The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.3.377
30(3), 377–390

Comparison of imputation methods for item nonresponses in a panel study

Hyejung Lee[a] · Juwon Song[a,1]

[a]Department of Statistics, Korea University

(Received February 16, 2017; Revised April 3, 2017; Accepted April 3, 2017)

Abstract

When conducting a survey, item nonresponse occurs if the respondent does not respond to some items. Since analysis based only on completely observed data may cause biased results, imputation is often conducted to analyze data in its complete form. The panel study is a survey method that examines changes of responses over time. In panel studies, there has been a preference for using information from response values of previous waves when the imputation of item nonresponses is performed; however, limited research has been conducted to support this preference. Therefore, this study compares the performance of imputation methods according to whether or not information from previous waves is utilized in the panel study. Among imputation methods that utilize information from previous responses, we consider ratio imputation, imputation based on the linear mixed model, and imputation based on the Bayesian linear mixed model approach. We compare the results from these methods against the results of methods that do not use information from previous responses, such as mean imputation and hot deck imputation. Simulation results show that imputation based on the Bayesian linear mixed model performs best and yields small biases and high coverage rates of the 95% confidence interval even at higher nonresponse rates.

Keywords: imputation, panel data, linear mixed model, ratio imputation, Korean Labor and Income Panel Study

1. 서론

설문조사에서 무응답은 응답해야 하는 대상이 참여를 거절하거나 접촉 불가, 부재, 변심, 이민 등의 이유로 조사에 실패한 경우 또는 조사에는 응하였으나 일부 문항에 대하여 응답하지 않는 경우를 모두 통틀어 응답이 얻어지지 않았다는 의미로 무응답이라고 정의한다. 설문조사의 무응답은 흔히 단위무응답(unit nonresponse)과 항목무응답(item nonresponse)으로 구분할 수 있는데 응답자가 설문조사 전체에 참여하지 않은 경우 발생하는 무응답을 단위무응답이라 하고, 응답자가 설문조사에는 참여하였으나 일부분의 문항(특히 소득, 자산과 부채 등과 같은 금액을 묻는 질문이나 사생활과 연관된 문항)에 응답하는 것을 거부하여 응답값을 알 수 없는 경우의 무응답을 항목무응답이라고 부른다. 보통 데이터에 항목무응답이 포함되어 있을 때, 이 무응답을 제외하고 측정된 값만을 가지고 데이터 분석을 실시하는 완

[1]Corresponding author: Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-Gu, Seoul 02841, Korea. E-mail: jsong@korea.ac.kr

전한 자료에 근거한 분석(complete-case analysis)은 무응답 분포가 응답값 분포와 다른 경우 분석결과에 편의(bias)가 발생할 수 있다.

일반적으로 무응답이 존재하는 응답자의 전체 응답이 분석에서 제외되기 때문에 이를 채워 넣어 완전한 형태의 자료를 분석하기 위해서 무응답 대체(imputation)가 흔히 사용되고 있다. 대체 방법에는 평균대체(mean imputation)처럼 간단한 방법부터 회귀대체(regression imputation)와 확률적 회귀대체(stochastic regression imputation) 등 명시적 모형에 근거한 대체가 있으며, 핫덱대체(hot deck imputation)와 콜드덱대체(cold deck imputation)처럼 내재적 모형에 근거한 대체 방법이 있다. 이러한 대체 방법들은 주로 횡단면 자료(cross-sectional data)에서 무응답이 발생할 때 사용하는 방법이지만 패널자료(panel data)에서도 적용하는 경우가 많다. 그러나 패널자료의 장점은 동일한 설문 문항에 대해 반복적으로 질문하므로 어떠한 시점에서 무응답이 발생한 경우, 이에 대한 무응답 대체 방법으로 이전 시점의 정보를 활용한다면 보다 정확하게 대체할 수 있을 것이다. 이 관점에서 국내외 패널조사에서는 이전 시점의 정보를 이용하여 대체를 실시하는 방법을 흔히 적용해 왔다. 이 방법은 이전 시점의 응답값을 그대로 가져와 적용하는 last observation carried forward (LOCF) 방법에서부터 비대체(ratio imputation), 회귀대체 및 패널조사 자료에 적합한 모형에 근거한 대체 등 다양한 방법을 포함한다 (Song, 2015).

여러 가지 대체 방법이 사용되고 있으나 대부분 횡단면 자료의 대체 방법에 중점을 두고 연구가 진행되어 왔다. 따라서 본 연구에서는 패널자료에서 대체를 실시하는 방법에 초점을 맞추어 이전 시점의 정보를 이용하여 대체하는 방법들 중에서 어느 대체 방법이 보다 적절한 대체를 제공하는지 살펴보았다. 과거 정보 및 현재 시점의 연관 정보를 이용하는 방법인 선형혼합모형을 이용한 대체와 선형혼합모형에 근거한 베이지안 대체 방법을 고려하였고, 이와 함께 이전 시점의 정보를 사용하는 비대체 방법도 함께 살펴보았다. 또한 이 방법들을 횡단면 자료의 대체 방법인 평균대체 및 핫덱대체 결과와 비교하였다.

평균대체 방법은 무응답값들을 응답한 개체들의 평균으로 대체하는 것으로 손쉽게 사용할 수 있다. 독일 German Socio-Economic Panel (GSOEP)에서는 무응답률이 낮은 변수는 평균대체를 사용하였다 (Frick과 Grabka, 2004). 한국노동패널조사에서의 취업시기와 퇴직시기 변수는 트리모형을 사용하여 대체군을 형성한 후 대체군 내 평균값으로 대체하였다 (Song, 2015).

핫덱대체 방법은 응답한 개체들만의 자료에 근거하여 특성 변수들이 유사한 값을 가지도록 대체군(imputation cell)을 형성한 후, 대체군 내에서 무작위로 기증자를 선택하여 기증자의 응답값으로 무응답을 대체한다. 미국 Survey of Income and Program Participation (SIPP)은 지난 차수 자료를 이용하여 유사한 특성을 가진 기증자의 값으로 대체하는 순차적인 핫덱대체(sequential hot deck imputation)를 사용하였다 (U.S. Census Bureau, 2016). 캐나다 Survey of Labor and Income Dynamics (SLID)에서도 소득의 횡단면 대체에서는 최근접이웃 핫덱대체(nearest neighbor hotdeck imputation)를 사용하는데 매칭(matching)을 위한 변수들을 설정하여 최근접이웃을 찾아낸 후 이 응답자의 값으로 대체를 실시하였다. 영국 British Household Panel Survey (BHPS)에서는 범주형 소득관련 변수를 대체하기 위하여, 그리고 한국노동패널조사에서도 금액관련 변수를 대체하는 방법 중 하나로 핫덱대체를 사용하고 있다.

비대체는 소득과 소비 활동과 같이 매년 일정 비율로 증감하는 변수에 대하여 직전 시점의 응답값에 일정한 증감 비율을 곱하여 대체를 실시하는 대체 방법이다. 비율은 대체 대상 시점과 이전 시점이 모두 측정된 패널들을 대상으로 추정하는 것이 일반적인데 패널에 속하는 사람들의 특성에 따라 증감 비율이 다른 경우 이를 감안하는 것이 바람직할 것이다. 한국노동패널조사 자료의 경우 비대체를 실시하는데 대체군을 형성하고 대체군 내에서 증가 비율을 추정하여 대체를 실시하는 방법을 적용하고 있다 (Song, 2015). 미국 Panel Study of Income Dynamics (PSID)에서 사용된 대체 방법 중에서 전년도 자료 이

월(carryover) 방법은 지난해의 소득이 존재하는 경우 전년도 자료에 소득 증가분을 고려하여 이월하는 방법으로 비대체라고 볼 수 있다 (Duffy, 2011). 영국 BHPS에서도 무응답 대체 중에서 cross-wave imputation 방법이 있는데 대체하고자 하는 변수의 지난 차수의 값에 랜덤하게 선택한 비슷한 특성을 가지는 케이스의 변화율을 적용하는 방법으로 비대체의 일종이라고 볼 수 있다 (Taylor 등, 2010).

선형혼합모형(linear mixed model)을 이용한 대체는 패널자료 분석 시 많이 사용하는 방법인 선형혼합모형을 무응답 대체 시 사용한다. 선형혼합모형은 EM 알고리즘(expectation-maximization algorithm)을 사용하여 최대가능도(maximum likelihood; ML) 또는 제한최대가능도(restricted maximum likelihood; REML) 함수를 최대화하는 모수의 추정치를 구하는데, 이 때 EM 알고리즘의 반복 작업을 통하여 최종적으로 수렴하는 추정치를 찾는다 (Laird와 Ware, 1982). 이 때 적합한 모형에 근거하여 구한 기댓값을 가지고 대체를 실시하는 것이 가능하다. 이 모형에서는 기댓값을 주변부 평균(marginal mean) 또는 조건부 평균(conditional mean)으로 계산할 수 있는데, 본 논문에서는 각 패널의 값을 보다 정확히 예측하기 위하여 조건부 평균으로 대체를 실시하였다.

선형혼합모형에 근거한 베이지안 대체는 Markov Chain Monte Carlo (MCMC) 기법을 이용하여 선형혼합모형 하에서 무응답을 대체하는 방법이다 (Schafer와 Yucel, 2002). 베이지안 추정을 통해 모수의 사후분포(posterior distribution)에서 모수들을 추출하고 이 추출된 모수들에 근거하여 무응답의 예측분포(predictive distribution)에서 추출한 값을 가지고 대체하는 방법이다. 모수의 분포와 무응답의 예측분포를 고려하여 대체를 실시하므로 모형의 불확실성(uncertainty)을 더 적절히 반영할 수 있다.

본문의 구성은 다음과 같다. 2장은 본 연구에서 고려한 대체 방법들을 소개하였고 3장은 모의실험과 실제 자료에 대하여 무응답 대체를 실시할 때 사용한 자료인 한국노동패널조사에 대해 간략히 소개하였으며, 4장은 모의실험을 통해 2장에서 다룬 대체 방법들의 성능을 비교하였고, 5장에서는 이 방법들을 실제 한국노동패널자료의 무응답 대체에 적용한 예를 보였다. 마지막 6장에서는 본 논문의 결과를 요약하고 추후 연구 방향을 논의하였다.

2. 대체 방법(imputation method)

본 연구에서는 평균대체, 핫덱대체, 비대체, 선형혼합모형을 이용한 대체 및 선형혼합모형에 근거한 베이지안 대체 방법을 고려하였다.

2.1. 방법1 - 평균대체(mean imputation)

평균대체 방법은 응답한 개체들만을 가지고 구한 평균값으로 무응답을 대체하는 것이다. 대체가 쉽고 간단하며, 만약 결측자료 메커니즘(missing data mechanism)이 missing completely at random (MCAR) (Little과 Rubin, 2002)이면 평균 추정치는 불편성을 만족한다. 그러나 무응답을 모두 동일한 값(평균)으로 대체하기 때문에, 대체된 자료는 표준편차(standard deviation)를 과소추정하게 되며 무응답이 모두 동일한 평균값으로 대체되어 대체된 자료는 분포가 왜곡될 수 있다.

2.2. 방법2 - 핫덱대체(hot deck imputation)

핫덱대체 방법은 응답자의 값으로 무응답을 대체하는 기법이다. 일반적으로 무응답과 특성이 비슷한 개체를 찾기 위해서 여러 관련 변수들을 가지고 대체군을 형성한 다음에, 그 대체군 내에서 기증자를 무작위로 선택하여 기증자의 응답값으로 대체를 실시한다. 앞에서 설명한 평균대체에 비해, 자료의 분포를 잘 유지해 주고 어떤 형태의 변수에도 적용할 수 있다는 장점을 가지고 있어서 대체 방법으로 많이 사용

하고 있다.

2.3. 방법3 - 비대체(ratio imputation)

비대체 방법은 회귀대체의 특별한 경우로, 총근로소득처럼 매년 일정한 비율을 가지고 증감하는 경향이 있는 변수에 흔히 적용되고 있다. 종단면 자료에서 해당 차수(예를 들어 t번째 차수)의 대체하고자 하는 변수의 예측값을 구하기 위해 먼저 직전 차수($t-1$차수)와 비교해서 t차수의 응답값의 평균 증감 비율을 구한다. 여기서, y_{it}를 i번째 개체의 t차수에서의 반응변수값이라 하고 y_{it-1}을 i번째 개체의 $t-1$차수에서의 반응변수값이라고 하고 R은 해당 차수와 이전 차수 모두에서 응답한 개체들만의 집합을 의미할 때 증감비율은

$$\text{ratio} = \frac{1}{n}\left(\sum_{i\in R}\frac{y_{it}}{y_{it-1}}\right) \tag{2.1}$$

으로 계산된다. n은 $t-1$차수와 t차수에서 모두 응답한 개체들의 개수를 의미한다.

즉, ratio는 두 차수에서 모두 응답한 i 개체($i \in R$)의 반응변수값 증가율 ratio$_i$들의 평균을 의미한다. 비대체에서는 대체하고자 하는 t차수의 반응변수가 무응답인 경우 직전 차수인 $t-1$차수 반응변수값에 ratio를 곱하여 대체한다 ($y_{it} = y_{it-1} \times \text{ratio}$).

비대체는 이전 차수와 현재 차수 간의 비율을 구할 수 있으면 쉽게 대체 가능한 방법 중 하나이다. 이 비율은 대체 대상 시점과 이전 시점이 모두 측정된 패널들을 대상으로 추정하는 것이 일반적인데 패널에 속하는 사람들의 특성에 따라 비율이 다른 경우 이를 감안하여 대체군을 형성하여 대체군 내에서 비율을 계산하는 것이 바람직할 것이다.

2.4. 방법4 - 선형혼합모형(linear mixed model)을 이용한 대체

선형혼합모형은 종단면 데이터를 분석할 때 많이 사용하는 분석 모형으로서 개체들이 공통적으로 설명하는 부분을 고정효과(fixed effect)로, 각 개체만의 특성을 설명하는 부분을 랜덤효과(random effect)로 구분하여 개체 내 상관관계(within subject correlation)를 고려하여 분석하는 분석 기법이다. 고정효과는 모형에 포함된 설명변수로 표현하고, 랜덤효과는 설명변수에 의해 표현되지 않는 각 개체의 변동을 설명하고 있다. 이 모형은 반응변수의 일부 차수가 무응답이더라도 모두 포함하여 분석을 실시할 수 있다는 장점을 지닌다. 고정효과는 β로, 랜덤효과는 b로 표현하면 모형은 다음과 같이 표현할 수 있다.

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{2.2}$$

여기서, $i = 1, 2, \ldots, N$(개체의 수), $n_i = i$번째 개체에 대한 반복 측정 횟수를 나타낼 때 y_i는 i번째 개체의 반응변수값 $(y_1, \ldots, y_{n_i})$로 구성된 $(n_i \times 1)$ 행렬이며, X_i는 공변량으로 구성된 $(n_i \times p)$ 행렬로 이 때 p는 고정효과에 포함된 변수의 개수를 의미하며, Z_i는 공변량으로 구성된 $(n_i \times q)$ 행렬로 이 때 q는 랜덤효과에 포함된 변수의 개수를 의미하며, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$는 모집단에 대한 고정효과를 나타내는 계수이며, $b_i = (b_{i1}, b_{i2}, \ldots, b_{iq})^T$는 i번째 개체에 대한 랜덤효과를 나타내는 계수로 일반적으로 정규분포($b_i \sim N(0, \Psi)$)를 가정하며, 그리고 ϵ_i는 i번째 개체에 대한 오차항으로 $\epsilon_i \sim N(0, \sigma^2 I)$를 가정한다.

EM 알고리즘을 사용하여 최대가능도 또는 제한최대가능도 함수를 최대화하는 모수의 추정치를 구한다. EM 알고리즘의 E(expectation) 단계에서는 응답된 자료와 현재 반복에서의 모수 추정값을 이용하여 완전한 자료의 기대 로그우도를 구하며, M(maximization) 단계에서는 E 단계에서 구한 완전한 자료의

기대 로그우도를 최대화하는 모수 추정치를 갱신한다. E 단계와 M 단계를 반복하여 최종적으로 수렴하는 추정치를 구한다 (Dempster 등, 1977). ML 추정량(estimator)은 고정 모수의 수가 증가할수록 편의가 발생하지만 REML 추정량은 불편추정량(unbiased estimator)을 만족하므로 선형혼합모형의 모수를 추정할 때 흔히 사용한다.

2.5. 방법5 - 선형혼합모형(linear mixed model)에 근거한 베이지안 대체

Schafer와 Yucel (2002)은 MCMC 기법을 이용하여 선형혼합모형 하에서 무응답을 대체하는 방법을 제안하였다. 선형혼합모형 식 (2.2)에 대하여 베이지안 추정을 통해 모수의 사후분포에서 모수들을 추출한다. 그리고 이 추출된 모수들에 근거하여 무응답인 y 변수의 예측분포 하에서 y_i의 무응답값을 대체하는 방법으로서 다음과 같이 3단계 과정을 반복하게 된다.

(단계 1) 무응답과 모수들(β, Σ, Ψ)의 값이 임의로 주어졌다는 가정 하에 $b_1, b_2, \ldots, b_N$의 분포에서 임의값(random values)을 추출한다.

(단계 2) (단계 1)에서 얻은 $b_1, b_2, \ldots, b_N$과 무응답의 값이 임의로 주어졌다는 가정 하에 모수들(β, Σ, Ψ)을 사후분포로부터 추출한다.

(단계 3) (단계 1)에서 구한 $b_1, b_2, \ldots, b_N$과 (단계 2)에서 구한 모수들(β, Σ, Ψ)을 기반으로 하여 무응답값을 예측분포에서 추출한다.

위의 3단계를 차례로 반복할 때 각 단계에서 추출하는 모수나 무응답의 분포는 이전 단계에서 추출된 값에 의존하지만 이를 반복적으로 실행하면 목적함수(target distribution)에서 추출할 수 있다. (단계 1)에서 무응답과 모수들의 초기값을 EM 알고리즘을 사용하여 추정한 최대우도추정치(mle)로 설정하면 목적함수로 수렴 시간을 절약할 수 있다. (단계 2)에서 공분산 행렬(Σ, Ψ)에 대한 베이지안 사전 분포(Bayesian prior distribution)를 정할 때 모형의 모수들에 대한 prior의 영향을 최소화하기 위하여 약한 사전 분포(prior distribution)를 사용하는 것이 일반적이다.

선형혼합모형을 이용한 대체와 선형혼합모형에 근거한 베이지안 대체는 모두 선형혼합모형을 사용한다는 점에서 동일하다. 선형혼합모형을 이용한 대체는 대체값을 선형혼합모형에서 적합된 모형에 근거하여 구한 기댓값으로 대체하는 반면에, 선형혼합모형에 근거한 베이지안 대체는 모수의 분포와 무응답의 예측분포를 고려하여 대체를 실시하므로 모형의 불확실성을 더 적절히 반영하여 대체를 실시하는 방법이라는 점에서 차이가 있다.

3. 한국노동패널조사

한국노동패널조사(Korean Labor and Income Panel Study; KLIPS)는 가구의 특성과 가구원들의 경제활동 및 노동시장 이동, 소득활동 및 소비, 교육 및 직업훈련, 사회생활 등에 대하여 매년 1회씩 추적 조사하는 종단면 조사이다(http://www.kli.re.kr/klips). 1998년 5,000가구의 13,321명 가구원을 대상으로 시작하여, 2009년(12차)에 표본을 추가하여 모집단을 전국단위로 확대하였다. 2014년에 실시된 제17차 KLIPS 조사에서는 원표본 3,451가구 및 분가한 2,101가구를 합하여 총 5,552가구에 대한 조사를 성공하였으며, 가구 기준 원표본 유지율은 69.0%로 전년도에 비하여 0.4%포인트 하락한 한편, 개인 응답자는 총 10,757명을 조사 성공하였다 (Lee 등, 2015). 최근 19차 조사(2016년)를 완료하여 국내에서 존재하는 패널 중에서 장수패널이라고 볼 수 있다. 2017년에는 패널 탈락, 표본의 고령화 문제와 표본 마모 등의 문제를 해결하기 위해 표본을 추가할 계획에 있다. 현재 1–18차 데이터가 홈페이지에 공개되어 있으며 회원가입을 하면 자료를 무료로 다운로드하여 사용할 수 있다.

한국노동패널자료도 문항에 무응답이 존재하여 무응답을 대체한 자료도 함께 제공하고 있다. 한국노동패널자료는 전반적으로 봤을 때 무응답 비율이 높지 않은 편이다. 문항 중에서 월세금은 1차 조사 때 22%로 나타나 가장 높은 무응답 비율을 가지는 문항이었으나, 조사가 진행되면서 무응답 비율은 1–5% 이내의 비율을 가지며 점점 낮아지는 추세를 보이며 17차 조사에서는 0.2%로 낮은 비율을 가진다. 문항별로 무응답 비율이 상이하지만 대지면적, 연건평 및 시가 등의 문항은 무응답 비율이 10% 내외이고, 임대보증금, 금융소득, 총근로소득과 부동산소득 등의 문항은 5% 내외로 나타나 대체적으로 조사 초기에는 무응답 비율이 높은 편이었다. 그러나 조사가 점점 안정될수록 무응답 비율은 점점 낮아지는 추세를 가지는데, 한국노동패널조사는 7–8차 조사 이후부터는 무응답 비율이 대체적으로 5% 내외 이거나 모두 응답한 경우도 많아지는 것으로 나타났다. 한국노동패널자료에서는 무응답 대체 대상을 주로 소득, 자산, 부채 등 금액과 관련된 문항, 주된 일자리에서의 종업원수, 고용형태 및 취업시기 등과 같이 주요 문항들을 중심으로 선정하였다. 대체 방법은 핫덱대체, 비대체, 회귀대체, 중위수대체, 최빈값대체, 평균대체, 확률 기반 대체, 논리적 대체 및 임의대체 등을 가지고 문항의 특성에 따라 적합한 대체 방법을 적용하였다. 예를 들면, 월평균 생활비는 항목별로 과거자료를 이용하여 비대체를 하는 것을 원칙으로 하였다. 이 때 비보정 승수는 대체군을 형성하여 대체군 내에서의 비율로 구하였다. 주거비는 이사여부에 따라 2가지 대체 방법을 고려하여 이사하지 않은 가구의 경우에는 비대체를, 이사한 가구는 핫덱대체를 하였다. 주된 일자리에서의 종업원수는 범주형 종업원수 정보가 있으면 음이항 회귀모형(negative binomial regression)을 바탕으로 한 Stochastic EM (SEM) 방법으로, 범주형 종업원수 정보가 없으나 과거자료가 존재하고 이직을 한 경우에는 음이항 회귀모형으로, 이직을 하지 않은 경우에는 종업원수 구간별 비대체를 하였고, 범주형 종업원수 및 과거 자료도 없으면 음이항 회귀모형을 사용하여 대체하였다. 노동패널 홈페이지에 관련 보고서를 제공하고 있어서 무응답 처리에 대한 자세한 내용을 볼 수 있다 (Song, 2015).

4. 모의실험

4.1. 모의실험 자료 생성

본 연구에서는 종단면 자료에 대한 대체 방법들의 성능을 비교하기 위해 3개 차수(wave)인 13–15차 자료(2010–2012년 조사 자료)를 고려하였고 그 중 대체 대상 변수로 무응답이 상대적으로 많이 나타나는 '작년 한 해 총근로소득(이하 총근로소득)'을 선택하였다. 모의실험을 위하여 13–15차에서 총근로소득을 모두 응답한 4,849가구만을 고려하였고 이 중 월평균 생활비를 응답하지 않은 5가구를 제외하여 총 4,844가구를 연구 대상 모집단으로 간주하였다. 그리고 무응답 자료의 생성을 위하여 결측자료메커니즘이 missing at random (MAR)을 따른다는 가정 하에서 10%, 20%, 그리고 40%의 무응답이 발생하였다고 가정하였다. 무응답을 생성하기 위하여 총근로소득과 연관성이 높은 교육과 월평균 생활비를 가지고 자료를 9개 그룹으로 나누었다. 13차년도 자료의 경우에는 무응답 비율이 낮아서 상대적으로 무응답 비율이 가장 높은 1차년도 자료를 가지고 각 그룹별 무응답 비율을 계산하였다. 학력 범주는 고졸 미만, 고졸–전문대졸 미만, 전문대졸 이상으로 3개로 구분하였고, 월평균생활비 범주는 70만 원 미만, 70만–200만 원 미만, 200만 원 이상으로 3개로 구분한 후 1차년도 자료에서 9개의 각 그룹별 무응답 비율을 구하였다. 이 비율에 비례하여 전체 무응답 비율이 10%, 20% 및 40%가 되도록 각 그룹별 무응답 가구수를 조정하였다. 참고로 1차에 근거한 월평균 생활비 범주를 13–15차에 적용할 경우에는 조사 시점의 영향으로 대부분의 가구 월평균 생활비가 200만 원 이상으로 나타났다. 따라서 13–15차에 1차년도의 범주값을 동일하게 사용하는 것은 적절하지 않으므로 200만 원 미만, 200–300만 원 미만, 300만 원 이상으로 다시 구분하여 무응답 자료 생성에 사용하였다. 학력 범주는 1차년도와 동일한 범주값을 적용하였다. 패널자료에서는 총근로소득을 매 차수에서 조사하므로 무응답이 조사대상 변수의 이전 값에 의

Table 4.1. The number of nonresponding households, by nonresponse ratio

Nonresponse rate	Number of households	Number of households below median	Number of households above median
10%	424	128 (30%)	296 (70%)
20%	847	254 (30%)	593 (70%)
40%	1,697	473 (28%)	1,224 (72%)

존하는 것이 가능하므로 조사 시점의 총근로소득이 많을수록 무응답이 많이 발생한다고 가정하였다. 이를 위하여 9개 각각의 그룹에서 중위수를 기준으로 중위수 미만에서는 그룹별 무응답 가구수의 30%, 중위수 이상에서는 70%의 무응답이 발생한다고 가정하였다. 그런데 무응답 비율이 40%인 경우 무응답의 비율이 높아 일부 그룹에서 필요한 무응답 가구수를 추출할 수 없어 이 경우 중위수 미만에서는 28%로, 중위수 이상에서는 72%의 가구에서 무응답이 발생하도록 조정하였다. Table 4.1은 무응답 비율에 따른 무응답 가구수를 보여주는 표이며, 무응답 비율이 10%일 때 무응답 가구수는 424가구이고 무응답 비율이 20%는 847가구이고 무응답 비율이 40%인 경우는 1,697가구로 확정하였다.

각 대체군 내에서 중위수는 13차 총근로소득을 기준으로 계산하였다. 각 가구의 13차 총근로소득이 중위수 미만인지 이상인지에 따라 두 집단으로 구분하였다. 다음 최종적으로 중위수 이상 및 미만인 각 집단에 대해 Table 4.1의 무응답 비율에 따라 무작위로 추출한 가구의 14차 총근로소득을 무응답으로 처리하였다. 그리고 독립적으로 동일한 무응답 비율에 따라 무작위로 추출한 가구의 15차 총근로소득을 무응답으로 처리하였다. MAR 가정을 만족하도록 첫 번째 차수인 13차 총근로소득은 무응답이 없이 모두 값을 가지고 있다고 가정하였다. 이 가정은 또한 14차 자료에 대한 비대체를 할 때 비의 분모가 결측되지 않도록 하는 효과를 지닌다. 전체 자료에 대하여 무응답 가구를 단순임의추출(simple random sampling)하는 과정을 100번 실시하여 모의실험을 위한 무응답 자료를 100개 생성하였다.

4.2. 모의실험 방법 및 결과

각 모의실험 자료에 대하여 다음의 5가지 대체 방법에 근거하여 대체를 실시하였다.

(1) 평균대체: 각 해당차수에서 응답한 총근로소득을 가지고 구한 평균값으로 무응답을 대체하였다.

(2) 핫덱대체: 먼저 대체군을 형성하기 위해서 가구주의 학력과 성별, 그리고 월평균 생활비 정보를 사용하였다. 가구주의 학력은 3개 범주(고졸 미만, 전문대졸 미만, 전문대졸 이상)로, 월평균 생활비는 4개의 범주(150만 원 미만, 250만 원 미만, 350만 원 미만, 350만 원 이상)로 구분하였다. 각 대체군내에서 기증자(donor)를 무작위로 추출하여 기증자의 응답값으로 무응답을 대체하였으며, 각 기증자는 한 번만 대체에 사용되었다.

(3) 비대체: 우선 비(ratio)를 구해야 하는데, 여기서도 핫덱대체에서 대체군을 형성할 때 사용한 가구주의 학력, 성별, 그리고 월평균 생활비를 가지고 그룹을 구분하여 각 그룹에서의 비를 계산하였다. 14차 무응답 가구의 총근로소득의 대체는 13차 총근로소득에 비를 곱하여 구했으며, 15차도 동일한 방법을 적용하여 대체하였다. 만약 14차와 15차가 모두 무응답인 가구는 14차 자료의 무응답을 우선 대체한 후 대체된 14차의 값에 비를 곱하여 15차를 대체하였다.

(4) 선형혼합모형에 근거한 대체: 먼저 선형혼합모형을 적합하기 위해 모의실험 자료 1에 대하여 여러 가지 후보 모형을 분석하여 AIC, AICC 및 BIC의 값이 가장 작은 모형을 최종 모형으로 선택하였다. 이 과정을 몇 개의 랜덤하게 선택한 모의실험 자료에 대하여 적용해 본 결과 모두 동일한 모형을 선택하였다. 최종 모형은 설명변수(고정효과)로 월평균생활비(로그 변환함), 가구주의 학력(고

Table 4.2. Bias and root mean square error (RMSE)

Wave	Imputation method	Nonresponse rate					
		10%		20%		40%	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
14th	complete	50.03	0	106.58	0	315.11	0
	mean	50.02	978.55	106.58	1,372.19	315.11	2,051.06
	hotdeck	19.62	1,111.06	38.91	1,530.28	118.40	2,342.99
	ratio	−61.71	764.63	−123.22	1,092.95	−232.94	1,647.23
	mixed	33.63	**618.95**	64.70	**883.65**	152.67	**1,401.86**
	pan	**1.41**	914.68	**1.87**	1,333.44	**−15.97**	2,080.82
15th	complete	35.49	0	81.48	0	252.41	0
	mean	35.49	831.40	81.49	1,194.06	252.41	1,717.46
	hotdeck	13.10	830.81	30.63	1,196.50	92.04	1,705.83
	ratio	−59.86	772.97	−127.87	1,245.91	−285.27	1,657.17
	mixed	26.78	**459.75**	61.16	**669.92**	158.05	**988.45**
	pan	**−10.21**	903.22	**−13.38**	1,290.38	**−42.50**	2,020.21

졸 미만, 전문대졸 미만, 전문대졸 이상의 3개 범주), 성별, 가구원수, 그리고 연속형 변수로 고려한 조사차수를 포함하도록 구성하며, 절편과 기울기에 해당하는 조사차수의 계수가 랜덤효과로 선정되었다. 랜덤효과의 분산의 구조는 특정한 패턴을 가지지 않는다고 가정(unstructured로 설정)한 후 모형을 적합하였다. 총근로소득의 무응답값은 식 (2.2)에 의해 추정된 총근로소득의 조건부 기댓값으로 대체하였다. 선형혼합모형의 적합은 대부분의 상용 통계 패키지 프로그램에서 구현 가능한데 본 연구에서는 SAS PROC MIXED를 사용하였다.

(5) 선형혼합모형에 근거한 베이지안 대체: 선형혼합모형에 근거한 베이지안 대체에서는 선형혼합모형을 이용한 대체와 동일한 모형을 사용하였다. R패키지 PAN을 사용하여 대체를 실시하였으며 MCMC 알고리즘의 반복수(iteration number)는 10,000번으로 설정하였다. PAN에서는 단일 대체(single imputation) 뿐만 아니라 다중 대체(multiple imputation) 방법을 이용하여 대체 자료를 생성할 수 있는데 본 연구에서는 단일 대체를 실시하였다.

5가지 대체 방법 외에 무응답을 제외하고 완전한 케이스만을 가지고 분석한 결과도 함께 제시하였다. 대체 방법의 성능은 다음의 지표를 사용하여 비교하였다.

(1) 대체 자료의 총근로소득에 대한 평균 추정치와 참값과의 차이(bias)

(2) 각 개체의 참값과 대체값 사이의 평균제곱근오차(root mean square error; RMSE)

(3) 평균에 관한 95% 신뢰구간의 참값의 포함율(coverage rate)

이후 보여주는 결과표의 내용 중 대체 방법을 간단하게 표기하기 위해서 다음과 같이 정의하였다. complete는 완전하게 응답한 케이스만 가지고 분석한 결과이며, mean은 평균대체를, hotdeck은 핫덱대체를, ratio는 비대체를, mixed는 선형혼합모형을 이용한 대체이고, pan은 선형혼합모형에 근거한 베이지안 대체 방법을 의미한다.

Table 4.2는 각 대체 방법에 근거한 추정량과 참값과의 차이 및 RMSE를 나타낸 것으로 가장 작은 편의를 보이는 방법을 진하게 표시하였는데 선형혼합모형에 근거한 베이지안 대체가 무응답 비율 및 조사차수에 상관없이 가장 작은 편의를 가지는 것으로 나타났다.

Table 4.3. Coverage rate of 95% CI

Wave	Imputation method	Nonresponse rate			Total
		10%	20%	40%	
14th	complete	**100**	26	0	100
	mean	**100**	26	0	100
	hotdeck	**100**	98	21	100
	ratio	**100**	3	0	100
	mixed	**100**	99	0	100
	pan	**100**	**100**	**99**	100
15th	complete	**100**	44	0	100
	mean	**100**	43	0	100
	hotdeck	**100**	**100**	20	100
	ratio	95	0	0	100
	mixed	**100**	98	0	100
	pan	**100**	**100**	**82**	100

무응답 비율이 점점 증가하게 되면 편의도 커지는 양상을 보였다. 조사차수별로 살펴보면, 14차 자료에서 선형혼합모형에 근거한 베이지안 대체는 무응답 10%와 20%에서는 참값과의 차이가 1.41과 1.87로 고려한 대체 방법들 중 가장 작을 뿐만 아니라 참값과도 거의 비슷하게 추정하였다. 이 방법은 무응답 40%에서도 참값과의 차이가 −15.97로 가장 작았다. 다음으로 참값과의 차이가 작은 대체 방법은 핫덱대체로, 14차 자료에서 참값과의 차이가 무응답 10%에서는 19.62이고 무응답 20%에서는 38.91이며 무응답 40%에서는 118.40인 것으로 나타났다. 선형혼합모형을 이용한 대체는 핫덱대체 다음으로 참값과의 차이가 작은데 14차 자료에서 무응답 10%에서는 33.63이고 무응답 20%에서는 64.70이며 무응답 40%에서는 152.67로 나타났다. 이와 다르게 비대체가 무응답 비율이 10%와 20%일 때 대체 방법 중에서 참값과의 차이가 가장 큰 것으로 나타났으나, 무응답 40%에서는 참값과의 차이가 −232.94로 완전한 케이스를 가지고 분석한 결과(315.11)나 평균대체(315.11)보다 작았다. 15차 자료에서도 비슷한 결과를 가지는데, 참값과의 차이가 선형혼합모형에 근거한 베이지안 대체가 가장 작았으며 비대체가 가장 크다는 것을 알 수 있다.

다음으로 참값과 대체값 사이의 RMSE를 살펴보면 조사차수 및 무응답 비율에 상관없이 모두 선형혼합모형을 이용한 대체에서 RMSE 값이 가장 작은 것으로 나타났다. 14차 자료에서 이 방법의 RSME는 무응답 40%일 때 1,401.86이고 15차 무응답 40%인 경우 988.45이다. 이 방법이 가장 작은 RMSE를 가지는 이유는 각 개체의 대체값을 조건부 기댓값으로 추정하기 때문인 것으로 추측된다. 나머지 대체 방법들의 RMSE는 조사차수와 무응답 비율에 따라 다르게 나타나 뚜렷한 형태를 찾아보기는 어렵다. 핫덱대체가 14차 자료에서 무응답 비율에 상관없이 RMSE(무응답 비율 10%에서 1,111.06, 무응답 비율 20%에서 1,530.28, 무응답 비율 40%에서 2,342.99)가 가장 큰 반면에, 15차에서는 선형혼합모형에 근거한 베이지안 대체가 무응답 비율에 상관없이 RMSE(무응답 비율 10%에서 903.22, 무응답 비율 20%에서 1,290.38, 무응답 비율 40%에서 2,020.21)가 가장 크게 나타났다. 선형혼합모형을 이용한 대체는 추정량의 표준오차를 작게 추정하여 RMSE의 값이 작은 반면에, 선형혼합모형에 근거한 베이지안 대체는 베이지안 추정으로 모수와 무응답의 불확실성을 고려하여 추정했기 때문에 RMSE 값이 상대적으로 크게 나타나는 것으로 보인다. 반면에 평균대체와 비대체는 평균 추정 시 편의가 상대적으로 크므로 이 편의가 RMSE에 영향을 준 것으로 보인다.

Table 4.3은 평균에 대한 95% 신뢰구간을 구한 후 이 구간이 참값을 포함하는 비율을 나타낸다. 대부분의 방법에서 무응답 비율이 높을수록 그리고 조사차수가 뒤로 갈수록 포함률은 낮아진다.

Table 5.1. Nonresponse rate of interest from savings in waves 15–17

	Wave		
	15th	16th	17th
Nonresponse	0	17	34
Response	591	801	812
Total	591	818	846
Nonresponse rate	0%	2.08%	4.02%

선형혼합모형에 근거한 베이지안 대체는 무응답 비율이 40%에서도 99%(14차) 및 82%(15차)를 포함하여 다른 방법에 비해 가장 우수한 결과를 보였다. 선형혼합모형을 이용한 대체는 20%가 무응답인 경우는 포함률이 14차 자료에서는 99%이고 15차 자료에서는 98%로 두 번째로 좋은 성능을 보였으나, 무응답의 비율이 40%에서는 모든 차수에서 참값을 포함하지 못하였다. 핫덱대체의 경우 무응답 비율 20%에서는 선형혼합모형을 이용한 대체와 거의 동일한 결과(14차 98%, 15차 100%)를 보였으며, 무응답 비율 40%에서도 14차 자료의 경우 21%와 15차 자료의 경우 20%를 포함하여 비록 낮은 비율이지만 선형혼합모형을 이용한 대체보다 포함률이 높게 나타났다. 즉, 무응답 비율이 낮은 경우에는 핫덱대체와 선형혼합모형을 이용한 대체의 결과의 포함률이 비슷하지만 무응답 비율이 높은 경우에는 핫덱대체의 포함률이 조금 더 우수한 것으로 나타났다. 이에 반해 평균대체와 비대체는 가장 좋지 않은 결과를 보였으며, 무응답 비율이 20%인 경우에도 불구하고 포함률이 매우 저조한 것으로 나타났다.

5. 실제 한국노동패널자료의 무응답 대체 적용

모의실험에서 비교한 5가지 대체 방법을 실제 자료의 대체에 적용해 보기 위하여 한국노동패널자료에서 무응답이 발생한 다른 변수인 '작년 한 해 금융소득-은행 등 금융기관 이자/투자소득(이하 이자소득)'을 선정하였으며, 조사차수는 시뮬레이션에서 사용한 13–14차 자료의 경우에는 무응답이 발생하지 않아서 15–17차 자료에 대하여 대체를 실시하였다.

Table 5.1을 보면 이자소득의 무응답 비율이 15차는 0%, 16차는 2.08%이며 17차는 4.02%로 높지 않은 편이다. 참고로 대체 방법을 적용할 때 '작년 한 해 월평균 저축액(이하 월평균 저축액)'을 설명변수 및 대체군을 형성하기 위하여 사용하는데 이 변수에 무응답이 발생하는 경우는 이 변수를 보조변수로 사용할 수 없어서 제외하였으며, 이에 따라 이자소득의 무응답 가구 중 16차 3가구와 17차 4가구가 분석에서 제외되었다.

이 자료에 대하여 본 논문에서 고려한 각 대체 방법을 다음과 같이 적용하였다. 평균대체는 각 해당차수에서 응답한 이자소득만을 가지고 구한 평균값으로 무응답을 대체하였다.

핫덱대체는 월평균 저축액에 근거하여 대체군을 형성하였는데, 이를 4개의 범주(50만 원 이하, 100만 원 이하, 200만 원 이하, 200만 원 초과)로 구분하여 대체군을 형성한 후 각 대체군 내에서 기증자를 무작위로 추출하여 기증자의 응답값을 가지고 무응답을 대체하였다. 기증자는 한 번만 대체에 사용하도록 하였다.

비대체는 비를 구하기 위해 가구주의 학력, 성별과 월평균 저축액을 가지고 대체군을 형성한 후 각 대체군 내에서의 비율을 계산하였다. 16차 무응답 가구의 이자소득의 대체는 15차 이자소득에 비를 곱하여 구하면 되는데, 16차 무응답 17가구가 모두 15차 때 이자소득이 없다고 대답하여 비대체를 하게 되면 비율과 상관없이 0으로 대체되었다. 17차도 3가구를 제외한 나머지 31가구는 0으로 대체되어 비대체는 적합하지 않다고 생각되어 비대체를 실시하지 않았다. 실제 한국노동패널조사에서 이자소득을 대

Table 5.2. Descriptive statistics of interest from savings in waves 16–17

Wave	Imputation method	N	Mean	Standard deviation	95% confidence interval	
					Lower bound	Upper bound
16th	complete	801	238.05	500.42	203.39	272.70
	mean	818	238.05	495.19	204.11	271.98
	hotdeck	818	239.22	500.04	204.96	273.49
	mixed	818	235.30	495.56	201.34	269.26
	pan	818	235.76	495.81	201.78	269.74
17th	complete	812	223.00	392.59	196.00	250.01
	mean	846	223.00	384.61	197.09	248.92
	hotdeck	846	221.22	387.97	195.07	247.36
	mixed	846	217.97	385.41	192.00	243.94
	pan	846	222.14	386.21	196.11	248.16

체한 방법은 작년도 이자소득이 존재하는 경우에는 비대체를, 작년도 이자소득이 존재하지 않는 경우 중 지난 한 달 이자소득이 존재하면 지난 한 달 이자소득에 12개월을 곱하여 대체를, 그 외의 경우에는 핫덱대체를 실시하였다 (Song, 2015).

선형혼합모형을 이용한 대체는 설명변수(고정효과)로 로그변환한 월평균 저축액과 조사차수를 고려하였고, 절편을 랜덤효과로 고려하여 최종 적합하였고 랜덤효과의 분산의 구조는 특정한 패턴을 가지지 않는다고 가정(unstructured로 설정)하여 SAS PROC MIXED로 분석하였다. 이자소득의 무응답은 식 (2.2)에 의해 추정된 이자소득의 조건부 기댓값으로 대체하였다.

선형혼합모형에 근거한 베이지안 대체는 선형혼합모형을 이용한 대체와 동일한 모형을 사용하였다. R패키지 PAN을 사용하여 대체를 실시하였으며 MCMC 알고리즘의 반복수는 10,000번으로 설정하여 단일 대체를 실시하였다. 대체를 실시한 결과를 비교하기 위해서 각 조사차수별 대체 자료의 이자소득의 평균, 표준편차와 평균에 대한 95% 신뢰구간을 구하였고 그 결과가 Table 5.2에 나타난다.

16차의 경우 선형혼합모형을 이용한 대체의 평균값은 235만 3천 원으로 가장 작았으며, 선형혼합모형에 근거한 베이지안 대체는 235만 8천 원으로 유사하게 추정하였다. 한편 핫덱대체가 239만 2천 원으로 대체 방법 중에서 가장 큰 값으로 추정하였고, 가장 작게 추정한 선형혼합모형을 이용한 대체와의 차이가 3만 9천 원으로 큰 편이었다. 표준편차는 평균대체가 495.19로 가장 작게 나타났고 선형혼합모형을 이용한 대체의 경우에는 495.56으로 작은 편에 속하며, 선형혼합모형에 근거한 베이지안 대체는 495.81이며 핫덱대체는 500.04로 완전히 응답된 자료 만에 근거한 표준편차인 500.42와 함께 가장 크게 나타났다. 평균에 관한 95% 신뢰구간을 보면 완전하게 응답된 케이스만을 가지고 분석한 결과의 신뢰구간 폭이 가장 넓었다(203.39–272.7). 대체 방법의 신뢰구간 폭은 평균대체(204.11–271.98), 선형혼합모형을 이용한 대체(201.34–269.26), 선형혼합모형에 근거한 베이지안 대체(201.78–269.74), 그리고 핫덱대체(204.96–273.49)의 순서로 넓은 편이나 차이는 크지 않았다. 17차도 16차와 비슷한 결과를 나타냈다. 선형혼합모형을 이용한 대체의 평균이 218만 원으로 역시 가장 작았으며 완전하게 응답한 케이스만을 가지고 분석한 결과와 평균대체의 평균이 223만 원으로 가장 크게 추정하였다. 표준편차의 경우 대체 방법 중에서 평균대체가 384.61로 가장 작고 완전하게 응답한 케이스만을 가지고 분석한 결과가 392.59로 가장 크게 나타났다. 평균에 관한 95% 신뢰구간의 폭은 완전하게 응답한 케이스만을 가지고 분석한 결과(196–250.01)가 가장 넓게 나타났다. 대체 방법들 중에는 평균대체(197.09–248.92), 선형혼합모형을 이용한 대체(192–243.94), 선형혼합모형에 근거한 베이지안 대체(196.11–248.16), 그리고 핫덱대체(195.07–247.36)의 순서로 넓으나 그 차이는 미약한 것으로 나타났다.

6. 결론

설문조사를 실시할 때 소득, 자산과 부채 등과 같은 금액과 관련된 문항에서 항목무응답이 종종 발생한다. 무응답이 존재하는 응답자의 자료는 분석에서 제외되어 추정량에 편의가 발생할 가능성이 있으므로 이를 대체한 후 완전한 형태의 자료로 제공하여 분석할 수 있도록 무응답을 대체하는 방법이 흔히 사용되고 있다. 본 연구에서는 패널자료에서 무응답 대체를 실시할 때, 패널자료의 특성을 이용한 대체 방법들의 성능을 살펴보았다. 패널자료는 횡단면 자료와 달리 무응답이 발생한 변수의 이전 시점에 측정한 정보를 가지고 있다는 장점이 있다. 그렇기 때문에 이전 시점의 정보를 이용하여 대체하는 방법들이 보다 적절한 대체를 실시하는지 살펴보고, 이전 시점의 정보를 이용하여 대체하는 방법들 중에서 어느 대체 방법이 보다 적절한 대체를 제공하는지 모의실험을 통해 조사해 보았다. 과거 정보를 이용하는 방법인 선형혼합모형을 이용한 대체와 선형혼합모형에 근거한 베이지안 대체 방법, 그리고 비대체를 고려하였고, 이와 함께 평균대체와 핫덱대체도 비교하여 살펴보았다.

모의실험 결과 선형혼합모형에 근거한 베이지안 대체 방법이 본 논문에서 고려한 다른 대체 방법들에 비해 무응답 비율이 높아지더라도 평균 추정량의 편의도 작으며 평균에 관한 95% 신뢰구간의 포함률도 높은 편으로 나타나서 가장 좋은 대체 방법인 것으로 확인되었다. 각 개별 개체의 응답값은 선형혼합모형에 근거한 대체가 가장 유사하게 추정하는 것으로 나타났다. 이는 이전 시점의 정보를 이용하여 대체하는 방법이 이전 시점의 정보를 고려하지 않는 다른 대체 방법에 비해 대체의 정확도가 높다고 볼 수 있다. 무응답 비율이 낮은 경우에는 핫덱대체, 선형혼합모형을 이용한 대체나 선형혼합모형에 근거한 베이지안 대체 모두 괜찮은 방법으로 나타났으나, 평균대체와 비대체의 성능은 낮게 나타났다.

모의실험에서는 무응답 비율을 최대 40%까지 설정해 본 결과 대체 방법들의 편의가 현격하게 다르고 일부 방법의 평균에 대한 95% 신뢰구간의 포함율은 현저히 낮아짐을 보여 주어 선형혼합모형에 근거한 베이지안 대체가 우수하다는 것을 확인할 수 있었다. 반면 한국노동패널자료의 실제 무응답 비율은 모의실험에서 고려한 결측의 비율보다 훨씬 낮은 5% 미만이다. 하지만 해외 유사 자료에서의 무응답률은 소득관련 변수에서 약 30% 정도로 상당히 높게 나타나는 것이 일반적이며 개인 사생활 중시 풍조 등과 맞물려 국내 조사 자료들에서의 무응답 비율도 증가하는 추세를 보이고 있다. 따라서 본 연구에서는 현재 한국노동패널자료의 무응답 비율보다 상당히 높은 실험 조건 하에서 모의실험을 실시하여 추후 무응답 비율이 높아지는 경우를 대비하고자 하였다. 한편, 5장에서는 실제 자료인 한국노동패널자료를 가지고 무응답 대체를 적용하였는데 무응답 비율이 낮으므로 대체 후 대체 방법 간의 편의가 크지 않게 나와서 현재 한국노동패널의 대체 방법으로 사용되는 비대체도 적용하기에 적절한 방법으로 생각된다. 하지만 추후 한국노동패널자료의 무응답 비율이 증가하거나 무응답 비율이 높은 다른 패널자료의 경우 비대체보다는 선형혼합모형에 근거한 베이지안 대체를 적용하는 것이 적절할 것으로 기대된다.

본 연구의 핫덱대체는 이전 시점의 정보를 사용하지 않고 진행하였다. 본 연구결과에 따르면 이전 시점 정보를 사용한 대체가 더 우수한 결과를 도출하므로 핫덱대체를 패널자료의 대체에 사용하는 경우 대체군을 형성할 때 이전 시점 정보를 포함한다면 더 나은 성능을 보일 수 있지 않을까 기대된다. 이에 관해서는 추후 연구가 진행되면 바람직할 것으로 생각된다.

본 연구에서는 3개 차수 만에 근거하여 모의실험을 실시하였는데 패널자료는 많은 경우 3개 이상의 시점을 포함하므로 이전 모든 시점의 정보를 활용하여 대체를 실시한다면 더 많은 정보에 근거하여 대체를 실시할 수 있다. 하지만 시점이 증가하게 되면 모형적합 시 시간에 따른 변화를 적절히 모형화해야 하며 이는 자료의 변화 추세에 따라 달라질 것으로 생각되고 이 방법의 성능도 달라질 수 있을 것이다. 추후 시점을 유연하게 확장해 가면서 보다 정확한 대체 방법은 무엇인지 살펴보는 연구가 진행된다면 더 적절한 대체를 실시할 수 있을 것으로 기대된다.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, **39**, 1–38.

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

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

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics*, **38**, 963–974.

Lee, K., Lee, J., Shin, S., Lee, H., and Kim, K. (2015). *The Economic Activity of Korean Individuals and Hou-eholds-2014 (Wave 17) Annual Report of the KLIPS Study*, Korea Labor Institute.

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

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values, *Journal of Computational and Graphical Statistics*, **11**, 437–457.

Song, J. (2015). *A Study of Improved Item Nonresponse Imputation Methods for KLIPS*, Korea Labor Institute.

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

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

패널자료에서의 항목무응답 대체 방법 비교

이혜정[a] · 송주원[a,1]

[a]고려대학교 통계학과

(2017년 2월 16일 접수, 2017년 4월 3일 수정, 2017년 4월 3일 채택)

요 약

설문조사를 실시할 때 응답자가 설문조사의 일부 문항에 대하여 응답하지 않는 경우 항목무응답이 발생한다. 무응답이 발생한 자료를 제외하고 완전하게 응답된 자료 만에 근거한 분석은 분석 결과에 편의가 발생할 수 있으므로, 이를 채워 넣어 완전한 형태의 자료로 분석하기 위해서 무응답 대체가 흔히 사용되고 있으며 여러 가지 무응답 대체 기법들을 비교하는 연구들도 많이 존재한다. 패널조사 연구는 연구 대상 패널에 대하여 정해진 시간에 따라 반복적으로 동일한 설문 문항에 대하여 응답을 조사하여 시간에 따른 변화를 살펴보는 조사 방법을 나타낸다. 패널조사 자료의 항목 무응답을 대체할 때 이전 시점의 응답 자료가 존재한다면 이를 포함하여 대체를 실시하는 것이 바람직한 것으로 여겨져 왔으나 이에 관한 직접적인 연구는 찾기 힘들다. 따라서 본 연구에서는 패널자료에서 이전 시점의 정보를 고려하지 않고 대체를 실시하는 방법과 이전 시점의 정보를 활용하여 대체하는 방법들 중에서 어느 대체 방법이 보다 적절한 대체를 제공하는지 살펴보았다. 특히 이전 시점의 응답 정보를 이용하는 방법인 비대체, 선형혼합모형을 이용한 대체와 선형혼합모형에 근거한 베이지안 대체 방법을 고려하였고, 이를 이전 시점의 정보를 고려하지 않는 대체 방법들 중 흔히 사용되는 평균대체, 핫덱대체 방법과 비교하였다. 모의실험 결과 선형혼합모형에 근거한 베이지안 대체 방법이 다른 대체 방법에 비해 무응답 비율이 높아지더라도 편의도 작으며 평균에 관한 95% 신뢰구간의 포함률도 높게 나타나서 가장 좋은 대체 방법으로 확인되었다.

주요용어: 무응답 대체, 패널자료, 선형혼합모형, 비대체, 한국노동패널

[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: jsong@korea.ac.kr