고령화연구패널조사 제1차 기본조사 데이터 무응답 대체방법 및 분석방법 소개

송주원 · 이혜정*

Ⅰ. 서 론

한국노동연구원은 급속하게 진행되고 있는 우리나라 고령화에 관한 정책 및 학술연구에 활용할 기초자료, 고령화연구패널조사(Korean Longitudinal Study of Ageing, 이하 KLoSA) 1차 기본조사 데이터를 2007년 3월에 베타(Beta)버전으로 고령연구패널조사 홈페이지(http://klosa.kli.re.kr)에서 배포하였다. 이후 제2차 데이터 클리닝, 무응답 보정, 새로운 변수 구성 등의 작업이 반영된 1.0버전 데이터를 11월 초 홈페이지에서 배포하였다. 또한 설문지, 코드북 및 사용자 안내서가 업데이트되어 있고, KLoSA와 관련된 여러 자료들을 보고 내려받을 수 있다.

여기서는 1.0버전에서 반영된 무응답 보정에 대하여 살펴볼 것이다. KLoSA는 대부분의 변수에서 결측값(Missing Value)의 비율이 5% 미만으로 작게 나타났으나 소득 및 자산영역의 일부 항목에서 결측값의 비율은 10~20% 내외까지 증가하였으며 일부 응답자가 많지 않은 항목의 경우 약 30% 정도까지 나타났다. 따라서 결측값을 포함한 변수에 대한 적절한 분석을 위하여 결측이 발생한 주요 항목에 대하여 다중대체(Multiple Imputation)가 실시되었는데, 특히 결측비율이 높은 소득 및 자산 항목의 대체에 중점을 두고 진행되었다. 각 변수별 결측비율이 거의 대부분의 변수에서 20% 미만으로 나타났으므로 Imputation의 수는 5개로 결정하였다. 이하에서는 적용된 무응답 대체방법에 대하여 설명하고 대체된 자료는 어떻게 구성되어 있는지 살펴보며 몇 가지 예를 통해 분석하는 방법을 설명하고자 한다.

* 송주원=고려대학교 통계학과 조교수(jsong@korea.ac.kr).
이혜정=한국노동연구원 고령자패널팀 연구원(snp625@kli.re.kr).

Ⅱ. 무응답 대체방법

1. 대체방법: 수정된 예측 평균에 근거한 핫덱방법

결측값을 대체하는 방법에는 여러 가지 종류가 있는데, KLoSA에서는 그 중 3가지 다중대체방법(Multiple Imputation Methods)에 대해 모의실험을 통하여 가장 좋은 결과를 보여주는 '수정된 예측 평균에 근거한 핫덱방법(Hotdeck based on a modified predictive mean matching)'을 최종 대체방법으로 결정하였다. 이는 Little(1988)이 제안한 일종의 핫덱 대체법(Hotdeck Imputation)으로서 미국 RAND의 Bell(1999)이 SAS Macro로 프로그램화하여 여러 가지 조사연구에 적용해 왔으며 우수한 결과를 보여 왔다. 이 방법은 결측이 발생한 자료값을 자료내 관찰된 값들 중 하나 또는 여러 개의 값을 가지고 대체시키는 일종의 핫덱 대체법이지만, 관찰값 중 하나 또는 여러 개의 값을 임으로 선택하는 랜덤핫덱(Random Hotdeck) 대신 자료를 비슷한 여러 개의 하위그룹(Subclass)으로 나누어 같은 하위그룹 내에서 핫덱 대체를 실시하는 것이다. 이 때 하위그룹은 결측이 발생한 변수에 대하여 관찰된 자료만을 대상으로 회귀모형(Regression Model)을 적합하여 결측이 포함된 모든 자료에 대한 예측값을 구한 후 예측값에 근거하여 층화(Stratification)를 하여 구성한다. 각 층 내에서 결측값은 같은 층의 관찰자 중에서 기증자(Donor)를 선택하여 기증자의 값으로 대체를 실시한다. 이 방법은 기증자를 선택하는데 있어서 임으로 한 명 또는 여러 명의 기증자를 선택하는 랜덤핫덱방법보다 회귀모형의 예측력이 클수록 좋은 결과를 기대할 수 있다.

KLoSA에서는 문항 및 응답의 특성에 따라 최적의 대체를 실시하기 위하여 그에 따라 대체방법이 적절히 변형되었으며 총 4가지 모형(모형 1(Hotdeck), 모형 2(Bounded Hotdeck), 모형 3(Hotdeck-GEE), 모형 4(Hotdeck & Regression)이 고려되었으며 내용은 다음과 같다.

가. 모형 1(Hotdeck)

이 모형은 원래 수정된 예측 평균에 근거한 핫덱모형이다. 결측이 발생한 각 변수에 대하여 관찰된 자료만을 대상으로 회귀모형을 적합한 다음 결측이 포함된 모든 자료에 대한 예측값을 구한 후 예측값에 근거하여 층화하는데 가능한 한 10명 이상의 구성원을 포함하도록 하위그룹을 구성한다. 각 하위그룹 내에서 각 결측값은 같은 하위그룹의 관찰자 중에서 기증자를 선택하고 기증자의 값으로 대체를 실시하였다.

나. 모형 2(Bounded Hotdeck)

이 모형은 모형 1의 확장된 모형으로 범주형 전환문항(Unfolding Bracket)에서 얻어진 정보를 포함하도록 확장한 핫덱모형이다. 서로 겹쳐지지 않는 5개의 범주형 전환문항이 시행되었으므로 각 응답값은 6개의 구간 중 하나의 구간으로 나타내 질 수 있다. 정확한 값 대신 범주형 전환문항에 응답한 경우 값이 어느 구간 안에 존재하는지에 관한 정보가 주어지므로 이 정보를 사용하여 대체가 실시되어야 한다. 예를 들면, 한 참여자의 임금소득에 대한 범주형 전환문항의 응답이 '2,400만 원 이상 6,000만 원 미만'이라면 이 응답자의 대체된 값은 이 구간 안에 속해야 제공된 정보와 일치하는 대체가 이루어지는 것이다. 따라서 우선 정확한 값이 응답된 참여자들의 관찰값들을 범주형 전환문항에서 선택한 6개의 구간으로 분리하였다. 각 구간 안에서 이 관찰값들은 범주형 전환문항에서 동일구간에 속하는 것으로 응답된 결측값에 대한 기증자의 후보집단(Pool)으로 사용된다. 각 구간별로 다수의 응답자가 있는 경우 회귀모형에 의한 예측값을 사용하여 하위 그룹을 나누고 동일한 하위그룹 내에서 결측값에 대한 기증자를 선택하여 기증자의 값을 가지고 대체를 실시하였다.

다. 모형 3(Hotdeck-GEE)

KLoSA에서는 문항 중 응답자가 루프(Loop) 숫자만큼 응답해야 하는 문항이 있다. 예를 들면, 자산영역에서 가지고 있는 보험의 개수를 물어보고 각 보험에 대해 가입한 날짜, 납부주기와 납부금액에 대한 문항들이 있다. 이와 같은 형태의 문항인 경우 이 모형을 적용하게 된다. 회귀모형을 적합할 때 각 변수값은 서로 독립이라 가정하지만, 한 응답자가 대답한 응답들은 서로 독립이 아니므로 모형 1을 적용할 수 없다. 그러므로 이 모형은 모형 1에서 회귀모형을 적합할 때 회귀모수가 변수값의 연관성을 포함하도록 GEE(Generalized Estimating Equation) 방법으로 모수를 추정하여 예측값을 계산하도록 변형하였다. 또한 범주형 전환문항을 포함한 문항은 모형 2에서와 같이 구간 정보를 이용하여 대체하였다.

라. 모형 4(Hotdeck & Regression)

결측값이 발생하였으나 동일구간 내에서 기증자가 존재하지 않는 경우에는 이 모형을 사용하여 대체한다. 이런 경우는 범주형 전환문항을 포함한 일부 문항의 극한구간(특히 금액이 높을수록)에서 발생한다. 예를 들면, 남자의 임금소득의 가장 상위구간으로

범주형 전환문항에 대하여 구분된 결측값은 존재하였으나 그 구간에 해당하는 관찰값은 존재하지 않으므로 모형 1을 적용할 수 없다. 그래서 모형 1을 약간 수정한 것으로 회귀모형에서 계산한 예측값을 결측값에 대체해 주었다.

모형 1부터 모형 4는 서로 독립적으로 적용하지 않고 혼합적으로 적용하였다. 예를 들면, 임금소득에 대한 응답이 구체적인 금액, 범주형 전환문항으로 응답하여 구간, 범주형 전환문항에서도 응답을 거절한 경우로 혼재되어 있다. 이 경우에는 범주형 전환문항에 응답한 사람의 임금소득은 모형 2를, 범주형 전환문항에 대한 응답조차 거부한 경우에는 모형 1을 이용하여 대체하였다.

2. 대체방법 순서

KLoSA는 CV. 커버스크린, A. 인구학적 배경, B. 가족, C. 건강, D. 고용, E. 소득, F. 자산과 G. 주관적 기대감 및 삶의 만족도 영역으로 총 8개로 구성되어 있다. 여러 영역 중에서 결측값의 대체는 다른 영역에 비해 결측값의 비율이 상대적으로 높은 소득과 자산영역이 중점적으로 시행되었으며, 대체할 때 필요한 몇 가지 관련변수들도 함께 대체한 후 사용하였다. 대체한 순서는 다음과 같다.

인구학적 배경영역 → 건강영역 → 고용영역 → 소득영역 → 자산영역 → 가족영역

기본 특성변수라 볼 수 있는 인구학적 배경영역에서 결측값을 포함하고 있는 응답자의 학력과 결혼상태를 결측값이 없는 나머지 인구학적 배경변수를 이용하여 5번 대체하였다. 이렇게 대체된 5개의 자료 각각에 대해 주요 변수 및 디자인 변수(Design Variables)들을 설명변수로 사용하여 건강영역의 주요 변수들을 대체하였다. 건강영역은 1% 미만의 비교적 낮은 결측비율을 보였다. 앞에서 대체한 변수들 및 관련변수를 설명변수로 이용하여 고용영역의 현재 고용 및 퇴직소득에 관한 대체를 하였고, 다음으로 소득영역의 주요 변수들을 대체하였다. 이 과정에서 자산영역의 자가소유 여부 및 금융자산 총액도 설명변수에 포함하여 소득과 자산의 연관성을 설명할 수 있도록 반영하였다. 대체된 각 개인당 총소득을 계산한 후 다른 관련변수들과 함께 설명변수로 설정하여 자산영역의 주요 변수들도 대체하였다. 마지막으로 자녀의 수, 자녀의 기본정보, 금전적인 지원(정기적/비정기적)에 대해 대체가 실시되었다.

Ⅲ. 자료구조 설명

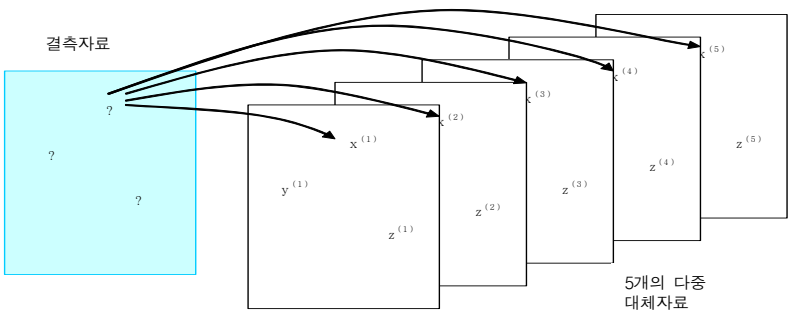
1. 대체된 자료의 형태

KLoSA에서는 각 결측값에 대하여 5개의 값을 대체하는 다중대체(Multiple Imputation)를 사용하였기 때문에 대체된 데이터는 5개가 생성된다. 5개의 대체된 데이터에서 관찰값(즉, 응답자가 실제로 응답한 값)은 모두 동일하지만 대체된 결측값들은 데이터마다 값이 같을 수도 있고 다를 수도 있다. [그림 1]에서 나타나는 것처럼 결측이 있는 데이터에 대해 5번 결측값을 대체하게 되면 결측값에 대해 5개의 값을 가지게 되는 것이다.

5개의 대체된 데이터에는 대체된 변수들만 포함되어 있어서 연구자가 사용할 때에는 원자료(A. 인구학적 배경영역부터 G. 주관적 기대감 및 삶의 만족도 영역까지, 데이터명: w01_v10k)와 결합하여 사용해야 한다. 대체된 변수명과 원자료에 있는 변수명이 동일하므로 데이터를 덮어씌운 다음 분석을 수행하면 된다. 대체된 데이터는 w01_i1_v10k부터 w01_i5_v10k까지 총 5개이며 대체된 자료의 번호는 데이터명의 중간에 있는 i 뒤의 숫자로 구분하면 된다. 예를 들면, SAS Macro를 이용하여 원자료와 대체된 데이터를 합치는 방법은 <표 1>과 같다.

합쳐진 데이터를 가지고 연구자는 원하는 분석을 각각 반복적으로 시행하면 된다. 각 자료에 대해 독립적으로 분석이 시행된 후 분석결과는 일반적으로 5개의 통계량 및 관련 분산・표준오차로 나타내는데 각각 다른 5개의 통계량이 아닌 하나의 통합된 통계량을 구해야 한다. 5개의 통계량을 하나의 값으로 통합하는 방법은 뒤에서 자세하게 소개하도록 하겠다.

〔그림 1〕 결측자료에 대하여 5개의 다중대체를 실시한 경우의 예



〈표 1〉 SAS Macro를 이용하여 5개의 대체된 데이터를 하나의 데이터셋으로 만들어 주는 프로그램의 예

```
%MACRO fulldata;
  %DO j=1 %TO 5;
      DATA w01_i&j._v10k;
      MERGE w01_v10k w01_i&j._v10k;
      BY pid;
   RUN;
  %END;
%MEND;

%fulldata;
```

2. 대체된 결측값의 구분방법

연구자가 합쳐진 데이터를 사용하여 분석할 경우 어느 값이 실제 값이고 대체된 값인지를 구분할 수 있다면 유용할 것이다. 또한 합쳐진 데이터를 이용하여 실제값만을 가지고 분석하는 것이 가능할 것이고, 대체된 값들이 실제값들과 비슷한지에 대한 분석도 가능할 수 있게 된다. 그래서 이런 정보를 알려주는 변수, 깃발변수(Flag Variable)를 데이터에 추가적으로 포함해 주었다. 깃발변수의 이름은 대체된 원래 변수명의 마지막에 _(Underbar)를 붙여 주었다. 예를 들면, 인구학적 배경영역에서 학력에 대한 문항인 w01A003에 대한 깃발변수명은 w01A003_이 된다.

<표 2>를 보면 깃발변수의 값과 각 값에 대한 의미를 보여준다. 깃발변수의 값은 0부터 3까지이며 '0'이면 해당 문항의 관찰값이 응답에 의하여 관찰된 값이라는 의미이며, '1'이면 관찰값이 결측되었으나 수정된 예측 평균에 근거한 핫덱방법으로 대체된 것을 의미한다. 그리고 '2'는 범주형 전환문항을 포함한 소득영역 및 자산영역의 변수에서 응답을 구간으로 하지 않고 대략적인 값으로 응답한 경우 그 값으로 대체되었음을 의미한다. '3'이면 가족대표자의 응답을 가지고 대체되었다는 의미이다. 일부 깃발변수에서 보이는 결측값은 이 문항이 앞의 문항에 부속되어 있고 앞 문항의 응답 때문에 이 문항이 응답되지 않았다는 것을 의미한다. 예를 들면, 월평균 임금소득액은 w01E001에서 임금

〈표 2〉 깃발변수의 변수값 및 변수설명

변수값	변수설명
0	응답자가 대답한 관찰값
1	관찰값이 모형을 통해 대체된 경우
2	범주형 전환문항에 구간이 아닌 값으로 응답한 경우
3	가족대표자의 응답을 가지고 대체한 경우
.	이 문항에 대한 응답대상자가 아닌 경우

소득이 있다고 응답한 연구대상자에게만 질문되었으므로 w01E001에서 임금소득이 없다고 응답한 경우 이 값은 결측값을 갖는다.

Ⅳ. 분석방법

1. 다중대체된 자료의 분석

다중대체된 자료의 경우 결측값이 없이 대체된 한 개 이상의 자료가 제공되며 이에 따른 분석은 다중대체된 각 자료의 분석 및 분석된 자료를 통합한 결과 도출의 두 단계로 나누어지게 된다.

다중대체된 자료 각각은 결측값이 대체되어 결측값이 없는 완전한 자료 형태를 가지고 있으므로 자료 각각에 대하여 연구목적에 알맞은 분석을 시행하면 된다. 예를 들어, 회귀분석(Regression Analysis)을 시행하고자 한다면 동일 관심변수에 대하여 동일 설명변수를 가지고 5개 자료 각각에 대하여 회귀분석을 실시하면 된다. 이렇게 분석을 실시하는 경우 추정된 회귀계수(Regression Coefficients), 표준오차(Standard Errors), 그리고 검정통계량(Test Statistics)은 5개 자료 각각으로부터 약간씩 다르게 나타나는데 이는 관심변수가 결측이 되어 참값을 알지 못하는 불확실성에 근거한 차이를 나타내는 것이다. 하지만 연구자의 분석 목적은 관심자료에 대한 5개의 서로 다른 결론이 아니라 한 개의 통합된 결론을 내리는 것이므로 5개 분석의 결과를 통합하여 한 개의 결론을 도출하기 위하여 아래의 통합과정을 거쳐야 한다.

2. 분석된 자료를 통합한 결과 도출

Multiple Imputation을 m번 시행한 자료 각각에 대하여 분석을 시행한 후 얻어진 모수의 추정값들을 $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_m$ 이라 하자. 또한 이 모수의 추정된 분산을 각각 W_1, W_2,…, W_m이라 가정하자. 예를 들어, 회귀분석을 실시하면 i번째 자료에 근거한 회귀분석에서 관심 설명변수의 회귀계수의 추정값이 $\hat{\theta}_i$ 이 되고 그 회귀계수의 표준오차의 추정값의 제곱이 W_i가 된다. 이 경우 통합된 모수의 추정값은

$$\bar{\theta}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

으로 표현될 수 있다. 즉, 추정된 모수들의 평균값이 통합된 모수의 추정값이 된다.

통합된 모수의 분산의 추정값은 다음 두 개의 분산 성분의 합으로 표현된다. 첫 번째 분산 성분은

$$\overline{W}_m = \frac{1}{m}\sum_{i=1}^{m} W_i$$

로서 각 모수의 추정된 분산들의 평균이다. 이 분산 성분은 대체내 분산(within imputation variance)으로 부른다. 두 번째 분산 성분은

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \overline{\theta}_m\right)^2$$

으로 표현되는데 이는 각 대체된 자료의 모수의 추정값들 사이의 분산을 나타내므로 대체간 분산(between imputation variance)이라 부른다. 통합된 모수의 분산의 추정값은

$$T_m = \overline{W}_m + \frac{m+1}{m} B_m$$

으로 구할 수 있다.

자료가 충분히 큰 경우, 이 모수에 대한 분포는 다음의 t-분포를 따른다.

$$\left(\theta - \overline{\theta}_m\right) T_m^{-1/2} \sim t_\nu ,$$

여기서 t-분포의 자유도 ν 는 $\nu = \left(\nu_0^{-1} + \hat{\nu}_{\text{obs}}^{-1}\right)^{-1}$ 로 계산되는데 ν_0 는

$\nu_0 = (m-1)\left(1 + \frac{1}{m+1}\frac{\overline{W}_m}{B_m}\right)^2$ 을 나타내고 $\hat{\nu}_{\text{obs}}$ 에서 ν_{com} 이 결측값이 없을 때 모수 θ에 대한 추정의 자유도를 나타낼 때 $\hat{\nu}_{\text{obs}} = (1-\hat{\gamma}_D)\left(\frac{\nu_{\text{com}}+1}{\nu_{\text{com}}+3}\right)\nu_{\text{com}}$ 을 나타낸다.

또한 여기서 $\hat{\gamma}_m$ 은 $\hat{\gamma}_m = (1+1/m)B_m/T_m$ 으로서 결측에 의하여 손실된 모수 θ에 대한 정보의 부분(fraction of information about θ missing due to nonresponse)이라 불린다. 모수의 분포가 t-분포를 따르므로 t-분포에 근거한 검정을 시행하거나 모수의 신뢰구간을 구할 수 있다. 또한 이 통합방법은 관심모수들에 대한 다변량 검정 및 신뢰구간의 계산 등으로의 확장도 가능하다(Rubin, 1987).

3. 예 제

Multiple Imputation을 시행하여 만들어진 m개의 자료들에 근거한 m개의 분석결과를 통합하는 과정은 연구자들이 직접 프로그램화하여 시행하기에 어려울 수 있으므로 현재 여러 가지 통계프로그램에서는 이 결과를 통합하는 Procedure를 제공하고 있다. 예를 들어, SAS의 PROC MIANALYZE는 위와 같이 분석된 자료의 모수들을 통합한 결과를 제공해 준다. 그 외에 무료 통계프로그램인 R도 다중대체된 자료를 분석한 후 통합하는 함수를 제공하고 있다. 또한 Schafer(1997)가 개발한 NORM은 통계프로그램이 필요 없이 독립적으로 시행되는 작은 크기의 프로그램으로서 위의 단계를 수행하고 통합된 결과를 제공하고 있는데 이 프로그램은 http://www.stat.psu.edu/~jls/misoftwa.html에서 무료로 다운받을 수 있다.

다음은 SAS에서 Multiple Imputation으로 대체된 여러 개의 자료를 이용한 단순평균계산, 층화평균계산 및 회귀분석계수의 계산에 대한 예제를 설명하고자 한다.

우선 SAS에서는 여러 개의 자료에 대하여 동일한 모형을 가지고 분석을 실시하고자 하는 경우에 여러 개의 자료를 한 개의 자료로 통합한 후 통합된 자료에 대하여 한 개의 Procedure를 이용하여 자료별 분석을 시행하는 것이 가능하다. 이를 위하여 5개의 대체된 자료를 한 개의 자료로 통합하고 각 대체된 자료를 나타내는 변수를 가지고 각 자료를 구분하면 된다(앞의 자료구조 설명에서 각 대체된 자료를 원자료와 결합하는 프로그램의 예를 설명함). 제공된 대체된 자료는 몇 번째로 대체된 자료인지 나타내는 구별변수인 w01imputation_을 가지고 있으므로 이 변수별로 분석을 시행하면 된다. 이를 위한 SAS 프로그램은 <표 3>과 같다.

〈표 3〉 5개의 대체된 데이터를 하나의 데이터로 통합하는 SAS 프로그램의 예

```
DATA total;
   SET w01_i1_v10k  w01_i2_v10k  w01_i3_v10k  w01_i4_v10k  w01_i5_v10k;
   _imputation_=w01imputation_;
RUN;
```

여기서 새롭게 생성된 변수 _imputation_은 w01imputation_과 동일한 변수로서 각각의 자료를 분석한 후 SAS의 PROC MIANALYZE를 이용하여 자료를 통합하는 과정에 사용하기 위하여 생성되었다.

〈예제 1〉 금융자산의 단순평균계산

〈표 4〉 SAS 프로그램의 예

```
* 각 대체 자료별 단순평균계산;
PROC SURVEYMEANS DATA=total;
    VAR w01F085;
    BY _imputation_;
    ODS OUTPUT STATISTICS=stat1;
RUN;

* 각 대체된 자료별로 계산된 단순평균을 통합하여 원자료의 단순평균 추정;
PROC MIANALYZE DATA=stat1;
    MODELEFFECTS mean;
    STDERR stderr;
RUN;
```

SAS Procedure SURVEYMEANS에서 BY 변수를 사용하여 각 대체된 자료별로 금융자산의 단순평균 및 표준오차를 계산한 후 이들 통계량들을 자료명 stat1에 저장하였다. 이 저장된 통계량들은 Procedure MIANALYZE를 사용하여 통합되었다. 이 때 MODELEFFECTS 문에는 통합할 통계량 $\hat{\theta}_i$ (여기서는 평균)을 나타내는 변수 mean을 써주고 STDERR문에는 W_m의 제곱근인 평균의 표준오차를 나타내는 stderr 변수를 써주면 된다. Procedure MIANALYZE는 <표 5>와 같은 결과를 제공한다.

이 표에서 보는 바와 같이 금융자산의 단순평균은 1023.78만 원으로 나타나고 표준편차는 42.46이다. 금융자산의 평균에 대한 95% 신뢰구간은 (940.49, 1107.06)으로 계산되어지며 금융자산이 0이라는 귀무가설은 t-통계량이 24.11, p-value가 <.0001로 5% 유의수준하에서 유의하게 나타난다.

〈표 5〉 SAS 결과

The MIANALYZE Procedure

Model Information

Data Set	WORK.STAT
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF
mean	75.883499	1711.837134	1802.897333	1568

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mean	0.053194	0.051716	0.989763

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
mean	1023.775585	42.460539	940.4902	1107.061	1568

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	ThetaO	t for H0: Parameter=ThetaO	Pr > |t|
mean	1012.992767	1033.328748	0	24.11	<.0001

〈예제 2〉 금융자산의 표본설계 가중치를 이용한 평균계산

이계오 외(2007)에 설명된 표본설계 가중치를 이용한 금융자산의 평균 추정치를 계산해 보았다.

〈표 6〉 SAS 프로그램의 예

```
* 각 대체 자료별 표본설계 가중치를 이용한 평균계산;
PROC SURVEYMEANS DATA=total;
        STRATA w01region1 w01region2 w01enu_type;
        CLUSTER w01enu;
        VAR w01F085;
        WEIGHT w01wgt;
        BY _imputation_;
        ODS OUTPUT STATISTICS=stat2;
RUN;

* 각 대체된 자료별로 계산된 표본설계 가중치를 이용한 평균을 통합하여 원자료의 층화평균 추정;
PROC MIANALYZE DATA=stat2;
   MODELEFFECTS mean;
   STDERR stderr;
RUN;
```

SAS Procedure SURVEYMEANS에서 BY 변수를 사용하여 각 대체된 자료별로 금융자산의 표본설계 가중치를 이용한 평균 및 표준오차를 계산한 후 이들 통계량들을 자료명 stat2에 저장하였다. 이 저장된 통계량들은 Procedure MIANALYZE를 사용하여 통합되었다. 이 때 MODELEFFECTS문에는 통합할 통계량 $\hat{\theta}_i$ (여기서는 표본설계 가중치를 이용한 평균)을 나타내는 변수 mean을 써주고 STDERR문에는 W_m의 제곱근인 평균의 표준오차를 나타내는 stderr 변수를 써주면 된다. Procedure MIANALYZE는 <표 7>과 같은 결과를 제공한다.

이 표에서 보는 바와 같이 금융자산의 표본설계 가중치를 이용한 평균은 1044.27만원으로 나타나고 표준편차는 59.73이다. 금융자산의 표본설계 가중치를 이용한 평균에 대한 95% 신뢰구간은 (927.19, 1161.34)로 계산되어지며 금융자산이 0이라는 귀무가설은 t-통계량이 17.48, p-value가 <.0001로 5% 유의수준하에서 유의하게 나타난다.

〈표 7〉 SAS 결과

The MIANALYZE Procedure

Model Information

Data Set	WORK.STAT2
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance: Between	Within	Total	DF
mean	38.539587	3521.403652	3567.651157	23804

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mean	0.013133	0.013046	0.997398

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
mean	1044.266383	59.729818	927.1921	1161.341	23804

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	ThetaO	t for HO: Parameter=ThetaO	Pr > |t|
mean	1035.807531	1050.103901	0	17.48	<.0001

〈예제 3〉 금융자산에 대한 회귀분석

금융자산과 성별, 연령의 관계를 나타내는 회귀모형을 적합한 분석을 시행하였다.

〈표 8〉 SAS 프로그램의 예

```
* 각 자료별 회귀분석 실시;
PROC REG DATA=total OUTEST=outreg COVOUT;
    MODEL w01F085 = w01gender1 w01a001_age;
    BY _imputation_;
RUN;

* 각 자료별 회귀분석 결과의 통합;
PROC MIANALYZE DATA=outreg;
    MODELEFFECTS Intercept w01gender1 w01a001_age;
RUN;
```

SAS Procedure REG에서 BY 변수를 사용하여 각 대체된 자료별로 회귀분석을 실시한 후 OUTEST문을 사용하여 자료명 outreg에 회귀계수 및 회귀계수의 표준오차 등을 저장한다. 여기에 저장된 통계량들을 Procedure MIANALYZE에서 통합하여 준다. 이 때 통합하고자 하는 통계량은 절편(Intercept) 및 나이, 성별을 나타내는 두 변수의 계수, 즉 세 개의 회귀모형 모수가 되며 이를 MODELEFFECTS문에 나타내준다. 여기서 Intercept는 변수명이 아니고 회귀모형의 절편을 의미한다.

<표 9>에서 보는 바와 같이 회귀모형의 절편(Intercept)은 2127.83, 성별(w01gender1)과 나이(w01A001_age)의 회귀계수는 각각 -103.22와 -13.24로 나타나며, 절편을 포함한 세 회귀모수의 표준오차는 각각 260.81, 21.51, 4.18로 추정된다. 절편 및 각 변수의 회귀계

〈표 9〉 SAS 결과

The MIANALYZE Procedure

Model Information

Data Set	WORK.OUTREG
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance: Between	Within	Total	DF
Intercept	890.0930253	66955	68024	16224
w01gender1	29.573442	427.272480	462.760610	680.15
w01a001_age	0.538980	16.850509	17.497264	2927.5

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
Intercept	0.015953	0.015823	0.996845
w01gender1	0.083057	0.079391	0.984370
w01a001_age	0.038383	0.037622	0.992532

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
Intercept	2127.827592	260.813283	1616.605	2639.050	16224
w01gender1	-103.215014	21.511871	-145.453	-60.977	680.15
w01a001_age	-13.240722	4.182976	-21.443	-5.039	2927.5

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	t for H0: Theta0	Parameter=Theta0	Pr > |t|
Intercept	2081.400933	2164.672953	0	8.16	<.0001
w01gender1	-111.891438	-97.829171	0	-4.80	<.0001
w01a001_age	-14.074498	-12.248854	0	-3.17	0.0016

수가 0인가를 검정하는 t-통계량은 각각 8.16, -4.80, -3.17로서 모두 5% 유의수준하에서 통계적으로 유의하게 나타난다.

Ⅴ. 맺음말

이상으로 KLoSA에 적용된 무응답 보정방법에 대하여 살펴보았다. 결측을 포함하고 있는 데이터를 사용하여 연구할 경우 결측으로 인해 누락되는 부분이 발생할 수 있다. 그리고 단순대체(Single Imputation)가 된 데이터를 사용할 경우에는 결측값이 알려지지 않았다는 불확실성을 반영하지 못하며 추정량의 표준오차를 과소추정하는 문제를 가지고 있다. 그러나 KLoSA의 다중대체된 데이터는 결측값에 대한 불확실성을 적절하게 반영하기 때문에 효과적인 통계적 추론을 할 수 있다.

다중대체방법에 대한 보다 구체적인 내용은 고령화연구패널조사 홈페이지(http://klosa.kli.re.kr)에서 내려받아 볼 수 있다(데이터 및 보고서와 사용자 안내서가 제공되어 있음). 향후 KLoSA 제1차 기본조사 데이터는 사용자의 편의성을 개선하기 위하여 다중대체된 데이터를 가지고 소득영역과 자산영역의 변수들을 서로 조합함으로써 임금소득, 가구소득, 금융자산, 부동산 자산 등의 '생성변수'를 추가하여 배포될 것이다. KLI

<참고문헌>

송주원 외(2007),「고령화연구패널조사 1차 조사 자료에서 발생하는 결측값에 대한 Multiple Imputation－소득 및 자산 관련 변수들의 대체 모형」, http://klosa. kli.re.kr

송주원 외(2007),「고령화연구패널조사 1차 조사 자료에서 발생하는 결측값에 대한 Multiple Imputation－대체된 자료의 분석」, http://klosa.kli.re.kr

신현구·부가청·이혜정(2006),「고령화연구패널조사 제1차 기본조사 소개」,『노동리뷰』9월호, 한국노동연구원.

이계오·김영원·장지연(2007),「고령화연구패널조사의 표본설계 연구」, 한국통계학회 2007년 춘계학술대회 발표문

Little R. J. A.(1988), "Missing data adjustments in large surveys," *Journal of Business and Economic Statistics* 6, pp.287~301.

Bell R.(1999), "Depression PORT Methods Workshop (I)", RAND: Santa Monica, CA.

Rubin D. B.(1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley: New York.

Schafer, J. L.(1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London, UK.