의실험은 5-겹 교차검증으로 실시하였다. 5-겹 교차검증은 10-겹 교차검증으로 대체한 결과와 비교했을 때 크게 다르지 않다는 것을 확인하였다.

5) 랜덤 포레스트 기반 대체

랜덤 포레스트 기반 대체를 위해 R패키지 caret 및 RandomForest를 사용하였다. 파라미터 튜닝은 〈부표 A-1〉 및 〈부표 A-2〉와 같이 설정하였다.

6) 서포트 벡터 머신 기반 대체

서포트 벡터 머신 기반 대체를 위해서 R패키지 e1071을 사용하였다. 파라미터 튜닝은 〈부표 A-3〉 및 〈부표 A-4〉와 같이 설정하였다.

7) 인공 신경망 기반 대체

인공 신경망 기반 대체를 위해 R패키지 caret 및 neuralnet을 사용하였다. 파라미터 튜닝은 〈부표 A-5〉 및 〈부표 A-6〉과 같이 설정하였다.

나. 대체 방법 평가지표

대체 방법의 평가지표는 편향 및 제곱근평균제곱오차, 참값의 포함률, 예측의 정확성 지표, 추정의 정확성 지표를 고려하였다. 예측의 정확성 지표는 평균절대편차와 제곱근평균제곱편차이고, 추정의 정확성 지표는 평균의 절대 상대 차이, 변동계수의 절대 차이, 표준화된 왜도의 절대 차이 그리고 표준화된 첨도의 절대 차이이다.

참값의 포함률(COV)을 제외한 나머지 평가지표의 값은 작을수록 대체 방법의 효과가 좋다고 볼 수 있다.

추가적으로 히스토그램(histogram) 및 상자그림(box plot)을 활용하여 대체 방법별 대체된 자료의 분포도 살펴보았다.

3. 결과

모의실험 결과표에서 대체 방법의 표기를 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 방법을 의미한다. 설명변수의 개수에 따라 구분하기 위해서 knn을 예로 들면 설명변수가 3개인 경우 knn_3으로, 6개인 경우에는 knn_6으로 표기하였다. 나

머지 대체 방법도 동일하게 적용하였다. 마지막으로 완전한 자료는 'complete'로 표기하였다.

평가지표도 간단하게 표기하기 위해 'BIAS'는 편향이고, 'RMSE'는 제곱근평균제곱오차, COV는 참값의 포함률이다. 예측의 정확성 지표의 경우 'MAD'는 평균절대편차, 'RMSD'는 제곱근평균제곱편차이고 추정의 정확성 지표의 경우 'ADB'는 평균의 절대 상대 차이, 'ADV'는 변동계수의 절대 차이, 'ADS'는 표준화된 왜도의 절대 차이 그리고 'ADK'는 표준화된 첨도의 절대 차이를 의미한다.

결과표는 무응답 비율에 따라 대체 방법별 평가지표 결과를 정리하여 효과를 살펴볼 수 있도록 하였다. 한국복지패널자료에 근거한 모의실험 결과는 아래와 같다. 본문에서는 무응답 비율 5%, 20%, 50%에 대해서만 설명하였고 10%와 30%는 부록으로 첨부하였다.

가. 무응답 비율 5%에 대한 결과

[그림 4-1]은 대체 자료의 경상소득 평균 추정치와 참값과의 편향을 나타낸 것으로 3개의 보조변수를 사용한 랜덤 포레스트 대체(rf_3)는 7000원으로 대체 방법 중 가장 작은 편향을 가졌을 뿐만 아니라 실제값과 거의 유사하였다. 대체 방법의 편향 크기는 절댓값으로 비교한다. 전반적으로 보면 평균, 핫덱, 비대체 방법보다 K-최근접 이웃, 랜덤 포레스트, 서포트 벡터 머신, 신경망 대체 방법에 대한 결과의 편향이 작은 편에 속하였다. 또한 제곱근평균제곱오차의 값도 작은 편인 것으로 나타났다([부도 A-1] 참조).

대체군을 활용하지 않은 핫덱대체(hotdeck)가 -87만 1000원, 평균대체(mean)가 -80만 9000원으로 가장 큰 편향을 가진다. 대체군을 활용

한 평균대체(mean_3)와 핫덱대체(hotdeck_3)는 -20만 원 및 -28만 2000원으로 대체군을 활용하지 않았을 때보다 크게 감소하였다. 이는 보조변수를 활용하여 대체군을 형성한 뒤 대체군 내에서 대체를 실시하면 대체의 정확도가 높아진다는 기존 연구와 부합하는 결과다.

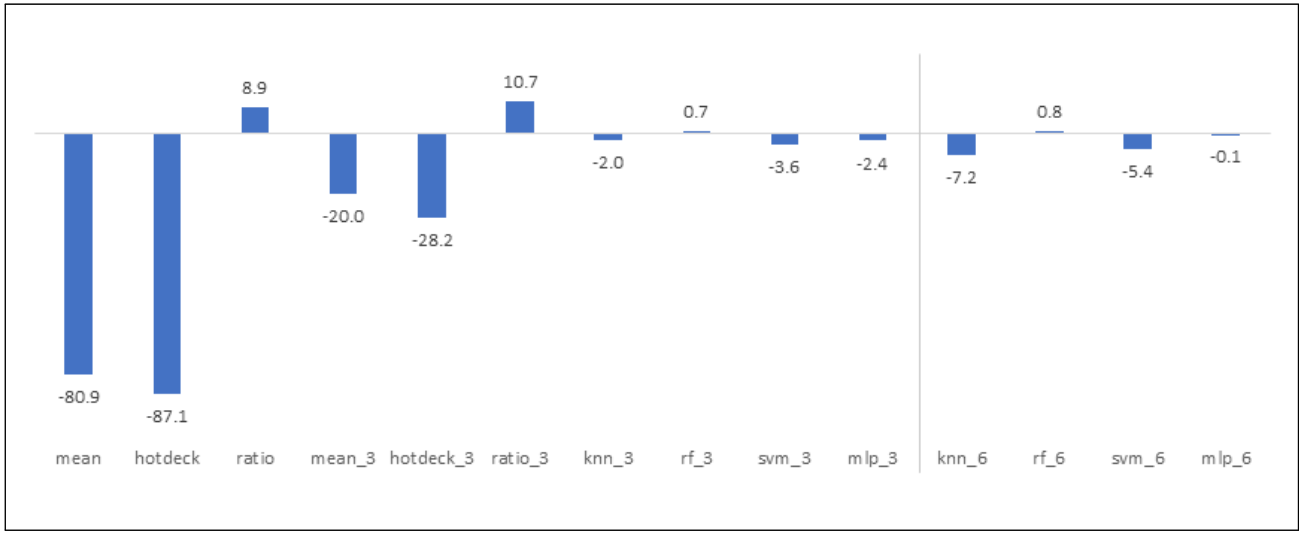
한편 대체군을 활용한 비대체(ratio)와 대체군을 활용하지 않은 비대체(ratio_3)는 평균 및 핫덱 대체에 비해 작은 편으로 나타났다. 비대체의 경우 이전 차수 경상소득과 해당 차수 경상소득 간 변동이 크지 않아서 비(ratio)가 약 1로 계산되어 대체 효과가 좋은 것으로 나타났다.

추가적으로 6개의 보조변수를 사용하여 무응답 대체를 실시한 결과는 다음과 같다. 신경망에 근거하여 대체(mlp_6)한 경우 편향이 1000원으로 실제값과 매우 유사하였다. K-최근접 이웃 대체(knn_6)의 경우는 편향이 -7만 2000원으로 가장 크게 나타났으며 3개의 보조변수를 사용한 경우(-2만 원)보다 크게 나타났다. 이처럼 K-최근접 이웃 대체의 경우 4차원 이상이 되면 '차원의 저주'로 과적합되어 선형회귀보다 못한 효과를 가질 수 있다.

[그림 4-2]는 평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률에 대한 것이다. 무응답 비율이 낮은 편인데도 hotdeck의 포함률은 86%로 나타났다. hotdeck과 mean(93%)을 제외한 나머지 대체 방법은 참값을 모두 포함하는 것으로 나타났다.

[그림 4-1] 무응답 비율 5%에서의 전체 평균에 대한 편향(한국복지패널)

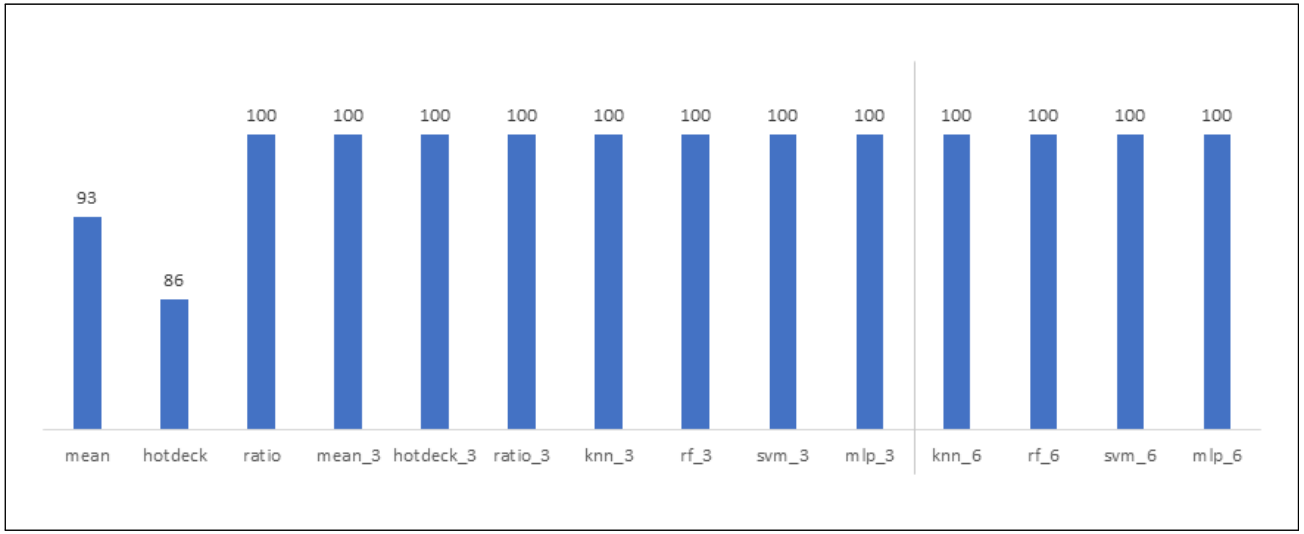
(단위: 만 원)



자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〔그림 4-2〕 무응답 비율 5%에서의 참값의 포함률(한국복지패널)

(단위: %)



자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〈표 4-9〉는 예측 및 추정의 정확성에 대한 결과이다. 예측의 정확성은 개체의 대체값과 실제값이 얼마나 근사한지를 평가하는 것이고, 추정의 정확성은 대체값이 실제값 분포의 적률들을 얼마나 잘 유지하는지를 평가하는 것이다.

〈표 4-9〉 무응답 비율 5%에서의 추정의 정확성 지표 결과(한국복지패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	160.3	1093.8	0.01931	0.01629	0.53454	18.39110
hotdeck	217.3	1418.7	0.02078	0.01655	0.34173	8.99175
ratio	64.7	748.8	0.00248	0.01497	0.42592	11.71610
mean_3	101.6	817.2	0.00478	0.01978	0.33284	11.75580
hotdeck_3	144.6	1083.9	0.00673	0.01093	0.27725	7.70607
ratio_3	65.2	756.1	0.00282	0.01535	0.44463	12.48710
knn_3	63.4	577.5	**0.00147**	0.01206	0.21374	6.19440
rf_3	72.6	659.1	0.00161	**0.00979**	**0.21236**	**5.75082**
svm_3	**61.1**	**576.0**	0.00149	0.01129	0.21259	6.14365
mlp_3	88.2	738.0	0.00196	0.01747	0.24303	8.41989
knn_6	58.8	581.8	0.00205	0.01291	0.22563	7.00750
rf_6	54.1	532.8	0.00145	0.00993	0.20252	5.37427
svm_6	51.1	525.2	0.00163	0.01063	0.21032	5.93953
mlp_6	61.8	604.6	0.00164	0.01268	0.20939	6.22020

주: 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

먼저 예측의 정확성을 보면 svm_3(MAD: 61.1, RMSD: 576)이 대체 방법 중에서 가장 우수하였고, 반면에 hotdeck(MAD: 217.3, RMSD: 1418.7)은 가장 큰 값을 가진다.

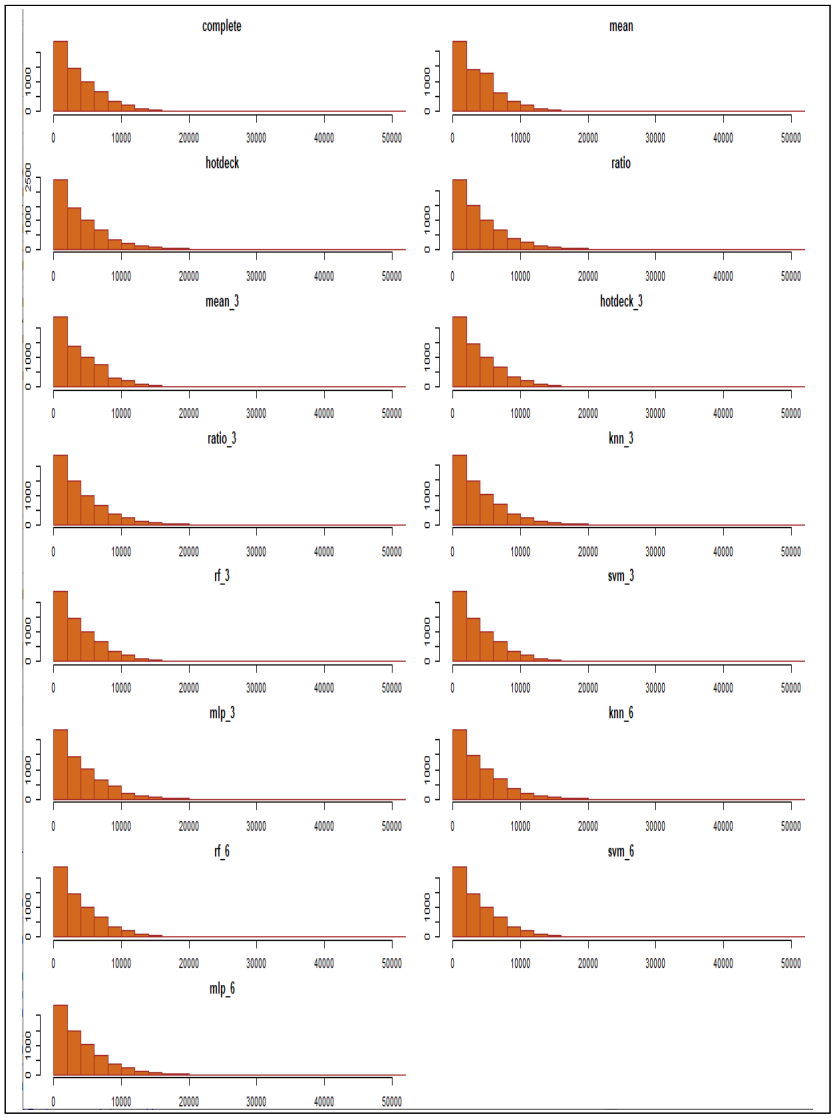
추가적인 분석에 대한 결과는 6개 설명변수를 사용한 rf_6, svm_6, mlp_6은 3개일 때보다 작은 편으로 나타났다.

추정의 정확성에 대한 〈표 4-9〉를 보면 4가지 지표값의 차이가 대체 방법에 따라 아주 큰 편은 아니지만 효과를 살펴보면 다음과 같다. rf_3과 svm_3은 4개의 지표값이 모두 우수한 편에 속하였다. 한편 mean은 가장 큰 값을 가져서 결과가 좋지 않은 것으로 나타났다.

설명변수 6개를 사용하여 대체한 rf_6, svm_6, mlp_6은 설명변수가 3개인 경우보다 작은 값을 가진다. 그러나 knn_6은 knn_3보다 약간 크게 나타났다.

다음은 히스토그램 및 상자그림을 활용하여 대체 방법별 대체된 자료의 분포를 살펴보았다([그림 4-3] 및 [그림 4-4] 참조). 모의실험에 사용된 100개 자료 중에서 8번째 자료를 사용하였다. 무응답 비율이 5%로 낮은 편이어서 전반적으로 보면 대체된 후 자료의 분포는 실제 자료와 비슷하게 유지되고 있다. 단, mean의 경우 히스토그램에서는 3번째 막대가 높은 편이며 상자그림에서는 구간의 길이가 약간 짧은 편으로 나타났다. hotdeck의 경우 편향과 제곱근평균제곱오차는 큰 편에 속하였으나 실제 분포와 유사하여 분포 유지 측면에서는 우수한 것으로 나타났다. 이는 핫덱대체는 분포를 유지하면서 무작위로 기증자를 추출하는 방법이기 때문이다.

〔그림 4-3〕 무응답 비율 5%에서 8번째 자료의 분포-1(한국복지패널)

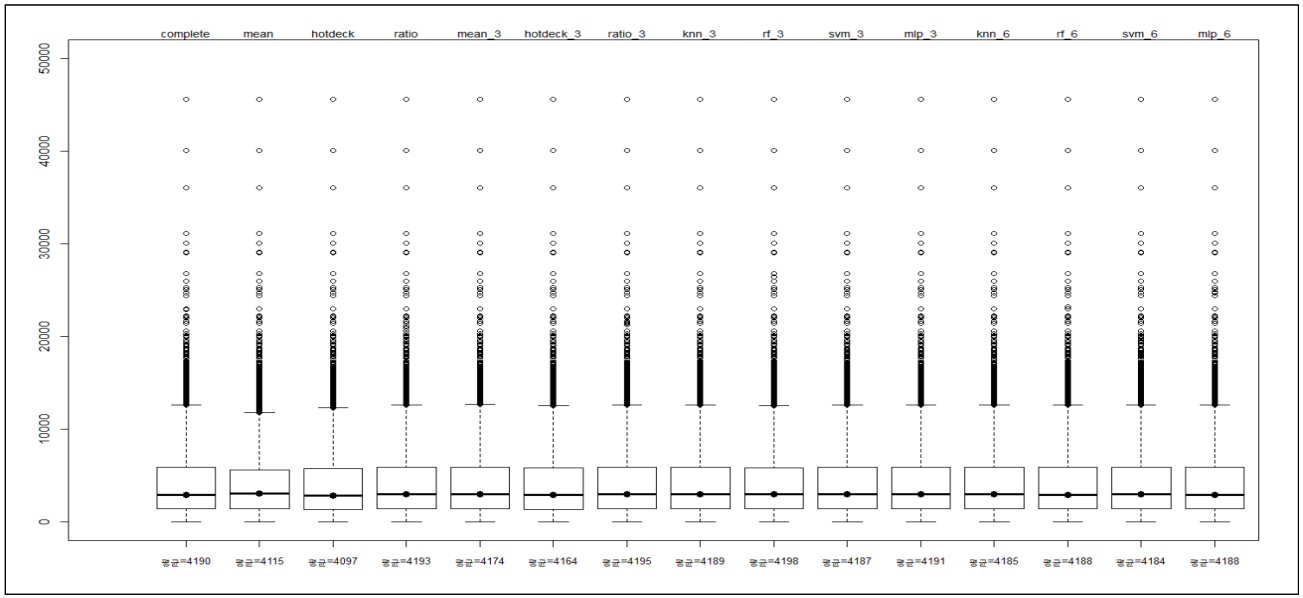


주: X축은 경상소득(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[그림 4-4] 무응답 비율 5%에서 8번째 자료의 분포-2(한국복지패널)



주: X축은 경상소득 평균(단위: 만 원)이고, Y축은 경상소득(단위: 만 원)을 나타냄.
자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

나. 무응답 비율 20%에 대한 결과

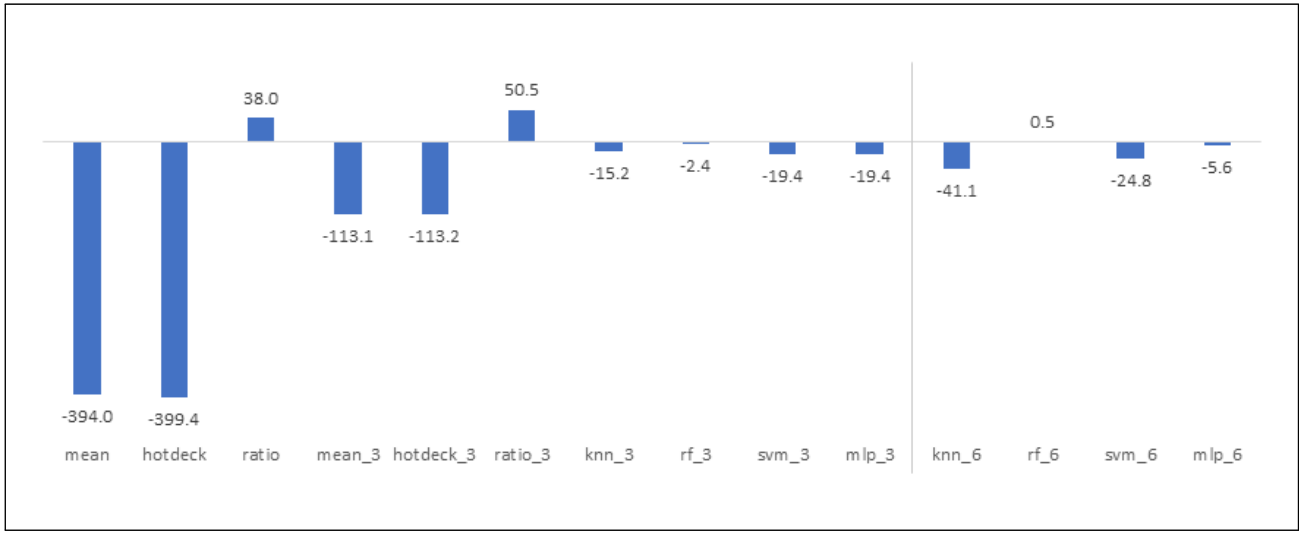
전반적으로 무응답 비율 5%일때 편향보다 값이 증가하였는데 rf_6은 5000원으로 약간 감소하였다(무응답 비율 5%의 경우 8000원). rf_3의 편향은 -2만 4000원으로 가장 작을 뿐만 아니라 무응답 비율 5%에서 20%로의 증가 폭도 가장 작았다. 전반적으로 기계학습 대체 방법의 편향이 작은 편에 속하였다. 더불어 제곱근평균제곱오차도 작은 편으로 나타나 안정적인 결과를 보였다([부도 A-7] 참조).

추가적인 분석을 보면 rf_6(5000원)은 rf_3(-2만 4000원)에 비해 편향이 감소하여 참값과 유사하게 추정되었다. 랜덤 포레스트 대체 방법은 하이퍼파라미터 튜닝 시 설명변수를 3~5개로 설정하였는데 대체적으로 3개가 최종 선택되었다. 그중 경상소득과 상관관계가 높은 생활비 변수가 선택되면서 대체 효과가 나아졌다고 볼 수 있다. mlp_6(-5만 6000원)도 mlp_3(-19만 4000원)보다 큰 폭으로 감소하여 우수한 결과를 보였다. 반면에 knn_6은 knn_3보다 큰 편으로 무응답 비율 5%의 결과와 동일하였다. 또한 knn_6의 제곱근평균제곱오차도 44.4로 knn_3(23.9)에 비해 1.9배 정도 큰 값을 가져 저조한 결과를 보였다([부도 A-7] 참조).

평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률의 경우 mean과 hotdeck은 하나도 포함하지 못하였으며 hotdeck_3과 mean_3의 경우도 33%와 23%로 낮은 편으로 나타났다([그림 4-6] 참조). knn_6의 경우 99%로 하나를 포함하지 못하였고, 나머지 대체 방법은 참값을 모두 포함하여 우수한 결과를 보였다.

[그림 4-5] 무응답 비율 20%에서의 전체 평균에 대한 편향(한국복지패널)

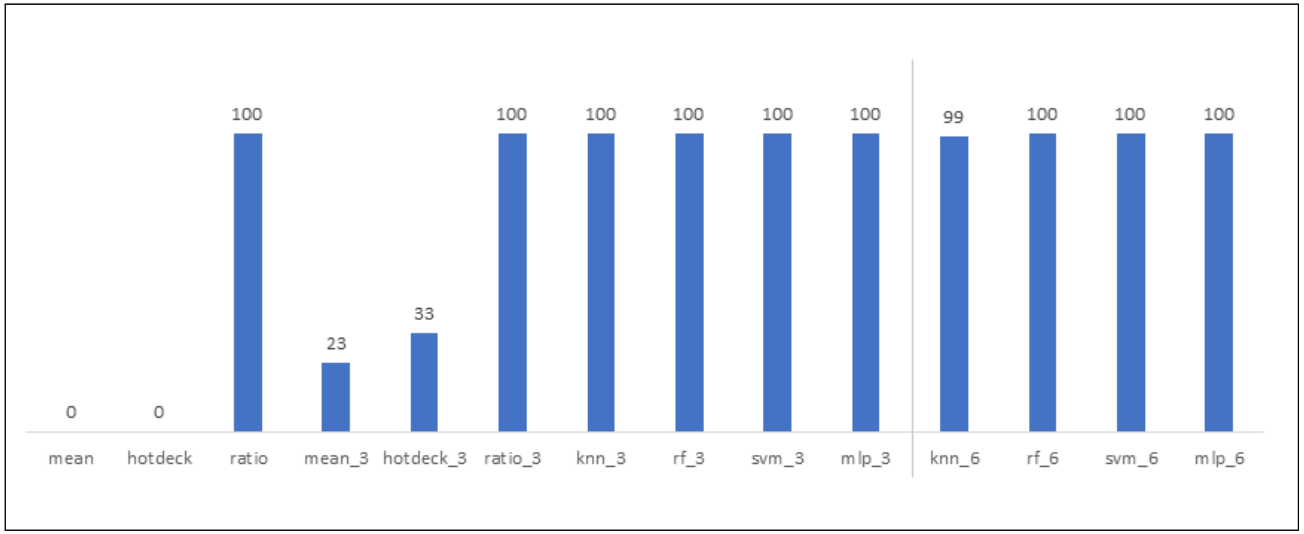
(단위: 만 원)



자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[그림 4-6] 무응답 비율 20%에서의 참값의 포함률(한국복지패널)

(단위: %)



자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〈표 4-10〉은 예측 및 추정의 정확성에 대한 결과이다. 먼저 예측의 정확성을 보면 svm_3(MAD: 247.5, RMSD: 1272.8)이 가장 작고, hotdeck(MAD: 872.2, RMSD: 2923.8)이 가장 크게 나타났다. 이는 무응답 비율 5%의 결과와 동일하였다.

〈표 4-10〉 무응답 비율 20%에서의 추정의 정확성 지표 결과(한국복지패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	657.1	2300.1	0.09404	0.06944	2.27554	89.61940
hotdeck	872.2	2923.8	0.09532	0.05597	1.61778	55.11270
ratio	263.6	1726.2	0.00907	**0.03483**	1.13027	31.86550
mean_3	409.4	1729.1	0.02698	0.08627	1.22637	54.48140
hotdeck_3	590.1	2384.0	0.02702	0.03982	1.42534	44.10200
ratio_3	266.7	1752.1	0.01205	0.03760	1.18582	34.83290
knn_3	259.0	1277.4	0.00446	0.05244	0.78579	25.34478
rf_3	295.4	1464.0	**0.00431**	0.03581	**0.76423**	**19.87692**
svm_3	**247.5**	**1272.8**	0.00513	0.04858	0.79360	24.26302
mlp_3	355.3	1569.7	0.00593	0.07470	0.84153	36.04684
knn_6	243.6	1296.1	0.00981	0.05628	0.82424	30.22904
rf_6	222.2	1200.6	0.00362	0.04259	0.80301	20.08513
svm_6	210.0	1186.6	0.00604	0.04536	0.79734	23.26994
mlp_6	249.9	1321.7	0.00415	0.05381	0.78611	24.68146

주: 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

rf_6, svm_6, mlp_6에 대한 예측의 정확성은 설명변수가 3개일 때보다 작은 편에 속하였다. 이 중에서 rf_6의 값이 가장 작은 것으로 나타났다.

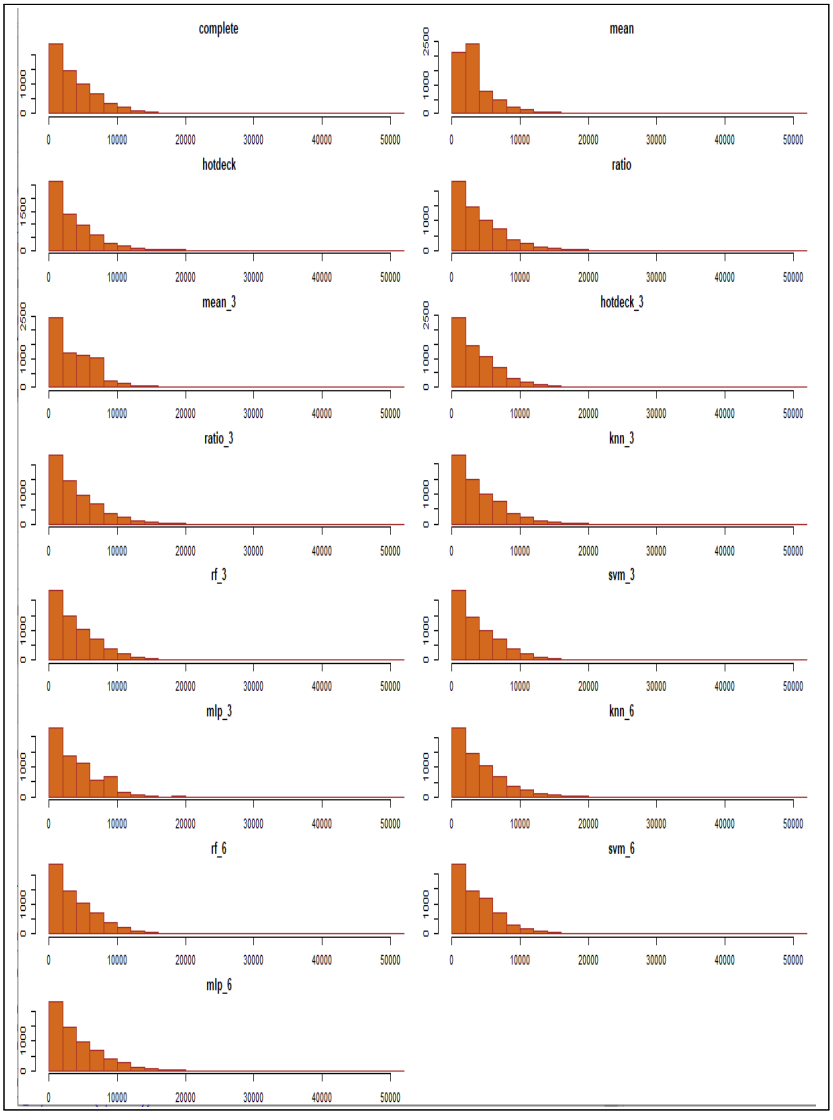
추정의 정확성에 대한 〈표 4-10〉을 보면 rf_3의 결과가 우수한 것으로 나타났다. 다음으로 knn_3, svm_3이 작은 편에 속하였다. 반면에 hotdeck은 가장 저조한 결과를 보였다.

설명변수 6개를 사용하여 대체한 경우 mlp_6은 mlp_3일 때보다 작은 값을 가진다. 반면에 knn_6은 knn_3보다 값이 소폭 증가하였는데, 이는 무응답 비율 5%의 경우와 동일한 결과이다.

[그림 4-7]의 히스토그램을 보면 mean, mean_3, mlp_3의 자료 분포는 실제 자료와 다른 형태를 보이고 있다. 나머지 대체 방법은 전반적으로 유사한 분포를 보였다.

[그림 4-8]의 상자그림을 보면 무응답 비율이 5%인 경우보다 대체 방법에 따른 분포의 변화가 있다. mean은 무응답이 모두 응답값의 평균으로 대체되었기 때문에 구간의 길이가 매우 짧게 나타났다. 더불어 hotdeck, hotdeck_3, svm_6도 약간 짧은 편으로 나타났다.

〔그림 4-7〕 무응답 비율 20%에서 8번째 자료의 분포-1(한국복지패널)

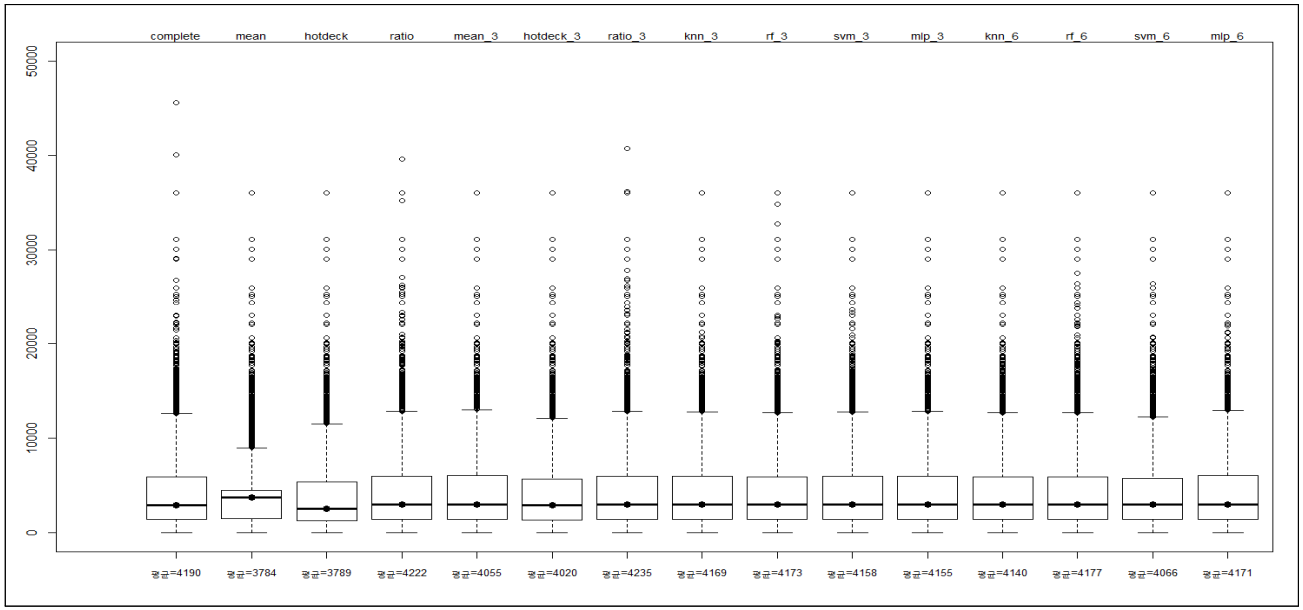


주: X축은 경상소득(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[그림 4-8] 무응답 비율 20%에서 8번째 자료의 분포-2(한국복지패널)



주: X축은 경상소득 평균(단위: 만 원)이고, Y축은 경상소득(단위: 만 원)을 나타냄.
자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

다. 무응답 비율 50%에 대한 결과

[그림 4-9]를 보면 rf_3의 편향은 -31만 1000원으로 가장 작고 다음으로 svm_3이 -63만 3000원, ratio가 132만 2000원으로 작았다. ratio는 ratio_3(263만 6000원)보다 절반 정도 작은 편향을 가지는 것으로 나타났다. 이는 비대체에서 비율을 사용하여 대체할 때 ratio의 비율은 1.008 정도였으나, ratio_3의 비율은 대체군마다 다른 비율을 가져서 최소 1.007에서 최대 1.012였다(모의실험 자료 10 기준). 대체군을 활용한 경우 대체군 내에 속한 개체의 개수가 적어져서 비의 변동이 커지게 되어 오히려 예측력이 좋지 않게 된 것으로 볼 수 있다. 한편 rf_3, svm_3과 ratio의 제곱근평균제곱오차도 다른 대체 방법에 비해 작은 편으로 나타나 안정적인 결과를 보였다([부도 A-13] 참조).

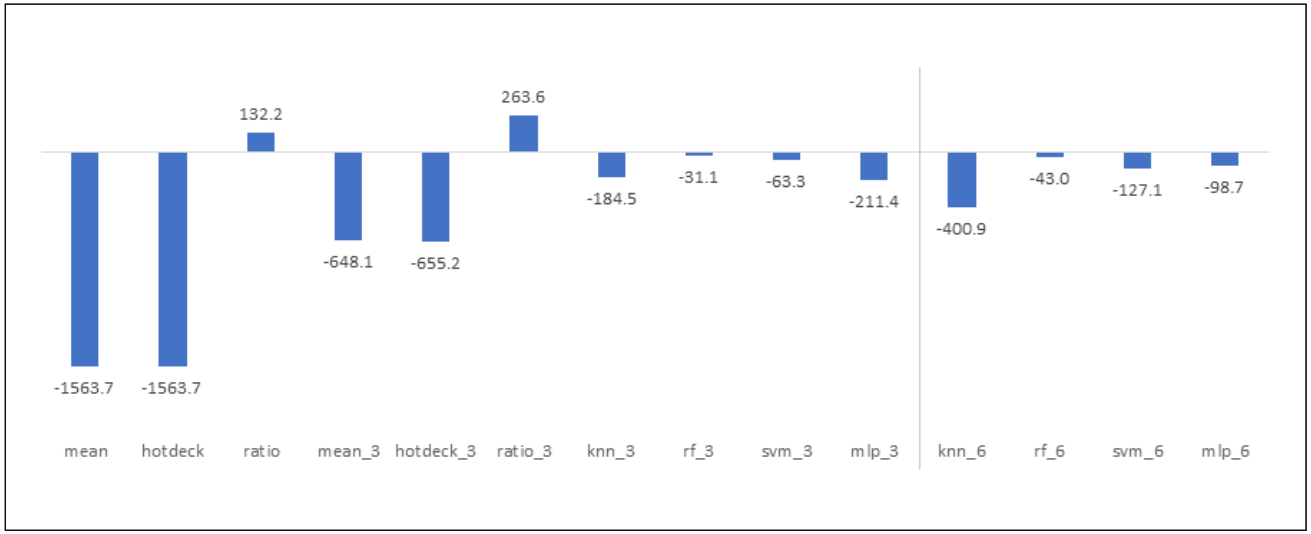
mlp_6의 편향은 -98만 7000원으로 mlp_3(-211만 4000원)에 비해 크게 감소하였다. 반면에 knn_6, rf_6, svm_6은 3개의 보조변수를 사용하여 대체한 결과보다 증가하여 좋지 않은 결과를 보였다. 제곱근평균제곱오차의 결과도 mlp_6은 우수하였으나 knn_6은 큰 폭으로 증가하여 저조한 결과로 나타났다([부도 A-13] 참조).

전반적으로 대체 방법에 대한 참값의 포함률은 저조한 편이나 rf_3과 svm_3의 경우 76%와 70%로 가장 높은 포함률을 보였다([그림 4-10] 참조). 평균, 핫덱, 비대체 중에서 ratio만이 포함률을 가지나 12%로 낮은 편에 속하였다.

rf_6의 포함률은 79%로 rf_3(76%)에 비해 약간 증가했고, mllp_6(35%)도 mlp_3(4%)보다 크게 증가하였다. svm_6(32%)은 svm_3(70%)의 절반에도 미치지 못하였다.

[그림 4-9] 무응답 비율 50%에서의 전체 평균에 대한 편향(한국복지패널)

(단위: 만 원)

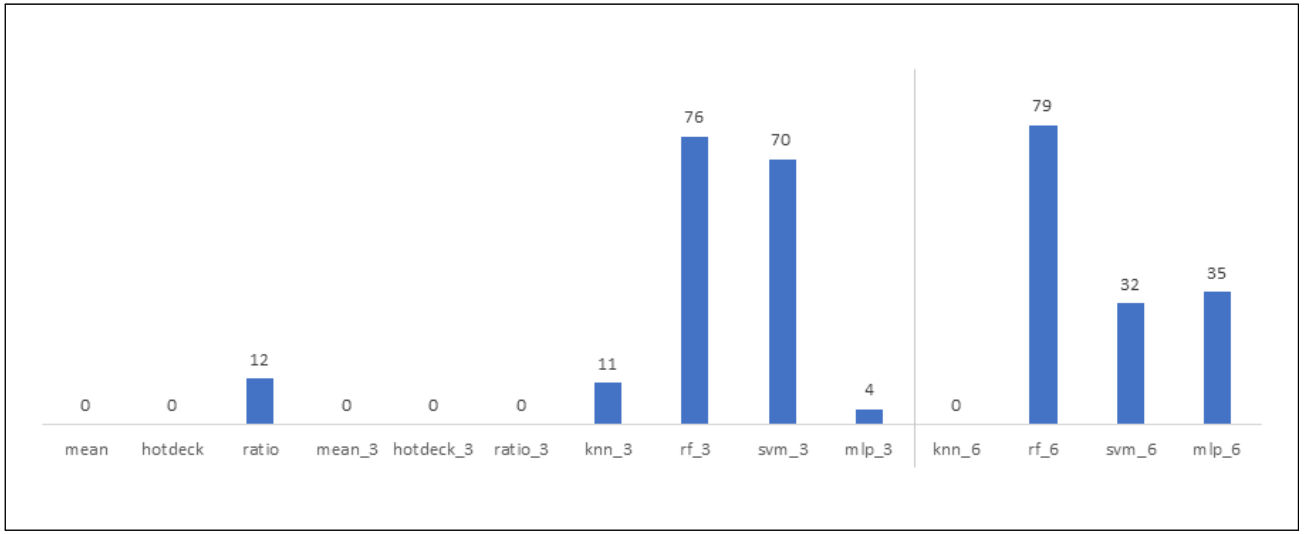


주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[그림 4-10] 무응답 비율 50%에서의 참값의 포함률(한국복지패널)

(단위: %)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〈표 4-11〉에서 예측의 정확성 지표는 svm_3이 가장 작고 hotdeck이 가장 큰 것으로 나타나 앞선 결과와 동일하였다. knn_3의 경우도 작은 편에 속하여 예측의 정확성이 우수하다고 볼 수 있다.

〈표 4-11〉 무응답 비율 50%에서의 추정의 정확성 지표 결과(한국복지패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	1835.5	4023.4	0.37323	0.22876	12.85180	751.87400
hotdeck	2136.3	4539.5	0.37323	0.16693	8.45286	353.82200
ratio	662.6	2829.6	0.03155	**0.06051**	**1.88230**	**49.28710**
mean_3	1123.9	2995.2	0.15469	0.28661	6.41505	318.00200
hotdeck_3	1463.3	3880.3	0.15638	0.16583	7.02469	233.52800
ratio_3	707.5	2977.5	0.06291	0.08287	1.91765	51.63770
knn_3	696.2	2208.1	0.04437	0.15901	2.42653	103.57398
rf_3	735.0	2297.3	**0.01395**	0.10569	2.03726	63.32083
svm_3	**624.3**	**2079.2**	0.01819	0.12509	2.12073	70.72690
mlp_3	882.3	2564.3	0.05075	0.22692	3.23306	143.05131
knn_6	734.9	2333.8	0.09569	0.18223	3.67192	167.31424
rf_6	569.8	2032.3	0.01474	0.12785	2.23116	65.69976
svm_6	581.7	2114.6	0.03058	0.14288	2.27187	82.45985
mlp_6	643.1	2183.5	0.02511	0.15900	2.23766	93.49614

주: 1) 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

rf_6은 6개의 설명변수를 사용하여 대체한 방법 중에서 효과가 가장 좋으며, rf_3의 지표보다도 작았다. 반면에 knn_6은 knn_3보다 큰 값을 가진다.

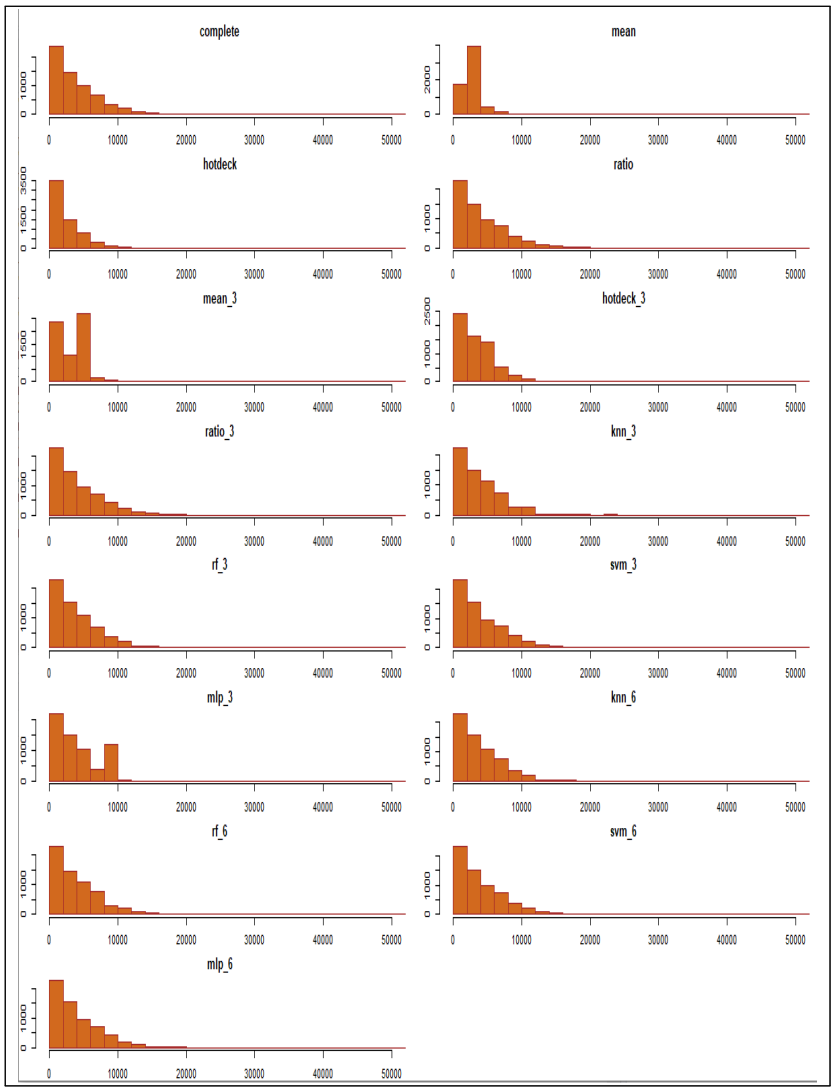
추정의 정확성은 ratio, ratio_3, rf_3의 결과가 우수한 편으로 나타났다. 다른 무응답 비율에서의 결과에 비해 ratio가 우수한 편으로 나타났다. 한편 svm_3도 rf_3의 결과와 유사하게 나타나 효과가 좋은 편에 속하였다.

mlp_6은 mlp_3보다 작은 값을 가져 효과가 향상되었으나 나머지는 3개일 때보다 모두 큰 편이었다. 특히, knn_6의 증가 폭이 가장 크게 나타났다.

무응답 비율이 50%로 높아지면서 mean, mean_3, hotdeck, hotdeck_3, mlp_3의 분포는 실제 분포와 다른 형태를 보였다([그림 4-11] 참조). 평균대체의 경우는 절반의 개체가 응답값의 평균으로 대체된 영향으로 볼 수 있다. 더불어 핫덱대체도 기증자 부족 문제로 동일한 개체를 가지고 중복 대체를 허용하였기 때문이다.

[그림 4-12]를 보면 mean의 상자그림은 실제와 매우 다른 형태를 가지는 것으로 나타났다. 이와 함께 hotdeck과 hotdeck_3은 구간의 길이가 짧은 편에 속하였다.

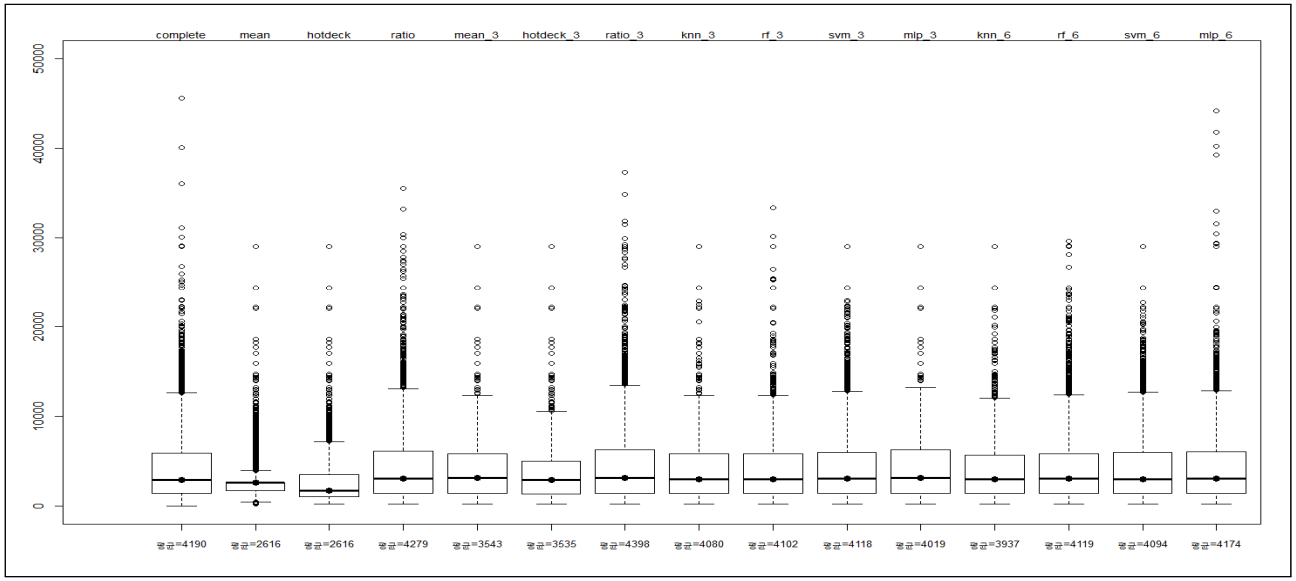
〔그림 4-11〕 무응답 비율 50%에서 8번째 자료의 분포-1(한국복지패널)



주: 1) X축은 경상소득(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[그림 4-12] 무응답 비율 50%에서 8번째 자료의 분포-2(한국복지패널)



주: 1) X축은 경상소득 평균(단위: 만 원)이고, Y축은 경상소득(단위: 만 원)을 나타냄.
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

4. 소결

이 절에서는 한국복지패널자료에 근거하여 '경상소득'을 대체하는 모의실험을 실시하였다. 다양한 무응답 비율(5%, 10%, 20%, 30%, 50%)에서 기계학습 통계방법을 이용한 대체 방법의 효과를 살펴보았다. 이때 무응답 대체 시 많이 사용하는 평균, 핫덱, 비대체 방법도 함께 비교하였다. 〈표 4-12〉는 각각의 무응답 비율에서 평가지표별 가장 우수한 효과를 가지는 대체 방법에 해당 무응답 비율을 표기하여 모의실험 결과를 정리하였다.

3개의 설명변수를 사용한 랜덤 포레스트 대체 방법의 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.

다음으로 3개 설명변수를 사용한 K-최근접 이웃 및 서포트 벡터 머신 대체 방법의 결과도 좋은 편에 속하였다. 특히 3개의 설명변수를 사용한 서포트 벡터 머신 대체 방법은 무응답 비율에 상관없이 개별 개체의 대체값과 실제값과의 유사 정도(예측의 정확성)가 가장 우수하였다. 이렇듯 기계학습 통계방법을 이용한 대체 방법의 효과는 좋은 편으로 나타났다.

한편 대체군을 활용하지 않은 비대체도 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 패널자료의 항목무응답 대체 시 많이 사용하고 있는 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 인지하였다. 더불어 대체 대상 변수의 특징을 고려하여 대체군을 보다 견고하게 구분한다면 대체 효과도 좀 더 향상될 수 있을 것이다.

〈표 4-12〉 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: COV는 91% 이상의 포함률을 만족하는 경우임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

다음은 추가적으로 실시한 모의실험의 결과를 정리하였다. 무응답 대체 시 무응답 대체 대상 변수와 관련 있는 여러 개의 설명변수를 더 포함하여 무응답 처리 효과를 살펴보았다. 〈표 4-12〉와 동일한 방법으로 4개의 대체 방법 중 평가지표별 가장 우수한 효과를 가지는 대체 방법에 무응답 비율을 표기하였다(〈표 4-13〉 참조).

6개 설명변수를 사용한 랜덤 포레스트 대체 방법은 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 방법은 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 대체의 효과를 향상시키는 중요한 요인은 다음과 같다. 무응답 대체 대상 변수와 관련 있는 설명변수를 탐색하여 선택하고, 적절한 설명변수 개수를 선정하는 등의 과정을 필수적으로 실시해야 한다고 생각한다.

마지막으로 6개 설명변수를 사용한 신경망 대체 방법의 효과는 3개일 때보다 향상된 결과를 보였으며, 무응답 비율이 증가할수록 효과는 더 좋아지는 것으로 나타났다.

〈표 4-13〉 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: COV는 91% 이상의 포함률을 만족하는 경우임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

제2절 한국의료패널자료[7]를 이용한 모의실험

1. 자료 설계

이 절에서는 작년에 공개된 한국의료패널자료 8~9차(2014~2015년)로 모의실험을 하였다. 무응답 대체 대상 변수는 '작년 한 해 월평균 생활비(이하 생활비)'를 사용하였다.

모의실험을 위한 연구 대상 모집단은 8~9차에서 생활비를 모두 응답하고, 무응답을 대체할 때 설명변수로 활용되는 변수들도 모두 응답한 가구만을 추출하여 총 6473가구로 확정하였다.

결측자료 메커니즘은 MAR을 따른다는 가정하에 무응답 자료를 생성하였고, 무응답 비율은 5%, 10%, 20%, 30%, 50%를 고려하였다. 1절에서 실시한 한국복지패널조사와 마찬가지로 한국의료패널조사도 무응답 비율이 높지 않은 편이다. 이를 반영하기 위해 무응답 비율이 가장 낮은 5%에서부터 점진적으로 증가하여 최고 50%까지의 무응답 비율에 따른 무응답 대체 방법 효과를 살펴보았다.

무응답 생성을 위한 자료 설계는 다음과 같다. 무응답을 발생시키기 위하여 생활비와 연관성이 있는 '가구주의 연령(이하 연령)'과 '가구주의 학력(이하 학력)'을 가지고 4개의 대체군 집단(이하 집단)으로 나누었다. 가구주 연령 범주는 60세 미만과 이상인 2개로, 학력 범주는 고졸 미만과 이상인 2개로 구분하였다. 집단별 생활비 평균은 집단 1은 197만 원, 집단 2는 305만 원, 집단 3은 115만 원 그리고 집단 4는 198만 원이었다.

---

7) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출한 자료 활용.
한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출한 자료 활용.

유의수준 5%하에서 (집단 3), (집단 1, 집단 4), (집단 2)로 묶이는 것으로 나타났다.

〈표 4-14〉 연령 및 학력에 따른 집단 구분(한국의료패널)

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

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

〈표 4-15〉 집단별 무응답 발생 개수의 비율(한국의료패널)

(단위: %)

	대졸 미만	대졸 이상
60세 미만	6	58
60세 이상	12	24

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

생활비가 많을수록 응답하지 않을 가능성이 높다는 점을 반영하여 무응답 비율에 따른 무응답 발생 개수의 비율을 셀마다 다르게 구성하였다. 예를 들어 무응답 비율이 10%라고 가정하면, 집단 1의 무응답 발생 개수는 추출해야 하는 무응답 개수에서 0.6%(=10×0.06), 집단 2는 5.8%(=10×0.58), 집단 3은 1.2%(=10×0.12), 집단 4는 2.4%(=10×0.24)가 해당된다.

추가적으로 고려한 가정은 다음과 같다. 생활비는 매해 조사하므로 무응답이 이전 차수의 생활비에 의존하는 것이 가능하여 생활비가 많을수록 무응답이 많이 발생한다고 가정하였다. 이에 따라 4개 집단에서 중위

수를 기준으로 중위수 미만에서는 집단별 무응답 가구 수의 30%, 중위수 이상에서는 70%의 무응답이 발생한다고 가정하였다. MAR 가정을 만족하기 위해서 8차 생활비는 무응답이 없이 모두 값을 가지고 있다고 가정하고 9차 생활비를 무응답으로 처리하였다. 이는 모의실험에서 대체 방법 중 하나인 비대체에서 비를 계산할 때 비의 분모가 결측이 되지 않도록 하는 효과도 가진다. 무응답 비율별 무작위로 추출한 가구의 9차 생활비를 무응답으로 처리하였다. 무응답 비율이 5%인 경우 무응답 발생 개수는 324가구, 10%인 경우 649가구, 20%인 경우 1296가구, 30%인 경우 1942가구, 50%인 경우 3236가구이다. 〈표 4-16〉부터 〈표 4-20〉은 무응답 비율별 생성된 무응답 가구 수를 보여 주고 있다.

〈표 4-16〉 무응답 비율 5%에서의 집단별 무응답 발생 개수(한국의료패널)

(단위: 가구)

	생활비 중위수		전체
	미만	이상	
집단 1	6	13	19
집단 2	56	132	188
집단 3	12	27	39
집단 4	23	55	78
전체	97	227	324

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

〈표 4-17〉 무응답 비율 10%에서의 집단별 무응답 발생 개수(한국의료패널)

(단위: 가구)

	생활비 중위수		전체
	미만	이상	
집단 1	12	27	39
집단 2	113	263	376
집단 3	23	55	78
집단 4	47	109	156
전체	195	454	649

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

〈표 4-18〉 무응답 비율 20%에서의 집단별 무응답 발생 개수(한국의료패널)

(단위: 가구)

	생활비 중위수		전체
	미만	이상	
집단 1	23	55	78
집단 2	225	526	751
집단 3	47	109	156
집단 4	93	218	311
전체	388	908	1296

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

〈표 4-19〉 무응답 비율 30%에서의 집단별 무응답 발생 개수(한국의료패널)

(단위: 가구)

	생활비 중위수		전체
	미만	이상	
집단 1	35	82	117
집단 2	338	788	1126
집단 3	70	163	233
집단 4	140	326	466
전체	583	1359	1942

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

〈표 4-20〉 무응답 비율 50%에서의 집단별 무응답 발생 개수(한국의료패널)

(단위: 가구)

	생활비 중위수		전체
	미만	이상	
집단 1	58	136	194
집단 2	563	1314	1877
집단 3	116	272	388
집단 4	233	544	777
전체	970	2266	3236

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

연구 대상 모집단 자료로부터 무응답 가구를 단순임의추출방법으로 추출하는 과정을 100번 반복하여 최종 모의실험용 무응답 자료 100개를 생성하였다.

2. 대체 방법 및 평가지표

한국복지패널자료와 동일한 대체 방법과 평가지표를 사용하였으며 내용은 다음과 같다.

가. 무응답 대체 방법

무응답 대체 방법은 1) 평균대체, 2) 핫덱대체, 3) 비대체, 4) K-최근접 이웃 대체, 5) 랜덤 포레스트 대체, 6) 서포트 벡터 머신 대체, 7) 신경망 대체로 총 7가지를 고려하였다.

1) 평균대체

대체군 활용 여부에 따라 2개의 대체값을 생성하였다. 먼저 대체군을 사용하지 않는 경우에는 응답값을 가지고 구한 평균값으로 무응답을 대체하였다.

다음으로 대체군을 형성하여 대체군 내에서의 대체를 실시하였다. 대체군은 연령 범주 2개(60세 미만·이상), 학력 범주 2개(고졸 미만·이상) 그리고 8차 생활비 범주 2개(중위수 미만·이상)를 고려하여 8개 대체군으로 구분한 다음 각 대체군 내에서의 평균값을 구하였다. 무응답은 연령, 학력, 8차 생활비 범주에 해당하는 대체군 내의 평균값으로 대체하였다. 평균대체, 핫덱대체 및 비대체도 동일한 대체군을 활용하여 무응답 대체를 실시하였다.

2) 핫덱대체

대체군이 없는 경우의 대체 방법은 응답한 모든 값을 기증자 후보로 사용하여 기증자 중에서 무작위로 추출한 후 기증자의 응답값으로 무응답을 대체한다.

대체군을 사용한 경우는 대체군 내에서 응답한 모든 값을 기증자 후보로 사용하여 기증자 중에서 무작위로 추출한 후 기증자의 응답값으로 무응답을 대체한다. 단, 무응답 비율 30%, 50%의 경우 해당 대체군 내에서의 기증자 부족으로 무응답이 채워지지 않아서 기증자를 중복 사용하였다.

3) 비대체

대체군이 없는 경우는 이전 차수와 해당 차수에 대해 모두 응답값을 가지고 있는 개체만을 대상으로 하여 비(ratio)를 계산하였다. 무응답은 이전 차수의 응답값에 해당 대체군 내의 비를 곱한 값으로 대체하였다.

대체군을 활용한 경우의 대체 방법은 연령 및 학력 범주로 4개의 대체군을 생성한 후 대체군 내에서의 비를 계산하였다. 무응답은 이전 차수의 응답값에 해당 대체군 내의 비를 곱한 값으로 대체하였다.

4) K-최근접 이웃 기반 대체

평균, 핫덱, 비대체에서 보조변수로 사용한 3개 변수를 활용하여 무응답 대체를 실시하였다. 이는 동일한 보조변수를 사용하여 무응답을 대체한 방법들의 효과를 비교하기 위해서이다.

또한 설명변수를 더 포함(3개 → 10개 증가)하여 무응답 대체를 실시하였다. 더 많은 정보를 사용했을 때 무응답 대체의 효과를 살펴보기 위해서이다. 랜덤 포레스트, 서포트 벡터 머신, 신경망 기반 대체 방법에도 동일하게 적용하였고, 〈표 4-21〉은 사용한 설명변수에 대한 목록이다.

〈표 4-21〉 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), 가구근로소득, 만성질환 여부, 저축 여부, 가구 내 근로자 수, 가구원 수

주: 생활비(8차)를 제외한 나머지 설명변수는 무응답 대체 대상 변수와 동일한 9차임.
자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

K-최근접 이웃 기반 대체를 위해 R패키지 caret을 사용하였다. 최고의 모형 적합을 찾기 위해 모델 옵션(option)을 조정하는 파라미터 튜닝(Parameter tuning)은 5-겹 교차 검증(5-fold cross-validation)으로 하였다.

5) 랜덤 포레스트 기반 대체

랜덤 포레스트 기반 대체를 위해 R패키지 caret 및 RandomForest를 사용하였다. 파라미터 튜닝은 〈부표 A-9〉, 〈부표 A-10〉과 같이 설정하였다.

6) 서포트 벡터 머신 기반 대체

서포트 백터 머신 기반 대체를 위해서 R패키지 e1071을 사용하였다. 파라미터 튜닝은 〈부표 A-11〉, 〈부표 A-12〉와 같이 설정하였다.

7) 인공 신경망 기반 대체

인공 신경망 기반 대체를 위해 R패키지 caret 및 neuralnet을 사용하였다. 파라미터 튜닝을 위해 〈부표 A-13〉, 〈부표 A-14〉와 같이 설정하였다.

나. 대체 방법 평가지표

대체 방법의 평가지표는 편향 및 제곱근평균제곱오차, 참값의 포함률, 예측의 정확성 지표 그리고 추정의 정확성 지표를 고려하였다. 예측의 정확성 지표는 평균절대편차와 제곱근평균제곱편차이고, 추정의 정확성 지표는 평균의 절대 상대 차이, 변동계수의 절대 차이, 표준화된 왜도의 절대 차이 그리고 표준화된 첨도의 절대 차이이다.

참값의 포함률(COV)을 제외한 나머지 평가지표의 값은 작을수록 대체 방법의 효과가 좋다고 볼 수 있다.

추가적으로 히스토그램 및 상자그림을 활용하여 대체 방법별 대체된 자료의 분포도 살펴보았다.

3. 결과

대체 방법 및 평가지표의 표기는 한국복지패널자료 모의실험에 대한 결과표의 표기와 동일하게 사용하였다(제1절 한국복지패널자료를 이용한 모의실험에서의 3. 결과 내용 참조).

결과표는 무응답 비율에 따라 대체 방법별 평가지표 결과를 정리하여 효과를 살펴볼 수 있도록 하였다. 한국의료패널자료에 근거한 모의실험의 결과는 아래와 같다. 본문에서는 무응답 비율 5%, 20%, 50%에 대해서만 설명하였고 10%와 30%는 부록으로 첨부하였다.

가. 무응답 비율 5%에 대한 결과

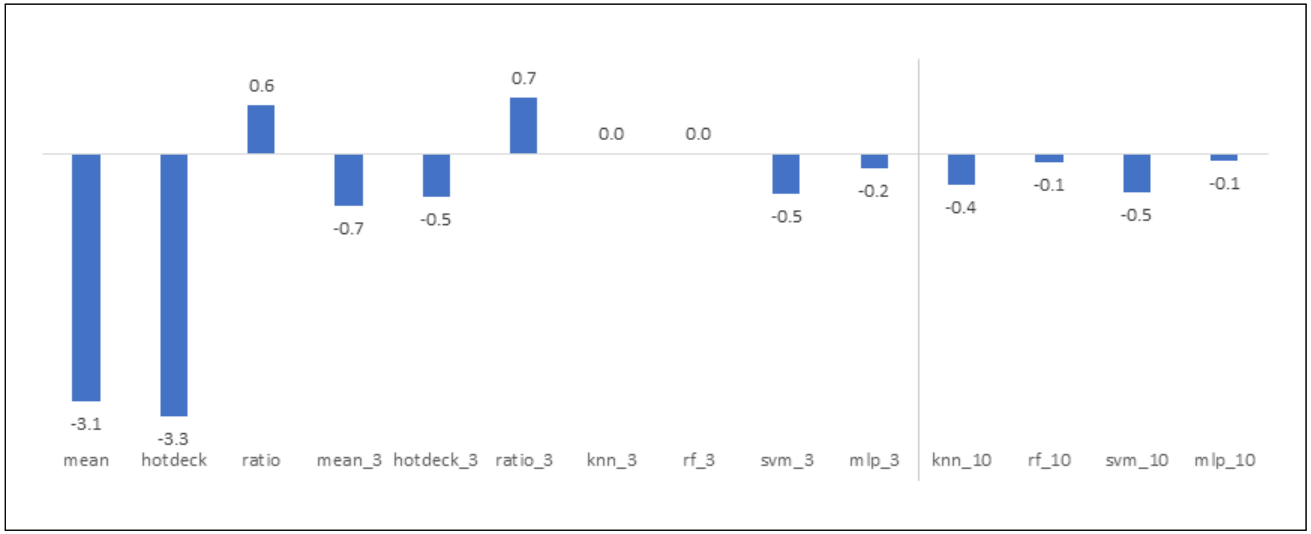
[그림 4-13]은 대체 자료의 생활비 평균 추정치와 참값과의 편향을 보여 주고 있다. 대체 방법의 편향 크기를 절댓값으로 비교하여 설명하면 다음과 같다. 3개의 보조변수를 사용한 K-최근접 이웃(knn_3) 및 랜덤 포레스트(rf_3) 대체의 편향은 0원으로 참값과 동일하게 나타나 우수한 결과를 보였다. 전반적으로 기계학습 대체 방법의 편향이 평균, 핫덱, 비대체 방법보다 작은 것으로 나타났다. K-최근접 이웃(knn_3), 랜덤 포레스트(rf_3)와 신경망(mlp_3) 대체 방법의 제곱근평균제곱오차의 값이 가장 작아서 안정적인 결과를 가진다([부도 A-1] 참조). 반면에 대체군을 활용하지 않은 핫덱대체(hotdeck)와 평균대체(mean)가 -3만 3000원과 -3만 1000원으로 가장 큰 편이었다. 그러나 대체군을 활용한 핫덱대체(hotdeck_3)와 평균대체(mean_3)의 편향은 큰 폭으로 감소하여 효과가 향상되었다. 비대체는 대체군 활용 여부와 상관없이 편향이 비슷하였다.

추가적으로 설명변수를 더 포함하여 대체한 결과를 보면 10개 설명변수를 추가한 랜덤 포레스트(rf_10)와 신경망(mlp_10) 대체 방법에 대한 편향은 1000원으로 가장 작은 값을 가진다. 이와 함께 10개 설명변수를 추가한 랜덤 포레스트(rf_10)와 신경망(mlp_10) 대체 방법의 제곱근평균제곱오차의 값도 3000원으로 안정적인 결과를 보였다([부도 A-1] 참조). 10개의 설명변수를 추가한 K-최근접 이웃 대체 방법(knn_10)의 편향은 4000원으로 3개의 설명변수(knn_3, 0원)일 때보다 증가하였다.

[그림 4-14]는 평균 추정치에 대한 95% 신뢰구간에서의 참값의 포함률로 무응답 비율이 5%로 높지 않은 편인데도 mean(87%)과 hot-deck(76%)의 결과는 좋은 편이 아니었다. 이를 제외한 나머지 대체 방법의 결과는 참값을 모두 포함하는 것으로 나타났다.

[그림 4-13] 무응답 비율 5%에서의 전체 평균에 대한 편향(한국의료패널)

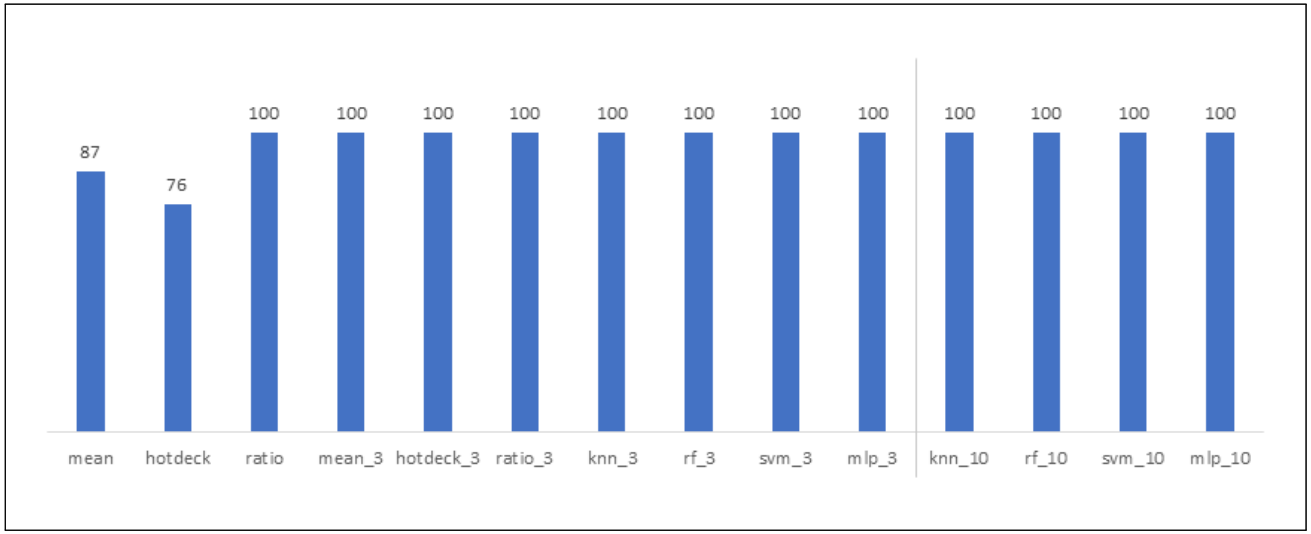
(단위: 만 원)



자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

[그림 4-14] 무응답 비율 5%에서의 참값의 포함률(한국의료패널)

(단위: %)



자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

〈표 4-22〉는 예측 및 추정의 정확성에 대한 결과이다. 전반적으로 기계학습 방법을 적용한 대체 방법이 예측의 정확성이 좋은 편으로 나타났다. svm_3, knn_3, rf_3이 작은 편으로 실제 값과 근사하게 추정하였다고 볼 수 있다. 이와 다르게 hotdeck은 대체 방법 중에서 가장 큰 값을 가지는 것으로 나타났다.

추가적으로 설명변수의 개수를 더 포함하여 대체를 실시한 경우, 10개 설명변수를 사용한 대체 방법은 3개일 때보다 예측의 정확성이 조금 향상되었다. 4개 대체 방법 간 지표의 값은 유사한 것으로 나타났다.

추정의 정확성 지표는 기계학습 대체 방법의 결과가 우수한 편에 속한 반면에 mean은 좋지 않은 것으로 나타났다. 한편 hotdeck은 예측의 정확성 결과와 다르게 추정의 정확성은 우수하였다. 이는 핫덱대체의 장점인 분포 유지를 반영한 결과로 볼 수 있다.

10개의 설명변수를 사용한 대체 방법은 3개인 경우와 지표 간 차이가 크지 않고 유사하게 나타났다.

〈표 4-22〉 무응답 비율 5%에서의 추정의 정확성 지표 결과(한국의료패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	6.3	37.4	0.01402	0.01244	0.05673	0.47694
hotdeck	8.7	50.1	0.01485	0.00484	**0.02601**	0.17559
ratio	4.0	25.8	0.00276	0.00193	0.03312	0.30211
mean_3	4.5	27.8	0.00296	0.01185	0.03407	0.20261
hotdeck_3	6.4	39.2	0.00265	0.00198	0.02974	0.24668
ratio_3	4.0	25.9	0.00316	**0.00184**	0.03242	0.28619
knn_3	**3.7**	23.1	**0.00110**	0.00832	0.04156	0.15011
rf_3	3.8	23.4	0.00116	0.00737	0.03310	0.15008
svm_3	**3.7**	**23.0**	0.00228	0.00744	0.03357	0.15164
mlp_3	4.0	24.9	0.00138	0.00916	0.04459	**0.14755**
knn_10	3.4	21.9	0.00180	0.00665	0.03116	0.14936
rf_10	3.1	19.9	0.00099	0.00626	0.03933	0.16024
svm_10	3.1	20.1	0.00218	0.00652	0.03844	0.15265
mlp_10	3.2	21.1	0.00103	0.00633	0.04518	0.17338

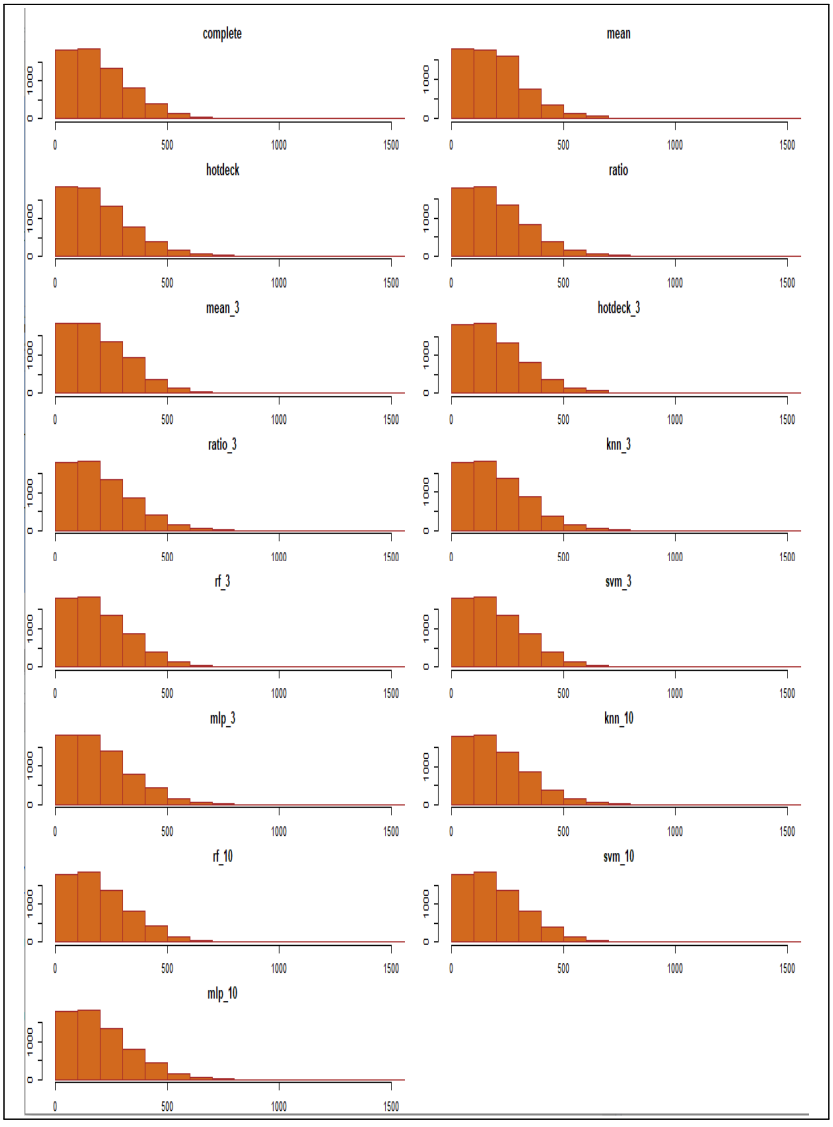
주: 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[그림 4-15] 및 [그림 4-16]은 자료에 대한 분포를 보여 주고 있다. 무응답 비율이 5%로 낮은 편이어서 전반적으로 봤을 때 실제 분포와 유사한 형태를 가지고 있다. 히스토그램의 경우 mean의 분포가 약간 다르게 나타났다. 상자그림의 경우 ratio는 구간의 길이가 약간 짧은 편이고, ratio_3은 이상치가 조금 더 많은 편으로 보인다.

〔그림 4-15〕 무응답 비율 5%에서 1번째 자료의 분포-1(한국의료패널)

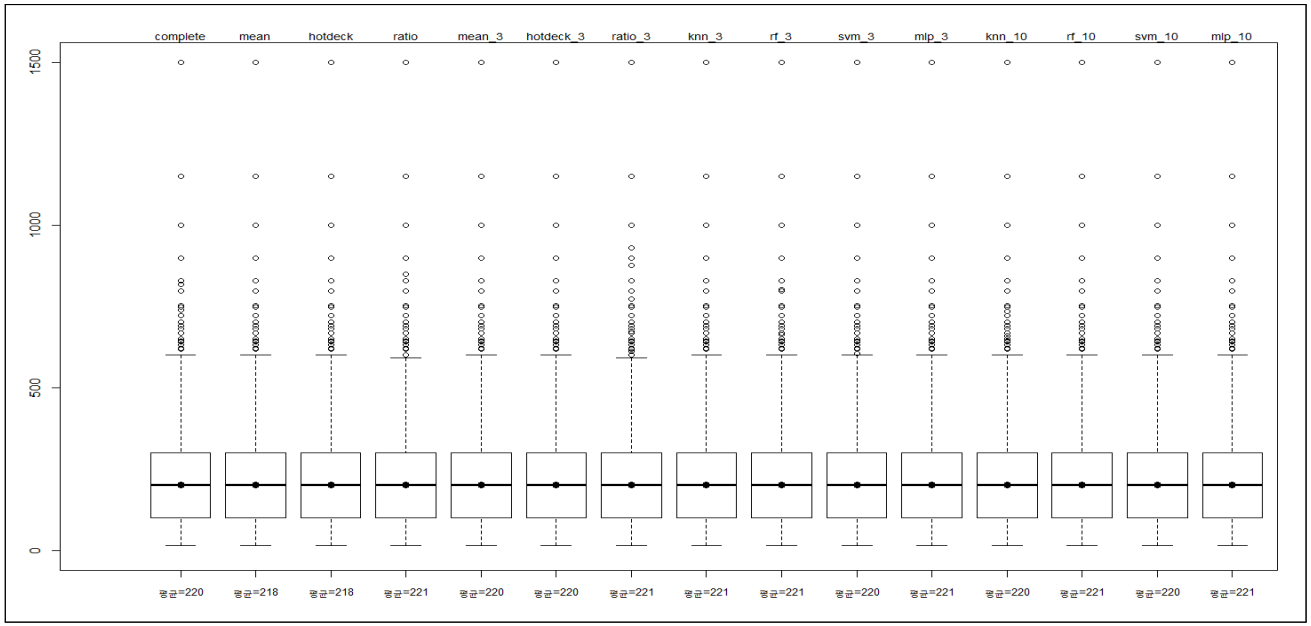


주: X축은 생활비(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[그림 4-16] 무응답 비율 5%에서 1번째 자료의 분포-2(한국의료패널)



주: X축은 생활비 평균(단위: 만 원)이고, Y축은 생활비(단위: 만 원)를 나타냄.
자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

나. 무응답 비율 20%에 대한 결과

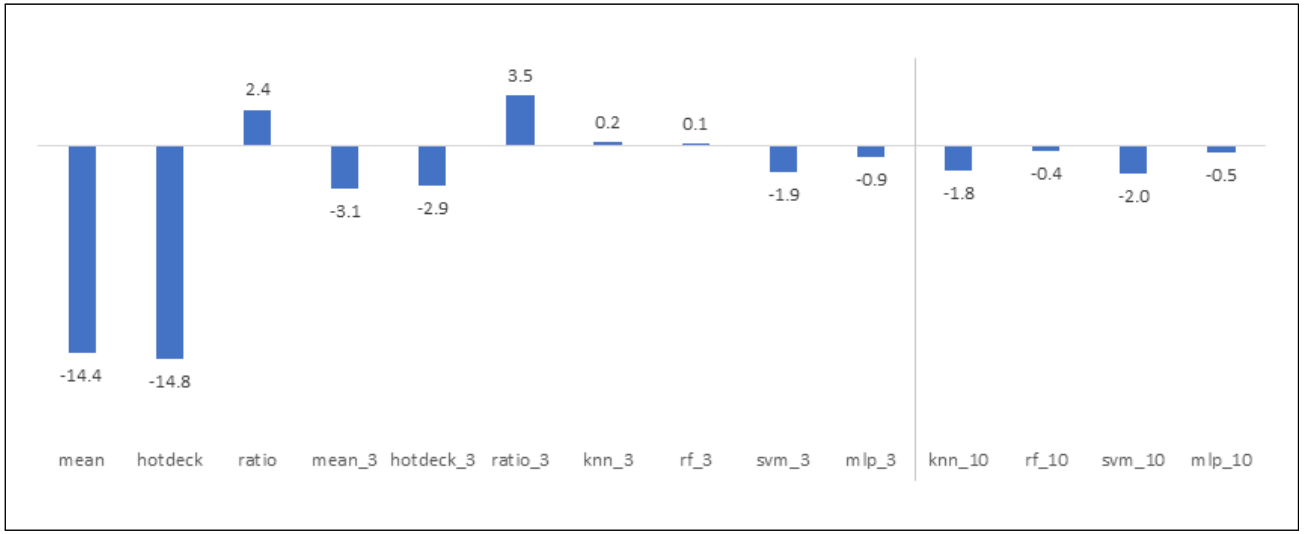
무응답 비율 5%에 비해 전반적으로 편향이 증가하였는데, rf_3과 knn_3은 편향의 증가 폭이 작을 뿐만 아니라 다른 대체 방법에 비해 편향도 가장 작아 참값을 유사하게 추정하였다([그림 4-17] 참조). 또한 rf_3과 knn_3의 제곱근평균제곱오차도 가장 작은 값을 가져서 안정적인 결과를 보였다([부도 A-20] 참조). 전반적으로 무응답 비율 5%일 때와 비슷한 결과를 보였다.

추가적으로 실시한 대체 방법 중에서 knn_10의 편향은 -1만 8000원으로 knn_3(2000원)에 비해 9배로 증가 폭이 가장 크게 나타났다. 참고로 knn_10의 제곱근평균제곱오차도 knn_3보다 약 2.4배 증가하였다([부도 A-20] 참조). 이는 설명변수 개수의 증가로 인한 과적합의 영향으로 볼 수 있다. 이에 반해 mlp_10은 -5000원으로 mlp_3(-9000원)에 비해 소폭 감소하였다.

[그림 4-18]은 평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률로 mean과 hotdeck은 참값을 하나도 포함하지 못하여 저조하였다. 또한 hotdeck_3(79%), ratio_3(64%)과 mean_3(57%)도 높지 않은 편이었다. 이를 제외한 나머지 대체 방법의 포함률은 높은 편에 속하였다.

[그림 4-17] 무응답 비율 20%에서의 전체 평균에 대한 편향(한국의료패널)

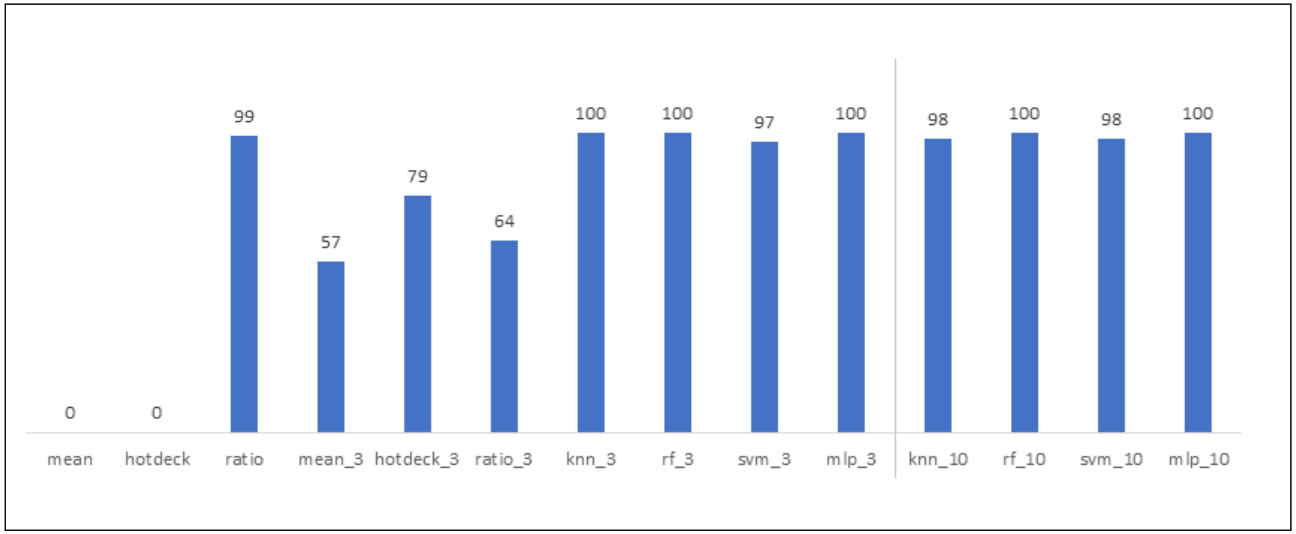
(단위: 만 원)



자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

[그림 4-18] 무응답 비율 20%에서의 참값의 포함률(한국의료패널)

(단위: %)



자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

먼저 〈표 4-23〉에서 예측의 정확성을 보면 svm_3의 값이 가장 작은 것으로 나타나 무응답 비율 5%의 결과와 동일하였다. 다음으로 knn_3과 rf_3도 좋은 편으로 나타났다.

〈표 4-23〉 무응답 비율 20%에서의 추정의 정확성 지표 결과(한국의료패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	25.5	76.7	0.06530	0.05256	0.24207	1.82838
hotdeck	34.7	99.9	0.06708	0.02138	0.09818	0.68360
ratio	15.7	51.4	0.01112	**0.00364**	**0.06581**	0.63995
mean_3	18.0	55.8	0.01392	0.05017	0.17790	0.65090
hotdeck_3	25.2	77.5	0.01294	0.00534	0.08307	0.77130
ratio_3	16.1	52.0	0.01573	0.00378	0.06747	**0.61979**
knn_3	14.9	46.2	**0.00281**	0.03374	0.19481	0.72268
rf_3	15.1	46.9	0.00285	0.03004	0.15901	0.68507
svm_3	**14.6**	**45.9**	0.00861	0.03053	0.16446	0.65957
mlp_3	15.9	49.9	0.00436	0.03719	0.21316	0.70885
knn_10	14.0	45.1	0.00815	0.02698	0.14697	0.65347
rf_10	12.3	40.2	0.00263	0.02539	0.18075	0.79467
svm_10	12.2	40.5	0.00915	0.02683	0.18366	0.74522
mlp_10	12.8	42.5	0.00300	0.02502	0.20055	0.88697

주: 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

10개 설명변수를 사용한 대체 방법은 3개일 때보다 예측의 정확성이 약간 향상되었다.

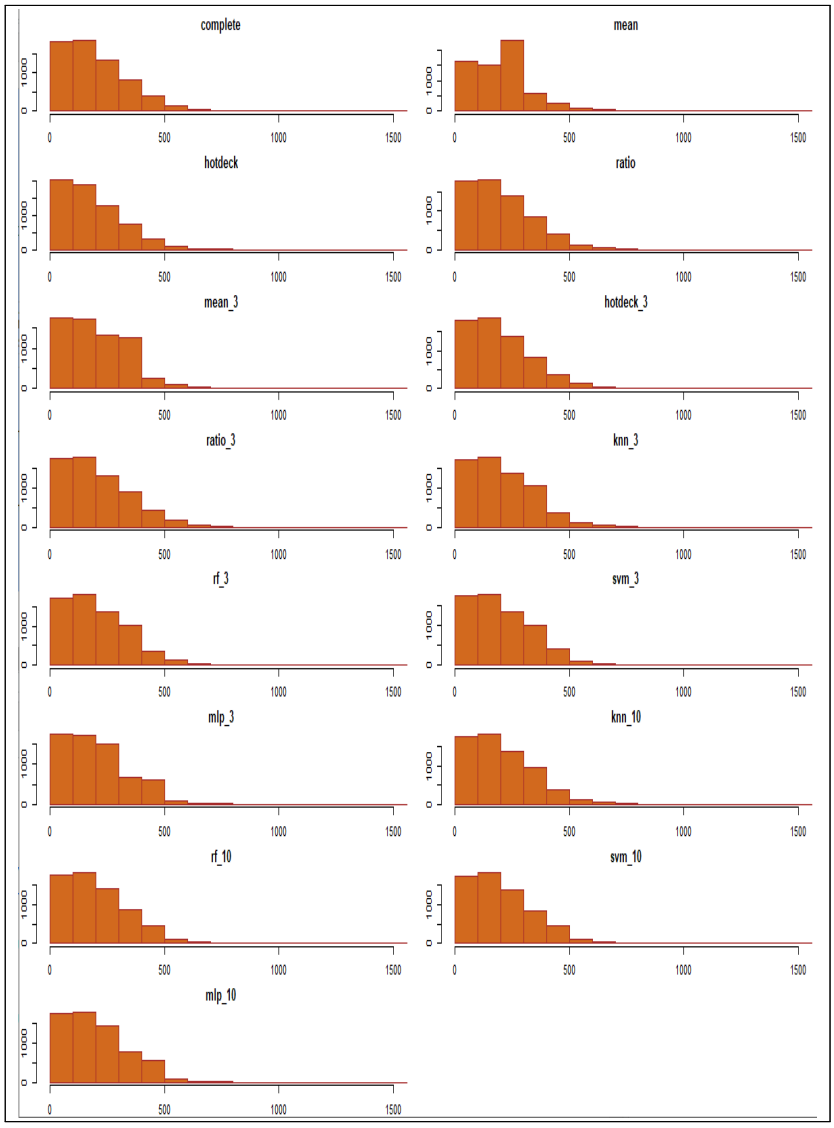
추정의 정확성을 보면 ratio의 효과가 가장 좋은 편으로 나타났다.

한편 10개의 설명변수를 사용한 대체 방법에 대한 추정의 정확성은 3개인 경우와 지표 간 값이 유사하여 뚜렷한 경향을 찾아볼 수 없었다.

[그림 4-19]의 히스토그램을 보면 hotdeck, ratio, hotdeck_3, ratio_3, rf_10, svm_10은 실제 분포와 유사한 형태를 보이고 있다. 이에 반해 mean과 mean_3은 어느 한 부분의 막대그래프가 높게 솟아 약간 다른 분포를 나타낸다.

[그림 4-20]의 상자그림을 보면 mean의 경우 구간의 길이가 다른 대체 방법에 비해 많이 짧고 다른 형태를 보이고 있다. 반면에 mean_3은 구간의 길이가 약간 긴 것으로 나타났다.

〔그림 4-19〕 무응답 비율 20%에서 1번째 자료의 분포-1(한국의료패널)

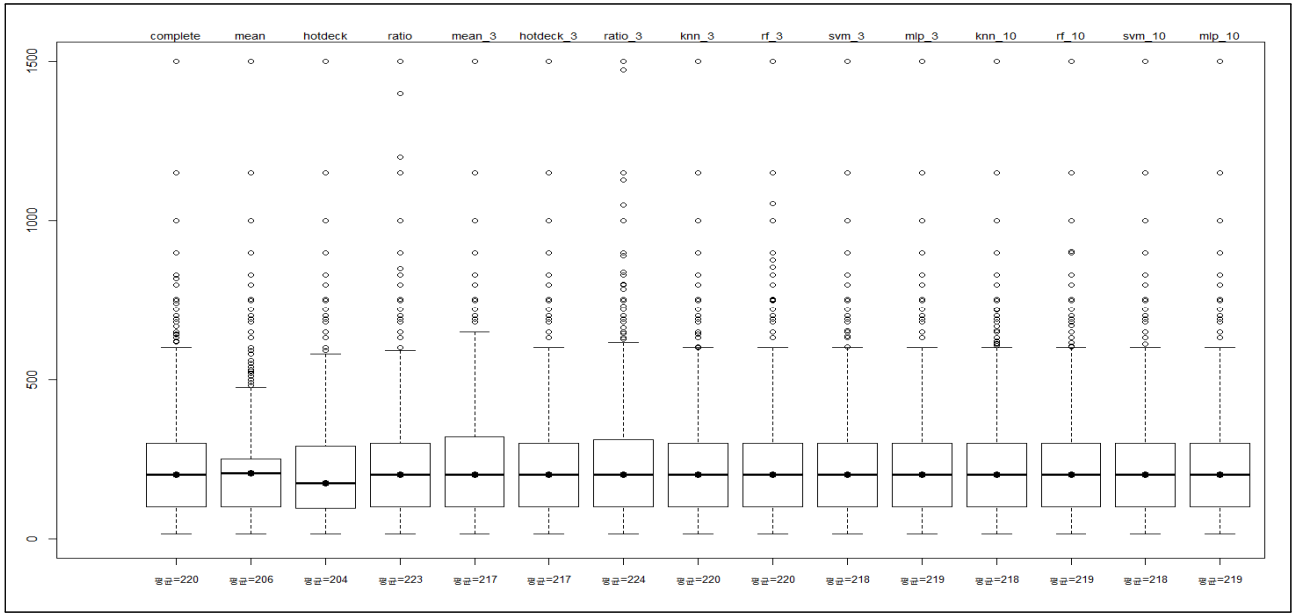


주: X축은 생활비(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

[그림 4-20] 무응답 비율 20%에서 1번째 자료의 분포-2(한국의료패널)



주: X축은 생활비 평균(단위: 만 원)이고, Y축은 생활비(단위: 만 원)를 나타냄.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

다. 무응답 비율 50%에 대한 결과

[그림 4-21]을 보면 rf_3의 편향이 2000원으로 가장 작으며 무응답 비율의 증가에 따른 편향의 증가 폭도 크지 않은 것으로 나타났다. 또한 제곱근평균제곱오차도 2만 9000원으로 가장 작아서 안정적인 결과를 보였다([부도 A-26] 참조). 한편 hotdeck_3(-17만 4000원)의 경우 대체군 내의 기증자 부족 문제를 위해 중복 추출로 대체하여 mean_3(-17만 2000원)과 비슷한 크기를 가진다. ratio_3(17만 6000원)은 ratio(6만 원)에 비해 편향이 크게 나타났다. ratio의 비율은 1이었으나 ratio_3의 비율은 대체군마다 다른 비율을 가지며 최소 0.996에서 최대 1.018이었다(모의실험 자료 10 기준). 대체군을 활용한 경우 대체군 내에 속한 개체의 개수가 적어져서 비의 변동이 커지게 되어 오히려 예측력이 좋지 않아진 것으로 볼 수 있다.

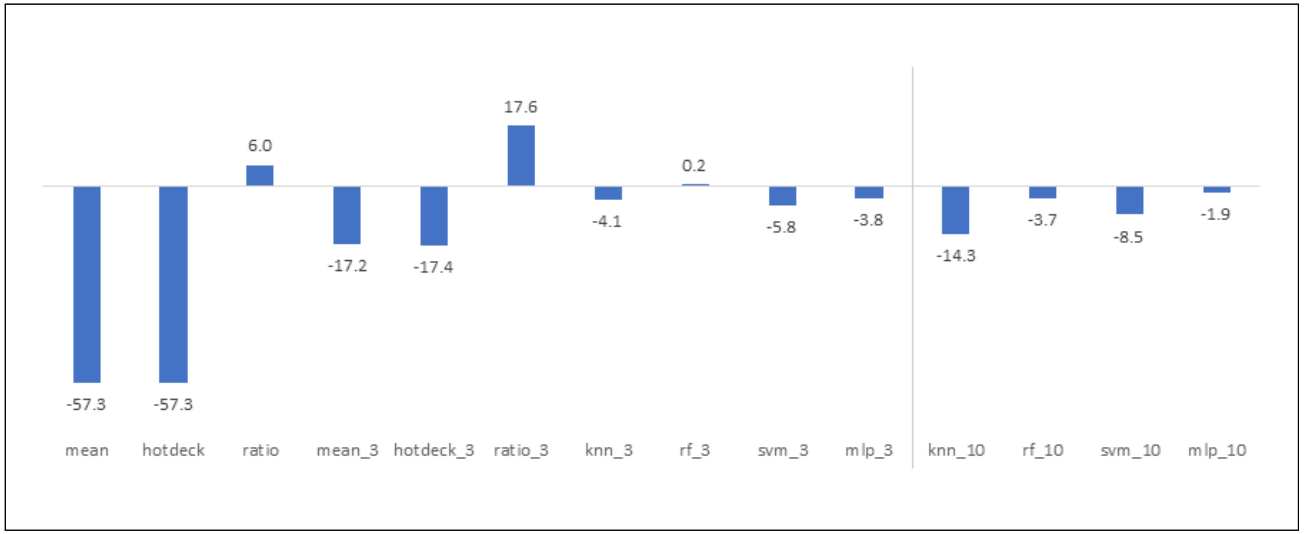
mlp_10(-1만 9000원)의 편향은 mlp_3(-3만 8000원)의 절반이었으며, 대체 방법 중에서 가장 작았다. 그리고 mlp_10의 제곱근평균제곱오차도 mlp_3에 비해 감소하였다([부도 A-26] 참조). 나머지 대체 방법은 증가하였는데, 특히 knn_10은 -14만 3000원으로 knn_3(-4만 1000원)보다 3.5배 정도 저조한 결과를 보였다.

평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률은 전반적으로 매우 저조한 것으로 나타났다([그림 4-22] 참조). 이는 50%나 되는 높은 무응답 비율이 반영된 결과이다. 그렇지만 rf_3은 74%의 가장 높은 포함률을 나타냈다. 나머지 기계학습 대체 방법도 낮은 수준이지만 포함률을 가지고 있다. 반면에 평균, 핫덱과 비대체는 전혀 포함하지 못하는 것으로 나타났다.

mlp_10과 rf_10의 경우만 포함률을 가지는데, 특히 mlp_10의 포함률은 60%로 mlp_3(37%)에 비해 증가하였다. 반면에 rf_10은 41%로 나타났으나 rf_3(74%)의 절반을 상회하고 있다.

〔그림 4-21〕 무응답 비율 50%에서의 전체 평균에 대한 편향(한국의료패널)

(단위: 만 원)

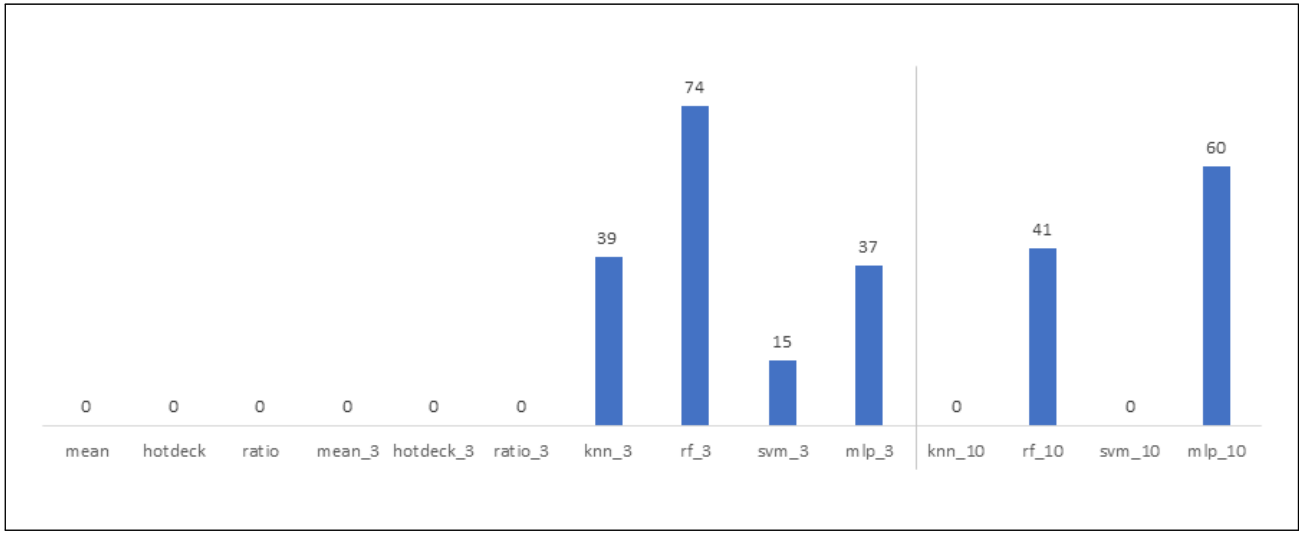


주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

[그림 4-22] 무응답 비율 50%에서의 참값의 포함률(한국의료패널)

(단위: %)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.
자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

〈표 4-24〉에서 예측의 정확성을 보면 svm_3이 가장 작으며 대체적으로 기계학습 대체 방법이 낮은 편으로 나타났다.

〈표 4-24〉 무응답 비율 50%에서의 추정의 정확성 지표 결과(한국의료패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	72.1	136.3	0.26026	0.16961	0.95605	8.47826
hotdeck	87.5	159.2	0.26026	0.03956	0.30894	1.63493
ratio	39.4	81.3	0.02744	**0.00531**	**0.07392**	**0.93821**
mean_3	46.4	93.2	0.07832	0.15603	0.79819	2.04627
hotdeck_3	62.4	121.2	0.07912	0.03270	0.20431	1.90219
ratio_3	43.5	89.0	0.07975	0.02329	0.14098	1.06864
knn_3	38.5	76.7	0.02093	0.08308	0.56613	2.54909
rf_3	38.8	76.4	**0.01016**	0.07213	0.50700	2.34002
svm_3	**36.8**	**73.5**	0.02690	0.07916	0.54919	2.38072
mlp_3	40.2	79.7	0.01823	0.09089	0.71343	2.99201
knn_10	39.2	80.5	0.06479	0.08108	0.47796	2.21549
rf_10	31.1	64.7	0.01678	0.06514	0.54069	2.66703
svm_10	31.1	66.5	0.03867	0.07869	0.64492	2.84138
mlp_10	33.5	76.6	0.01425	0.06826	1.24918	24.76851

주: 1) 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

추가적인 분석에 대한 결과는 다음과 같다. knn_10은 knn_3보다 약간 큰 값을 가지나, 나머지 대체 방법은 3개의 설명변수를 사용한 경우보다 작아서 예측의 정확성이 향상되었다고 볼 수 있다.

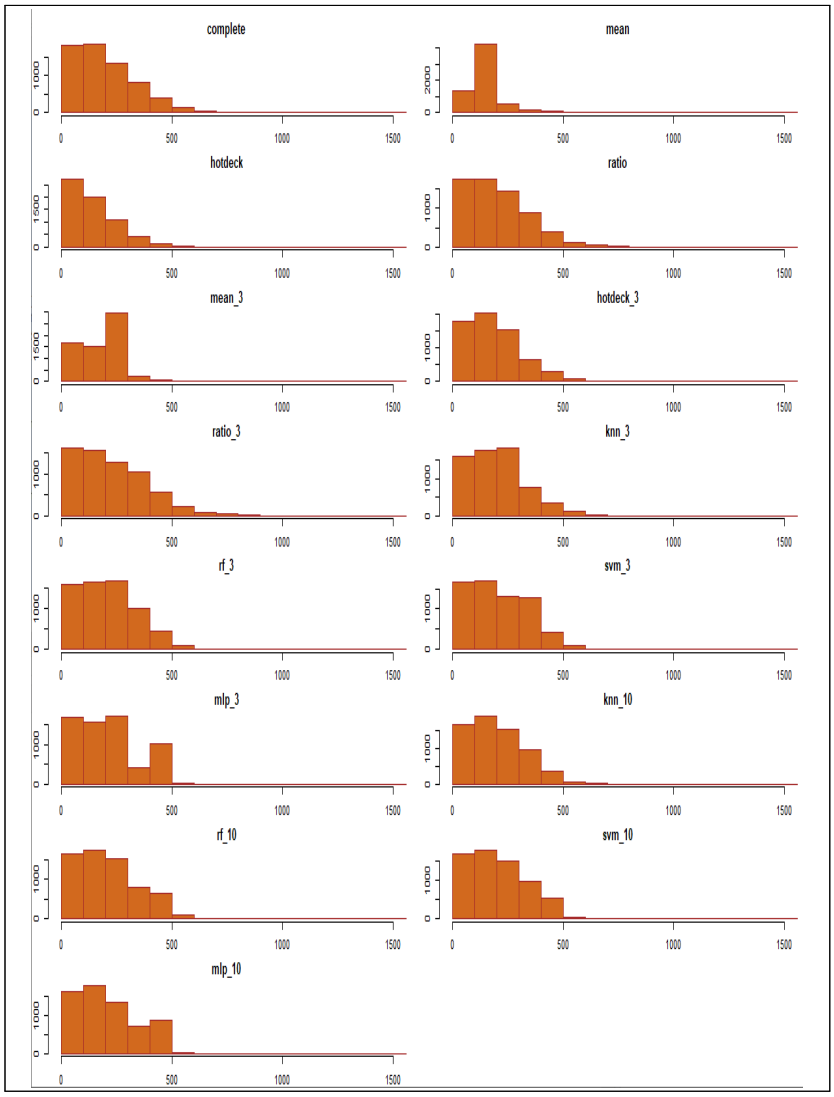
추정의 정확성의 경우에는 뚜렷한 특징을 찾을 수 없으나, ratio의 경우 다른 무응답 비율에 비해 효과가 좋은 편에 속하였다.

추정의 정확성은 앞의 결과와 비슷하게 뚜렷한 경향을 찾을 수 없었다. 단, mlp_10의 경우 ADS는 1.24918이고 ADK는 24.76851로 아주 큰 값을 가졌다. mlp_10의 모의실험 자료 중 8개의 값이 크기 때문이며, 특히 38번 모의실험 자료는 ADS가 22.83354이고 ADK가 88.5446으로 나타났다.

히스토그램 및 상자그림도 높아진 무응답 비율의 영향으로 대체 방법에 따른 자료 분포가 실제 분포와 다른 형태를 보이고 있다([그림 4-23] 참조). 특히 mean과 mean_3은 특정 막대가 높게 솟아 있으며, mlp_3도 역시 다른 형태를 보이고 있다.

[그림 4-24]를 보면 svm_3, mlp_3, rf_10, svm_10, smlp_10이 실제 분포와 전반적으로 유사한 편이라고 볼 수 있다. mean의 경우 구간의 길이가 매우 짧으며 상·하한에서의 이상치 또한 많이 분포되어 있는 것으로 나타났다. hotdeck, mean_3, hotdeck_3, knn_3, rf_3, knn_10은 구간의 길이가 짧은 편에 속하는 반면 ratio_3은 긴 편으로 나타났다.

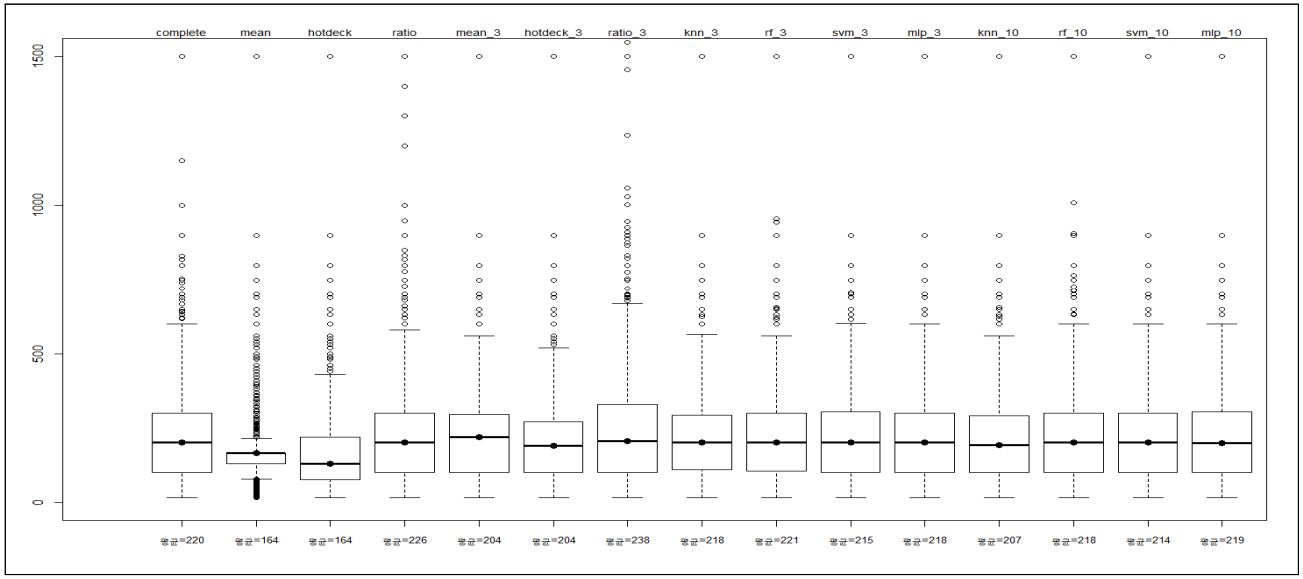
〔그림 4-23〕 무응답 비율 50%에서 1번째 자료의 분포-1(한국의료패널)



주: 1) X축은 생활비(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[그림 4-24] 무응답 비율 50%에서 1번째 자료의 분포-2(한국의료패널)



주: 1) X축은 생활비 평균(단위: 만 원)이고, Y축은 생활비(단위: 만 원)를 나타냄.
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

4. 소결

이 절에서는 한국의료패널자료에 근거하여 '생활비'를 대체하는 모의 실험을 실시하였다. 다양한 무응답 비율(5%, 10%, 20%, 30%, 50%)에서 기계학습 통계방법을 이용한 대체 방법의 효과를 살펴보았다. 무응답 대체 시 많이 사용하는 평균, 핫덱, 비대체 방법도 함께 비교하였다. 〈표 4-25〉는 각각의 무응답 비율에서 평가지표별 효과가 가장 우수한 대체 방법에 해당 무응답 비율을 표기한 결과를 정리한 것이다.

3개의 설명변수를 사용한 랜덤 포레스트 대체 방법의 효과가 우수한 것으로 나타났다. 무응답 비율과 상관없이 생활비에 대한 평균 추정량은 참값과의 차이(편향)가 가장 작았고, 평균 추정치에 대한 95% 신뢰구간에서 참값의 포함률(포함률)도 높았다.

한편 3개의 설명변수를 사용한 서포트 벡터 머신 대체 방법은 무응답 비율에 상관없이 개별 개체의 대체값과 실제값 사이의 유사 정도(예측의 정확성)가 가장 우수하였다. 다음으로 3개의 설명변수를 사용한 K-최근접 이웃 대체 방법의 결과도 좋은 편에 속하였다.

대체군을 활용하지 않은 비대체 방법도 기계학습 통계방법의 결과와 유사하였으며, 특히 무응답 비율 20% 이상에서는 자료의 분포를 잘 유지(추정의 정확성)하는 측면에서 우수한 결과를 가졌다.

대체군을 활용한 비대체의 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 활용하지 않은 평균 및 핫덱 대체는 무응답 비율이 낮은 5%

에서도 나머지 대체 방법에 비해 저조한 결과를 나타냈다. 그러나 대체군을 사용하여 평균 및 핫덱 대체를 실시하면 효과가 개선될 수 있다는 점을 다시 한번 확인하였다. 또한 대체 대상 변수의 특징을 고려하여 대체군을 보다 견고하게 구분한다면 대체 효과가 보다 더 향상될 것으로 생각된다.

〈표 4-25〉 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: COV는 91% 이상의 포함률을 만족하는 경우임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

다음은 부가적으로 실시한 모의실험으로 무응답 대체 시 무응답 대체 대상 변수와 관련 있는 여러 개의 변수를 더 포함하였을 때의 무응답 처리 효과를 살펴보았다. 〈표 4-26〉은 4개의 대체 방법 중에서 평가지표별 가장 우수한 효과를 가지는 대체 방법에 무응답 비율을 표기하였다.

4개의 대체 방법 중에서 10개의 설명변수를 사용한 랜덤 포레스트 대체 방법이 편향도 작은 편이고, 포함률도 높은 편이며 참값에 더 가깝게 대체하였다. 이와 더불어 10개의 설명변수를 사용한 랜덤 포레스트 대체 방법은 무응답 비율과 상관없이 3개의 설명변수를 사용할 때보다 예측의 정확성에 대한 효과가 향상되었다.

한편 10개의 설명변수를 사용한 신경망 대체 방법은 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 정확성에 해당하는 표준화된 왜도의 절대 차이와 표준화된 첨도의 절대 차이가 일부 대체된 자료에서 큰 값을 가져서, 무응답 대체 시 주의가 필요하다.

10개의 설명변수를 사용한 K-최근접 이웃 대체 방법은 3개의 설명변수를 사용했을 때보다 평가지표의 결과가 좋지 않은 것으로 나타났다. 설명변수 개수가 증가함에 따라 나타나는 차원의 저주로 인한 과적합의 결과로 볼 수 있다. 무응답 대체 대상 변수와 관련 있는 설명변수 탐색 및 선택 그리고 적절한 설명변수 개수 선정 등의 과정은 무응답 대체의 효과를 향상시키는 중요한 요인이라고 볼 수 있다. 그러므로 무응답 대체 전에 필수적으로 실시해야 하는 것으로 생각된다.

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

〈표 4-26〉 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: COV는 91% 이상의 포함률을 만족하는 경우임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

제 5 장 결론 및 향후 과제

5 결론 및 향후 과제

이 연구는 보건복지 분야 패널자료 품질 개선 연구로 항목무응답 대체 방법을 중심으로 살펴보았다. 이 장에서는 주요 연구 결과를 정리하고 향후 연구 과제를 제안하고자 한다.

보건복지 분야 패널자료에 근거하여 실시한 모의실험은 무응답 비율을 최저 5%에서부터 최고 50%까지 극대화하여 무응답 비율의 변화에 따른 대체 방법의 효과를 살펴보았다. 또한 빅데이터 분석기법 중 하나인 기계학습 통계방법을 무응답 대체에 적용해 봄으로써 활용 가능성도 가늠해 보았다. 대체 방법의 효과는 평균 추정치와 참값과의 일치 정도를 확인할 수 있는 '편향' 및 평균 추정치를 일관되게 추정하는지 알 수 있는 '제곱근평균제곱오차', 평균 추정치에 대한 95% 신뢰구간에서의 참값의 '포함률', 개별 개체의 대체값과 실제값 간의 유사 정도를 보여 주는 '예측의 정확성', 대체된 자료의 적률이 실제 자료의 적률을 잘 유지하는지 볼 수 있는 '추정의 정확성'으로 평가하였다.

대체 대상 변수는 한국복지패널자료의 경우 '경상소득'이었고, 한국의료패널자료는 '생활비'로 선정하였다. 2개 패널자료에 대한 모의실험 결과는 전반적으로 비슷하나 세부적으로는 다르게 나타났다.

먼저 한국복지패널자료의 결과는 3개의 설명변수를 사용한 랜덤 포레스트 대체 방법의 효과가 가장 우수한 것으로 나타났다(〈표 4-12〉 참조, p.116). 무응답 비율과 상관없이 경상소득에 대한 평균 추정량의 편향도 작고 포함률도 높은 편이었다. 또한 무응답 비율에 따라 약간 차이가 있지만 추정의 정확성도 잘 유지하는 편에 속하였다. 다음으로 3개 설명변

수를 사용한 K-최근접 이웃 및 서포트 벡터 머신 대체 방법의 결과도 좋은 편에 속하였다. 특히 3개의 설명변수를 사용한 서포트 벡터 머신 대체 방법은 무응답 비율에 상관없이 예측의 정확성에서 가장 우수하였다. 이렇듯 기계학습 통계방법에 기반한 대체 방법의 결과는 좋은 편으로 나타났다.

대체군을 활용하지 않은 비대체도 무응답 비율 30% 이하에서는 포함률이 높아 효과가 좋은 편이었다. 특히 무응답 비율 50%의 경우 편향, 포함률과 추정의 정확성이 K-최근접 이웃 및 신경망 대체 방법보다 효과가 우수하였다. 비대체 방법은 대체 대상 변수의 이전과 해당 차수 간 변동이 크지 않을 때 효과적이므로, 실제 패널자료에서 항목무응답을 대체할 때 많이 사용하는 방법 중 하나이다.

무응답 비율이 낮은 경우에는 대체군을 사용하지 않은 평균 및 핫덱 대체를 제외한 나머지 대체 방법에 따른 효과 차이가 비슷하였다. 그리고 대체군을 사용하여 평균 및 핫덱 대체를 실시하면 대체군을 사용하지 않은 경우보다 평가지표의 결과가 모두 향상되는 효과를 보였다.

무응답 대체 대상 변수와 관련 있는 정보를 더 포함한 경우에 대한 대체 방법의 효과를 평가하기 위하여 부가적으로 모의실험을 실시하였다. 설명변수를 더 포함(총 6개)하여 대체한 결과는 무응답 비율에 상관없이 랜덤 포레스트 대체 방법이 예측의 정확성을 제외한 나머지 지표의 결과에서 모두 우수하였다(〈표 4-13〉 참조, p.118). 신경망 대체 방법은 기존보다 결과가 향상되었으며, 무응답 비율이 증가할수록 효과는 더 좋아지는 것으로 나타났다. 한편 K-최근접 이웃 대체 방법은 기존보다 효과가 좋지 않은 것으로 나타났다.

이 부가적인 모의실험은 기계학습 방법에 기반한 대체 방법에만 적용하였다. 평균, 핫덱, 비대체의 경우에는 대체군으로 활용될 변수의 개수

가 증가하게 되면 대체군 내의 개체의 개수가 부족한 현상이 발생할 수 있어서 대체하기 어려울 수 있다는 한계점을 가지기 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.

다음으로 한국의료패널자료의 모의실험 결과는 3개의 설명변수를 사용한 랜덤 포레스트 대체 방법의 효과가 우수한 것으로 나타났다(〈표 4-25〉 참조, p.151). 무응답 비율과 상관없이 생활비에 대한 평균 추정량의 편향이 가장 작았고 포함률도 높았다. 다음으로 K-최근접 이웃 대체 방법 및 서포트 벡터 머신 대체 방법도 편향이 작고 포함률이 높은 것으로 나타났다. 서포트 벡터 머신 대체 방법의 경우 무응답 비율과 상관없이 예측의 정확성이 아주 우수하였다.

대체군을 활용하지 않은 비대체 방법도 기계학습 통계방법의 결과와 유사했으며, 특히 무응답 비율 20% 이상에서는 추정의 정확성이 우수한 것으로 나타났다.

대체군을 활용한 비대체의 효과도 무응답 비율 10% 이하에서는 포함률과 추정의 정확성이 우수하게 나타났다. 그러나 무응답 비율 20%에서부터는 효과가 좋지 못한 결과를 가져왔다. 이는 무응답 대체 시 대체군에 대한 정보 부족 때문이므로, 대체군 내의 개체 개수는 충분히 보장되어야 할 것이다.

무응답 비율이 낮은 5%에서도 대체군을 사용하지 않은 평균 및 핫덱 대체는 나머지 대체 방법에 비해 저조한 결과를 나타냈다. 그러나 대체군을 사용하여 평균 및 핫덱 대체를 실시하면 효과가 개선된 결과를 보였다.

설명변수를 더 포함(총 10개)하여 대체한 결과는 랜덤 포레스트 대체

방법이 편향도 작은 편이었고, 포함률도 높은 편이며 참값에 더 가깝게 대체하여 4가지 대체 방법 중에서 가장 우수하였다(〈표 4-26〉 참조, p.153). 신경망 대체 방법은 무응답 비율과 상관없이 편향이 작아 참값과 유사하였고 포함률도 높은 편이었다. 단, 무응답 비율이 50%일 때 추정의 정확성에 해당하는 표준화된 왜도의 절대 차이 및 표준화된 첨도의 절대 차이가 일부 대체된 자료에서 큰 값을 가지는 것으로 나타나, 무응답 대체 시 주의가 필요하다. 서포트 벡터 머신 대체 방법의 경우 예측의 정확성은 무응답 비율과 상관없이 우수한 편이었으나, 편향은 무응답 비율 10%에서부터 큰 편에 속하였다. K-최근접 이웃 대체 방법은 4가지 대체 방법 중에서 효과가 가장 저조했을 뿐만 아니라 3개 설명변수를 사용한 경우보다 대체 효과도 좋지 않은 것으로 나타났다.

두 개의 모의실험에 대한 종합적인 결과 및 활용 방안은 3가지로 설명할 수 있다. 첫째, 패널자료에서의 항목무응답 대체 시 기계학습 통계방법의 적용 가능성을 확인하였다. 대체 방법의 선정은 어떠한 평가지표에 중점을 두는지에 따라 달라질 수 있다. 무응답 비율에 상관없이 개별개체의 대체값과 실제값 간의 유사 정도(예측의 정확성)에 초점을 둔다면 서포트 벡터 머신 대체 방법을, 평균 추정치와 참값과의 일치 정도(편향)라면 랜덤 포레스트 대체 방법을 추천한다. 특히 랜덤 포레스트 대체 방법의 경우에는 편향뿐만 아니라 다른 평가지표도 우수한 결과를 보여 실무에서 활용해 볼 수 있다고 생각된다.

둘째, 대체군 활용 여부에 따라 대체 효과가 크게 다르다는 결과를 통해서 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 보다 견고하게 형성할수록 대체 효과는 더 향상될 수 있다고 생각한다.

셋째, 보조변수로 활용하는 설명변수 개수의 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 값이지만 최소 길이가 점차 길어져 평균 길이와 비슷해지므로 최소 길이라는 의미가 사라지기 때문이다. 실무에서 활용할 때 차원 저주의 문제는 정규화(regularization), 군집화(clustering), 차원 축소, 설명변수 선택 방법 등으로 해결하는 것을 추천한다. K-최근접 이웃, 랜덤 포레스트 대체 방법의 결과를 통하여 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 패널자료의 무응답 특징, 범위 등을 자료의 형태에 따라 다르게 고려해야 하므로 이 연구의 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 하며 모집단 분포로부터 표본 분포가 왜곡되지 않고 비슷하게 유지될 수 있어야 한다. 이 점을 인지한다면 보다 정확하고 신뢰성 있는 무응답 대체 자료를 제공할 수 있을 것이다.

향후 연구 과제 시 3가지를 고려해 볼 수 있다. 첫째, 무응답 자료를 설계할 때 패널자료의 고유 특성을 반영하는 것이다. 이 연구에서의 모의실험 자료는 일반적인 상황에 근거하여 설계하였다. 그러나 한국복지패널

자료의 경우 저소득층 가구를 많이 포함하고 있으므로 균등화 소득에 따른 가구 구분 변수(일반 및 저소득층 가구)를 활용해 보는 것이다. 둘째, 패널가구의 표본이탈(sample attrition) 정보를 항목무응답 대체 시 고려해 보는 것이다. 표본이탈 유형(pattern) 분석 등을 통하여 항목무응답 발생과의 연관성을 살펴보고, 필요하다면 항목무응답 대체 시 표본이탈 정보를 활용해 볼 필요가 있다. 셋째, 다른 빅데이터 분석기법을 적용해 보는 것이다. 이 연구에서는 이전 시점(t-1 조사 차수)의 정보만을 사용하여 항목무응답 대체를 실시하였다. 그러나 패널자료는 2시점 이상(t-1, t-2, … 조사 차수)의 정보를 가지므로, 이러한 시간 정보를 이용하는 순환신경망(RNN: recurrent neural network)을 항목무응답 대체 시 적용 보는 것이다.

참고문헌

김영원, 조선경. (1996). 표본조사에서 항목 무응답 대체 방법. **한국통계학회논문집, 3**(3), 145-159.

김유빈, 이지은, 최승주, 신선옥, 이혜정, 정현상. (2018). **제20차(2017)년도 한국 가구와 개인의 경제활동-한국노동패널 기초분석보고서**. 세종: 한국노동연구원.

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

김남순, 서제희, 정연, 이정아, 배정은, 이나경, . . . 김진영. (2018). **2016년 한국의료패널 기초분석보고서(Ⅱ) - 질병 이환, 만성질환, 건강 행태와 건강수준**. 세종: 한국보건사회연구원.

김태완, 오미애, 이주미, 신재동, 정희선, 이병재, . . . 김정욱. (2018). **2018년 한국복지패널 기초분석보고서**. 세종: 한국보건사회연구원.

박창이, 김용대, 김진석, 송종우, 최호식. (2018). **R을 이용한 데이터마이닝 (개정판)**. 서울: 교우사.

손세림. (2019). **기계학습방법을 이용한 순서형 결측자료 대체의 성능 비교**. (석사학위논문. 고려대학교 대학원. 서울). http://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=d5baad97336d02eeffe0bdc3ef48d419에서 인출.

송주원. (2017). **노동패널 항목 무응답 처리 방법 보완**. 세종: 한국노동연구원.

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

이혜정. (2018a). **한국노동패널조사 2017년 예비표본 구축 개요 및 현황**. 패널브리프, 16.

이혜정. (2018b). **함수적 모형을 이용한 연속형 패널자료의 항목무응답 대체 방법.** (박사학위논문, 고려대학교 대학원. 서울). http://www.riss.kr/sear

ch/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=cc891e2985667e5dffe0bdc3ef48d419에서 인출.

한국고용정보원. (2019a). **2018년 고령화 연구패널 이용자 가이드(User's Guide)**. https://survey.keis.or.kr/klosa/klosaguide/List.jsp에서 2019. 10. 12. 인출.

한국고용정보원. (2019b). **고령화연구패널 1~6차 자료 무응답 대체값 산출을 위한 다중대체 안내서**. https://survey.keis.or.kr/klosa/klosaguide/List.jsp에서 2019. 10. 12. 인출.

한국노동패널조사. (2019). **항목 무응답 flag 변수 정보(20190418)**. https://www.kli.re.kr/klips/selectBbsNttList.do?bbsNo=96&key=497에서 2019. 9. 16. 인출.

한국보건사회연구원. (2014). **한국의료패널조사(8차)[데이터파일]**. https://www.khp.re.kr에서 2019. 6. 7. 인출.

한국보건사회연구원. (2015). **한국의료패널조사(9차)[데이터파일]**. https://www.khp.re.kr에서 2019. 6. 7. 인출.

한국보건사회연구원. (2017). **한국복지패널조사(12차)[데이터파일]**. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

한국보건사회연구원. (2018). **한국복지패널조사(13차)[데이터파일]**. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

한국보건사회연구원. (2019a). **2008-2016 연간데이터 한국의료패널 코드북(버전 1.5) 및 변수표**. https://www.khp.re.kr:444/web/data/board/view.do?bbsid=55&seq=2092에서 2019. 6. 7. 인출.

한국보건사회연구원. (2019b). **한국복지패널 13차년도 조사자료 User's Guide**. https://www.koweps.re.kr:442/data/guide/list.do에서 2019. 5. 14. 인출.

Agency for Healthcare Research and Quality(AHRQ). (2019a). *Medical Expenditure Panel Survey(MEPS)*. Retrieved from https://www.meps.ahrq.gov/ 2019. 10. 12.

Agency for Healthcare Research and Quality(AHRQ). (2019b). *MEPS HC-201 2017 Full Year Consolidated Data File[Data]*. Retrieved fr

om https://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-201 2019. 10. 12.

Chambers, R. L. (2000). *Evaluation criteria for statistical editing and imputation, Working Paper for the Euredit Project on the Development and Evaluation of New Methods for Editing and Imputation*. Southampton: University of Southampton.

Duffy, D. (2011). (2007). *PSID income and wage imputation methodology*. In PSID Technical Series Paper# 11-03.

German Institute for Economic Research(DIW Berlin). (2019). *Socio-Economic Panel (SOEP)*. Retrieved from https://www.diw.de/en/soep 2019. 10. 12.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). *The german socio-economic panel (SOEP)*. Jahrbücher für Nationalökonomie und Statistik, 239(2), 345-360.

Grabka, M. M., & Westermeier, C. (2015). *Editing and multiple imputation of item non-response in the wealth module of the German Socio-Economic Panel*. SOEP Survey Papers Series C, 272.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Institute for Social Research, University of Michigan. (2019). *PSID Main Interview User Manual: Release 2019*. Retrieved from https://psidonline.isr.umich.edu/data/Documentation/UserGuide2017.pdf.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: springer.

Jerez, JM., Molina, I., Garcia-Laencina, PJ., Alba, E., Ribelles, N., Martin, M., Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine 50*(2), 105-115.

Kara, S., Zimmermann, S., & SOEP Group. (2018). *SOEPcompanion (v34)*. SOEP Survey Papers 588:Series G. Berlin: DIW/SOEP.

Little, R. J., & Su, H. L. (1989). Item non-response in panel surveys. *Panel surveys*, 400-425.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Lu, C. J., Lee, T. S., & Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision support systems, 47*(2), 115-125.

Machlin, S. R., & Dougherty, D. D. (2007). Overview of Methodology for Imputing Missing Expenditure Data in the Medical Expenditure Panel Survey. *Methodology Report, 19*. Agency for Healthcare Research and Quality, Rockville, Md. Retrieved from http://www.meps.gov/mepsweb/data_files/publications/mr19/mr19.pdf.

Marcia Freed, T(ed)., Brice, J., Buck, N., & Prentice-Lane, E. (2018). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices. Colchester: University of Essex*. Retrieved from http://www.iser.essex.ac.uk/bhps/documentation 2019. 10. 12.

Mitchell, T. M. (1997). *Machine Learning*. New York: NY, USA McGraw-Hill.

Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M., & Cubiles-de-la-Vega, M.-D. (2011). Missing value imputation on

missing completely at random data using multilayer perceptrons. *Neural Networks, 24*(1), 121-129.

University of Michigan Institute for social research, (Survey research center). (2019). *Panel Study of Income Dynamics(PSID)*. Retrieved from https://psidonline.isr.umich.edu/default.aspx 2019. 10. 12.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

Watson, N. & Starick, R. (2011). Evaluation of alternative income imputation methods for a longitudinal survey. *Journal of Official Statistics, 27*(4), 693-715.

Welniak, E. J. & Coder, J. F. (1980). A measure of the Bias in the March CPS Earnings Imputation System. *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods*, 421-425.

Wikipedia. (2019). *About RMSE*. Retrieved from https://en.wikipedia.org/wiki/Root-mean-square_deviation 2019. 10. 12.

부 록

1. 한국복지패널자료에 근거한 모의실험

가. 기계학습 방법론에 기반한 대체 방법 - 파라미터 튜닝

〈부표 A-1〉 랜덤 포레스트 설명변수 3개의 경우 파라미터 튜닝(한국복지패널)

파라미터	설정한 값
mtry	3
ntree	300, 500

〈부표 A-2〉 랜덤 포레스트 설명변수 6개의 경우 파라미터 튜닝(한국복지패널)

파라미터	설정한 값
mtry	3, 4, 5, 6
ntree	300, 500

〈부표 A-3〉 서포트 벡터 머신 기반 대체 설명변수 3개의 경우 파라미터 튜닝(한국복지패널)

파라미터	설정한 값
cost	1, 10, 100
gamma	0.0625, 0.125, 0.25

〈부표 A-4〉 서포트 벡터 머신 기반 대체 설명변수 6개의 경우 파라미터 튜닝(한국복지패널)

파라미터	설정한 값
cost	1, 10, 100
gamma	0.015625, 0.03125, 0.0625

〈부표 A-5〉 인공 신경망 기반 대체 설명변수 3개의 경우 파라미터 튜닝(한국복지패널)

파라미터	설정한 값
layer1	6, 8
layer2	0, 9, 10

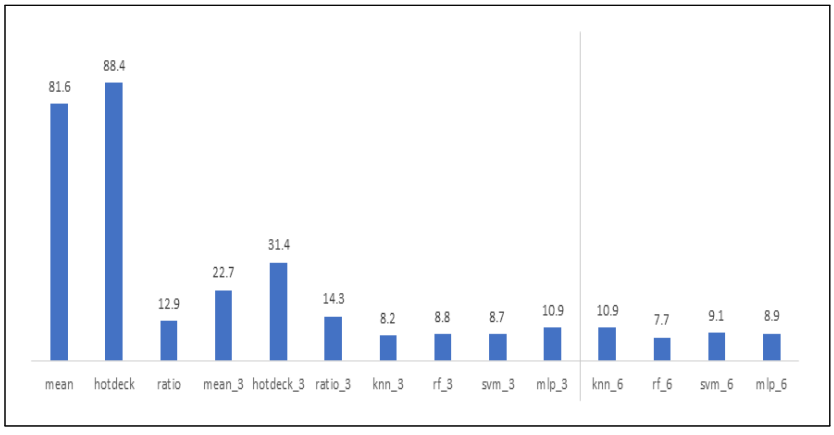
〈부표 A-6〉 인공 신경망 기반 대체 설명변수 6개의 경우 파라미터 튜닝(한국복지패널)

파라미터	설정한 값
layer1	1, 2, 3
layer2	0, 2, 4

나. 무응답 비율 5%에 대한 결과

[부도 A-1] 무응답 비율 5%에서의 전체 평균에 대한 제곱근평균제곱오차(한국복지패널)

(단위: 만 원)

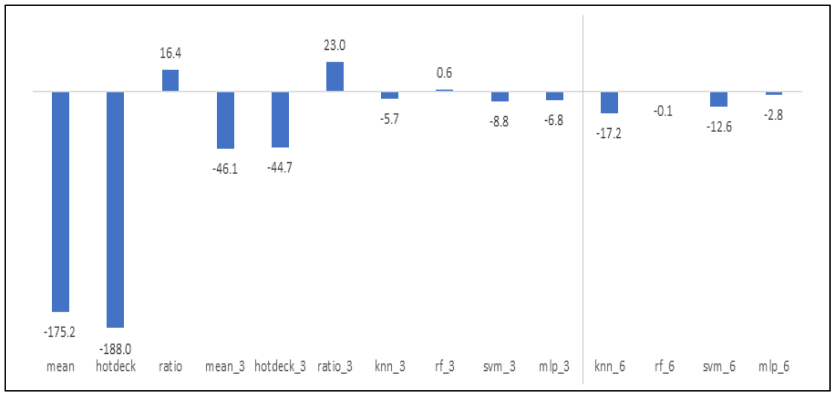


자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

다. 무응답 비율 10%에 대한 결과

[부도 A-2] 무응답 비율 10%에서의 전체 평균에 대한 편향(한국복지패널)

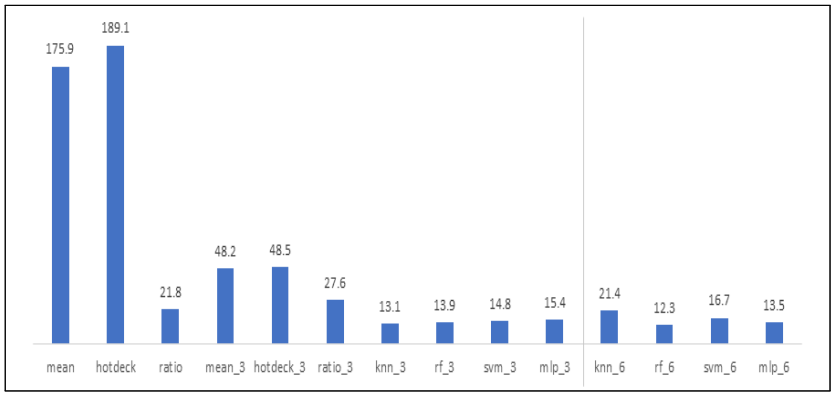
(단위: 만 원)



자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[부도 A-3] 무응답 비율 10%에서의 전체 평균에 대한 제곱근평균제곱오차(한국복지패널)

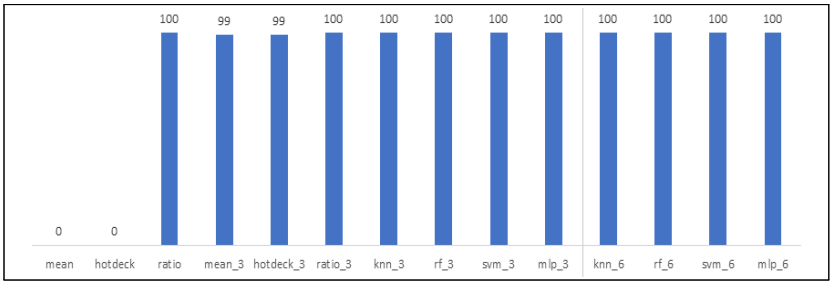
(단위: 만 원)



자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[부도 A-4] 무응답 비율 10%에서의 참값의 포함률(한국복지패널)

(단위: %)



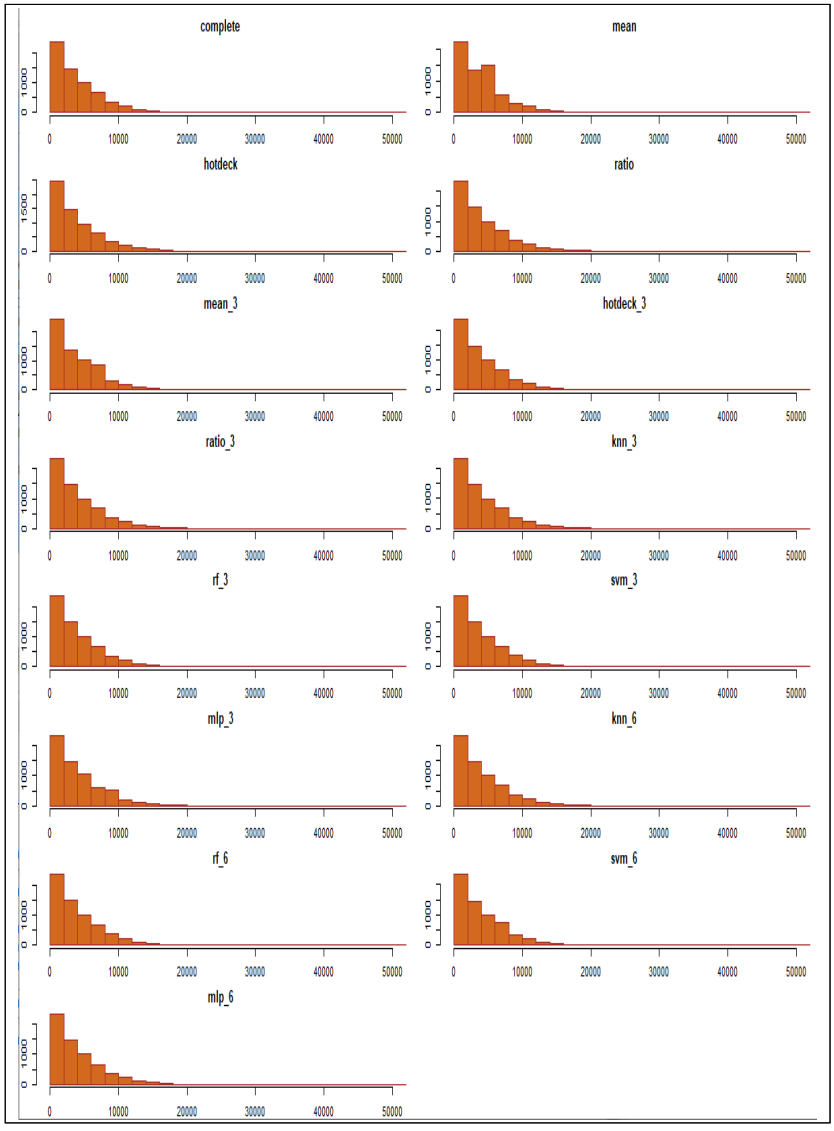
자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

〈부표 A-7〉 무응답 비율 10%에서의 추정의 정확성 지표 결과(한국복지패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	324.3	1601.2	0.04181	0.03526	1.14224	41.45470
hotdeck	436.0	2020.4	0.04487	0.02776	0.73408	22.85880
ratio	131.7	1209.5	0.00429	0.02736	0.84376	23.09010
mean_3	205.0	1200.2	0.01100	0.04122	0.60891	25.00630
hotdeck_3	299.3	1759.2	0.01071	0.02805	0.88884	23.95780
ratio_3	131.6	1174.6	0.00560	0.02806	0.87341	24.52450
knn_3	128.4	**865.2**	**0.00241**	0.02568	**0.39001**	12.01546
rf_3	148.1	1029.1	0.00267	**0.01973**	0.40108	**10.11041**
svm_3	**122.9**	869.8	0.00280	0.02410	0.39372	11.99405
mlp_3	178.0	1088.7	0.00282	0.03641	0.45838	17.07609
knn_6	119.6	877.1	0.00432	0.02685	0.41100	13.74198
rf_6	110.3	809.4	0.00231	0.02067	0.37410	9.82055
svm_6	103.7	798.1	0.00322	0.02218	0.38501	11.42187
mlp_6	124.6	902.4	0.00244	0.02667	0.38561	12.19817

주: 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).
자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[부도 A-5] 무응답 비율 10%에서 8번째 자료의 분포-1(한국복지패널)

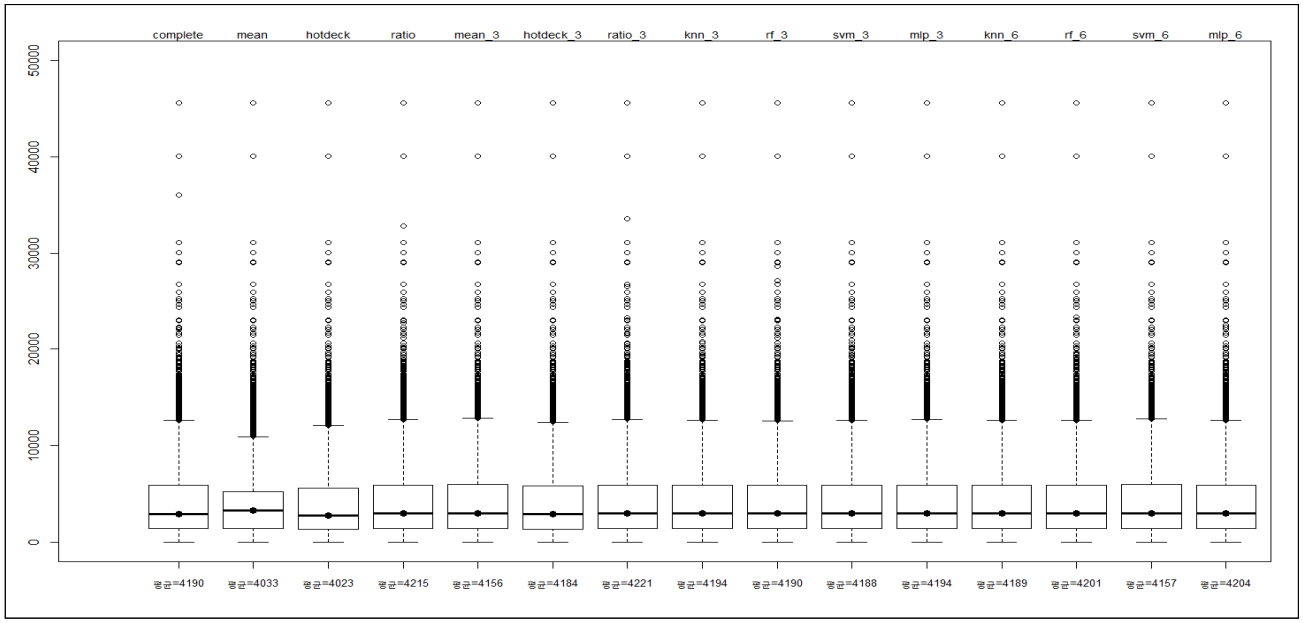


주: X축은 경상소득(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[부도 A-6] 무응답 비율 10%에서 8번째 자료의 분포-2(한국복지패널)



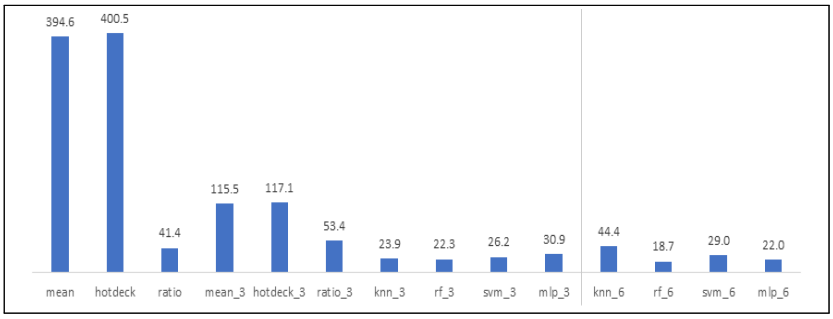
주: X축은 경상소득 평균(단위: 만 원)이고, Y축은 경상소득(단위: 만 원)을 나타냄.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

라. 무응답 비율 20%에 대한 결과

[부도 A-7] 무응답 비율 20%에서의 전체 평균에 대한 제곱근평균제곱오차(한국복지패널)

(단위: 만 원)

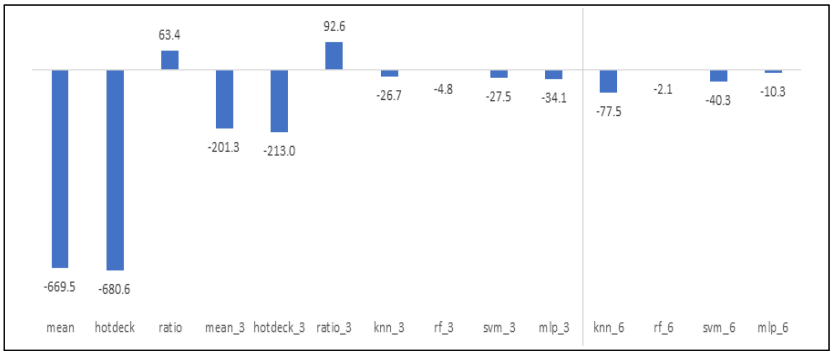


자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

마. 무응답 비율 30%에 대한 결과

[부도 A-8] 무응답 비율 30%에서의 전체 평균에 대한 편향(한국복지패널)

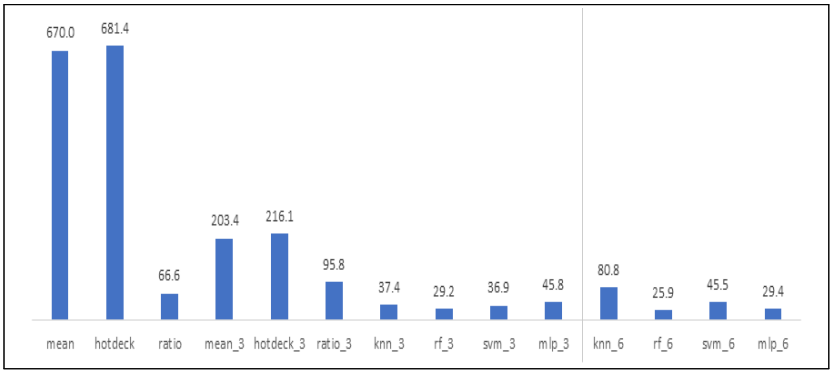
(단위: 만 원)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.
자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[부도 A-9] 무응답 비율 30%에서의 전체 평균에 대한 제곱근평균제곱오차(한국복지패널)

(단위: 만 원)



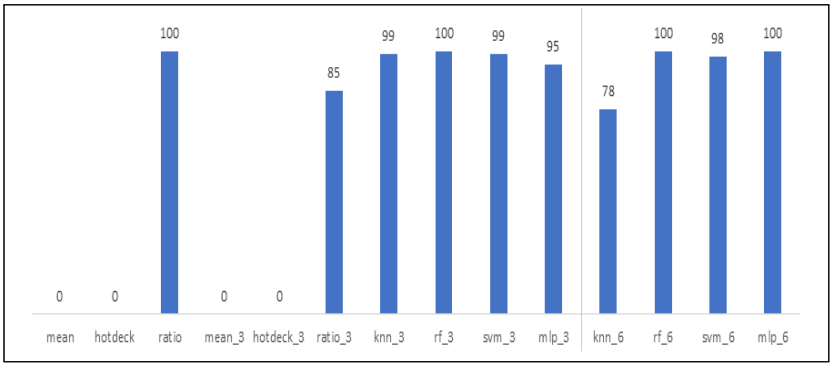
주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[부도 A-10] 무응답 비율 30%에서의 참값의 포함률(한국복지패널)

(단위: %)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

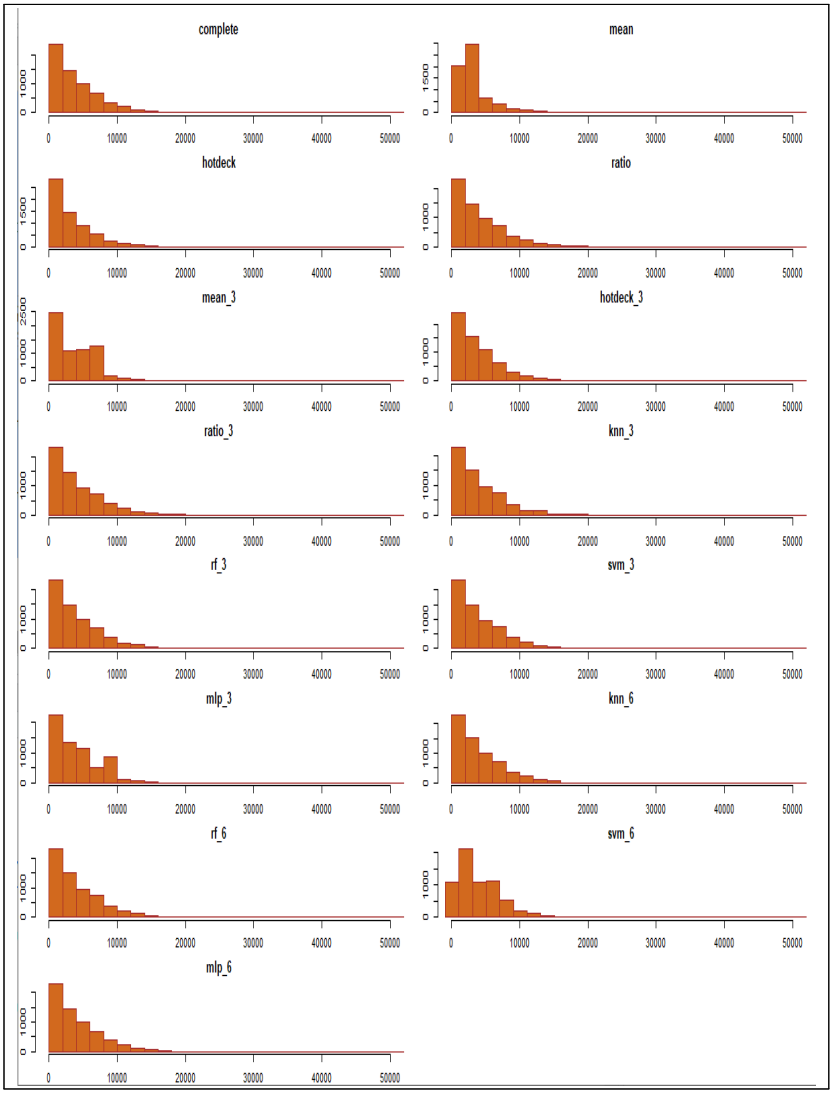
〈부표 A-8〉 무응답 비율 30%에서의 추정의 정확성 지표 결과(한국복지패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	1002.7	2869.3	0.15980	0.10566	3.96833	171.01300
hotdeck	1299.7	3520.2	0.16244	0.07565	2.35971	84.99770
ratio	393.3	2136.2	0.01513	**0.04636**	1.49739	40.95860
mean_3	616.3	2133.9	0.04806	0.13552	2.03844	94.35380
hotdeck_3	888.6	3029.5	0.05084	0.06262	2.71109	85.68990
ratio_3	400.9	2182.9	0.02211	0.05264	1.56277	44.69390
knn_3	388.2	1570.5	0.00746	0.08015	**1.06496**	38.46391
rf_3	438.0	1755.2	**0.00576**	0.05237	1.06567	**28.21540**
svm_3	**368.9**	**1563.0**	0.00741	0.07183	1.09723	34.73549
mlp_3	530.1	1918.2	0.00893	0.11370	1.13730	55.73325
knn_6	372.3	1610.5	0.01850	0.08571	1.10378	47.94395
rf_6	331.9	1476.8	0.00483	0.06444	1.13481	29.12940
svm_6	317.9	1477.3	0.00970	0.06936	1.09931	34.30363
mlp_6	375.9	1629.1	0.00560	0.08075	1.09078	36.39668

주: 1) 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

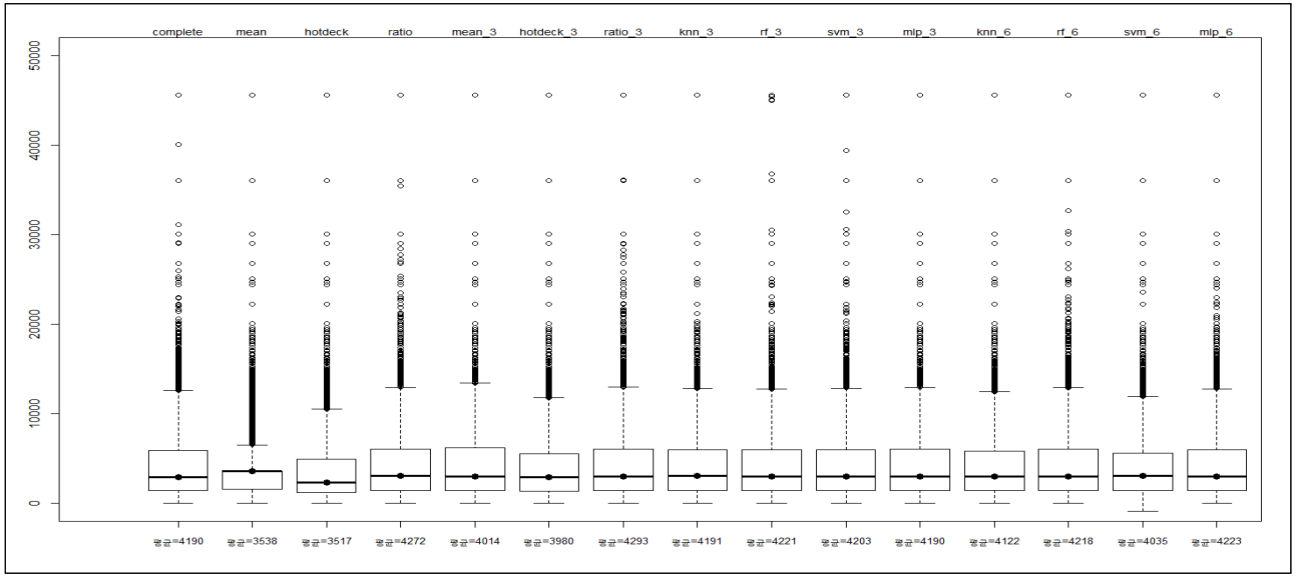
〔부도 A-11〕 무응답 비율 30%에서 8번째 자료의 분포-1(한국복지패널)



주: 1) X축은 경상소득(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

[부도 A-12] 무응답 비율 30%에서 8번째 자료의 분포-2(한국복지패널)



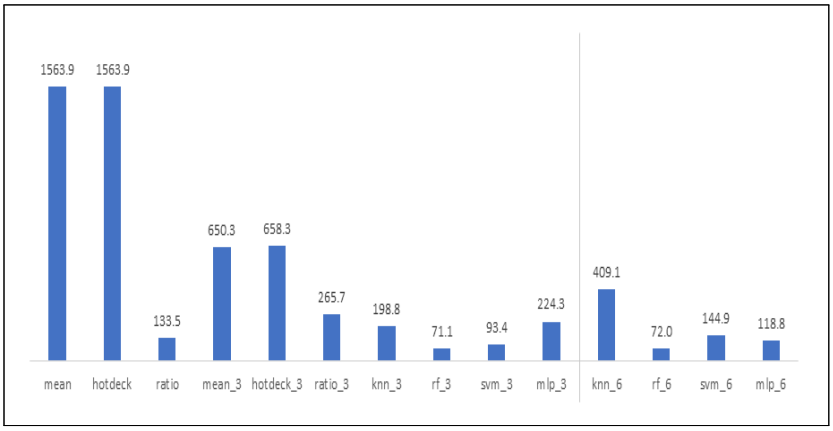
주: 1) X축은 경상소득 평균(단위: 만 원)이고, Y축은 경상소득(단위: 만 원)을 나타냄.
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

바. 무응답 비율 50%에 대한 결과

〔부도 A-13〕 무응답 비율 50%에서의 전체 평균에 대한 제곱근평균제곱오차(한국복지패널)

(단위: 만 원)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.
자료: 1) 한국보건사회연구원. (2017). 한국복지패널조사(12차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.
2) 한국보건사회연구원. (2018). 한국복지패널조사(13차)[데이터파일]. https://www.koweps.re.kr에서 2019. 5. 14. 인출.

2. 한국의료패널자료에 근거한 모의실험

가. 기계학습 방법론에 기반한 대체 방법 - 파라미터 튜닝

〈부표 A-9〉 랜덤 포레스트 설명변수 3개의 경우 파라미터 튜닝(한국의료패널)

파라미터	설정한 값
mtry	3
ntree	300, 500

〈부표 A-10〉 랜덤 포레스트 설명변수 10개의 경우 파라미터 튜닝(한국의료패널)

파라미터	설정한 값
mtry	3, 4, 5
ntree	300, 500

〈부표 A-11〉 서포트 벡터 머신 기반 대체 설명변수 3개의 경우 파라미터 튜닝(한국의료패널)

파라미터	설정한 값
cost	1, 10, 100
gamma	0.0625, 0.125, 0.25

〈부표 A-12〉 서포트 벡터 머신 기반 대체 설명변수 10개의 경우 파라미터 튜닝(한국의료패널)

파라미터	설정한 값
cost	1, 10, 100
gamma	0.03125, 0.0625, 0.125

〈부표 A-13〉 인공 신경망 기반 대체 설명변수 3개의 경우 파라미터 튜닝(한국의료패널)

파라미터	설정한 값
layer1	6, 8
layer2	9, 10

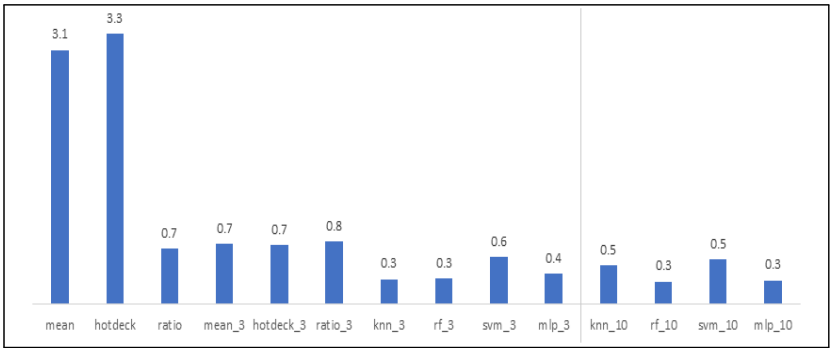
〈부표 A-14〉 인공 신경망 기반 대체 설명변수 10개의 경우 파라미터 튜닝(한국의료패널)

파라미터	설정한 값
layer1	2, 3
layer2	6, 8

나. 무응답 비율 5%에 대한 결과

[부도 A-14] 무응답 비율 5%에서의 전체 평균에 대한 제곱근평균제곱오차(한국의료패널)

(단위: 만 원)

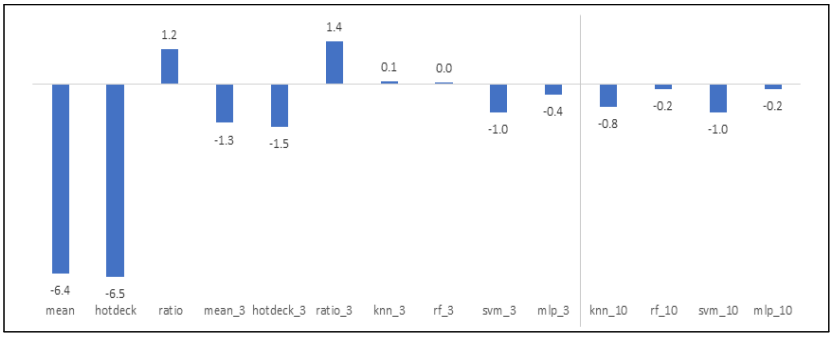


자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

다. 무응답 비율 10%에 대한 결과

[부도 A-15] 무응답 비율 10%에서의 전체 평균에 대한 편향(한국의료패널)

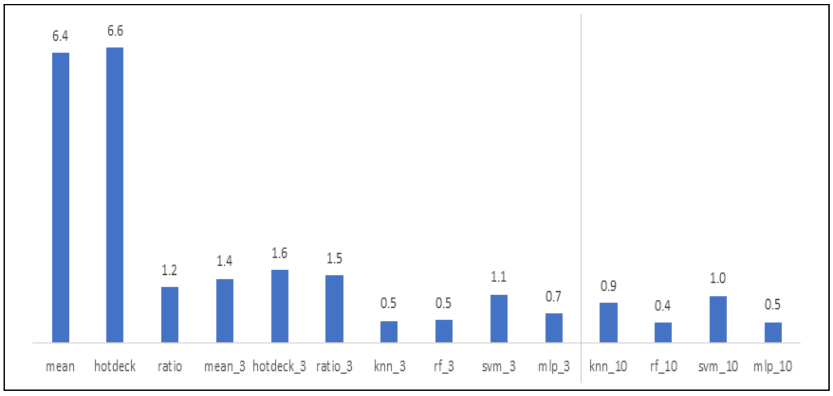
(단위: 만 원)



자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[부도 A-16] 무응답 비율 10%에서의 전체 평균에 대한 제곱근평균제곱오차(한국의료패널)

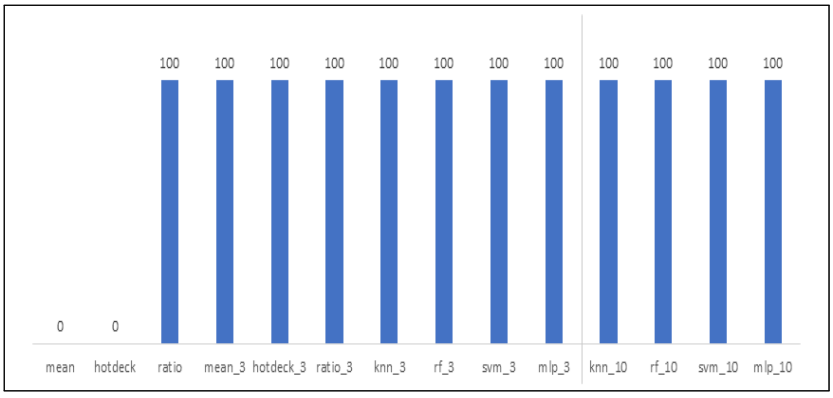
(단위: 만 원)



자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[부도 A-17] 무응답 비율 10%에서의 참값의 포함률(한국의료패널)

(단위: %)



자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

〈부표 A-15〉 무응답 비율 10%에서의 추정의 정확성 지표 결과(한국의료패널)

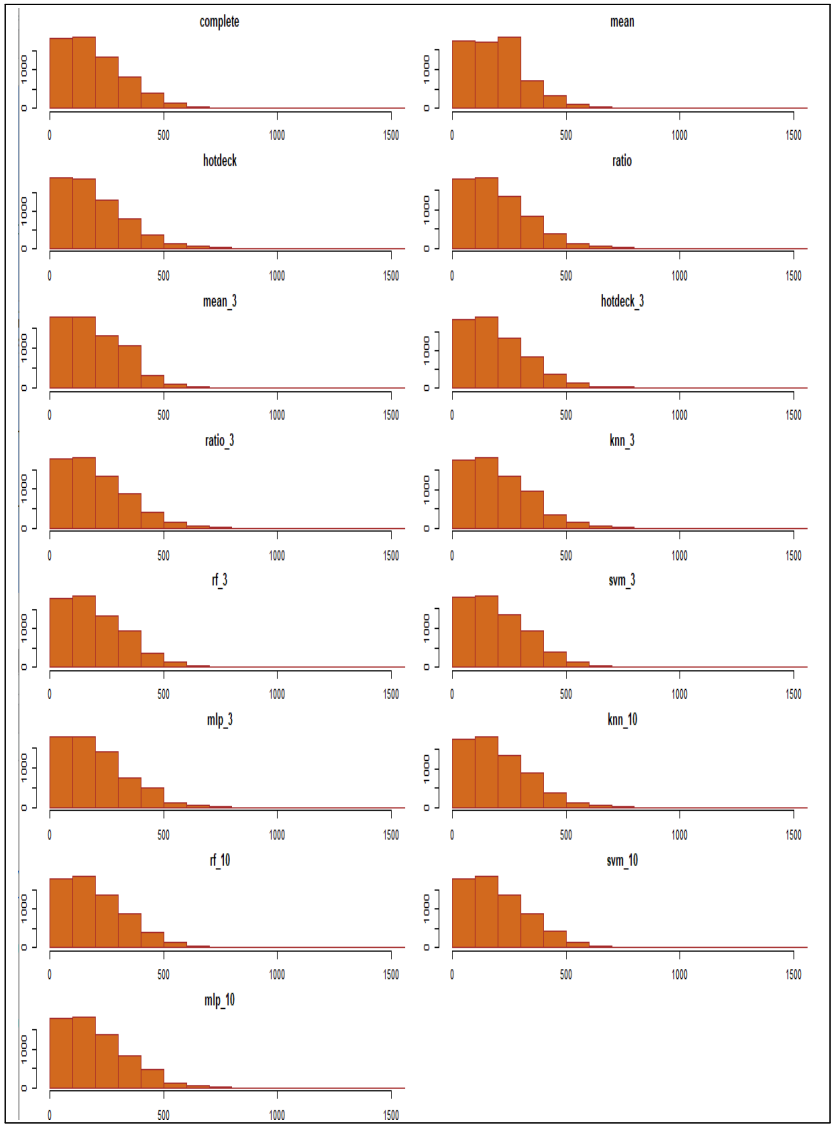
대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	12.6	53.0	0.02908	0.02499	0.11192	0.86035
hotdeck	17.3	70.5	0.02955	0.01036	0.05905	0.45159
ratio	7.9	36.2	0.00528	0.00235	0.04521	0.43674
mean_3	9.0	39.1	0.00598	0.02376	0.07321	0.40011
hotdeck_3	12.6	54.6	0.00669	0.00306	0.05346	0.49670
ratio_3	8.0	36.3	0.00648	**0.00208**	**0.04482**	0.42190
knn_3	7.4	32.5	**0.00178**	0.01641	0.08583	**0.32130**
rf_3	7.5	33.0	0.00181	0.01453	0.06831	0.33170
svm_3	**7.3**	**32.3**	0.00439	0.01477	0.07062	0.32074
mlp_3	7.9	35.0	0.00241	0.01813	0.09250	0.32233
knn_10	6.8	31.0	0.00361	0.01317	0.06477	0.31793
rf_10	6.1	28.1	0.00169	0.01229	0.08101	0.35067
svm_10	6.1	28.2	0.00439	0.01284	0.07985	0.33085
mlp_10	6.4	29.6	0.00177	0.01236	0.09156	0.38649

주: 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[부도 A-18] 무응답 비율 10%에서 1번째 자료의 분포-1(한국의료패널)

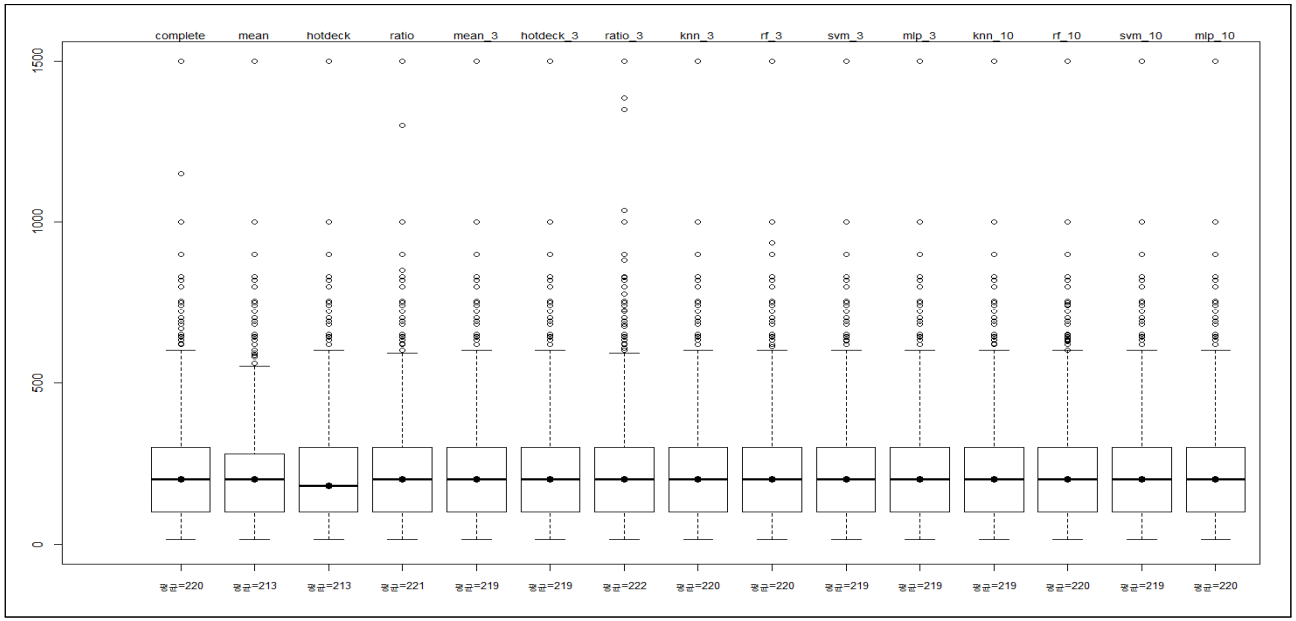


주: X축은 생활비(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[부도 A-19] 무응답 비율 10%에서 1번째 자료의 분포-2(한국의료패널)



주: X축은 생활비 평균(단위: 만 원)이고, Y축은 생활비(단위: 만 원)를 나타냄.

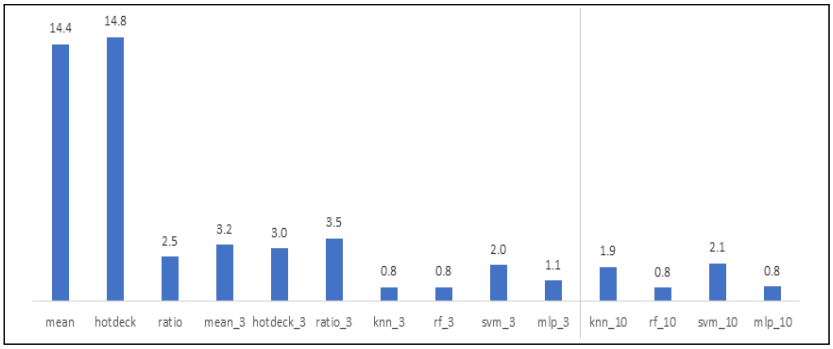
자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

라. 무응답 비율 20%에 대한 결과

[부도 A-20] 무응답 비율 20%에서의 전체 평균에 대한 제곱근평균제곱오차(한국의료패널)

(단위: 만 원)

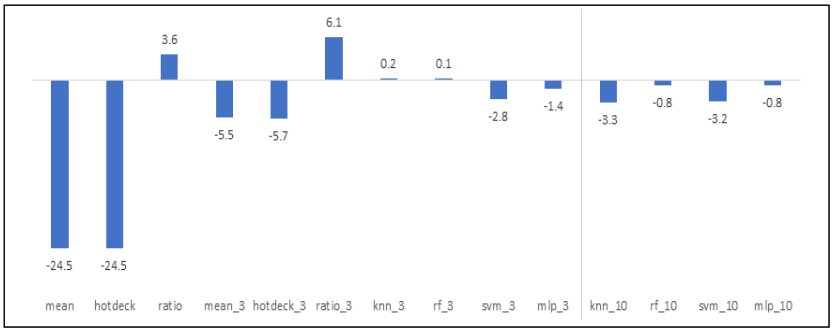


자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

마. 무응답 비율 30%에 대한 결과

[부도 A-21] 무응답 비율 30%에서의 전체 평균에 대한 편향(한국의료패널)

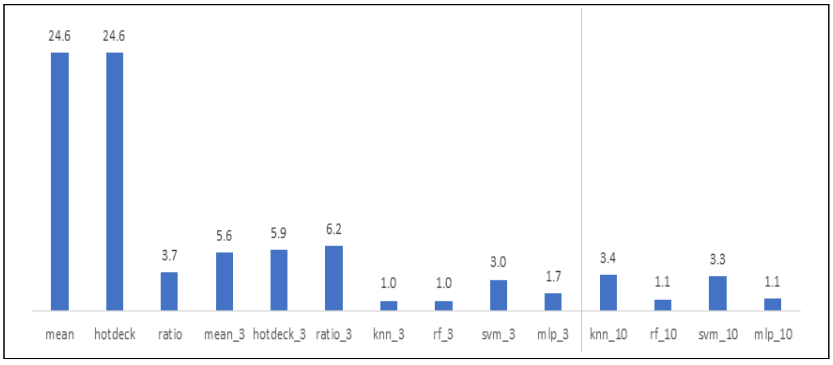
(단위: 만 원)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.
자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[부도 A-22] 무응답 비율 30%에서의 전체 평균에 대한 제곱근평균제곱오차(한국의료패널)

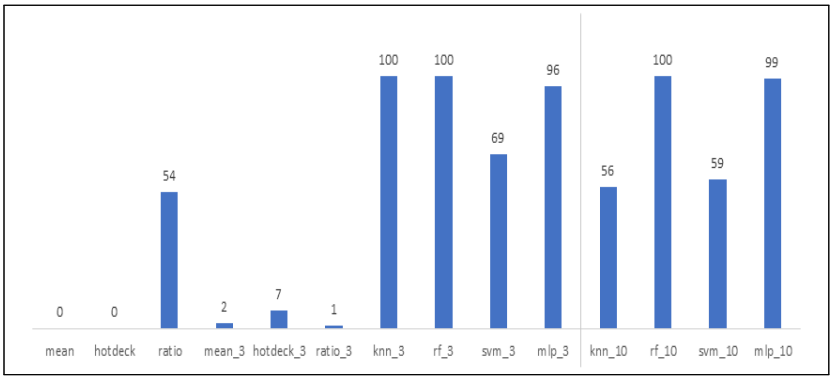
(단위: 만 원)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.
자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[부도 A-23] 무응답 비율 30%에서의 참값의 포함률(한국의료패널)

(단위: %)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.
자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

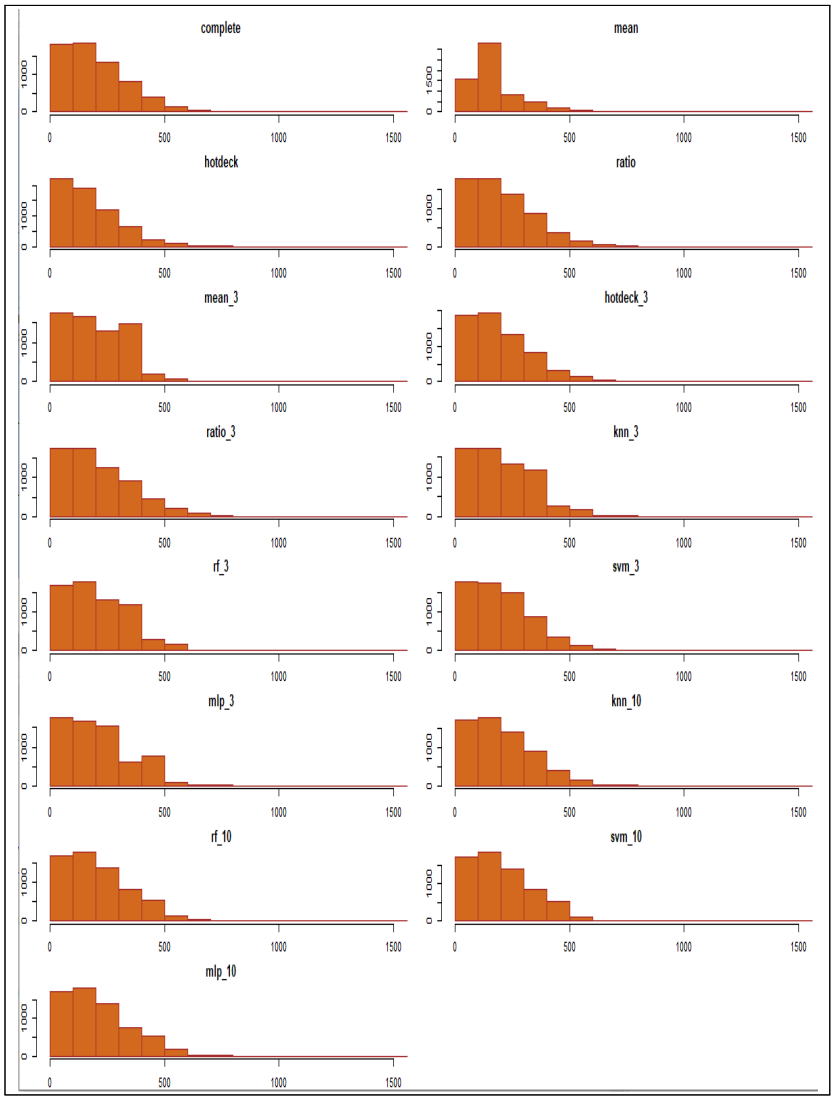
〈부표 A-16〉 무응답 비율 30%에서의 추정의 정확성 지표 결과(한국의료패널)

대체 방법	예측의 정확성		추정의 정확성			
	MAD	RMSD	ADB	ADV	ADS	ADK
mean	39.0	96.0	0.11148	0.08194	0.44307	3.52189
hotdeck	52.0	122.3	0.11146	0.03356	0.16838	1.08203
ratio	23.6	62.9	0.01648	**0.00413**	**0.07252**	0.78174
mean_3	27.0	68.5	0.02501	0.07817	0.29752	0.78274
hotdeck_3	37.6	94.5	0.02597	0.00835	0.11680	1.10695
ratio_3	24.5	64.5	0.02787	0.00582	0.08111	**0.73633**
knn_3	22.4	56.5	**0.00343**	0.05108	0.30634	1.06991
rf_3	22.7	57.6	0.00354	0.04482	0.24163	0.95108
svm_3	**21.9**	**56.1**	0.01296	0.04542	0.25473	0.95797
mlp_3	23.8	60.8	0.00656	0.05551	0.34233	1.10025
knn_10	21.3	55.9	0.01484	0.04102	0.22807	0.94431
rf_10	18.4	49.0	0.00421	0.03749	0.27158	1.16036
svm_10	18.3	49.5	0.01459	0.04080	0.29249	1.14447
mlp_10	19.2	51.9	0.00414	0.03624	0.30252	1.34158

주: 1) 평가지표에서 가장 작은 값을 가지는 대체 방법을 볼드체로 표시하였음(mean에서부터 mlp_3까지 해당).
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

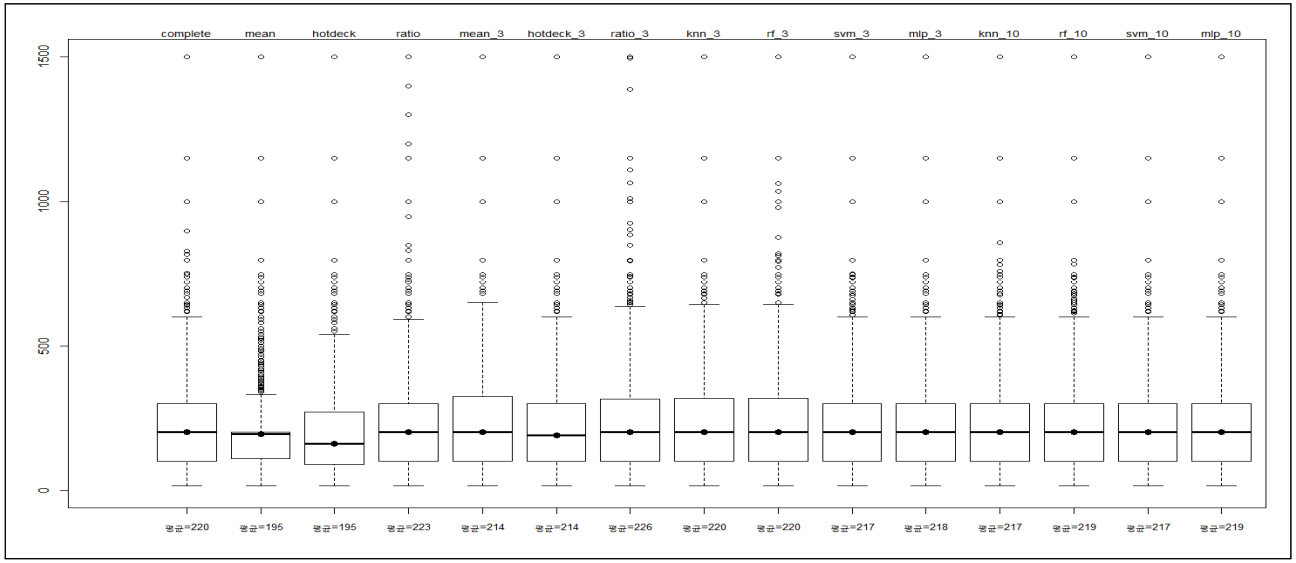
[부도 A-24] 무응답 비율 30%에서 1번째 자료의 분포-1(한국의료패널)



주: 1) X축은 생활비(단위: 만 원)이고, Y축은 빈도(단위: 가구)를 나타냄.
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

[부도 A-25] 무응답 비율 30%에서 1번째 자료의 분포-2(한국의료패널)



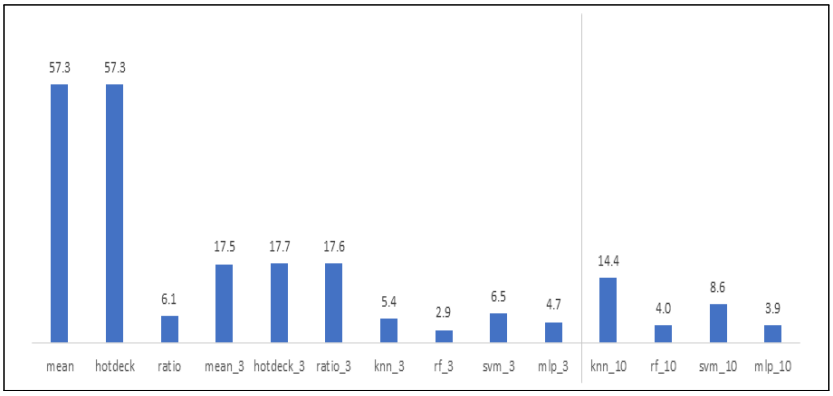
주: 1) X축은 생활비 평균(단위: 만 원)이고, Y축은 생활비(단위: 만 원)를 나타냄.
2) hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.
2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr에서 2019. 6. 7. 인출.

바. 무응답 비율 50%에 대한 결과

〔부도 A-26〕 무응답 비율 50%에서의 전체 평균에 대한 제곱근평균제곱오차(한국의료패널)

(단위: 만 원)



주: hotdeck_3은 기증자 부족으로 복원 추출로 대체한 결과임.

자료: 1) 한국보건사회연구원. (2014). 한국의료패널조사(8차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

2) 한국보건사회연구원. (2015). 한국의료패널조사(9차)[데이터파일]. https://www.khp.re.kr 에서 2019. 6. 7. 인출.

간행물 회원제 안내

회원에 대한 특전

- 본 연구원이 발행하는 판매용 보고서는 물론 「보건복지포럼」, 「보건사회연구」도 무료로 받아보실 수 있으며 일반 서점에서 구입할 수 없는 비매용 간행물은 실비로 제공합니다.
- 가입기간 중 회비가 인상되는 경우라도 추가 부담이 없습니다.

회원 종류

전체 간행물 회원	보건 분야 간행물 회원
120,000원	75,000원
사회 분야 간행물 회원	**정기 간행물 회원**
75,000원	35,000원

가입방법

문의처

- (30147) 세종특별자치시 시청대로 370 세종국책연구단지 사회정책동 1~5F
  간행물 담당자 (Tel: 044-287-8157)

KIHASA 도서 판매처

- 한국경제서적(총판) 737-7498
- 교보문고(광화문점) 1544-1900
- 영풍문고(종로점) 399-5600
- 서울문고(종로점) 2198-2307
- Yes24 http://www.yes24.com
- 알라딘 http://www.aladdin.co.kr