일반논문 「행정논총」 제54권 제4호(2016. 12): 291~319

설문자료의 결측치 처리방법에 관한 연구:
다중대체법과 재조사법을 중심으로

고길곤*
탁현우**

〈目 次〉

Ⅰ. 서 론
Ⅱ. 이론적 배경
Ⅲ. 연구설계
Ⅳ. 분석결과
Ⅴ. 결 론

〈요 약〉

결측치에 대한 고민 없이 분석에서 제외하는(완전제거법) 경우, 기술통계 뿐만 아니라 상관관계나 회귀계수 같은 변수 간 관계에 대한 분석결과에도 영향을 미친다는 사실이 알려져 있다. 하지만 재조사법과 다중대체법과 같은 통계적 방법 등 '결측치를 어떻게 처리할 것인가'에 대한 연구에 비해 이들의 실질적인 활용방법에 관한 연구는 부족한 실정이다.

본 논문은 재조사법과 통계적 대체방법의 장점을 결합한 분석방법을 제시하였다. 단순히 재조사법과 다중대체법 간의 상대적 우월성 평가에 맞추기보다는, 이들을 적절히 조화한 비용효과적인 결측치 보정 가능성을 제시하였다. 또한, 그 성능을 평가하기 위해 가상데이터에 대한 시뮬레이션 자료와 실제 설문자료를 이용하여 분석 결과의 타당성을 교차 검증하였다.

분석 결과 시뮬레이션과 실제 데이터를 이용한 분석 모두에서 결측 메커니즘이나 변수 간 관계 등과 상관없이 결측률이 약 30% 이하인 경우 통계적 다중대체법이 통계적 편의를 유발하지 않는 것으로 나타났다. 또한, 특정 변수의 결측된 원인을 설명할 수 있는 변수인 종속변수나 분석모형에 포함되지 않더라도 결측된 변수를 설명할 수 있는 관측된 변수(보조변수)를 대체모형에 포함하는 경우, 통계적 대체방법의 성능은 더욱 향상되는 것으로 나타났다. 따라서, 결측률이 높은 자료에서도 재조사법을 통해 결측률을 일정 수준 이하로 낮추는 경우 통계적 대체방법을 통해 분석결과의 타당성을 높일 수 있음을 확인하였다.

【주제어: 결측치, 설문조사, 재조사법, 다중대체법】

* 주저자, 서울대학교 교수, 한국행정연구소 겸무연구원(kilkon@gmail.com)
** 교신저자, 국회예산정책처 사업평가관(takhw@nabo.go.kr)
논문접수일(2016.7.16), 수정일(2016.10.24), 게재확정일(2016.11.2)

Ⅰ. 서 론

결측치(missing value)를 처리하는 최적의 해법은 알려져 있지 않기 때문에(Berglund & Heeringa, 2014), 이를 어떻게 처리할지 결정하기는 쉽지 않다(Burgess 외, 2013). 그럼에도 불구하고 결측치로 인한 정보의 손실이 분석결과에 미치는 영향을 고려하면, 결측치를 적절하게 처리하는 것은 여전히 중요하다. 결측치는 발생 메커니즘과 결측의 비율에 따라 분석모형에서 제3의 요인으로 작용하게 되며, 분석결과의 편의(bias)나 통계적 검정력(statistical power)의 문제를 일으키게 된다(Rubin, 1987:1). 특히, 결측으로 인해 체계적인 정보의 손실이 있는 경우 그 문제는 더 심각해지는데, 이는 연구결과가 모집단의 특성을 대표하지 못하는 자료에 의존하게 되기 때문이다(McKnight 외, 2007:6). 또한, 어떤 처리방법을 사용하느냐에 따라 분석결과에 미치는 영향은 긍정적일 수도 혹은 부정적일 수도 있으므로[1], 연구결과의 타당도와 신뢰도를 높이기 위해서는 결측치에 대한 이해와 처리방법에 대한 논의가 필요하다.

고길곤 외(2014)는 실제 설문자료를 분석하여 결측치를 완전히 제거하고 분석하는 경우에 평균 및 분산과 같은 기술통계치에 문제가 발생할 뿐 아니라 변수 간 관계를 분석하는 회귀분석의 결과에도 영향을 미친다는 것을 확인하였다. 다만, 기존에 주로 사용되고 있는 완전제거법 대신 다중대체법을 사용하는 경우 재조사를 통해 결측을 대체하는 것과 유사한 결과를 얻을 수 있다고 주장한다. 다중대체법은 보건의료통계나 역학(epidemiology), 인구센서스(population census) 등의 분야에서 활발히 연구가 진행되고 있으며(Burgette & Reiter, 2010; 김서영·박라나, 2013; 한혜은·김영원, 2015), 높은 통계의 품질을 확보하기 위해서 조사비용의 증가와 시간의 제약 조건에도 불구하고 조사과정에서의 무응답 축소를 위한 노력의 중요성이 강조되고 있다(김서영·박라나, 2013; 한혜은·변종석, 2014).

본 논문은 재조사법과 통계적 대체방법의 장점을 결합한 비용 효과적인 분석방법을 탐색하고자 한다. 연구의 초점을 단순히 재조사와 다중대체 간의 상대적 우월성 평가에 맞추기보다는, 양자를 적절히 조화하여 결측치를 보정하는 방법의 가능성을 제시하는데 맞추고자 한다. 이를 위해, 가상데이터에 대한 시뮬레이션 분석과 실제 설문자료를 활용한 분석을 함께 살펴본 후 분석결과의 타당성을 교차 검증하였다. 마지막으로 분석결과를 바탕으로 적절한 재조사 수준은 어느 정도인지에 대한 문제와 다중대체의 대체모형

1) 결측치 처리방법 중 평균대체나 단일대체법 등은 결측 메커니즘에 따라 오히려 분석결과의 왜곡을 유발할 수도 있다(Baraldi & Enders, 2010).

에 포함할 관측된 변수의 활용방법 등을 함께 논의하였다.

Ⅱ. 이론적 배경

1. 결측치 발생의 메커니즘

일반적으로 데이터에서 특정 정보가 누락된 것을 결측치(missing data)라고 하며(McKnight 외, 2007), 특히 설문조사에서 발생하는 경우 무응답(non-response)이라고 한다. 무응답은 크게 항목 무응답(item non-response)과 단위 무응답(unit non-response)으로 나뉜다(Graham, 2012:4-5). 항목 무응답은 응답자가 설문에 참여는 하였으나 일부 질문에는 응답하지 않는 경우를 의미하며, 응답대상자가 설문 자체에 참여하지 않는 것을 단위 무응답이라고 한다. 또한, 단위 무응답의 특수한 경우로 패널조사와 같이 동일 응답자를 대상으로 반복 측정하는 설문에 참여했다가 일정 시점 이후 탈락하는(drop out) 경우를 의미하는 웨이브 무응답(wave non-response)이 있다(Schafer & Graham, 2002). 자료의 결측이 발생하는 원인은 주로 참여자, 연구설계, 그리고 참여자와 연구설계의 상호작용 등으로 설명되는데(McKnight 외, 2007:54-57), 예를 들면, 어떤 참여자가 특정 질문에 불쾌감을 느끼거나(참여자 요인), 어떤 설문이 너무 많은 참여자의 시간을 요구하거나(설계 요인), 특정 성향이 있는 사람이 답변하지 못하는 질문을 포함하는 경우(참여자와 설계의 상호작용 요인) 등이 이에 해당한다. 이외에도 부적절한 측정도구, 자료수집 과정의 영향, 데이터 관리의 부실 등의 요인이 존재한다.

결측치의 발생원인에 관하여 변수 간의 관계의 관점에서 구조적으로 이해하는 결측발생 메커니즘에 대한 논의가 보편적이다(Little & Rubin, 2014). 결측 발생 메커니즘은 결측치의 처리방법이 작동할 수 있는 전제조건이 되며, 그 성능과도 밀접한 관계를 갖고 있다(Baraldi & Enders, 2010).

Little & Rubin(2014)은 결측여부를 나타내는 매트릭스 R[2]과 결측된 변수의 관계를 의미하는 ϕ를 이용하여 결측 메커니즘을 3가지로 분류하고 있다. 첫 번째는 완전임의결측(missing completely at random, MCAR)이다. MCAR은 어떤 변수의 결측여부가 순전

2) 어떤 변수에 대해서 특정 응답자의 응답값이 관측된 경우는 1, 결측된 경우를 0의 값을 가지는 벡터를 생각해 보자. 이와 같은 방식으로 설문자료에 포함된 변수들의 결측 여부에 관한 정보를 포함하는 매트릭스를 의미한다.

히 확률적으로 결정되는 경우를 말한다. 즉, 동전 던지기와 같은 확률과정(random process)에 의해 결측이 정해지는 것으로 결측 매트릭스(R)와 결측된 변수의 관측된 값(Y_{obs}) 혹은 관찰되지 않은 결측값(Y_{mis}) 간에 아무런 상관관계가 없는 경우에 해당한다. 응답자가 특정 문항을 실수로 건너뛰고 응답하거나, 우연히 특정 문항의 응답이 누락된 경우에 해당한다. MCAR은 결측이 체계적 오차에 의한 것이 아니므로 결측된 케이스를 제외한 데이터셋은 완전한 데이터로부터 임의추출한 표본집단의 부분집합으로 이해할 수 있으나, 이러한 가정이 실제 자료에서 만족하기는 힘들다(Baraldi & Enders, 2010).

두 번째는 Y_{obs} 혹은 이와 관련된 특성에 의하여 체계적인 결측이 발생하는 경우로 임의결측(missing at random, MAR)이라고 한다. 이 경우 R의 값이 결정되는 확률은 오직 Y_{obs}에 의해 설명되는 것이다. 하지만, 이때 R과 결측된 변수, 특히 결측된 값[3] 간에는 아무런 관계가 없어야 한다. 나아가 Y_{mis}를 설명하는 잠재변수(underlying values)에 의해 결측이 설명되어서도 안된다(Baraldi & Enders, 2010). 하지만 Y_{mis}는 관측되지 않았기 때문에 MAR 가정을 만족하는지를 검증할 수도 없다는 한계가 있다(Enders, 2010:13).

마지막으로 R이 Y_{mis}와 관계되어 있는 비임의결측(not missing at random, NMAR)이 있다. NMAR은 결측이 결측된 변수 자체의 요인에 영향을 받는다고도 설명된다. 즉, R이 Y_{mis} 또는 Yobs와 관계를 가지는 경우를 의미하는데, 중요한 것은 Y_{mis}와의 관계가 존재해야 하며, 이는 NMAR을 MAR과 구분하는 기준이 된다. 하지만, 실제 자료에서 결측된 자료는 관찰할 수 없으므로 R과 Y_{mis}가 관계가 있는지를 확인하는 것은 불가능하다. 결국, R과 Y의 관계 유무에 따라 MCAR과 나머지 메커니즘(MAR, NMAR)은 구별이 가능할 수 있으나, MAR과 NMAR은 구분이 어렵다(McKnight 외, 2007:45-50).

MCAR의 경우는 결측을 무작위 오차로 가정하면 표본 크기의 문제를 제외하면 큰 문제가 되지 않기 때문에 본 논문에서는 NMAR과 MAR, 그리고 이들이 혼합된 형태를 가정하여 이들이 결측치 처리방법의 성능에 미치는 영향을 살펴보았다.

2. 결측 메커니즘과 처리방법

앞서 살펴본 결측의 발생 메커니즘은 결측치의 처리방법과 밀접한 관계가 있다. 완전제거법과 같은 전통적인 결측치 대체방법은 MCAR 가정하에서 활용 가능한 방법인 데 비해, 최근의 다중대체법은 대부분 보다 완화된 가정인 MAR에서 불편추정량을 얻을 수 있는 방법으로 알려져 있다(Baraldi & Enders, 2010)[4]. 결측치 처리방법은 크게 전통적

3) 결측된 변수의 결측된 값은 관측되지 않았기 때문에 실제로는 알 수 없다.

인 방법과 다중대체법으로 나뉠 수 있는데, 이하에서는 결측의 메커니즘과의 관계와 그 성능의 관점에서 결측치 처리방법을 서술하고자 한다[5].

우선, 전통적인 대체방법으로는 완전제거법, 한쌍제거법(pairwise deletion), 평균대체법(mean substitution), 회귀대체법(regression imputation), 확률적 회귀대체(stochastic imputation) 등이 있다. 완전제거법이나 한쌍제거법[6]은 대부분의 통계 패키지 소프트웨어에서 기본옵션(default)으로 활용되는 결측치 처리방법으로 결측을 포함하는 응답자를 분석에서 제외하고, 남아 있는 관측치에 대해 통계분석을 시행한다. 이들 방법은 결측 메커니즘이 MCAR인 경우에 적용가능한데, MAR이나 NMAR인 경우 모수의 추정량에 편의가 발생한다. 또한, MCAR인 경우라고 할지라도 불완전 관측치가 분석에서 제거되면서 분석에 사용되는 표본의 수가 줄어들게 되므로 통계적 검정력이 감소하게 된다 (Baraldi & Enders, 2010).

평균대체와 회귀대체법 등은 단일대체법(single imputation)으로 분류되는데, 각각의 결측치를 일정한 과정을 통해 생성된 하나의 값으로 대체한다. 대체를 통해 완전한 데이터셋을 구성할 수 있다는 점에서 흥미로운 방법이지만, 대부분의 단일대체 값이 편의추정량을 발생시키는 것으로 알려져 있다. 추정량의 편의뿐만 아니라 통계적 검정력의 관점에서도 문제가 되는데, 결측값을 확정된 하나의 값으로 대체하기 때문에 표준오차(standard error)가 과소추정 되기 때문이다(Baraldi & Enders, 2010). 한편, 평균대체는 결측된 변수의 관측된 값들의 평균으로 결측치를 대체하는 방법으로 해당 변수의 분산을 작게 하는 문제와 함께 다른 변수와의 상관관계를 낮추는 등 분석결과의 편의를 가져오는 것으로 알려져 있다(Baraldi & Enders, 2010).

회귀대체의 경우 결측된 변수의 관측된 값을 종속변수로 하고, 나머지 변수를 설명변수로 하여 추정한 회귀식을 활용하여 결측된 변수의 결측치를 추정하여 대체한다. 평균대체와 달리 설명변수의 조건부평균으로 결측을 대체하기 때문에 더욱 발전된 방법으로 생각되지만, 이 경우에도 단일대체가 가지는 한계를 그대로 가진다. 이와 같이 단일대체법이 결측된 변수의 분산을 과소추정하는 문제를 줄이기 위하여, 확률적 회귀대체법

4) 하지만, NMAR 가정 하의 결측 데이터를 제대로 처리할 방법은 없으며, 결측치 처리방법이 널리 활용되기 위해서 지속적인 연구가 필요한 부분이다.

5) 보다 자세한 결측치 발생 메커니즘과 처리방법에 대한 논의는 '고길곤, 탁현우, 이보라. (2014). 설문조사 연구에서 결측치의 영향과 대체방법의 적절성에 대한 실증연구. 정책분석평가학회보. 24(3). pp.49-75.'를 참조하기 바란다.

6) 상관관계 분석과 같이 두 변수 간의 관계를 분석하는 경우 분석의 대상이 되는 두 변수 모두 관측된 경우만 분석에 포함되는 방식이다. 이에 반해 완전제거법은 분석에 포함되는 변수 전체에 대해 하나라도 결측치가 존재하는 경우 분석에서 제외된다.

(stochastic regression imputation)이 소개되었는데, 결측변수의 결측치의 추정에 일정한 확률오차항(random error term)[7]을 포함시켜 변동성(variability)을 고려하는 방식인데, MAR 가정하에서 불편 추정치를 유도하지만, 여전히 표본오차를 과소추정하여 1종 오류를 야기하는 문제가 있다(Baraldi & Enders, 2010).

이러한 노력에도 불구하고 여전히 기존의 대체방법이 가지는 문제는 완전히 해결되지 않았다. 이에, 한 번의 대체값으로 결측치를 다루는 방법에서 벗어나 대체를 여러 번 수행하여 이들의 평균으로 분석을 수행하는 다중대체가 논의되기 시작하였다. 다중대체법은 3단계로 구성되는데 가능한 대체 값의 분포에서 추출된 서로 다른 값으로 결측치를 처리한 복수의 데이터셋을 생성한 뒤(imputation phase), 이들 데이터셋에 대하여 각각 분석을 수행하고(analysis phase), 그 결과 얻은 모수의 추정량과 표본오차를 통합하여(pooling phase) 하나의 분석결과를 제시하는 방법이다.

다중대체의 단계를 구체적으로 살펴보면 다음과 같다. 첫 번째 단계는 각각의 결측치를 일정한 알고리즘에 따라 다른 대체값으로 대체한 m개의 데이터셋을 생성한다. 다음으로 m개의 완전한 데이터셋을 각각 분석하고 각 데이터셋의 분석결과에서 모수의 추정치와 표준오차를 확보한다. 마지막으로 각 데이터 셋의 결과를 Rubin's rule[8](Rubin, 1987)에 의해 결합한다(Graham, 2012). 다양한 다중대체방법 중에서도 본 연구에서는 완전조건부 대체법(fully conditional specification, FCS)을 분석에 활용하였다.[9] FCS 외에도 MCMC(Markov chain monte carlo)와 같은 방법이 널리 사용되는데, 두 방법은 결측치 대체모형에 대해 분포가정을 사전에 하는지를 기준으로 구분된다. FCS는 분포에 대한 가정이 없이 연속된 회귀방정식을 통해 값을 대체해 나가는 방식이다. FCS 방법은 사전 채워넣기(filled-in) 단계와 이어지는 대체(imputation) 단계로 나누어진다(표 1). 우

---

7) $e \sim N(0, \sigma^2)$

8) 각 데이터셋 별로 구한 추정치($\overline{Q}$)와 표준오차($\sqrt{T}$)를 결합하는 과정은 다음과 같다. 추정치의 결합은 각 데이터셋으로부터 구한 추정치의 평균으로 정의된다. $\overline{Q}$의 분산 T는 대체내 분산(within-imputation variance) W와 대체간 분산(between-imputation variance) B의 결합값으로 정의된다. SAS에서는 여러 개의 다중대체된 데이터셋의 결과를 결합할 때, 'PROC MIANALYZE' 프로시저를 사용한다.

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\widehat{Q}_i$$

$$\overline{W} = \frac{1}{m}\sum_{i=1}^{m}\widehat{W}_i,\ B = \frac{1}{m-1}\sum_{i=1}^{m}(\widehat{Q}_i - \overline{Q})^2,\ T = \overline{W} + (1+\frac{1}{m})B$$

9) 본 연구에서는 MCMC와 FCS를 모두 사용하여 분석을 하였으나, 두 방법 간 분석결과의 차이가 거의 없어 분포가정이 없는 FCS를 중심으로 결과를 제시하였다.

선 채워넣기 단계에서는 모든 변수의 결측치를 변수의 순서대로 채우게 되는데, 앞서 채워진 변수가 다음 채워지는 변수의 독립변수(covariate)로 활용되는 방식이다. 대체단계는 앞서 채워진 값들을 변수의 순서대로 대체하는 과정이다. 이러한 과정을 충분히 길게 하여 대체된 데이터셋에서 결측치가 독립적인 추출이 될 때까지 시행한다.

〈표 1〉 FCS 대체방법의 상세

채워넣기 단계(filled-in)	대체단계(imputation)
$\theta_1^{(0)} \sim P(\theta_1 \mid Y_{1,obs})$	$\theta_1^{(t+1)} \sim P(\theta_1 \mid Y_{1,obs}, Y_2^{(t)}, \ldots, Y_p^{(t)})$
$Y_{1*}^{(0)} \sim P(Y_1 \mid \theta_1^{(0)})$	$Y_{1*}^{(t+1)} \sim P(Y_1 \mid \theta_1^{(t+1)})$
$Y_1^{(0)} = (Y_{1,obs}, Y_{1*}^{(0)})$	$Y_1^{(t+1)} = (Y_{1,obs}, Y_{1*}^{(t+1)})$
…	…
$\theta_p^{(0)} \sim P(\theta_p \mid Y_1^{(0)}, \ldots, Y_{p-1}^{(0)}, Y_{p,obs})$	$\theta_p^{(t+1)} \sim P(\theta_p \mid Y_1^{(t+1)}, \ldots, Y_{p-1}^{(t+1)}, Y_{p,obs})$
$Y_{p*}^{(0)} \sim P(Y_p \mid \theta_p^{(0)})$	$Y_{p*}^{(t+1)} \sim P(Y_p \mid \theta_p^{(t+1)})$
$Y_p^{(0)} = (Y_{p,obs}, Y_{p*}^{(0)})$	$Y_p^{(t+1)} = (Y_{p,obs}, Y_{p*}^{(t+1)})$

주. θ는 모수를 의미하며, obs는 이미 관측된 값을 *는 결측치를 의미한다. 즉, $Y_{p,obs}$는 p번째 변수의 관측된 값을 Y_{p*}는 p번째 변수의 결측치를 의미한다.

이상의 논의와 같이 결측치의 대체방법은 더 보편적인 결측 메커니즘에 적용할 수 있는 방법을 찾기 위해 발전해 왔으며, 분석결과의 불편성과 효율성을 제고시키고자 노력해 왔음을 알 수 있다. 하지만, 대부분의 결측 메커니즘을 따르는 가상의 데이터를 활용한 시뮬레이션 분석(Long & Johnson, 2015; Eekhout, 2015; Burgress 외, 2013; Daniel & Kenward, 2012; Peyre 외, 2011, Young & Johnson, 2010)에 의존해 왔으나, 대체방법의 실제 적용가능성에 대한 논의와 함께 최근 실제 데이터를 활용한 분석(고길곤 외, 2014; Lieberman-Berz 외, 2014; Cox 외, 2014)이 증가하는 추세이다. 본 연구에서도 결측치 대체방법의 실제 적용가능성의 관점에서 가상의 데이터를 활용한 시뮬레이션 분석과 더불어 재조사법과의 혼합 활용을 통해 실제 데이터에서 활용가능한 결측치 대체방법을 제시하고자 한다.

3. 결측치 처리방법 적용의 쟁점들

전통적인 결측치에 관한 많은 연구는 결측이 분석결과에 미치는 영향을 탐구하는 기능적인 접근(functional approach)보다는 구조적인 접근에 주목해 왔다(McKnight 외, 2007:9). 하지만 결측치의 처리에 관한 연구에서 결측치의 발생에 대한 구조적인 접근은

이상적인 성격이 강하며, 실제 결측치가 분석결과에 미치는 영향과 이를 바로 잡을 방법에 대한 기능적이고 실용적인 접근이 필요하다.

사실 결측치에 대한 가장 좋은 해결책은 그것이 발생하지 않도록 하는 것이다. 이는 설문의 설계, 모집단과 표본의 설정, 자료수집방법, 측정도구 선정 등을 통해 설문대상자가 응답하는데 드는 부담을 줄여주는 것과 관련이 있다(McKnight 외, 2007:65-87). 결측치에 대한 바람직한 대안은 무응답에 대해 재조사(callbacks)라고 할 수 있다. 일반적으로 재조사법은 동일 응답자에게 결측이 있는 질문을 다시 물어보고 답을 얻어 내는 방식으로 진행된다. 측정시점의 차이로 인한 내적타당성 위협요인이 존재할 수 있으나 동일 응답자로부터 응답결과를 얻을 수 있으므로 참값에 가까운 결과를 얻을 수 있다.

그러나 재조사법은 비용적인 측면의 문제가 크게 발생하여 실제 사용에 제약이 따른다. 본 연구에서는 설계단계에서의 예방책을 더 이상 사용할 수 없는 상황에서 통계적 대체법과 재조사법이라는 방법에 대해 초점을 맞추고 있다. 즉, 이미 설문이 진행된 상황에서 연구자가 취할 수 있는 결측치의 처리방법을 탐색하고자 하는 것이다. 실용적인 결측치의 처리를 위해서는 비용효과적인 접근이 필요한데(McKnight 외, 2007:10), 다중대체법이 적은 비용과 노력으로 보다 많은 표본을 확보할 수 있게 하는 방법이기 때문이다. 따라서 상당한 수준의 결측이 존재하고 있는 상황에서 조사법과 통계학적 다중대체법을 혼합하여 사용할 경우 이러한 효과를 얻을 수 있는지가 중요한 연구문제가 될 수 있다.

이를 검증하기 위해 본 논문에서 중요하게 고려하고 있는 쟁점은 다음과 같다. 첫째, 결측 메커니즘의 관점에서 통계적 대체방법의 가정이 실제 데이터에서도 충족되어야 한다. 만일 결측을 포함한 자료가 MAR을 만족하지 못할 때는 편의추정량이 발생할 수 있다(Azur 외, 2011). 또한, 결측값이 관찰된 값에 설명된다고 하더라도 이 효과는 결측을 설명하는 효과와 결측을 설명하는 다른 변수와의 고유한 관계 효과가 혼합되어 발생하는 문제가 존재한다. 예를 들어 3개의 변수 X_1, X_2, Y 중 X_1의 결측이 관찰된 변수 X_2에 의해서 설명된다고 하면, X_1의 결측은 관측된 자료에 의한 것으로 MAR 가정을 따른다. 하지만, X_1과 X_2가 Y변수를 설명하고 있는 상황에서 X_1은 X_2와 Y 모두와 상관관계가 존재할 가능성이 크고, 결국 R은 X_2에 의해서만 설명되는 것이 아니라 X_2와 상관관계를 가지고 있는 관측되지 않은 값을 갖는 X_1 자체에 의해서도 설명될 수 있게 된다.

반대로 실제 자료의 결측이 NMAR 가정을 따르는 경우 관측된 값으로 결측된 값을 추정하는 기존의 통계적 대체방법이 작동하지 않는다고 한다. 하지만, 분석자료의 결측이 NMAR인지 MAR인지를 판별할 방법이 없기 때문에(Baraldi & Enders, 2010), 사전에

통계적 대체방법의 적용가능성을 판단하는 것은 어려웠다. 그러나 이 경우에도 앞서 MAR의 경우와 같이 변수 간 관계에 따라 다른 결론이 가능하다. 예를 들어 X_1, X_2, Y의 변수가 존재하고 X_1에 NMAR을 따르는 결측이 존재한다고 하자. 이때, X_1과 X_2의 상관관계에 따라 두 가지 경우를 예상해 볼 수 있다. 먼저, X_1과 X_2가 상관관계가 없는 독립($\rho_{X_1X_2}=0$)인 경우이다. 이 경우, NMAR 가정에 따라 X_1의 결측 매트릭스인 R은 X_1에 의해서만 설명되고, X_2와는 관계가 없다. 따라서, NMAR 메커니즘에 의한 결측 데이터의 분석결과의 문제를 그대로 가진다. 다음으로, X_1과 X_2가 어느 정도 상관관계가 있는 경우이다. 이 경우, 결측이 X_1에 의해서만 설명된다고 보기 어려우며, 결국 완전한 NMAR 가정을 따르지 못하고 MAR과 NMAR이 섞여 있는 경우가 된다. 따라서 실제 결측을 포함한 데이터셋에서는 완벽한 MAR 혹은 NMAR 상황보다는 이 두 가지가 혼합된 상황이 일반적이라고 할 수 있다.

두 번째로, 자료의 결측률에 대한 관점이 고려되어야 한다. 통계적 대체방법은 관측된 정보를 통해 관측되지 않은 정보를 추정하는 방법이므로, 결측률이 높아지면 그 만큼 제한된 정보를 사용하게 되므로 대체의 정확성이 떨어질 수밖에 없다. 관측된 정보가 결측치를 추정하는데 얼마나 적절한가에 따라 달라질 수 있으나, 결측률과 대체방법의 정확성은 반비례할 것이다. 결측률이 5% 미만이면 결측 메커니즘과 관계없이 완전제거법을 사용해도 문제가 되지 않는다는 주장도 존재한다(Graham, 2009; Schafer, 1999). 또한, 완전임의결측인 경우는 관측치의 감소로 인한 검정력이 낮아지는 문제를 제외하면 무응답이 존재하는 관측치를 제외하더라도 나머지 관측치들이 전체 표본을 대표할 수 있으므로 분석의 편의가 발생하지 않는 주장도 있다(Graham, 2012; Lee 외, 2011). 따라서 재조사법과 다중대체법을 혼합에서 어느정도의 자료가 결측되었는지에 따라 성능이 달라질 수 있다.

세 번째는 보조변수(auxiliary variable)의 문제이다. 실제 자료에서는 MAR 결측이라고 가정하더라도 관측된 값 중 어느 변수에 의해 결측이 영향을 받는지에 대해 사전적으로 알 수 없다. 한편, NMAR 가정하에서도 관찰되지 않은 변수와 상관관계를 가진 관찰된 변수가 있다면, 이 변수를 대체모형에 포함시켜 결측치를 추정하는 것은 간접적으로만 가능하다. 따라서, 결측된 변수와 상관관계를 가질 수 있다고 판단되는 보조변수를 포함시킨 대체방법을 고려할 수 있다(Allison, 2000). 즉 실제 자료가 NMAR 가정을 따른다고 하더라도, 결측된 변수의 결측된 값은 측정할 수 없지만, 여기에 영향을 주는 잠재적인 요인을 대체모형에 포함시키는 것으로 NMAR 가정이 완화되는 효과를 얻을 수 있다.

한편 다중대체를 활용하여 결측을 처리한 후 회귀분석을 수행한다면 대체모형에서도

종속변수를 보조변수로 포함하는 것을 고려할 필요가 있다(Schafer, 1997). 종속변수를 대체모형에 포함하지 않을 때 독립변수에 대체된 값은 종속변수와 관련이 없어지며, 그 결과 회귀계수가 편의를 갖게 된다(Landerman 외, 1997). 다만, 종속변수를 포함하는 때도 회귀계수가 증가되는 문제가 발생한다는 주장도 있지만, 다중대체를 통해 대체값에 임의적 요인(random component)을 포함하는 다중대체 방법으로 이 문제를 피할 수 있다고 한다(Schafer, 1997).

이와 같은 결측치 처리의 다양한 쟁점에 대한 해답을 제시하기 위해 많은 연구들이 다중대체법의 적용가능성을 분석하고 있으나 대부분 MAR 가정을 구현한 가상의 데이터 셋에 대한 시뮬레이션을 통해 적용가능성(Burgess 외., 2013; Burgett & Reiter, 2010; Daniel & Kenward, 2012)을 논하고 있을 뿐 실제 데이터에 적용한 결과에 대한 연구는 부족하다. 이에 본 연구에서는 기존 연구와 같이 가상 데이터셋을 이용하여 이상 소개한 쟁점을 검토하고, 재조사법을 이용하여 결측치를 보정한 실제 설문자료를 활용하여 재조사와 다중대체를 어느 정도 조합 했을 때 타당한 결과를 얻는지 살펴보고자 한다.

Ⅲ. 연구설계

1. 가상데이터를 이용한 분석모형의 설계

결측 메커니즘, 결측률, 그리고 변수 간 관계에 따라 결측치 처리방법이 분석결과에 미치는 영향을 살펴보기 위해 이상적인 상황에서 가상데이터를 생성하여 분석을 실시하였다. 결측치를 포함한 설문조사 자료를 활용한 회귀분석을 가정하고, 3개의 변수(Y, X_1, X_2)를 포함하고 그중 하나의 변수(X_1)가 결측을 포함하도록 하였다. 결측 메커니즘을 구현하기 위하여 MAR과 NMAR 가정에 따라 결측 매트릭스를 생성하도록 하였으며, 결측 매트릭스에서 결측이 될지를 결정하는 확률변수의 분포를 조정하여 결측의 비율을 조정하였다. 또한, 변수 간 관계는 X_1과 X_2가 어느 정도의 상관관계를 가지는 경우($\rho_{X_1X_2} = 0.3$)와 전혀 상관관계가 존재하지 않는 독립인 경우($\rho_{X_1X_2} = 0$)의 두 가지 형태로 구현하였다.

시뮬레이션 분석 절차를 구체적으로 살펴보면 다음과 같다(그림 1). 분석에 활용한 통계패키지인 SAS의 난수 발생함수를 이용하여 정규분포를 따르는 100만 개의 관측치를 가지는 모집단을 생성하였다. 이때, 일반적인 설문자료의 응답값인 리커트 5점 척도의

형태를 가지도록 X_1은 평균 3, 분산 1의 분포를 따르고 X_2는 평균 3.3, 분산 1의 분포를 따르도록 하였으며, 각 관측치를 반올림하여, 정수화하고 1보다 작은 값은 1로 5보다 큰 값은 5로 보정하여 두 개의 변수를 생성하였다. 또한, 이들 두 변수의 선형결합과 오차항으로 변수 Y를 생성하였으며, Y 또한 리커트 5점 척도의 응답이 될 수 있도록 선형결합의 계수를 표준화하였다.[10] 가상의 데이터는 이상의 규칙으로 생성되었기 때문에, 이를 활용한 분석결과 회귀계수의 모수는 X_1과 X_2 모두 0.66 정도의 값을 가질 것으로 예상할 수 있다. 같은 방식으로 X_1과 X_2가 일정한 상관관계를 가지는 모집단을 구현하였는데, X_1과 X_2의 평균과 분산은 앞서 생성한 모집단과 같지만, '*PROC IML*' 프로시저를 활용하여 두 변수 간 상관계수가 0.3이 될 수 있도록 공분산 구조를 설정해 주었다.

다음 단계는 실제 설문조사와 같이 표본집단을 추출하고 응답을 받는 과정이다. 표본은 SAS의 '*PROC SURVEYSELECT*' 프로시저를 활용하여 모집단에서 1,000개를 비복원 추출로 생성하였다. 다음은 MAR과 NMAR 결측 메커니즘에 따라 X_1의 결측치를 발생하는 과정이다. 앞서 살펴 본 바와 같이 실제 상황에서 순수한 MAR과 NMAR은 기대하기 어려우며, 이들이 섞여 있는 형태로 결측 메커니즘이 적용될 가능성이 크다. 따라서 본 연구에서는 두 가지 메커니즘이 특정 비율 q에 따라 결정되도록 결측의 발생과 관련된 확률변수 M을 구현하였다.[11] 즉 q가 0이면, X_1에 의해서만 결측이 결정되므로 순수한 NMAR 메커니즘을 따르게 되고, q가 1이면 결측이 관측된 변수인 X_2의 영향만 받으므로 순수한 MAR 결측이 된다[12]. q가 0과 1 사이의 값을 갖게 되면, MAR과 NMAR 메커니즘 모두의 영향을 받는 형태가 된다. 결측 매트릭스 R을 구현하기 위해, 앞서 만든 확률변수 M의 분포에서 하위 p%에 해당하는 임계치 α_p를 구한 뒤, M의 값이 이 임계치보다 작으면 R은 0의 값을 갖고, 크면 1을 값을 갖도록 하였다.[13] 그 결과 R은 X_1 또는 X_2에

---

10) $Y = \frac{1}{\sqrt{2^2+2^2+1^2}}(2\times X_1 + 2\times X_2 + e), e \sim N(0,1)$

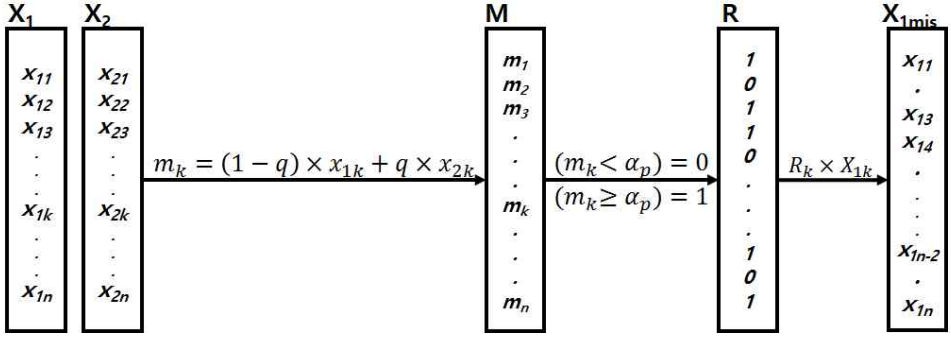
11) $M = (1-q)\times X_1 + q\times X_2 + e, e \sim N(0,1)$

12) 단순하게 두 개의 변수 X, Y에 대하여, Y 변수에 결측을 생성한다고 할 때, "X〈0"과 같이 관측된 다른 변수에 간단한 조건을 주고 이를 만족하는 경우 Y 변수에 결측이 되도록 하여 MAR 결측을 생성하고, Y 변수의 결측이 "Y〈0"과 같은 Y 변수 자체의 조건에 의해 결정되도록 하여 NMAR 결측을 생성하기도 한다(Allison, 2000). 본 연구에서는 MAR과 NMAR 결측에 대해 보다 실제 자료의 결측 메커니즘과 유사한 형태로 결측치를 생성하기 위해 본문과 같은 형태로 결측을 구성하였다.

13) $E(M) = (1-q)\times E(X_1) + q\times E(X_2), E(X_1) = 3, E(X_2) = 3.3$
$VAR(M) = (1-q)^2 \times VAR(X_1) + q^2 \times VAR(X_2) + VAR(e),$
$VAR(X_1) = VAR(X_2) = VAR(e) = 1$
$\therefore M \sim N(E(M), VAR(M))$

영향을 받는 M의 값에 따라 생성되었으며, 이렇게 구한 R과 X_1을 관측치별로 곱하여 0과 곱해지는 관측치는 결측이 되도록 구현하였다. 이러한 방식으로 X_1과 X_2가 상관관계를 가지지 않는 모집단과 일정 상관관계(0.3)를 가지는 모집단으로부터 각각 결측률 p가 0%에서 70%의 값을 가지는 변수 X_{1mis}를 X_1으로부터 생성하였다.

〈그림 1〉 가상데이터의 결측변수(X_{1mis}) 생성 방법



이렇게 독립변수 간 상관관계와 결측률, 그리고 결측치 발생 메커니즘에 따라 생성된 각각의 데이터셋에 대하여 대체방법에 따라 회귀분석을 수행한 후, 각 독립변수의 회귀계수를 비교하였다. 대체방법으로는 완전제거법과 다중대체법(FCS)이 사용되었다. 이때, 최초 결측률 70% 데이터셋에서 결측률이 10%씩 감소할 때, 감소한 결측률은 실제 상황에서는 재조사법을 사용한 결측치의 대체가 이루어진 것으로 가정할 수 있으며, 나머지 결측치는 통계적 대체방법인 다중대체법을 활용하여 대체하여 100% 관측된 데이터셋을 구성하여 분석하는 것과 같다. 완전제거법을 사용하는 경우와 다른 점은 완전제거법의 경우 모든 변수에서 관측된 관측치만 활용하므로 분석에 사용되는 관측치의 수가 다중대체법과 재조사법을 사용한 경우보다 적어진다. 이처럼 독립변수 간 상관관계에 따라 구성된 두 종류의 모집단에 대해 결측률을 변화시켜 가면서 서로 다른 대체방법을 적용한 회귀분석의 회귀계수에 대하여, 구간추정법을 활용하여 95% 신뢰수준에서 통계적으로 유의미한 차이가 있는지를 확인하고 그 결과를 시각적으로 확인할 수 있도록 그래프를 통해 대체방법별 차이를 제시하였다. 이상의 시뮬레이션을 활용한 분석방법의 논의를 정리하면 다음의 흐름도(그림 2)와 같다.

〈그림 2〉 가상자료를 이용한 시뮬레이션 분석 흐름도

2. 실제 설문데이터를 이용한 분석모형의 설계

본 연구는 가상의 데이터셋을 활용한 시뮬레이션 분석과 더불어 상당한 수준의 결측치를 포함하는 실제 설문조사를 이용한 분석도 수행하였다. 본 연구에서는 서울대학교 행정대학원 서베이조사센터에서 실시한 "2012년 삶의 질과 정부 역할에 관한 조사"를 사용하였다. 이 설문조사는 정부수준별 만족도와 행복수준을 주요 설문문항으로 하는 설문조사이다(표 2). 1차 조사 후 주요변수에 대하여 상당한 수준의 결측이 존재하여 재조사를 실사하였으며 대부분의 결측치가 대체될 수 있었다.

〈표 2〉 1차 조사와 2차 조사의 주요변수의 기초통계

변수	1차 조사			2차 조사			관측치 차이
	관측치수	평균	표준편차	관측치 수	평균	표준편차	
기초자치단체 만족도(Q1_1)	7752	3.238	1.130	8240	3.237	1.119	488
광역자치단체 만족도(Q1_2)	7402	2.664	1.005	8240	2.670	0.971	838
중앙정부 만족도(Q1_3)	7732	2.511	0.987	8240	2.511	0.955	508
행복정도(Q4)	8209	2.318	0.992	8240	2.312	0.967	31

주. 상기 변수 이외에 인구통계학적 변수(연령, 성별, 학력, 직업, 소득, 결혼상태, 부양가족 수, 종교)의 결측도 고려하면, 1차 조사결과 결측치가 하나라도 존재하는 관측치는 3994개

하지만, 재조사에도 불구하고 여전히 남아 있는 결측치를 포함한 관측치는 재조사를 통해 완전한 데이터셋이 만들어진 상황을 가정하기 위해 분석에서 제외했다. 결측치를 포함한 자료에 대하여 완전제거법으로 분석한 결과와 다중대체법과 재조사법을 사용한 결과의 차이를 비교하기 위하여 1차 조사 후 결측치가 없는 데이터셋에 재조사를 통해 대체된 자료를 일정 비율씩 추가하여 재조사를 통한 대체를 구현하였으며, 나머지 결측치는 다중대체법을 이용하여 처리하였다. 이를 알기 쉽게 나타내면 다음의 〈그림 3〉과 같다.

〈그림 3〉 실제 설문조사자료를 이용한 분석데이터 생성 방법

주. cb는 재조사, mis는 결측, obs는 관측된 값을 의미함

실제 데이터를 활용한 분석을 위한 데이터셋 구성의 구체적인 방법은 다음과 같다. 우선, 사용될 주요 변수 중 결측치가 하나라도 존재하는 경우 재조사를 시행하였다. 전체 응답자는 10,000명인데, 그 중 1차 조사에서 결측치가 존재하지 않는 응답자 중 500명을 비복원 임의추출하였으며, 마찬가지로 결측치가 존재하는 응답자(재조사를 시행한 응답자) 중 500명을 비복원 임의추출하였다. 재조사 집단은 결측률을 조정하기 위하여 뽑힌 500명의 1차 조사 응답을 100명 단위로 재조사 응답으로 대체하였다. 만약 (50-p)% 결측률을 가지는 데이터셋을 구성한다면, 1차 조사에서 추출한 500명과 재조사를 실시한 응답자 500명을 추출하고, 재조사에서 추출된 응답자 중 (10*p)명 만큼을 재조사한 응답으로 대체하는 방식이다. 예를 들어, 결측률 30%(p=20)의 데이터라면, 재조사를 실시한 집단에서 뽑힌 500명 중 200(=10*20)명은 재조사 응답으로 대체하고, 나머지 300명은 다중대체법을 활용하여 처리하는 방식이다. 이렇게 구성한 1000명의 응답자를 포함하는 설문자료는 결측률 50%에서 0%까지 10% 단위로 데이터셋이 구성되었으며, 이와 같은 방식으로 30번 데이터셋을 구성하여 각각 회귀분석을 수행한 후 회귀계수의 평균으로 결측률과 대체방법에 따른 차이를 비교하였다. 일정한 결측을 포함하지만 재조사와 다중대체법(FCS)이 혼합하여 결측을 처리한 경우, 그리고 완전제거법을 사용한 경우의 분석결과가 재조사를 통해 완전히 결측치를 대체한 데이터셋의 분석결과의 신뢰구간에 포함되는지를 확인하였다. 이를 통해 안정적으로 불편성을 만족하는 결측률과 결측치 처리방법의 조합을 탐색하고자 하였다.

Ⅳ. 분석결과

1. 가상데이터를 이용한 시뮬레이션 분석

설문조사를 활용한 사회과학 연구는 연구문제의 선정과 함께 그에 맞는 모집단을 선정하는 것에서 출발한다. 이렇게 선정한 모집단으로부터 표본추출방법을 통해 표본프레임을 구성하고, 이들에 대해 전화, 인터넷, 컴퓨터를 이용한 조사원 직접 면접(computer assisted personal interviewing, CAPI) 등의 방법으로 응답을 받는다. 본 연구에서는 이러한 설문조사의 절차에 맞추어 가상 데이터셋을 구성하였다. 우선, 보통의 설문문항에서 활용하는 5점 리커트 척도에 맞추어 Y, X_1, X_2의 3개의 변수를 포함하는 100만개의 모집단을 두 독립변수의 상관관계에 따라 상관관계가 없는 경우와 어느 정도(0.3)의 상관관계를 가지는 경우로 두 종류의 모집단을 구성하였으며, 이들 가상 데이터에 포함된 변수의 기술통계는 다음과 같다(표 3).[14]

〈표 3〉 모집단의 주요변수의 기술통계

변수	N	모집단 ($\rho_{X_1X_2}=0$)				모집단 ($\rho_{X_1X_2}=0.3$)			
		μ	SD	MIN	MAX	μ	SD	MIN	MAX
Y	1,000,000	4.086	0.886	1	5	4.048	0.954	1	5
X_1	1,000,000	3.002	1.008	1	5	2.999	1.009	1	5
X_2	1,000,000	3.288	1.003	1	5	3.287	1.001	1	5

이렇게 구성된 모집단에서 임의추출(random sampling)을 통해 1000개의 표본을 뽑았다. 실제 설문조사와 같이 이 과정에서 결측이 발생한다고 보고 추출된 1000개의 표본 중 결측 메커니즘에 따라 X_1에 결측을 발생시켰다. 결측이 관측된 변수인 X_2에 의한 경우 MAR 결측으로, 결측된 변수인 X_1에 의한 경우 NMAR 결측으로, 두 변수의 영향을 동일하게 받는 경우 혼합(mixed) 결측으로 설정하였으며, 10~70% 결측률을 가지는 데이터셋을 구현하였다(표 4, 5). 이렇게 구성된 결측 데이터셋에 포함된 주요 변수 간 상

14) 구체적으로 X_1은 $N(3,1^2)$에서, X_2는 $N(3.3,1^2)$에서 추출된 값으로 정의하였으며, Y는 "$Y=0.67X_1+0.67X_2+\epsilon$"의 관계를 가진다. 이들의 값을 설문조사의 응답으로 변환시키는 과정에서 1보다 작은 값은 1로, 5보다 큰 값은 5로, 그리고 1과 5사이의 값은 반올림하여 정수화하였다. 정수화와 변수 범위의 조정(truncated)의 결과로 각 독립변수의 모회귀계수가 0.67이 아닌 다른 값을 가지게 되었다.

관관계와 관측치 수는 다음 〈표 4〉에 정리하였다.

〈표 4〉 X_1과 X_2가 상관관계를 가지지 않는 경우 주요변수 간 상관관계

DATASET	Y와 MX_1		Y와 X_2		MX_1와 X_2		MX_1과 AUX		X_1과 R		X_2과 R	
	r	n	r	n	r	n	r	n	r	n	r	n
MAR 10%	0.612	888	0.631	1000	-0.017	888	0.981	888	0.012	1000	0.433	1000
MAR 20%	0.63	797	0.631	1000	-0.005	797	0.98	797	0.008	1000	0.528	1000
MAR 30%	0.633	686	0.631	1000	0.015	686	0.981	686	-0.003	1000	0.543	1000
MAR 40%	0.635	602	0.631	1000	0.034	602	0.982	602	-0.031	1000	0.574	1000
MAR 50%	0.641	487	0.631	1000	-0.007	487	0.984	487	0.008	1000	0.576	1000
MAR 60%	0.611	403	0.631	1000	-0.015	403	0.982	403	0	1000	0.6	1000
MAR 70%	0.592	309	0.631	1000	0.058	309	0.978	309	-0.065	1000	0.52	1000
NMAR 10%	0.573	884	0.631	1000	-0.003	884	0.978	884	0.421	1000	-0.012	1000
NMAR 20%	0.577	787	0.631	1000	0.018	787	0.976	787	0.488	1000	-0.029	1000
NMAR 30%	0.547	680	0.631	1000	0.006	680	0.976	680	0.583	1000	-0.002	1000
NMAR 40%	0.507	601	0.631	1000	-0.01	601	0.974	601	0.58	1000	0.019	1000
NMAR 50%	0.523	465	0.631	1000	0.029	465	0.971	465	0.585	1000	-0.051	1000
NMAR 60%	0.494	404	0.631	1000	-0.013	404	0.971	404	0.578	1000	-0.008	1000
NMAR 70%	0.537	315	0.631	1000	0.088	315	0.975	315	0.557	1000	-0.046	1000
MIXED 10%	0.569	891	0.631	1000	-0.071	891	0.978	891	0.245	1000	0.282	1000
MIXED 20%	0.567	801	0.631	1000	-0.055	801	0.981	801	0.286	1000	0.24	1000
MIXED 30%	0.559	680	0.631	1000	-0.07	680	0.983	680	0.311	1000	0.29	1000
MIXED 40%	0.545	580	0.631	1000	-0.125	580	0.98	580	0.316	1000	0.36	1000
MIXED 50%	0.566	483	0.631	1000	-0.087	483	0.981	483	0.317	1000	0.33	1000
MIXED 60%	0.479	415	0.631	1000	-0.123	415	0.979	415	0.354	1000	0.285	1000
MIXED 70%	0.446	300	0.631	1000	-0.232	300	0.979	300	0.308	1000	0.337	1000

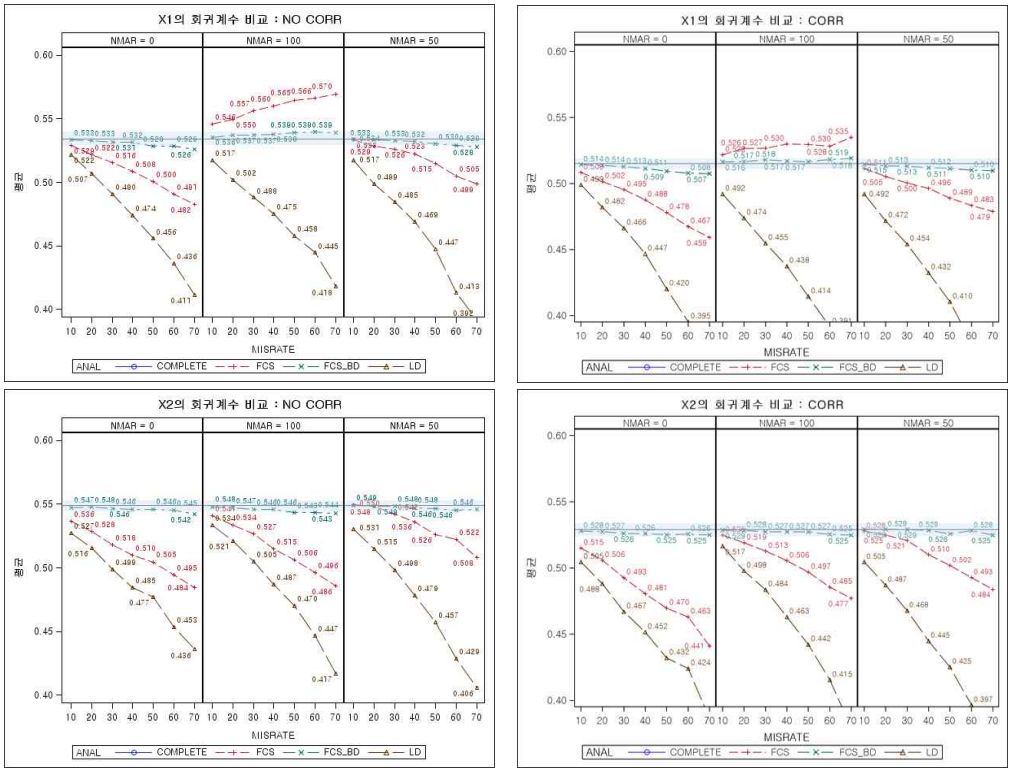
주. 음영부분 숫자는 p〈.1에서 유의하지 않음. 소숫점 셋째자리까지 표현함.
주. 데이터셋은 "결측메커니즘+결측률"로 표현함. MIXED는 MAR과 NMAR이 함께 적용됨
주. 상관관계가 존재하는 데이터셋의 상관관계는 모든 변수 간 관계가 p〈.1에서 유의미함.

실제 분석에서는 이와 같은 과정을 30번 반복하였다. 즉, 가상의 모집단에서 표본을 추출하여 완전제거법과 다중대체법을 적용하는 과정을 30번 반복하여 각각의 회귀분석 결과의 평균을 통해 결측치 대체방법에 따른 영향을 살펴보았다. MAR, NMAR, MIXED의 3가지 결측 메커니즘 및 결측률에 따라 Y와 MX_1(결측된 X_1) 간의 상관관계에 변화가 있는 것을 확인할 수 있으며, 가상 데이터셋에서는 Y와 X_2의 상관관계는 결측률에 따라 변하지 않고 있다. 결측 매트릭스인 R과 X_1 및 X_2의 관계는 MAR인 경우 X_2와, NMAR인 경우 X_1, 그리고, MIXED인 경우 X_1과 X_2 모두와 상관관계를 보이고 있는 것을 확인할 수 있다. 추가적으로 결측된 변수 MX_1과 이를 설명할 수 있는 보조변수(AUX)가 관찰되었다는 가정 하에 이를 대체모형에 포함시킬 예정이다[15]. 상관관계의 유무(NO CORR,

15) 관찰된 변수에 의해 결측이 설명되는 경우 MAR 결측이라고 하지만, 이 경우 결측된 변수에 의해서

CORR), 결측 메커니즘(MAR, NMAR, MIXED), 그리고 결측률(10~70%, 10% 단위)에 따라 구성된 총 42개의 데이터셋에 대하여 완전제거법(LD)과 재조사법 및 다중대체법(FCS)을 활용하여 회귀분석을 수행한 결과 X_1과 X_2에 대한 분석결과는 다음 〈그림 4〉와 같다. 추가적으로 X_1과 상관관계를 가지는 관찰된 제 3의 변수(AUX)를 대체모형에 포함한 보조변수의 논의(FCS_BD)를 함께 분석하였다.

〈그림 4〉 대체방법에 따른 회귀계수의 비교



최초 생성된 100만개의 관측치를 가지는 모집단으로부터 1000개의 표본을 뽑아 결측 데이터셋을 만드는 일련의 과정을 30번 반복하여 분석한 회귀계수의 평균을 시각화 하였으며, 그래프의 푸른 색 띠는 완전한 데이터셋을 가지고 분석한 회귀계수의 평균과

결측이 설명되지만, 결측된 변수가 관찰된 변수와 상관관계를 가지게 되는 경우를 가정하고 있다. 즉, 어떤 변수의 결측을 직접적으로 설명하지는 못하지만, 간접적으로 설명할 수 있는 변수의 존재를 가정한 것이다. 이를 통해 실제 설문조사에서 NMAR 결측이 의심되는 상황에서 관찰된 변수 중 이 결측된 변수와 상관관계가 높은 변수가 있다면 분석모형과는 상관없이 이를 대체모형에 포함시키면 다중대체의 성능이 향상될 수 있는지 여부를 확인하고자 한다.

표준오차를 통해 구한 회귀계수의 95% 신뢰수준에서의 신뢰구간으로, 결측치 처리방법(LD, FCS, FCS_BS)을 통해 구한 회귀계수가 이 범위에 들어가면 결측치가 없는 데이터셋을 분석한 결과와 통계적으로 같은 결과라고 할 수 있다. FCS를 활용한 다중대체의 대체모형에서는 MX_1을 관측된 변수인 Y와 X_2를 통해 대체하였고, FCS_BD 모형에서는 X_1과 상관관계를 가지고 있는 AUX를 추가적으로 포함하였다. Y는 회귀분석에서 종속변수로, X_1과 X_2는 독립변수로 사용되었다. AUX는 대체모형에만 활용되었을 뿐 회귀분석에서는 포함되지 않았다.

구체적으로 분석결과를 결측률, 대체방법, 대체모형, 변수 간 관계의 관점에서 살펴보면, 우선 독립변수 간 상관관계가 없는 모집단을 활용한 표본의 분석결과, 결측률이 줄어들면 대체방법이나 대체모형과 상관없이 완전한 데이터셋을 이용한 X_1과 X_2의 회귀계수 추정량과 그 신뢰구간으로 수렴하는 것을 확인할 수 있다. 결측률이 높다는 것은 관찰된 정보의 양이 적다는 것을 의미한다. 완전제거법(LD)의 경우 관측된 정보만을 가지고 분석을 수행하기 때문에, 관측된 정보가 모집단을 대표하지 못하면 이를 통해 분석한 회귀계수는 모집단의 회귀계수와는 차이를 가지게 된다. 다중대체법을 활용하는 경우(FCS) 완전제거법 보다 회귀계수가 참값에 가까워지는 것을 알 수 있으나 여전히 신뢰구간에는 포함되지 못한다. 다만, 결측률이 줄어들수록, 즉 보다 많은 관측치 정보가 활용될수록 다중대체법을 통한 분석결과의 편의가 줄어드는 것을 확인할 수 있다. 즉, 완전제거법에 비해 완전한 데이터를 활용한 분석결과에 보다 가까운 결과를 나타내고 있지만, 10% 이하의 결측률에 이르러서야 회귀계수의 신뢰구간에 포함되는 것을 확인할 수 있다. 보조변수를 포함한 대체모형을 활용한 다중대체법(FCS_BD)의 분석결과는 결측 메커니즘에 관계없이 우수한 성능을 보이고 있으며, 최대 50% 정도의 결측률을 가진 데이터에 대해서도 완전한 데이터의 분석결과로 수렴하고 있다.

FCS와 같은 다중대체법은 MAR 결측을 가정하고 있다. 따라서, 가상 데이터를 이용한 분석 중 MAR(NMAR=0)인 경우에만 효과를 나타낼 것으로 예상할 수 있다. 하지만, 실제로는 X_1과 X_2과 Y를 설명하고 있는 상황에서 결측이 결측된 변수 X_1에 의해서만 설명되는 경우는 드물다. 본 연구에서의 혼합 결측(NMAR=50)은 물론 완전한 NMAR의 경우에도 대체모형에 종속변수인 Y를 포함시키면 결측치 대체를 위한 정보를 어느 정도 확보할 수 있다. 즉, Y가 일종의 보조변수로 활용될 수 있음을 알 수 있다. 다만, 분석모형에 포함되지 않는 AUX와는 달리 분석모형에 종속변수로 포함되는 Y를 대체모형에 포함시킨 경우 특히 완전한 NMAR에서 X_1의 결측을 설명할 유일한 변수가 Y이므로, 대체된 값이 Y와 높은 상관관계를 가지게 되어 결국 회귀계수가 과대평가되는 결과를 보인다.

이는 대체모형에 종속변수를 포함시켜야 한다고 주장하는 선행연구에서도 이미 지적한 바 있는 것으로, 회귀계수가 과대평가되는 문제에도 불구하고 NMAR 결측의 경우 결측치를 대체할 유일한 정보라는 점에서 이를 포함시키는 것이 필요하다는 것을 알 수 있다. 또한, 분석모형에서는 중요하지 않은 변수이지만, 결측된 변수를 설명할 수 있는 관측된 변수가 존재한다면 이를 대체모형에는 포함시켜 분석하는 것이 필요하다는 것 또한 확인할 수 있다.

독립변수의 상관관계가 존재하지 않는 경우와 존재하는 경우를 비교하면, 다중공선성의 문제로 인해 상관관계가 존재하는 경우 상관관계가 존재하지 않는 경우에 비해 회귀계수의 점추정치가 다르게 나타나는 것을 확인할 수 있다. 하지만, 다중공선성의 문제는 회귀계수의 불편성의 문제보다 표준오차의 과대추정으로 인한 회귀계수의 신뢰구간을 넓히는 것이 문제가 된다(고길곤, 2014:426-427). 독립변수 간 상관관계가 존재하여 나타나는 회귀계수의 차이를 제외하면 상관관계의 유무에 관계없이 결측치의 대체방법은 작동하고 있음을 알 수 있다. 이는 가장 실제 설문자료와 유사한 상황인 독립변수 간 상관관계가 존재하고, 결측의 메커니즘이 혼합 결측인 경우에도 다중대체가 적용가능하다는 것을 보여준다. 다만, 가상 데이터의 경우와 같이 결측된 변수와 상관관계가 높은 보조변수를 실제 상황에서는 찾기 어려우므로, 대체모형에 보조변수를 포함시킨다는 가정 하에 재조사법을 통해 약 20% 정도의 결측률까지는 자료의 관측률을 높이는 노력이 필요하다고 할 수 있다. 물론 이는 보다 좋은 보조변수를 찾는 다면 가상 데이터의 분석결과와 같이 50% 이상의 결측률에서도 FCS_BD 모형이 작동하는 것을 알 수 있으나, 실제 상황에서 결측된 변수의 결측된 값을 알지 못하는 상황에서 최적의 보조변수를 찾는 것은 쉽지 않기 때문에 FCS와 FCS_BD 모형의 결과를 함께 고려할 필요가 있으므로, 약 20%의 결측률까지 재조사법의 활용이 필요하다고 주장할 수 있다. 이와 같은 가상 데이터를 활용한 분석결과의 예상을 토대로 실제 결측치를 포함한 설문자료를 통해 이러한 통찰이 적용가능한지를 확인하고자 한다.

2. 실제 설문자료를 활용한 분석

실제 설문조사의 경우 무응답의 참값이 무엇인지 알 수 없으므로, 완전한 데이터셋의 결과와 결측치 대체모형의 분석결과를 비교하는 연구는 불가능하다. 하지만, 본 연구에서는 약 1만 개의 관측치를 가지는 대표본 설문조사에서 약 40% 정도의 항목 무응답이 발생한 1차 조사 데이터와 이러한 결측치를 재조사를 이용하여 대체한 2차 조사 데이터를 확보하였다. 이처럼 재조사법을 통해 대체된 값을 참값으로 가정하고 통계적 대체방

법 간의 비교를 통해 앞서 수행한 가상 데이터의 결과를 검증해 보았다. 특히, 결측률에 따른 분석결과인 각 독립변수의 회귀계수를 비교하는 것을 통해 어느 정도의 결측이 재조사로 대체되었을 때, 통계적 대체방법이 제대로 작동하는지를 탐색하였다.

설문자료는 응답자의 정부수준(기초, 광역, 중앙)별 정부기관 만족도와 행복수준을 주요변수로 하고 있다. 실제 설문에서는 이들 주요변수와 함께 성별, 연령, 학력, 지역, 종교, 가족의 수 등 인구통계학적 변수도 관측되었다. 이를 활용하여 일반적인 행정학의 연구에서 관심이 있는 연구문제로서 "정부에 대한 만족도가 시민의 행복을 높여주는가?"에 대한 회귀분석을 시행한다고 가정하였다. 일반적인 설문조사의 표본의 수와 유사한 상황을 만들기 위하여 표본의 수를 1000개로 제한하였다. 전체 1만개의 관측치를 가진 데이터셋으로부터 1차 조사에서 완전한 응답이 이루어진 관측치에서 500명, 2차 조사를 통한 결측치의 대체를 통해 새롭게 완전한 응답이 이루어진 관측치에서 500명을 추출하여 1,000명의 샘플을 만들고, 2차 조사에서 대체된 응답자를 100명 단위로 포함시켜, 결측률을 50%-0%까지 변화시킨 데이터셋을 구성하였다.[16] 분석에 사용된 주요변수의 기술통계는 〈표 5〉와 같다.

〈표 5〉 결측률에 따른 주요 변수의 기술통계

데이터셋	행복수준 (Q4)			기초자치단체만족도 (Q1_1)			광역자치단체만족도 (Q1_2)			중앙정부만족도 (Q1_3)		
	N	평균	SD	N	평균	SD	N	평균	SD	N	평균	SD
50% 결측	991	2.344	1.002	846	2.475	0.953	728	2.628	0.984	838	3.259	1.119
40% 결측	991	2.344	1.002	882	2.481	0.947	793	2.636	0.969	875	3.246	1.110
30% 결측	992	2.344	1.001	912	2.493	0.951	845	2.647	0.957	904	3.230	1.108
20% 결측	995	2.346	1.002	941	2.504	0.948	901	2.657	0.945	938	3.230	1.098
10% 결측	996	2.346	1.001	970	2.512	0.949	950	2.667	0.937	966	3.218	1.094
0% 결측	1000	2.347	1.000	1000	2.511	0.943	1000	2.681	0.935	1000	3.207	1.089

이렇게 추출된 결측률별 분석데이터를 살펴보면, 종속변수로 사용될 행복수준(Q4)는 결측이 거의 없는 것을 알 수 있다. 또한, 나머지 변수도 약 150~300개 정도의 결측을 나타내고 있는데, 항목 무응답의 경우 이렇게 개별 변수의 결측은 50%가 되지 않지만,

16) 임의추출된 재조사 응답자 500명 중 100명 단위로 뽑는 방법은 계통적 표집방법으로 하였으며, 각 관측자의 id를 5로 나눈 나머지에 따라 "$\mathrm{mod}(id, 5) \leq k, k$는 $0,1,2,3,4$"의 수식을 만족하는 경우로 하였다. 예를 들어 30% 결측자료를 만들 때 1차 조사 500명에 재조사된 응답자 500명 중 응답자의 id가 5로 나누어 나머지가 1 이하인 경우(0 또는 1) 추출하는 방식으로 200명을 결정하여 추가하는 방식이다. 물론 여기서 추출된지 않은 300명의 응답은 1차 조사의 결과가 사용된다.

회귀분석과 같이 여러 변수를 활용하는 분석에서는 하나의 변수라도 결측이 발생하면 해당 응답자의 응답은 일반적으로 제거되기 때문에(완전제거법이 기본으로 사용되기 때문에) 개별 변수의 결측률보다 전체 분석데이터 관점에서의 결측률은 더 크게 나타나게 된다. 평균과 표준편차는 어느 정도의 변화는 나타나고 있지만, 1 표준편차 이내의 변화에 불과한 것으로 나타나고 있다. 이들 결측률의 변화에 따른 주요변수와 개별 변수의 결측 매트릭스 간의 상관관계를 분석한 결과는 다음 〈표 6〉과 같다.

〈표 6〉 결측률에 따른 변수 간 상관관계

데이터셋	변수	Q4	Q1_1	Q1_2	Q1_3	RQ1_1	RQ1_2	RQ1_3	RQ4
50% 결측	Q4	1	0.195	0.203	0.207	-0.133	-0.131	-0.047	
	Q1_1		1	0.597	0.353		-0.025	0.001	-0.075
	Q1_2			1	0.434	-0.016		0.014	-0.065
	Q1_3				1	-0.071	-0.013		0.032
40% 결측	Q4	1	0.195	0.207	0.198	-0.103	-0.088	-0.039	
	Q1_1		1	0.564	0.346		-0.031	0.017	-0.073
	Q1_2			1	0.421	-0.035		0.007	-0.063
	Q1_3				1	-0.067	-0.034		0.030
30% 결측	Q4	1	0.188	0.200	0.196	-0.129	-0.078	-0.071	
	Q1_1		1	0.557	0.343		0.002	0.025	-0.071
	Q1_2			1	0.414	-0.031		0.026	-0.061
	Q1_3				1	-0.055	-0.020		0.029
20% 결측	Q4	1	0.182	0.192	0.194	-0.123	-0.069	-0.037	
	Q1_1		1	0.546	0.327		-0.006	-0.002	-0.051
	Q1_2			1	0.393	-0.037		0.008	-0.041
	Q1_3				1	-0.061	-0.052		0.010
10% 결측	Q4	1	0.186	0.191	0.186	-0.057	-0.016	-0.022	
	Q1_1		1	0.547	0.332		-0.004	-0.002	
	Q1_2			1	0.387	0.007		0.000	-0.016
	Q1_3				1	-0.029	-0.066		0.009

주. 음영부분 숫자는 글씨는 p〈.05 유의수준에서 통계적으로 유의하지 않은 상관계수
주. RQ#은 각 변수의 결측 매트릭스를 의미함.

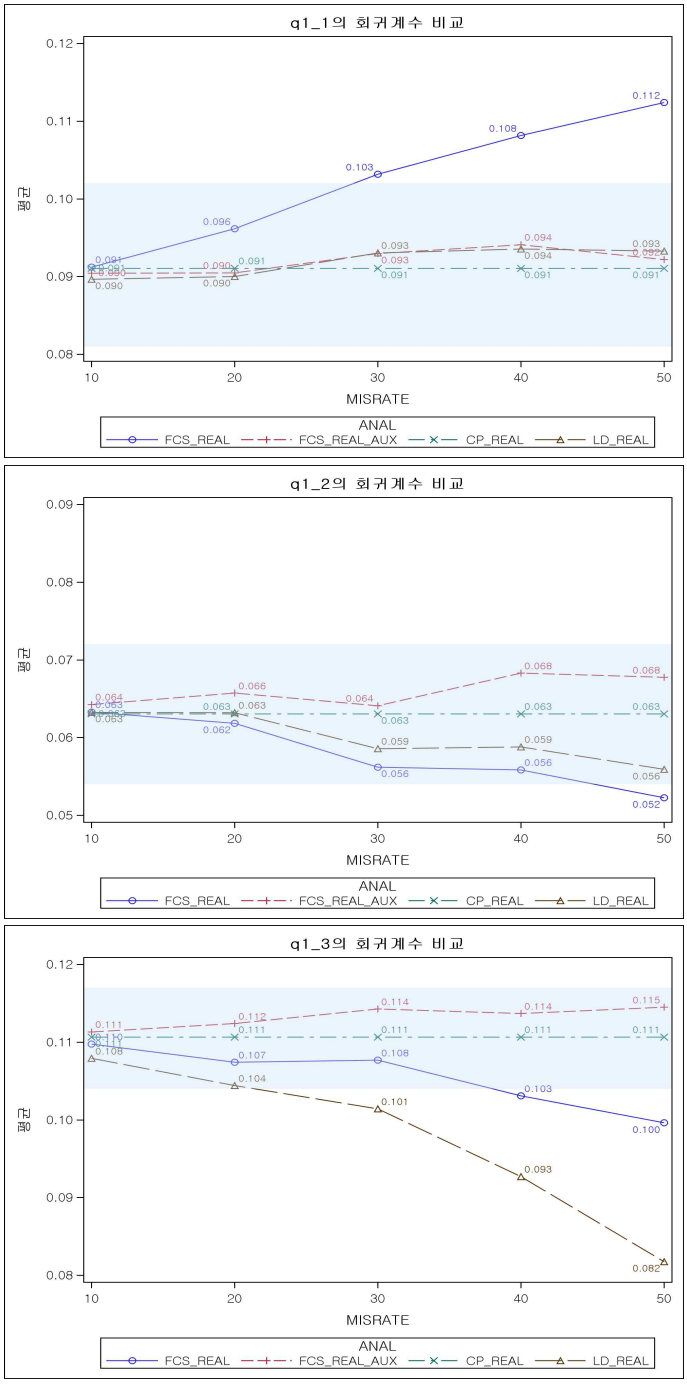
주요 변수 간에는 95% 신뢰수준에서 유의미한 상관관계가 존재하며, 결측률에 따라 변화가 존재한다. 결측 메커니즘을 설명할 수 있는 관찰된 주요변수와 결측 매트릭스의 관계를 보면, 유의미한 상관관계를 나타내는 경우와 그렇지 않은 경우가 혼재하고 있음

을 알 수 있다. 이는 MAR과 NMAR 결측이 변수의 특성에 따라 혼재하여 나타나고 있다는 점을 보여준다. 구체적으로, 만약 관찰된 변수와 특정 변수의 결측 매트릭스가 상관관계를 가지고 있다는 것은 MCAR의 가정이 배제된다는 것을 의미한다. 하지만, 이들이 MAR인지 NMAR인지는 알 수 없다. 왜냐하면, 이들을 구분하는 기준은 결측된 변수의 결측된 값이 결측 매트릭스를 설명하는지의 여부인데, 이 값은 관찰되지 않았으므로 분석할 수 없기 때문이다. 또한, 위의 표에서 결측 매트릭스와 상관관계가 존재하지 않는 경우 일단 관찰된 변수로 결측 매트릭스가 설명되지 않기 때문에 일단 MAR 가정은 배제되지만, 마찬가지로 MCAR인지 NMAR인지를 결정하는 것은 불가능하다. 결국 실제 데이터셋에서 이들 결측 메커니즘에 대해 판단하는 것은 어렵다.[17] 또한, 앞서 살펴본 가상 데이터의 가정과 같이 변수들 간의 상관관계가 존재하는 상황에서 결측의 메커니즘이 이상적인 조건에 맞게 성립하는 것도 생각하기 힘들다. 이러한 상황에서 본 연구의 목적인 재조사법과 통계적 접근을 통해 가장 비용효과적인 방법을 모색하는 것은 결측치로 인한 문제가 존재하고, 이를 해결하는 다양한 방법이 존재하는 상황에서 이를 어떻게 잘 활용할 것인가를 탐구하는 필수적인 과정이다.

앞서 가상 데이터를 이용하여 분석한 결과에서 사용한 완전제거법과 FCS, 그리고 보조변수를 이용한 FCS_BD 모형을 실제 데이터셋에 적용하여 분석한 결과 세 가지 정부수준별 만족도에 대한 만족도 회귀계수는 〈그림 5〉와 같다. 가상 데이터와는 달리 결측이 없는 데이터는 원래의 참값이 아니라 재조사를 통해 각 변수의 항목 무응답을 완전히 제거한 데이터셋을 활용하였다(CP_REAL). 그 결과는 그림에서 로 표시되어 있다. 완전한 데이터셋의 회귀계수에 대하여, 완전제거법(LD_REAL), 분석모형에 포함된 변수들로만 대체모형을 구성한 다중대체법(FCS_REAL), 그리고 분석모형에는 포함되지 않지만 보조변수로서 인구통계학적 변수를 대체모형에 포함시킨 다중대체법(FCS_REAL_AUX)의 회귀계수의 결과를 결측률(10%~50%)에 따라 그래프로 나타내었다. 구체적으로 분석결과는 모집단(1만개의 데이터셋)에서 1,000개의 표본을 뽑고 다중대체법과 완전제거법을 사용하여 각 변수의 회귀계수를 구하는 과정을 30번 반복한 회귀계수의 평균으로 나타나고 있으며, 푸른색 음영은 이 과정에서 완전한 데이터셋(결측률 0%)의 회귀계수의 95% 신뢰구간을 의미한다.

17) 물론 상관관계는 선형관계를 전제로 하므로, 비선형의 관계를 포함하면 상기의 논의보다 복잡한 설명도 가능하다.

〈그림 5〉 결측률 및 대체방법에 따른 회귀계수의 변화



분석결과를 보면, 대부분 CP_REAL에서 LD_REAL이 FCS_REAL이나 FCS_REAL_AUX 보다 더 멀리 떨어져 있는 것을 알 수 있다. 즉, 완전제거법보다는 다중대체법을 활용하는 것이 더 낫다는 것을 보여주는 결과다. 하지만, Q1_2의 경우는 FCS_REAL이 LD_REAL보다 더 큰 편의를 보이는데, 이를 통해 변수의 특성에 따라 다중대체법이 오히려 편의를 증가시키기도 한다는 것을 할 수 있다. 다만, 그 편의가 CP_REAL의 95% 신뢰구간 내에 있으므로, FCS_REAL의 작동하지 않았다기보다 우연히 LD_REAL이 편의를 나타내지 않은 것으로 해석함이 타당하다. 중요한 점은 인구통계학적 변수를 대체모형에 포함시킨 FCS_REAL_AUX 모형이 변수와 결측률에 관계없이 CP_REAL의 결과와 통계적인 차이를 보이지 않는다는 점이다. 이로써 분석모형에서는 필요하지 않은 변수라고 할지라도 대체모형에 포함시키는 경우 통계적 대체방법의 성능을 향상시킨다는 것을 확인할 수 있다. 또한, 결측된 변수의 결측원인을 가장 잘 설명할 수 있는 변수를 연구자의 경험이나 이론을 통해 파악할 수 있다면, 분석모형과는 관계없는 변수라고 할지라도 대체모형에 포함하는 것이 효과적이라는 것을 확인할 수 있다. 마지막으로, 분석모형에 포함된 변수로만 대체모형을 구성하는 경우, 즉 FCS_REAL의 결과도 세 변수 모두에서 결측률 30% 이하에서 CP_REAL의 신뢰구간에 포함되는 것을 알 수 있다. 이는 가장 보수적인 관점에서 통계적 대체방법을 적용하더라도 결측률 30% 이하에서는 분석결과의 편의를 유발하지 않는다는 것을 보여준다. 설문자료의 결측률[18]을 30% 이하로 만들어주는 노력이 필요하며, 이를 재조사법을 통해 수행해야 한다는 주장이 가능하다. 이처럼 재조사법과 통계적 다중대체법의 혼합 활용을 통해 결측 메커니즘과 같은 이상적인(ideal) 조건과 관계없이 실제 데이터에 대하여 기존의 분석방법인 완전제거법의 문제를 보완할 수 있을 것으로 기대된다.

V. 결 론

본 연구는 결측 메커니즘을 검증하기 어려워 다중대체법의 사용이 쉽지 않거나, 높은 비용으로 재조사법을 활용하기 어려운 경우 실제 연구에서 어떻게 결측치를 실질적으로 다루어야 할지를 살펴보았다. 기존의 연구방법론 교과서에서는 무응답이 존재하는 경우 2차 조사, 즉 재조사를 이용하여 항목별 무응답이 존재하는 관측치(혹은 응답자)로부터의 응답을 얻어내는 것이 좋다고 한다(김병섭, 2008). 하지만, 이러한 방법은 시간과 비

18) 개별 변수의 항목 무응답이 존재하는 경우 해당 응답자를 제거한 경우의 결측률

용의 문제에서 자유로울 수 없으며, 이로 인해 실제로 재조사법이 활용되는 경우는 매우 드물다. 한편 다중대체법은 적은 비용으로 활용할 수 있지만 실제 결측 데이터에 적용하는 경우 타당성에 대한 의문으로 인해 보편적으로 활용되고 있지 못한 실정이다. 따라서 어느 정도의 결측을 재조사로 보정을 하고 통계적 다중대체법은 어느 정도의 결측이 있을 때 불편추정치를 얻을 수 있는지를 살펴보는 것은 현실적으로 중요한 연구문제이다.

시뮬레이션 데이터를 이용한 분석과 실제 데이터를 이용한 분석 모두에서 결측 메커니즘, 변수 간 관계 등과 상관없이 결측률이 약 30% 이하인 경우 통계적 다중대체법이 통계적 편의를 유발하지 않는 것으로 나타났다. 또한, 특정 변수의 결측된 원인을 설명할 수 있는 변수를 대체모형에 포함시키는 경우, 종속변수 혹은 분석모형에 포함되지 않는 관측된 변수, 통계적 대체방법의 성능은 더욱 향상되었다. 물론, 이 결과는 설문의 종류, 문항의 특성, 그리고 응답자의 특성에 따라 다른 결과가 나타날 수 있다는 반론에서 자유롭지 못하다. 하지만 본 연구의 결과를 보수적으로 해석한다고 하더라도 결측을 포함한 설문자료에 대하여 통계적 다중대체법을 일차적으로 활용하고 재조사법을 보조적으로 사용하여 결측률을 줄인다면 분석 결과의 편의를 줄이는 것이 가능하다는 것을 보여준다. 또한 다중대체법을 사용할 때 결측 발생에 영향을 줄 수 있는 가능한 많은 변수를 설문조사 결과를 통해 얻어 결측치 대체에 사용하면 성능의 향상을 기대할 수 있는 것으로 나타났다.

행정학과 정책학 분야의 이론검증과 정책평가, 그리고 정책현상에 대한 설명의 많은 부분이 설문조사를 통해 이루어지고 있다. 설문조사를 통한 분석은 표본집단 선정과 표본추출의 방법에 따라 그 결과가 달라질 수 있으므로 연구자들은 이들의 결정에 신중해지기 마련이다. 하지만, 응답된 설문자료를 분석하는 단계에서는, 무응답으로 인해 앞서 공들여 준비한 표본프레임과 추출방법의 논의가 무의미해 질 수도 있음에도 불구하고, 해당자료에 포함된 무응답에 대한 관심은 미약하다. 전체 표본 중 설문에 응답한 응답자 수에 대한 정보에는 관심을 가지면서도, 주요 분석변수에 포함된 무응답에는 관심이 없어 보인다. 실제 설문조사에서 발생하는 항목 무응답은 연구자가 실제 분석에 포함되는 정보의 양에 대한 왜곡된 판단을 하게 할 가능성이 있다. 왜냐하면, 통계패키지의 기본 설정인 완전제거법을 통해 분석하는 경우가 대부분이고, 실제 분석에 사용된 관측치 수에 관심을 가지지 않는다면 실제 사용된 정보보다 많은 정보를 사용한 것으로 오인할 수 있다. 이보다 더 중요한 사실은 완전제거법에 의해 제거된 정보로 인하여 분석결과에 편의가 발생한다는 점이다. 따라서, 무응답에 대한 적절한 처리가 수행되지 않았거나, 최소한 연구자가 사용한 자료의 결측률에 대한 언급조차 없는 분석결과의 편의에 대한

우려할 만한 상황이다.

시뮬레이션과 실제 데이터에 대한 분석을 통해 연구의 외적타당성을 높이려는 노력에도 불구하고, 여전히 본 연구의 분석결과가 보편적으로 적용가능한 것인가라는 의문에서 완전히 자유로울 수는 없다. 하지만, 이는 이 분야에 대한 지속적인 관심과 후속연구를 통해 귀납적으로 접근해 가야 할 문제라고 판단된다. 이와 같은 관점에서 향후 종속변수에 결측이 존재하는 경우 나타나는 문제와 결측 메커니즘에 대한 검증방법에 대한 문제 등 중요한 이슈들에 관한 활발한 연구와 지속적인 관심이 행정학계에도 필요하다.

참고문헌

강민아·김경아. (2006). 행정학 및 정책학 조사연구에서 결측치 발생과 처리 방법에 대한 고찰.「한국행정학보」, 40(2): 31-52.

고길곤. (2014).「통계학의 이해와 활용」. 고양: 문우사.

고길곤·탁현우·이보라. (2014). 설문조사 연구에서 결측치의 영향과 대체방법의 적절성에 대한 실증연구.「정책분석평가학회보」, 24(3): 49-75.

김서영·박라나. (2013). 무응답가구의 특성 분석 사례연구.「조사연구」, 14(1): 31-67.

한혜은·변종석. (2014). 재조사의 무응답 편향 감소 효과.「조사연구」, 15(1): 21-45.

한혜은·김영원. (2015). 가구방문조사에서 무응답 보정을 위한 파라데이터 활용 - 국제성인역량조사 사례분석을 중심으로.「조사연구」, 16(1): 227-251.

Allison, P. D. (2000). Multiple Imputation for Missing Data A Cautionary Tale. *Sociological Methods & Research, 28*(3): 301-309.

Azur, M. J., Stuart, E. A., Frangakis, C., &Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research, 20*(1): 40-49.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1): 5-37.

Berglund, P., & Heeringa, S. G. (2014). *Multiple Imputation of Missing Data Using SAS*: SAS Institute.

Burgess, S., White, I. R., Resche-Rigon, M., & Wood, A. M. (2013). Combining multiple imputation and meta-analysis with individual participant data. *Statistics in medicine, 32*(26): 4499-4514.

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential

regression trees. *American Journal of Epidemiology*, kwq260.

Cox, B. E., McIntosh, K., Reason, R. D., & Terenzini, P. T. (2014). Working with missing data in higher education research: A primer and real-world example. *The Review of Higher Education*, 37(3): 377-402.

Daniel, R. M., & Kenward, M. G. (2012). A method for increasing the robustness of multiple imputation. *Computational Statistics & Data Analysis, 56*(6): 1624-1643.

Eekhout, I. (2015). Don't Miss Out!: Incomplete data can contain valuable information. URI: http://hdl.handle.net/1871/52171

Enders, C. K. (2010). *Applied missing data analysis*: Guilford Publications.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology, 60*: 549-576.

_____. (2012). *Missing data: Analysis and design*: Springer Science & Business Media.

Landerman, L. R., Land, K. C., & Pieper, C. F. (1997). An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological Methods & Research, 26*(1): 3-33.

Lee, M., Cronin, K. A., Gail, M. H., Dignam, J. J., & Feuer, E. J. (2011). Multiple imputation methods for inference on cumulative incidence with missing cause of failure. *Biometrical Journal, 53*(6): 974-993.

Lieberman-Betz, R. G., Yoder, P., Stone, W. L., Nahmias, A. S., Carter, A. S., Celimli-Aksoy, S., & Messinger, D. S. (2014). An Illustration of Using Multiple Imputation Versus Listwise Deletion Analyses: The Effect of Hanen's "More Than Words" on Parenting Stress. *American journal on intellectual and developmental disabilities*, 119(5): 472-486.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*: John Wiley & Sons.

Long, Q., & Johnson, B. A. (2015). Variable selection in the presence of missing data: resampling and imputation. *Biostatistics*, kxv003.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*: Guilford Press.

Peyre, H., Leplège, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, 20(2): 287-300.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*: John Wiley & Sons.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*: CRC press.

_____. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1):

3-15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods, 7*(2): 147.

Young, R., & Johnson, D. R. (2010, May). 'Imputing the Missing Y's: Implications for Survey Producers and Survey Users. *In Proceedings of the AAPOR Conference Abstracts* (pp. 6242-6248).

ABSTRACT

The Treatment of Missing Values using the Integrated Multiple Imputation and Callback Method

Kilkon Ko & Hyunwoo Tak

Even though many studies warn of the impact of missing values on analytical results, in practice, researchers simply rely on the listwise deletion method for their own convenience. Some argue that the multiple imputation method is inferior to the callback method even if they want to utilize it.

This paper tries to integrate the multiple imputation method with the callback method. As the callback method is costly, the appropriate ratio is a practically and theoretically important question. This survey tries to suggest a ratio using survey and simulation data.

According to the results of the analysis, the multiple imputation method does not cause significant statistical bias regardless of missing mechanisms or correlation among variables if the missing rate is less than 30%. In particular, including more auxiliary variables related to the missing structural mechanism can increase the performance of the multiple imputation. Hence, the callback approach can be used to reduce the missing rate below 30% and the multiple imputation to improve the validity of the results of analysis.

【Keywords: missing value, multiple imputation, callback, survey】